Proceedings of the 2023 Winter Simulation Conference
C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.

DYNAMIC STRATIFICATION AND POST-STRATIFIED ADAPTIVE SAMPLING FOR SIMULATION OPTIMIZATION

Pranav Jain
Sara Shashaani

Edward P. Fitts Department of Industrial and Systems Engineering
NC State University
915 Partners Way
Raleigh, NC 27606, USA

ABSTRACT

Post-stratification is a variance reduction technique that groups samples in respective strata only after collecting the samples randomly. We incorporate this technique within an adaptive sampling procedure in simulation optimization. We use concomitant variables to increase the accuracy of our proposed post-stratified adaptive sampling. Concomitant variables are auxiliary variables in simulation that approximate the boundaries of the optimal strata at each visited solution during the optimization procedure. A linear relationship between the concomitant variable and the output is desirable but not necessary for the effectiveness of the proposed methodology. In numerical experiments, we observe that performing post-stratified adaptive sampling with dynamically updated strata boundaries robustifies the algorithm in the sense that it reduces the algorithm's sensitivity to the initial solution and solver input parameters.

1 INTRODUCTION

Simulation optimization is a problem of determining the values for decision variables of a simulation model that optimize one or more of its performance measures. Simulation-optimization methods often rely on optimizing estimates of performance measures. Hence, their success depends on the estimated values' accuracy. A crude Monte-Carlo estimation via sample-average-approximation (SAA) (Kim et al. 2015) renders the estimator's variance inversely proportional to the square root of the sample size. Thus running many simulation replications can increase accuracy at a slow rate of $\mathcal{O}(n^{-1/2})$, which is limiting for many computationally costly simulations. Variance reduction techniques such as control variates (Lavenberg et al. 1982), importance sampling (Glynn and Iglehart 1989), and stratified sampling aim for gain in accuracy through careful consideration for the distribution of simulation inputs or outputs. Among them, stratified sampling with concomitant variables (Wilson and Pritsker 1984) utilizes other random variables produced during simulation to stratify the simulation outputs that will estimate the performance measure being optimized. It is best used within a post-stratification routine where the random simulation runs are still independent and identically distributed (iid) but grouped and weighted based on the stratification rule. The advantage of using concomitant variables has been explored for estimation of a single target value. But in an optimization task, the performance measure is estimated at many points raising the question of whether changing the stratification structure repeatedly would enhance the convergence behavior of the solver. In this work, our investigation of this question is within a class of stochastic optimization solvers that utilize adaptive sampling. With adaptive sampling, the sample size is not fixed and adapts to how much accuracy would be needed at a point. We explore the post-stratification in the adaptive setting while allowing the stratification itself to vary based on the trajectory of the search.

Post-stratification has been widely used as a variance reduction technique in simulations of queuing networks (Wilson and Pritsker 1984; Sabuncuoglu et al. 2008), focusing on a fixed stratification structure. In an optimization, however, depending on the variable used to stratify the outputs, the best way of splitting the data could depend on the solution at which the objective function is being evaluated. Intuitively, if the stratification variable has a high dependence on the output, it is likely that the distribution of the output conditional on that variable will look different at different points. So our goal in this paper is to investigate the effect of changing the stratification structure dynamically within the optimization framework. Dynamic stratification has been explored for Monte Carlo simulations and simulation optimization. However, most of these approaches stratify the outputs using greedy routines or heuristics like clustering or regression trees (Ross and Lin 2001; Zhao and Zhang 2014; Pettersson and Krumscheid 2021; Liu et al. 2022; Jain et al. 2021; Jain et al. 2022). The structure generated by these methods tends to depend on the data used. A small sample can result in a skewed structure and poor estimates. Thus they often require a large pilot sample size. This can be harmful from the perspective of optimization under a finite budget. A large initial sample size leaves less budget for exploration, which is of the essence for finding an optimal region. In this work, we use a closed-form approach for dynamic stratification, allowing us to use a relatively small initial sample size without sacrificing the accuracy of the stratification structure. We integrate this with adaptive post-stratified sampling to improve the performance of a derivative-free trust-region-based method. This integration makes the total sample size and the stratification structure part of the solver's decision at each iteration during optimization. Given the exploratory nature of this study, we choose to illustrate the proposed methods on an M/M/1 queue, where the relationships between variables under study is well-understood. We leave the rigorous analysis of effect on convergence rates, and impact in real-world settings to a future work.

Consider a simulation model generating outputs $Y(\theta) \in \mathbb{R}$ for some decision variable $\theta \in \mathbb{R}^d$. The aim is to determine the decision value that minimizes $Y(\theta)$ in expectation. For the M/M/1 queue, $Y(\theta) := S(\theta) + c_0\theta^2$ where θ is the exponential service rate, λ is the exponential interarrival rate, $S(\theta)$ is the mean sojourn time, and c_0 is the additional cost to penalize increase in service rate. The problem can thus be formulated as

$$\min_{\theta\in(\lambda,\theta_{\max}]} \quad f(\theta) := \mathbb{E}[Y(\theta)], \tag{1}$$

a box-constrained stochastic optimization problem with $\theta_{\max}$ as the upper bound of θ. We assume that the function $f(\theta)$ is bounded from below and has L-Lipschitz continuous gradients in $\mathbb{R}$. We can estimate the expectation in (1) via SAA using $\hat{f}(\theta, n) = n^{-1}\sum_{j=1}^{n} Y_j(\theta)$ where n is the number of simulation replications at θ. Let $X(\theta)$ be a random variable that is generated besides $Y(\theta)$ in one simulation run at θ. We will review stratified sampling as a variance reduction technique and the use of concomitant variables for stratification in Section 2. Section 3 presents the trust-region optimization equipped with stratified adaptive sampling where the strata randomly change in addition to the sample size. We explore different approaches for choosing concomitant variables. Numerical results for the M/M/1 problem are presented in Section 4 with conclusions in Section 5.

2 STRATIFIED SAMPLING

Stratified sampling involves dividing data into groups or strata so that data behaves more similar within each group. Instead of using a single distribution, stratified sampling uses different distributions for each group, which helps exploit data heterogeneity leading to variance reduction (Ross 2013). Sampling more points from a stratum with a higher variance will increase efficiency. Efficient allocation of the computational budget between strata can reduce the variance of the estimators and expedite the optimization.

