Screening Simulated Systems for Optimization

Jinbo Zhao [1] Javier Gatica [2] David J. Eckman [1]

[1]Texas A&M University, ISEN

[2]Pontificia Universidad Católica de Chile
Instituto de Ingeniería Matemática y Computacional

Winter Simulation Conference
December 11, 2023

This work is supported by the National Science Foundation under grant CMMI-2206972.

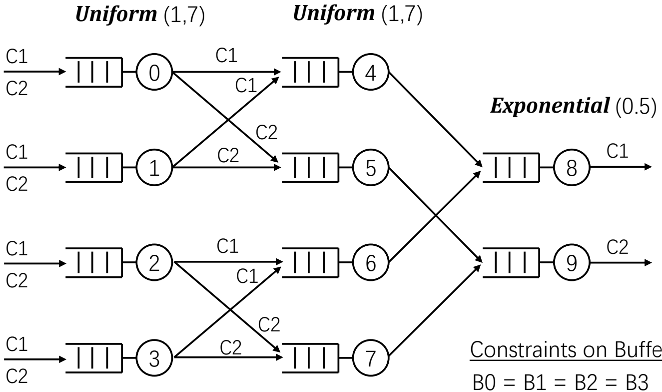
Ranking and Selection (R&S)

Considering a **finite** number of simulated systems for some decision.

Interested in systems with good (high) expected performance.

Expected performance can only be observed with error via simulation.

Case: Buffer Allocation Problem



C1: ***Uniform*** (2,18)
C2: ***Exponential*** (0.12)

<u>Constraints on Buffer size</u>

B0 = B1 = B2 = B3
B4 = B6
B5 = B7
B0 + B1 + ⋯ +B9 = 20

1001 systems under consideration with constraints.
10,015,005 systems under consideration without constraints.

Screening (a.k.a. Subset Selection)

Typical R&S approach:

1. Treat systems categorically.
   - Relationship between expected performances and decision variables are ignored.
2. Simulate all systems.
3. Perform statistical comparisons.
4. Return a subset of systems with some guarantee.

Exception to #1 and #2: Plausible Screening (PS)

- PS simulates a subset of all systems and exploits structure in the decision space to screen out even unsimulated systems.

Motivation for Screening

For problems with large numbers of systems, most systems have poor performance.

Goal: Eliminate inferior systems without expending significant simulation effort.

Decision-making contexts:

1. Decision ultimately based on secondary (possibly qualitative) considerations.
2. A one-time decision, in preparation for running an expensive selection procedure which recommends a single system.
3. For adopting different solutions in different scenarios.
   - Decisions may be made over time depending on the changing situation.
   - Screening offers an assortment of high-quality systems from which to choose.
4. Decision making with multiple objectives (not in this tutorial).

One-shot vs. Adaptive Screening

One-shot: Sample sizes for each system are fixed in advance.

- Total sampling budget is controlled.

Adaptive: Sequentially sample systems based on observed statistics.

- Total sampling budget is sometimes controlled.
- Avoid excessively sampling clearly inferior systems.
- Expect to return a smaller subset given the same budget as a one-shot procedure.

This Tutorial

A collection of ideas (some old, some new), not a comprehensive review of screening procedures. Bring a fresh perspective to an old problem.

Today's agenda:

- Section 2 – Problem setting, frequentist guarantees and their relative merits.
- Section 3 – Three screening procedures and numerical experiments.
- Section 4 – Screening procedures in parallel computing environments.
- Section 5 – Sample efficiency of procedures that combine screening and selection.

Section 2: Screening and Screening Guarantees

(Setup, An Impossible Triangle, 4 Frequentist Guarantees)

Setup

System i has expected performance μ_i for $i = 1, 2, \ldots, k$.

- Problem instance: $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)$.

System i receives n_i samples. Total budget: $N = \sum_{i=1}^{k} n_i$.

Normality Assumption

For each System i, $i = 1, 2, ..., k$, the outputs $X_{i1}, X_{i2}, ..., X_{in_i}$ are independent and normally distributed with unknown mean μ_i and unknown variance σ_i^2.