For ease of exposition, let us fix θ and drop it from the rest of this section. Let $\mathcal{D}$ be the support (range) of X and suppose we have m disjoint strata $\mathcal{D}_j,\ j = 1, 2, \cdots, m$ on this support, such that $\bigcup_{j=1}^{m} \mathcal{D}_j = \mathcal{D}$. Define $p_{X,j} = \Pr\{X \in \mathcal{D}_j\}$, the true probability of X falling inside $\mathcal{D}_j$, $\mathbb{P}_j$, its probability distribution function in stratum j, and $\sigma_j^2 = \text{Var}(Y \big| X \sim \mathbb{P}_j)$. With $f_j := \mathbb{E}[Y \big| X \sim \mathbb{P}_j]$, the mean in stratum j, one can

then write $f = \sum_{j=1}^{m} p_{X,j} f_j$, and devise the unbiased estimator $\check{f}(n_1, n_2, \ldots, n_m) = \sum_{j=1}^{m} p_{X,j} \hat{f}_j(n_j)$, where $\hat{f}_j(n) = n^{-1} \sum_{i=1}^{n} Y_{j,i}$ and $Y_{j,i}$ is the i-th i.i.d. simulation output with its input sampled from $\mathbb{P}_j$ defined on the support $\mathcal{D}_j$ with sample size n_j. Then

$$\text{Var}(\check{f}(n_1, n_2, \ldots, n_m) = \sum_{j=1}^{m} n_j^{-1} p_j^2 \sigma_j^2. \tag{2}$$

is the variance of the new estimator with stratified sampling, in comparison with that of the original Monte-Carlo-based estimator, i.e., $\text{Var}(\hat{f}(n)) = n^{-1}\sigma^2$, where $\sigma^2 = \text{Var}(Y)$. The sample variance in (2) is always smaller than the sample variance without stratification (Ross 2013) and its value depends on the sample size of each stratum n_j. If this sample size is selected appropriately, we can achieve the maximum reduction in the sample variance.

2.1 Splitting and Budget Allocation

Using stratified sampling best involves two questions: how to split $\mathcal{D}$ into m strata and how to choose n_j? Answering the first question requires finding a splitting structure such that each stratum has similar observable characteristics, like the variability of the objective function. Farias et al. (2020) propose a splitting technique based on similarity functions for classification, which demands accurate modeling of the available dataset's true distribution. Mulvey (1983) present a computationally expensive optimal cluster analysis. Tipton (2013) use k-means clustering and compare the algorithm with $1, 2, \cdots, k$ strata to determine the optimal k. Comparison is in terms of the ratio of the between-strata variance to the sum of within-strata and between-strata variance, which tends to increase with more strata. Jain et al. (2022) suggest more strata may hinder the optimization process rather than assist it.

The second question depends on the sampling strategy. *Proportional allocation* chooses the sample size of a stratum based on $p_{X,j}$. The proportional allocation of stratum j admits $n_j = p_{X,j} n$. *Optimal allocation* (Neyman 1934) chooses the sample size also based on the variance σ_j^2. The sample size of a stratum j in optimal allocation is determined as $n_j = w_j n$, where $w_j := p_{X,j}\sigma_j / (\sum_{i=1}^{m} p_{X,i}\sigma_i)$ is the weight of stratum j. Theoretically, optimal allocation maximizes the estimator's variance; however, this reduction depends on knowing $p_{X,j}$ and σ_j. When unknown, inaccurate estimates of these two can reduce or reverse the effectiveness of the optimal allocation. Jain et al. (2021) illustrate two approaches to estimate the weights: static and dynamic. Static weights are calculated at the start of optimization using inputs whose distribution remains fixed as θ changes. Dynamic weights use the outputs and are updated at the end of each iteration. Although dynamic weights capture the behavior of the objective function for a given θ, they are prone to estimation errors as they will need many replications to have a good enough estimate of σ_j, imposing more computational burden.

Other allocation strategies that minimize the variance within each stratum have widely been studied (Etoré and Jourdain 2010; Kawai 2010). Chaddha et al. (1971) determine the optimal allocation based on specific graphical procedures. Huddleston et al. (1970) use convex programming whereas Bretthauer et al. (1999) choose the optimal sample size of each stratum via branch and bound methods. Glynn and Zheng (2021) suggest using the delta method. Pettersson and Krumscheid (2021) leverage a hybrid allocation scheme, a combination of proportional and optimal allocation greedily dividing the input domain with hyperrectangles or simplices. Tipton et al. (2014) explore randomized experimental design with an inference and an eligibility population, where the inference population stratifies the support with proportional allocation, albeit not uniformly but based on the within-stratum distance function of the propensity score.

Although an optimization routine involves a sequence of θ's, most procedures above leads to a fixed sample size for all θ's in the search trajectory. Zhao and Zhang (2014) analytically prove better convergence properties in stochastic gradient algorithm using stratified sampling with fixed strata and fixed sample size. However, adaptive sampling advocates efficiency gain by changing the sample size during

the optimization (Shashaani et al. 2018; Bollapragada et al. 2018; Curtis and Scheinberg 2020). Espath et al. (2021) and Liu et al. (2022) use adaptive sample size with stratification during the optimization.

2.2 Optimal Stratification and Concomitant Variables

Unlike traditional stratified sampling settings, in this work we assume the stratification structure is not known a priori. Instead, stratification is part of the optimization algorithm that can change based on progress to provide efficiency in the search. However, changing the stratification multiple times during the optimization could cause additional variance and worsen the solver's performance without care. Besides, poor stratification can result in misleading estimates, slowing the optimization process, and even worse solutions with fixed computation budget (Jain et al. 2022). Dalenius (1950) propose a closed-form expression for the optimal stratification structure, which requires the distribution of Y, which is unknown. As a result, many researchers suggest simplifying assumptions to approximate optimal strata boundaries (Dalenius and Hodges Jr 1959; Ekman 1959).

Another approach is to determine the stratification structure with concomitant variables' distribution. In the context of the M/M/1 queue problem, the average waiting time, the normalized sum of service times, or the fraction of customers in the system are some options that serve as concomitant variables whose distribution is known or can be approximated. If the concomitant variables positively correlate with the simulation output, splitting based on them can reduce the estimator's variance. This is reminiscent of the control variates, although the concomitant variables may still affect the estimator even without correlation with the outputs.

Let $x(0)$ and $x(m)$ be the meaningful smallest and largest values of X ($-\infty$ and $+\infty$ if X is unbounded). Then optimal stratification involves determining the stratification boundaries $x(1), \cdots, x(m-1)$, for $\mathcal{D}_j = [x(j-1), x(j)]$, that minimize the variance of the estimator (2). Assuming a linear relation between X and Y, i.e. $Y = \alpha + \beta X + \epsilon$, then the boundaries that minimize the variance of optimal allocation, i.e.,

$$\min_{x(0)<x(1)<\cdots<x(m)} \quad \frac{1}{n}\sum_{j=1}^{m} p_{X,j}^2 \text{Var}\left(Y \middle| X \in [x(j-1), x(j)]\right),$$

can be determined by solving the set of equations

$$\frac{\beta^2[(x(j)-\mu_{X,j})^2+\sigma_{X,j}^2]+2\sigma_{\epsilon,j}^2}{\beta\sigma_{X,j}\sqrt{1+\sigma_{\epsilon,j}^2/(\beta^2\sigma_{X,j}^2)}} = \frac{\beta^2[(x(j+1)-\mu_{X,j+1})^2+\sigma_{X,j+1}^2]+2\sigma_{\epsilon,j+1}^2}{\beta\sigma_{X,j+1}\sqrt{1+\sigma_{\epsilon,j+1}^2/(\beta^2\sigma_{X,j+1}^2)}},\ j=1,2,\ldots,m-1, \tag{3}$$

where $\mu_{X,j} := \mathbb{E}[X|X \in \mathcal{D}_j]$ and $\sigma_{X,j}^2 := \text{Var}(X|X \in \mathcal{D}_j)$ are the mean and variance of X in stratum j, with $\sigma_{\epsilon,j}^2$ being the variance of the error term, ϵ, in stratum j (Cochran 1977; Singh and Sukhatme 1969). We admit this applies only to the case where the concomitant variable is one-dimensional, and for a multi-dimensional space, such an optimization would be cumbersome. Finding the concomitant variable's mean and variance in every stratum can be difficult, but an approximate stratification structure can be determined by substituting these exact quantities with their estimates. Solving (3) with the estimated amounts with nonlinear equality constraint is trivial with deterministic optimizers and allows us to estimate the approximate optimal boundaries consistently. Finding a variable X with an exact linear relationship with Y can be difficult. Instead, one can use a function $g(X)$ (e.g., X^3 or e^X) that appears to linearly depend on Y as the concomitant variable to determine the stratification structure. Even if the dependence between X or $g(X)$ and Y is not precisely linear, this approach will still be helpful as long as they are correlated. As we will see later in the experiments, finding another output variable in the simulation that increases or decreases with the increase or decrease of Y is often easier. In the M/M/1 queue, even without any knowledge of the closed-form relationships, one could use the intuitive knowledge that the sojourn time would likely be large if the waiting time was large.

The earliest use of concomitant variables to determine optimal stratification structure was proposed by Dalenius and Gurney (1951) and Taga (1967). Sethi (1963) and Cochran (1977) proposed iterative numerical procedures to determine the optimal strata boundaries. Singh and Sukhatme (1969) provided approximate optimal stratification boundaries based on (3). These approximate methods are easy to implement but are based on restrictive assumptions of X following a well-known probability distribution. Recent work on optimal stratification boundaries is focused on using optimization techniques (Brito et al. 2010; de Moura Brito et al. 2017) and dynamic programming (Khan et al. 2008) to solve (3). Though these methods are more accurate, they are also computationally expensive.

If the concomitant variable is independent of the decision variable θ, e.g., interarrival time in the M/M/1 queue, we can expect that this approach will not result in drastic changes in the boundaries during optimization. Meanwhile, a θ-dependent concomitant variable, e.g., random service times, leads to an infinite loop to find the optimal boundaries since estimating the unknown quantities in (3) depend on the sample size n_j which itself depends on the stratification structure. This motivates post-stratification, to draw samples independently before forming the stratification structure and grouping the outputs.

2.3 Post-stratification

Post-stratification, typically used in survey sampling, accommodates cases with no known strata. Suppose that we have n iid copies $\{(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)\}$ are available. Post-stratified sampling then allocates each Y_i to a stratum j based on the value of X_i such that n_j are the number of points that fall in stratum j. The variance of the post-stratified estimator for proportional allocation is

$$\text{Var}_{\text{post}}(\check{f}(n_1, n_2, \ldots, n_m)) = \frac{1}{n}\sum_{j=1}^{m} p_{X,j}\sigma_j^2 + \frac{1}{n^2}\sum_{j=1}^{m}(1 - p_{X,j})\sigma_j^2, \tag{4}$$

where σ_j^2 is estimated with $\hat{\sigma}_j^2 = \sum_{i:X_i \in \mathcal{D}_j}(Y_i - \hat{f}_j(n_j))^2/(n_j - 1)$ letting the SAAs in each stratum be $\hat{f}_j(n_j) = \sum_{i:X_i \in \mathcal{D}_j} Y_i/n_j$. The $p_{X,j}$ is estimated with $\hat{p}_{X,j} = \#\{X_i \in \mathcal{D}_j\}/n$ from the n replications. Note, the square root of the variance in (4) is the standard error of the estimator $\check{f}(n_1, n_2, \ldots, n_m)$, which we will denote with $\text{se}(n_1, n_2, \ldots, n_m)$.

If the estimates $\hat{p}_{X,j}$'s are accurate and the sample size of each stratum n_j is large, then post-stratified sampling is almost as accurate as stratified sampling with proportional allocation. Though the variance reduction with post-stratification may not reach that of one where samples are directly drawn from the known strata, it is more stable, removing another layer of randomness in the process. Compared to standard simulations, Wilson and Pritsker (1984) have achieved efficiency using concomitant variable-based post-stratification for queuing simulations. We apply the idea of concomitant variable-based post-stratification to a derivative-free trust-region-based simulation-optimization framework.

3 ASTRO-DF WITH POST-STRATIFICATION

Adaptive Sampling Trust-Region Optimization for Derivative-Free stochastic oracles (ASTRO-DF) is an almost sure convergent algorithm for stochastic non-convex problems (Shashaani et al. 2018; Ha and Shashaani 2023). ASTRO-DF uses adaptive sampling within a trust-region framework to boost its efficiency.