- Useful for delivering finite-sample guarantees.
- Common for R&S: apply batching and appeal to Central Limit Theorem (CLT).

Setup

Systems whose expected performances are within $\delta \geq 0$ of the best are "good."

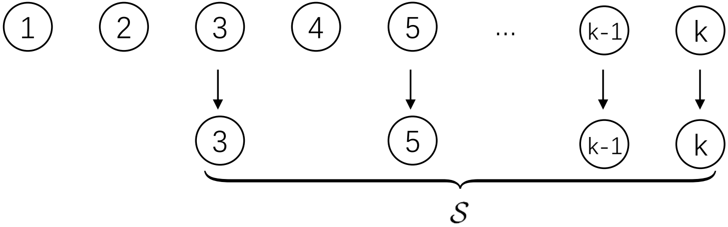
- $\delta = 0$ corresponds to **correct** selection.

$\mathcal{G} = \{i : \mu_i \geq \mu^* - \delta\}$: δ-optimal systems, i.e., good systems.

- We may have multiple good systems, even when $\delta = 0$.

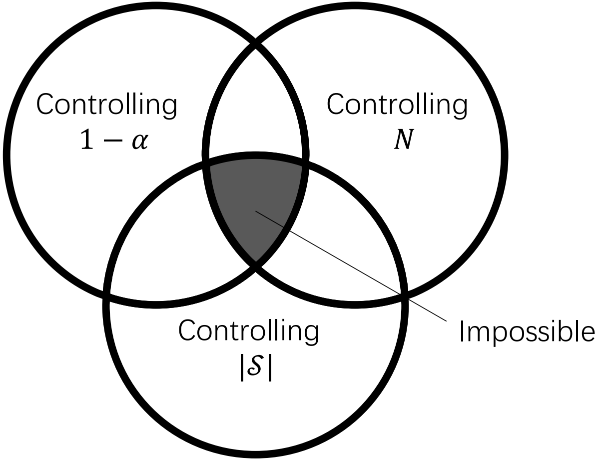
$\mathcal{S} \subseteq \{1, 2, ..., k\}$: returned subset.

- Based on sample means and sample variances which are sufficient statistics.



An Impossible Triangle for Screening Procedures

A procedure can control at most two of the three ($1-\alpha$, $|\mathcal{S}|$, N). The stricter the demands, the more the uncontrolled aspect suffers.



An Impossible Triangle for Screening Procedures

Controlling α and $|\mathcal{S}| = 1$.

- Selection procedures (Kim and Nelson, 2001).

Controlling α and $|\mathcal{S}| = m > 1$.

- Restricted subset selection (Koeing and Law, 1985; Sullivan and Wilson, 1989).

Controlling N and $|\mathcal{S}| = 1$.

- Fixed-budget knockout-tournament (FBKT) (Hong et al., 2022).

Controlling N and $|\mathcal{S}| = m > 1$.

- Easier when adopting the Bayesian perspective (Eckman et al., 2020).

Controlling α and N.

- Classical subset selection (Gupta, 1965; Nelson et al., 2001). $\Leftarrow$ **This talk.**

Fixed-Confidence Frequentist Guarantees

Frequentist perspective: $\boldsymbol{\mu}$ and $\mathcal{G}$ are regarded as fixed; $\mathcal{S}$ is random.

Set-wise Probability of Good Selection (Set-wise PGS) guarantee

For any $\boldsymbol{\mu} \in \mathbb{R}^k$, $\mathbb{P}(\mathcal{G} \subseteq \mathcal{S}) \geq 1 - \alpha$.

System-wise Probability of Good Selection (System-wise PGS) guarantee

For any $\boldsymbol{\mu} \in \mathbb{R}^k$, $\mathbb{P}(i \in \mathcal{S}) \geq 1 - \alpha$ for all $i \in \mathcal{G}$.