Trust-region methods form a local model, an approximation of the true objective function, around the current iterate and minimize this model to suggest an incumbent solution. Let θ_k be the current incumbent at iteration k, then the trust region is $\mathcal{B}_k = \{\theta : \|\theta - \theta_k\|_2 \leq \Delta_k\}$, a closed ball around θ_k, where Δ_k is the trust region radius. During iteration k, the local model $M_k(\theta)$ is generated within $\mathcal{B}_k$ using interpolation on several adjacent points to θ_k, hence being a derivative-free solver. A candidate for the next incumbent, denoted by $\tilde{\theta}_{k+1}$, reduces this model sufficiently while remaining within $\mathcal{B}_k$. $\tilde{\theta}_{k+1}$ is accepted as θ_{k+1}, and the trust region expands if the reduction in the function value is also sufficient. Otherwise, the candidate solution is rejected, the trust-region radius shrinks, and a new model is formed in a smaller neighborhood

around $\theta_{k+1} = \theta_k$. Since $\Delta_k \to 0$ as $k \to \infty$ almost surely, and the model gradient is in tandem with the gradient of the objective function, the algorithm is guaranteed to converge with probability 1.

3.1 Post-stratified Adaptive Sampling

Adaptive sampling determines the optimal sample size for each iteration as one that maintains the estimation error below a changing threshold, instead of using a fixed sample size. This makes the sample size stochastic, signified by a capital letter throughout the rest of the paper. Regulating the sample size based on the estimator's variance and how far the solution under consideration may be from optimality helps the efficiency of the optimization. In ASTRO-DF, the adaptive sample size at θ_k is determined by ensuring that the estimation error is less than the square of the trust-region radius that approximates the optimality gap, i.e., $N_k = \min\left\{n \geq \lambda_k : \frac{\hat{\sigma}_k}{\sqrt{n}} \leq \kappa\frac{\Delta_k^2}{\sqrt{\lambda_k}}\right\}$ with $\kappa > 0$ a constant, λ_k a deterministic increasing sequence, and $\hat{\sigma}_k$ the standard deviation estimate of the objective function value. The closer to the optimal solution, the larger the sample size, improving the estimates' accuracy.

At the beginning of iteration k, the total sample size at θ_k is unknown. This complicates the implementation of stratified sampling to ASTRO-DF as both stratification and budget allocation would need the total sample size for efficient variance reduction. To implement post-stratification, we first change the crude Monte Carlo variance and use (4) to obtain the standard error and use in

$$N_k = \min\left\{\sum_{j=1}^{m} N_{k,j} \geq \lambda_k : \widehat{\text{se}}_k(N_{k,1}, N_{k,2}, \ldots, N_{k,m}) \leq \kappa\frac{\Delta_k^2}{\sqrt{\lambda_k}}\right\}. \tag{5}$$

We first generate n_0 iid replications, determine the boundaries of the strata (explained next) and then update the standard error in (5). If the condition is not met, we add one more replication without changing the stratification and repeat.

3.2 Adaptive Splitting

How are the strata chosen for each θ_k? If the concomitant variable's distribution is known, we can use numerical iterative methods (Sethi 1963; Cochran 1977) that we will now explain. For the M/M/1 problem, suppose we want to form the stratification structure based on the mean service times of all the customers served till time t. Consider an instance of simulation where θ_k is the current service rate and $Z_{k,i}$ is the service time of i^{th} person in the queue – an exponential random variable with mean $1/\theta_k$. Let $Q_k(t)$ be the total number of customers served by time t. Then the mean service time is $\sum_{i=1}^{Q_k(t)} Z_{k,i}$. Since the total number of customers served till time t is a random variable that depends on θ_k, the variance of mean service cannot be bounded as t increases (Wilson 1979). We can consider the standardized mean service time $X_k = (Q_k(t))^{-1/2}\sum_{i=1}^{Q_k(t)} \theta_k(Z_{k,i} - 1/\theta_k)$. This standardized mean service time asymptotically follows the standard normal distribution (Wilson and Pritsker 1984; Chung 2001) with closed-form values for each of the quantities in (3), leading to exact boundaries as listed in Table 1. Even though the stratification structure for this case does not change throughout the optimization process, the concomitant variable's dependence on θ_k can be recovered by destandardizing, as the generated mean service times are standardized to group the outputs. For example, for $m = 4$ strata, the value of 0.2 for the concomitant variable after running a simulation places it on the third stratum, which updates its estimated mean and variance.

The above approach with Table 1 is not applicable when the concomitant variable distribution is unknown or its conditional probabilities, means, and variances are inconvenient to compute and hence it cannot be standardized, e.g., the mean waiting time. In this case, an approximate stratification structure, $\{x_k(0), x_k(1), \cdots, x_k(m)\}$, can be determined by solving (4) such that $x_k(0) < x_k(1) < \cdots < x_k(m)$ and (3) is satisfied for all $x_k(1), x_k(2), \cdots, x_k(m-1)$. The $\hat{p}_{k,X,j}$ probabilities can be estimated with the approximate strata boundaries. To implement this approach, n_0 iid simulations enable fitting a linear

Table 1: When the concomitant variable follows a standard normal distribution, the strata boundary is the optimal upper limit (independent of θ_k) and stratum probability is the weight of that stratum.

Number of strata (m)	Strata boundaries ($x(j)$'s)				Strata probabilities ($p_{X,j}$'s)			
	1	2	3	4	1	2	3	4
2	0.000	∞			0.500	0.500		
3	-0.612	0.612	∞		0.270	0.459	0.271	
4	-0.982	0.000	0.982	∞	0.163	0.337	0.337	0.163

regression between Y_k and X_k. As mentioned earlier, after n_0 simulations, more may be need to fulfill the adaptive sampling criteria but we do not update the strata boundaries with them as that could affect the variability of the updated standard error.

Algorithm 1 summarizes the working of S-ASTRO-DF, (Stratified ASTRO-DF). See (Shashaani et al. 2018) for model construction details. The steps for post-stratified adaptive sampling are listed in Algorithm 2. If the stratification structure can be estimated beforehand (e.g. the standardized mean service time), the post-stratified adaptive sampling skips the nonlinear optimization part and directly jumps to the allocation step. The nonlinear program is solved only once for each θ_k, and the computational cost of solving it is negligible compared to running the simulations. The worst case complexity of solving the nonlinear program to determine optimal stratification structure is $\mathcal{O}((m-1)^3 m n_0)$, where m is the number of strata and n_0 is the pilot sample size. This also gives a way to determine the two parameters m and n_0. A small n_0 may result in inaccurate estimates and a sub-optimal stratification structure, slowing down the progress. This can result in a substandard final solution given a finite budget. A large n_0 will leave less budget to explore and increase the cost of solving the nonlinear problem. Thus, the pilot sample size must be selected keeping in mind the trade-off between exploration and exploitation. On the other hand, as the total number of strata m increases, the variance of the post-stratified sampling estimator reduces. However, increasing m beyond a certain point may not significantly reduce the estimator's variance. Extensive studies have shown that for $m > 7$, the reduction in the estimator's variance is negligible (Cochran 1977). Increasing m also increases the computational cost of solving the nonlinear problem.

4 NUMERICAL RESULTS

We validate the proposed methods by running numerical experiments for the M/M/1 queue. Use of four concomitant variables: (i) expected utilization, (ii) mean waiting time, (iii) standardized mean service time and (iv) standardized mean interarrival time, will be compared with no stratification. Table 2 summarizes the concomitant variables used for stratification. Amongst the four concomitant variables considered, only the mean waiting time is known to have an exact linear relationship with mean sojourn time. For the other cases, there is some correlation between the concomitant variable and the mean sojourn time. We use the stratification boundaries for the standard normal distribution for the two inputs. The boundaries for two outputs are estimated by fitting a linear regression model and solving the nonlinear program.

Table 2: Summary of the concomitant variables used for numerical experiments.

Concomitant Variable	Type	Known linear relationship with mean sojourn time
Expected Utilization	Output	No
Mean Waiting Time	Output	Yes
Standardized Mean Service Time	Input	No
Standardized Mean Interarrival Time	Input	No

Algorithm 1 S-ASTRO-DF

Input: Initial solution θ_0 and TR radius Δ_0, maximum budget $b_{\max}$, total minimum sample size n_0, number of strata m, and success threshold $\eta_1 > 0$.
initialization: Set the total number of replications, $W_k = 0$ and iteration $k = 0$.
while $W_k < b_{\max}$ **do**
  Generate $\Theta_k = \{\theta_k^0, \theta_k^1, \dots, \theta_k^p\}$, a poised interpolation set within $\mathcal{B}_k$, where $\theta_k^0 := \theta_k$.
  Estimate $\check{f}_k^i(N_{k,1}^i, N_{k,2}^i, \dots, N_{k,m}^i)$ using Algorithm 2 for the i-th points in Θ_k.
  Set $W_k = W_k + \sum_{i=0}^{p} \sum_{j=1}^{m} N_{k,j}^i$.
  Generate a surrogate model $M_k(\cdot)$ by interpolation.
  If the model gradient $\nabla M_k(\theta_k)$ is small relative to Δ_k, shrink the TR and go to step 4.
  Minimize $M_k(\cdot)$ within $\mathcal{B}_k$ to obtain a candidate solution $\tilde{\theta}_{k+1}$.
  Estimate $\check{f}_k^s(\widetilde{N}_{k+1,1}, \widetilde{N}_{k+1,2}, \dots, \widetilde{N}_{k+1,m})$, the function value at $\tilde{\theta}_{k+1}$, using Algorithm 2.
  Set $W_k = W_k + \sum_{j=1}^{m} \widetilde{N}_{k+1,j}$.
  Compute the success ratio $\hat{\rho}_k = \frac{\check{f}_k^0(N_{k,1}^i, N_{k,2}^i, \dots N_{k,m}^i) - \check{f}_k^s(\widetilde{N}_{k+1,1}, \widetilde{N}_{k+1,2}, \dots, \widetilde{N}_{k+1,m})}{M_k(\theta_k^0) - M_k(\tilde{\theta}_{k+1})}$.
  if $\hat{\rho}_k > \eta_1$ **then**
    Set $\theta_{k+1} = \tilde{\theta}_{k+1}$ and $\Delta_{k+1} > \Delta_k$.
  else
    Set $\theta_{k+1} = \theta_k$ and $\Delta_{k+1} < \Delta_k$.
  end if
  Set $k = k + 1$ and go to step 4.
end while
output: Final calibrated wake parameter θ_k and its estimated loss $\check{f}_k(N_{k,1}, N_{k,2}, \dots, N_{k,m})$.

Algorithm 2 Post-Stratified Adaptive Sample Size Selection

input: TR radius Δ_k, deflation factor λ_k, solution of interest θ_k^i, strata boundaries $x_k^i(0), x_k^i(1), \cdots, x_k^i(m)$ if available, and strata probabilities $\hat{p}_{k,X,1}^i, \hat{p}_{k,X,2}^i, \cdots, \hat{p}_{k,X,m}^i$ if available.
Run n_0 iid simulations.
if strata boundaries are not known **then**
  Fit a linear regression model $Y_k^i = \alpha_k^i + \beta_k^i X_k^i + \epsilon_k^i$ for $(Y_{k,l}^i, X_{k,l}^i)$ for all $l = 1, 2, \cdots, n_0$.
  Set $x_k^i(0) = 0$ and $x_k^i(m) = \infty$.
  Determine the strata boundaries by minimizing (4) such that (3) is satisfied for all $x_k^i(j)$.
end if
Allocate the n_0 points to strata based on the strata boundaries and determine $N_{k,j}^i$ for all $j = 1, 2, \cdots, m$.
if strata probabilities are not known **then**
  Set $\hat{p}_{k,X,j}^i = N_{k,j}^i / N_k^i$.
end if
Calculate the sample mean $\check{f}_k^i(N_{k,1}^i, N_{k,2}^i, \dots, N_{k,m}^i)$ and sample variance $\widehat{\text{se}}_k^i(N_{k,1}^i, N_{k,2}^i, \dots, N_{k,m}^i)$.
while $\widehat{\text{se}}_k^i(N_{k,1}^i, N_{k,2}^i, \dots, N_{k,m}^i) > \frac{\kappa}{\sqrt{\lambda_k}} \Delta_k^2$ **do**
  Run a single iid simulation.
  Allocate this point to a stratum j and increase $N_{k,j}^i$ by 1.
  Update $\widehat{\text{se}}_k^i(N_{k,1}^i, N_{k,2}^i, \dots, N_{k,m}^i)$.
end while
output: Estimated loss $\check{f}_k^i(N_{k,1}^i, N_{k,2}^i, \dots, N_{k,m}^i)$ and sample sizes $N_{k,j}^i,\ j = 1, 2, \cdots, m$.