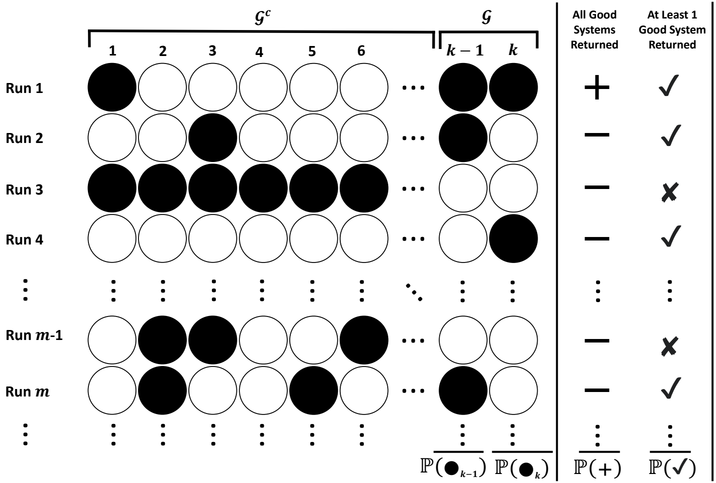
Expected False Elimination Rate (EFER) guarantee

For any $\boldsymbol{\mu} \in \mathbb{R}^k$, $\mathbb{E}[|\mathcal{S} \cap \mathcal{G}|]/|\mathcal{G}| \geq 1 - \alpha$.

Probability of Good Inclusion (PGI) guarantee

For any $\boldsymbol{\mu} \in \mathbb{R}^k$, $\mathbb{P}(|\mathcal{S} \cap \mathcal{G}| \geq 1) \geq 1 - \alpha$.

Fixed-Confidence Frequentist Guarantees



Set-wise PGS: $\mathbb{P}(+) \geq 1 - \alpha$

System-wise: $\min_{i \in \mathcal{G}} \mathbb{P}(\bullet_i) \geq 1 - \alpha$

EFER: $\frac{\sum_{i \in \mathcal{G}} \mathbb{P}(\bullet_i)}{|\mathcal{G}|} \geq 1 - \alpha$

PGI: $\mathbb{P}(\checkmark) \geq 1 - \alpha$

● **Returned System**
○ **Eliminated System**

Figure: Illustration of four fixed-confidence frequentist guarantees for screening on a problem instance with k systems, two of which are good ($k-1$ and k). $\mathbb{P}(\bullet_i)$ denotes the probability that System i is returned.

Fixed-Confidence Frequentist Guarantees

All four guarantees are equivalent when $|\mathcal{G}| = 1$.

Proposition

For any fixed confidence level $1 - \alpha \in (0, 1)$,

$$\text{Set-wise PGS} \quad \Rightarrow \quad \text{System-wise PGS} \quad \Rightarrow \quad \text{EFER} \quad \Rightarrow \quad \text{PGI.}$$

Other Guarantees

Controlling the Type II error:

- For any $\boldsymbol{\mu} \in \mathbb{R}^k$, $\mathbb{P}(\mathcal{S} \subseteq \mathcal{G}) \geq 1 - \alpha$ (Desu, 1970).
- For any $\boldsymbol{\mu} \in \mathbb{R}^k$, $\mathbb{P}(i \notin \mathcal{S}) \geq 1 - \alpha$, for all $i \notin \mathcal{G}$.

Controlling the average optimality gap (a.k.a. the expected opportunity cost) (Gao and Chen, 2015a).

Other definitions of "good":

- Top-*m* systems (Gao and Chen, 2015b).
- Expected performance $\geq \min\{\mu^*, \mu^\dagger\}$ (Pei et al., 2022).

Section 3: Procedures and Their Guarantees

(Procedure delivering the guarantees, Experiments showing the **Screening Power)**

Procedures Delivering Different Guarantees

The PGI guarantee can be obtained by prematurely terminating any sequential selection procedures that delivers the PGS guarantee (Ma and Henderson, 2017).

- Applies to indifference-zone-free procedures for $\delta = 0$ case (Fan et al., 2016; Wang et al., 2023).
- Can terminate based on wall-clock time, CPU time or total sample size.
- Motivates further study of the **rate** at which procedures eliminate systems.