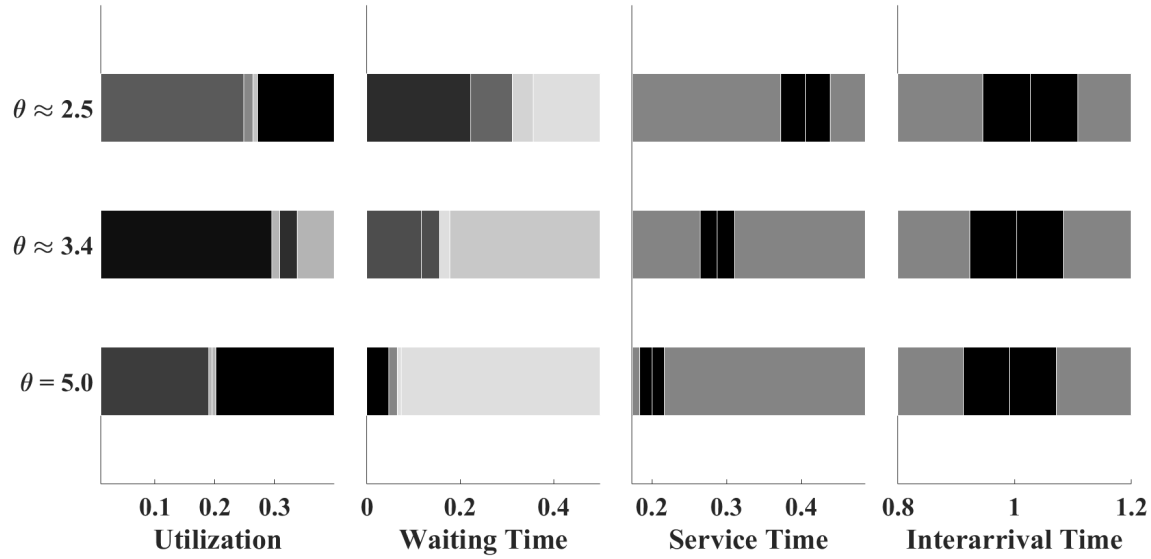


Figure 1: Change in stratification structure with θ during one macroreplication of the optimization routine at the beginning (bottom row), with 50% budget is complete (middle row) and at completion (top row). The gray-scale indicate the probability mass in each stratum, darker for more probability.

The standardized mean service and interarrival times are input to the simulation with a known distribution. We can thus use the stratification boundaries from Table 1. We determine the approximate stratification boundaries for the first two cases by solving the nonlinear problem. Figure 1 shows how the stratification structure changes with θ_k during a single simulation for each of the four different concomitant variables. As expected, the boundaries only slightly change with θ_k for interarrival times as X. They only shift and scale for the service time as X, given the fixed values in the table and standardization. But both waiting time and utilization set boundaries that looks largely different across θ's.

To see how each case would perform in terms of the quality of the terminal solution and its rate of progress, we run 20 independent macroreplications with common random numbers. Each macroreplication starts at the same initial point (θ_0) and has a total budget of $b_{\max} = 10,000$ simulations. Each simulation has a length of $t = 200$ with a warm-up period of 50 and the interarrival rate is 1. The constant, c_0, that penalizes high service rates in the objective function value is set to 0.1. The initial sample size n_0 plays an important role in determining the stratification structure; we have set it to 40. For the no stratification case, we do not need a large n_0 value and hence we have set it to 5. During optimization, we report the intermediate solutions and evaluate the objective function value at these intermediate solutions by running 200 independent post-replications (Eckman et al. 2023). To compare the stratified algorithms, we compare their performance for different number of strata.

Figure 2 plots various solver performance measures evaluated at intermediate budget points. The plot on the left shows the trajectory of the mean squared error (MSE). MSE considers bias and variance revealing the algorithm's robustness (Jain et al. 2022). The middle and right plots depict the trajectory of the mean and variance of the objective function value, respectively. Overall the concomitant variable-based post-stratification improves the algorithm's performance even when the concomitant variable is not highly correlated with the output (e.g., standardized mean interarrival time). Amongst the various concomitant variables explored, mean waiting time performs best as it is linearly correlated with the mean sojourn time. Using outputs (expected utilization and mean waiting time) initially shows slow convergence but achieves better solutions. Increasing the number of strata improves the performance of S-ASTRO-DF but also increases the computational cost of getting the optimal strata boundaries.

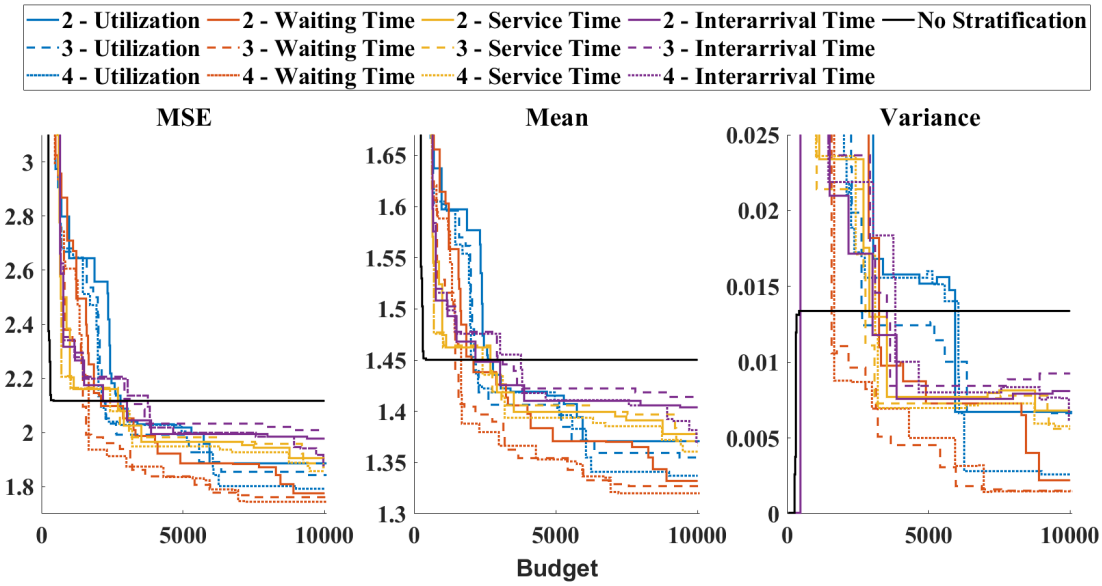
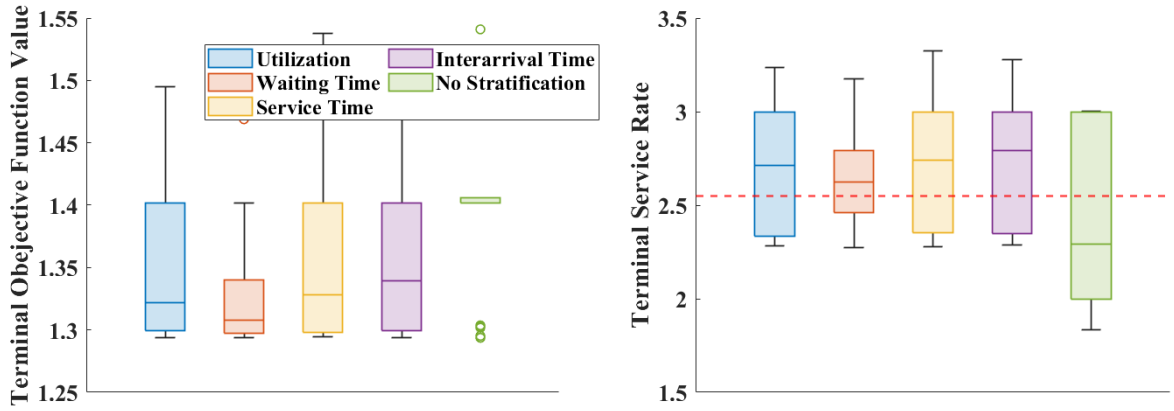