We will focus on three procedures designed for different guarantees:

- System-wise PGS: Screen-to-the-Best (STTB).
- Set-wise PGS: Decoupled STTB (DSTTB).
- EFER: Bisection Parallel Adaptive Survivor Selection (bi-PASS).

Screen-to-the-Best (STTB)

STTB is based on **pairwise comparisons**:

$$\mathcal{S}^{\mathrm{STTB}} = \left\{ i : \widehat{\mu}_i + \delta \geq \widehat{\mu}_j - t_{\beta,\nu}\sqrt{\frac{\widehat{\sigma}_i^2}{n} + \frac{\widehat{\sigma}_j^2}{n}} \text{ for all } j \neq i \right\}$$

- $t_{\beta,\nu}$ is the β quantile of the Student's t-distribution with ν degrees of freedom.
- For $\beta = (1-\alpha)^{1/(k-1)}$ and $\nu = n-1$, STTB delivers the system-wise PGS guarantee.

An extended version allows unequal sample sizes (Boesel et al., 2003), but we let $n_1 = n_2 = \cdots = n_k = n$.

Implemented in Simio and Arena.

Transitive Elimination

Transitive Elimination

For any Systems i, j, and ℓ, if System i eliminates System j, and System j eliminates System ℓ, then System i eliminates System ℓ.

- Allows screening procedures to be implemented in parallel using a **divide-and-conquer** scheme.

$$\mathcal{S}^{\mathrm{STTB}} = \left\{ i \colon \widehat{\mu}_i + \delta \geq \widehat{\mu}_j - t_{\beta,\nu}\sqrt{\frac{\widehat{\sigma}_i^2}{n} + \frac{\widehat{\sigma}_j^2}{n}} \text{ for all } j \neq i \right\}$$

- This formula for STTB is different from the original procedure, which was designed for PCS under an IZ formulation.
- The original procedure claims to have the transitive-elimination property when $\delta > 0$, but there are known counter-examples.

Decoupled STTB (DSTTB)

$$\mathcal{S}^{\mathrm{DSTTB}} = \left\{ i \colon \widehat{\mu}_i + t_{\beta,\nu}\sqrt{\frac{\widehat{\sigma}_i^2}{n}} + \delta \geq \widehat{\mu}_j - t_{\beta,\nu}\sqrt{\frac{\widehat{\sigma}_j^2}{n}} \text{ for all } j \neq i \right\}.$$

Delivers the system-wise PGS guarantee when $\beta = (1-\alpha)^{1/k}$ and $\nu = n-1$.
Delivers the set-wise PGS guarantee when $\beta = (1+(1-\alpha)^{1/k})/2$ and $\nu = n-1$.

The time complexity of DSTTB is $O(k)$ vs $O(k^2)$ for STTB.

We will compare the screening power of STTB and DSTTB a little later.