Figure 2: Comparison of the S-ASTRO-DF's performance with different choices for concomitant variables and number of strata. The x-axis represents the number of objective function evaluations.



(a) Terminal objective function value for the three cases.

(b) Terminal service rate for the three cases.

Figure 3: Comparing terminal solutions' distribution across configurations. The case with no stratification appears sensitive to varying configurations. The red line indicates the optimal service rate θ.

Next we compare the sensitivity of S-ASTRO-DF to the initial point θ_0. We consider three different starting points, $\theta_0 = \{3, 5, 10\}$ for $m = 4$. Figures 3(a) and 3(b), reveal that dynamic stratification results in more consistent results. ASTRO-DF with no stratification is sensitive to varying configurations. S-ASTRO-DF with the mean waiting time as X with approximate stratification structure shows same consistency as using optimal stratification structure with the standardized mean service time. The range of terminal values for S-ASTRO-DF with dynamic stratification is smaller.

5 CONCLUDING REMARKS

Stratified sampling is a well-known variance reduction technique for estimation. In optimization, however, rather than using fixed strata, changing the strata to best capture the objective function's local variability at each iteration can enhance robustness. We propose updating the stratification structure during optimization before performing a post-stratification. For updating the strata, we leverage information about other concomitant variables that are simulated, even if their distributions or properties are unknown. We integrate this dynamic post-stratification throughout optimization with a formerly established adaptive sampling strategy in a trust-region method for efficiency. Exploring the impact of different concomitant variables in a simple queuing model suggests that the effectiveness crucially depends on choosing the most linearly correlated variable with the output. In our experiments, we use a fixed number of strata during optimization. Choosing the best number of strata and using multiple concomitant variables is left for future research.

ACKNOWLEDGMENTS

This work was partially supported by National Science Foundation Grant CMMI-2226347.

REFERENCES

Bollapragada, R., R. Byrd, and J. Nocedal. 2018. "Adaptive Sampling Strategies for Stochastic Optimization". *SIAM Journal on Optimization* 28(4):3312–3343.

Bretthauer, K. M., A. Ross, and B. Shetty. 1999. "Nonlinear Integer Programming for Optimal Allocation in Stratified Sampling". *European Journal of Operational Research* 116(3):667–680.

Brito, J., N. Maculan, M. Lila, and F. Montenegro. 2010. "An Exact Algorithm for the Stratification Problem with Proportional Allocation". *Optimization Letters* 4:185–195.

Chaddha, R., W. Hardgrave, D. Hudson, M. Segal, and J. Suurballe. 1971. "Allocation of Total Sample Size when Only the Stratum Means are of Interest". *Technometrics* 13(4):817–831.

Chung, K. L. 2001. *A Course in Probability Theory*. San Diego: Academic Press.

Cochran, W. G. 1977. *Sampling Techniques*. New York: John Wiley & Sons.

Curtis, F. E., and K. Scheinberg. 2020. "Adaptive Stochastic Optimization: A Framework for Analyzing Stochastic Optimization Algorithms". *IEEE Signal Processing Magazine* 37(5):32–42.

Dalenius, T. 1950. "The Problem of Optimum Stratification". *Scandinavian Actuarial Journal* 1950(3-4):203–213.

Dalenius, T., and M. Gurney. 1951. "The Problem of Optimum Stratification. II". *Scandinavian Actuarial Journal* 1951(1-2):133–148.

Dalenius, T., and J. L. Hodges Jr. 1959. "Minimum Variance Stratification". *Journal of the American Statistical Association* 54(285):88–101.

de Moura Brito, J. A., G. S. Semaan, A. C. Fadel, and L. R. Brito. 2017. "An Optimization Approach Applied to the Optimal Stratification Problem". *Communications in Statistics-Simulation and Computation* 46(6):4419–4451.

Eckman, D. J., S. G. Henderson, and S. Shashaani. 2023. "Diagnostic Tools for Evaluating and Comparing Simulation-Optimization Algorithms". *INFORMS Journal on Computing* 35(2):350–367.

Ekman, G. 1959. "An Approximation Useful in Univariate Stratification". *The Annals of Mathematical Statistics* 30(1):219–229.

Espath, L., S. Krumscheid, R. Tempone, and P. Vilanova. 2021. "On the Equivalence of Different Adaptive Batch Size Selection Strategies for Stochastic Gradient Descent Methods". *arXiv preprint arXiv:2109.10933*. https://doi.org/10.48550/arXiv.2109.10933, accessed 23rd May 2022.

Etoré, P., and B. Jourdain. 2010. "Adaptive Optimal Allocation in Stratified Sampling Methods". *Methodology and Computing in Applied Probability* 12(3):335–360.

Farias, F., T. Ludermir, and C. Bastos-Filho. 2020. "Similarity Based Stratified Splitting: An Approach to Train Better Classifiers". *arXiv preprint arXiv:2010.06099*. https://doi.org/10.48550/arXiv.2010.06099, accessed 3^{rd} May 2022.

Glynn, P. W., and D. L. Iglehart. 1989. "Importance Sampling for Stochastic Simulations". *Management Science* 35(11):1367–1392.

Glynn, P. W., and Z. Zheng. 2021. "Efficient Computation for Stratified Splitting". In *Proceedings of the 2021 Winter Simulation Conference*, edited by S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, 1–8. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ha, Y., and S. Shashaani. 2023. "Iteration Complexity and Finite-Time Efficiency of Adaptive Sampling Trust-Region Methods for Stochastic Derivative-Free Optimization". *arXiv preprint arXiv:2305.10650*. https://doi.org/10.48550/arXiv.2305.10650, accessed 18th May 2023.

Huddleston, H., P. Claypool, and R. Hocking. 1970. "Optimal Sample Allocation to Strata using Convex Programming". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 19(3):273–278.

Jain, P., S. Shashaani, and E. Byon. 2021. "Wake Effect Calibration in Wind Power Systems with Adaptive Sampling based Optimization". In *Proceedings of the IISE Annual Conference*, edited by A. Ghate, K. Krishnaiyer, and K. Paynabar, 43–48. Red Hook, NY: Institute of Industrial and Systems Engineers.

Jain, P., S. Shashaani, and E. Byon. 2022. "Robust Simulation Optimization with Stratification". In *Proceedings of the 2022 Winter Simulation Conference*, edited by B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C. Corlu, L. Lee, E. Chew, T. Roeder, and P. Lendermann. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kawai, R. 2010. "Asymptotically Optimal Allocation of Stratified Sampling with Adaptive Variance Reduction by Strata". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 20(2):1–17.

Khan, M. G. M., N. Nand, and N. Ahmad. 2008. "Determining the Optimum Strata Boundary Points using Dynamic Programming". *Survey Methodology* 34(2):205–214.

Kim, S., R. Pasupathy, and S. G. Henderson. 2015. "A Guide to Sample Average Approximation". In *Handbook of Simulation Optimization*, edited by M. C. Fu, 207–243. New York: Springer.

Lavenberg, S. S., T. L. Moeller, and P. D. Welch. 1982. "Statistical Results on Control Variables with Application to Queuing Network Simulation". *Operations Research* 30(1):182–202.

Liu, B., X. Yue, E. Byon, and R. A. Kontar. 2022. "Parameter Calibration in Wake Effect Simulation Model with Stochastic Gradient Descent and Stratified Sampling". *The Annals of Applied Statistics* 16(3):1795–1821.

Mulvey, J. M. 1983. "Multivariate Stratified Sampling by Optimization". *Management Science* 29(6):715–724.

Neyman, J. 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection". *Journal of the Royal Statistical Society* 97(4):558–625.

Pettersson, P., and S. Krumscheid. 2021. "Adaptive Stratified Sampling for Non-Smooth Problems". *arXiv preprint arXiv:2107.01355*. https://doi.org/10.48550/arXiv.2107.01355, accessed 4th April 2022.

Ross, S. 2013. "Chapter 9 - Variance Reduction Techniques". In *Simulation* (Fifth ed.)., edited by S. Ross, 153 – 231. New York: Academic Press.

Ross, S. M., and K. Y. Lin. 2001. "Applying Variance Reduction Ideas in Queuing Simulations". *Probability in the Engineering and Informational Sciences* 15(4):481–494.

Sabuncuoglu, I., M. M. Fadiloglu, and S. Çelik. 2008. "Variance Reduction Techniques: Experimental Comparison and Analysis for Single Systems". *IIE Transactions* 40(5):538–551.

Sethi, V. 1963. "A Note on Optimum Stratification of Populations for Estimating the Population Means". *Australian Journal of Statistics* 5(1):20–33.

Shashaani, S., F. S. Hashemi, and R. Pasupathy. 2018. "ASTRO-DF: A Class of Adaptive Sampling Trust-Region Algorithms for Derivative-Free Stochastic Optimization". *SIAM Journal on Optimization* 28(4):3145–3176.

Singh, R., and B. Sukhatme. 1969. "Optimum stratification". *Annals of the Institute of Statistical Mathematics* 21:515–528.

Taga, Y. 1967. "On Optimum Stratification for the Objective Variable based on Concomitant Variables using Prior Information". *Annals of the Institute of Statistical Mathematics* 19(1):101–129.

Tipton, E. 2013. "Stratified Sampling using Cluster Analysis: A Sample Selection Strategy for Improved Generalizations from Experiments". *Evaluation Review* 37(2):109–139.

Tipton, E., L. Hedges, M. Vaden-Kiernan, G. Borman, K. Sullivan, and S. Caverly. 2014. "Sample Selection in Randomized Experiments: A New Method using Propensity Score Stratified Sampling". *Journal of Research on Educational Effectiveness* 7(1):114–135.

Wilson, J. R. 1979. *Variance Reduction Techniques for the Simulation of Queueing Networks*. Ph. D. thesis, Purdue University, West Lafayette, Indiana. https://www.proquest.com/docview/302964264?pq-origsite=gscholar&fromopenview=true, accessed 12th March 2023.

Wilson, J. R., and A. A. B. Pritsker. 1984. "Variance Reduction in Queueing Simulation using Generalized Concomitant Variables". *Journal of Statistical Computation and Simulation* 19(2):129–153.

Zhao, P., and T. Zhang. 2014. "Accelerating Minibatch Stochastic Gradient Descent using Stratified Sampling". *arXiv preprint arXiv:1405.3080*. https://doi.org/10.48550/arXiv.1405.3080, accessed 12th March 2022.

AUTHOR BIOGRAPHIES

PRANAV JAIN is a fourth-year Ph.D. student in the Edward P. Fitts Department of Industrial and Systems Engineering at North Carolina State University. His main research is in data-driven simulation optimization. His email address is pjain23@ncsu.edu.

SARA SHASHAANI is an Assistant Professor in the Edward P. Fitts Department of Industrial and System Engineering at North Carolina State University. Her research interests are probabilistic data-driven models and simulation optimization. She is a co-creator of SimOpt. Her email address is sshasha2@ncsu.edu and her homepage is https://shashaani.wordpress.ncsu.edu/.