DSTTB for Set-wise PGS

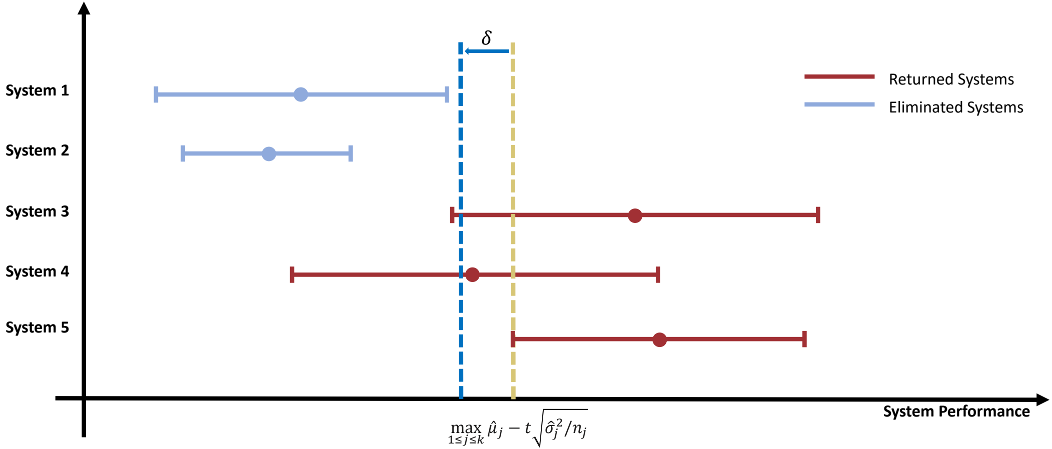


Figure: Set-wise PGS guarantee

DSTTB for System-wise PGS

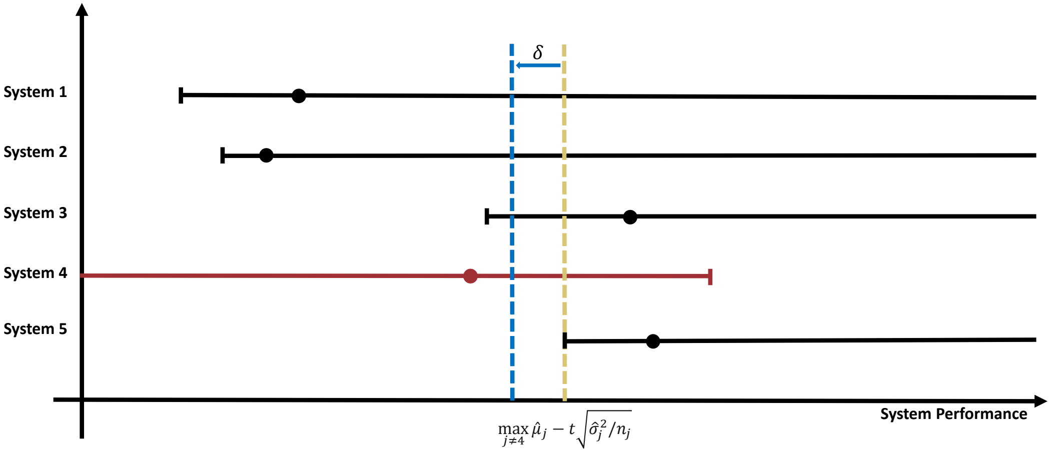


Figure: System-wise PGS guarantee (Screening System 4)

Bisection Parallel Adaptive Survivor Selection (bi-PASS)

bi-PASS sequentially samples surviving systems and eliminates systems over time.

A system is eliminated if its estimated expected performance is sufficiently lower than an adaptive standard $\hat{\mu}^*$.

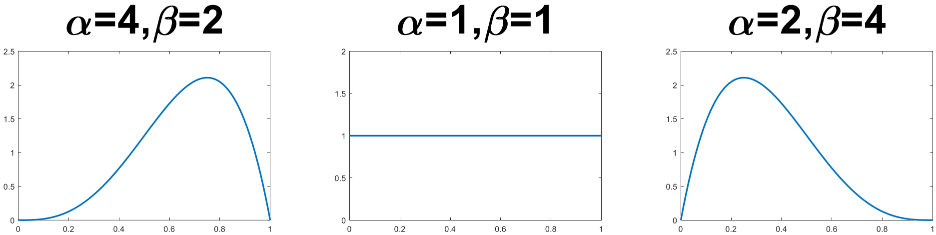
- A boundary function $g(n_i/\widehat{\sigma}_i)$ controls the allowed difference between $\hat{\mu}^*$ and $\hat{\mu}_i$.
- The adaptive standard is designed to converge to the desired standard as the sample size increases.

At any time, the set of surviving systems delivers the EFER guarantee for $\delta = 0$.

Stopping criteria can be flexible.

Comparing STTB and DSTTB

Problem instances with expected performances described by $\mu_i = F^{-1}(i/k)$ for $i = 1, 2, \ldots, k$, where F^{-1} is the inverse cdf of a certain Beta distribution.



- 40 macro replications, $k = 1000$, $\delta = 0$ and $1 - \alpha = 0.95$.
- Common, unknown variance $\sigma^2 = 0.5$.

Compare STTB, DSTTB set-wise and DSTTB system-wise.

Comparing STTB and DSTTB

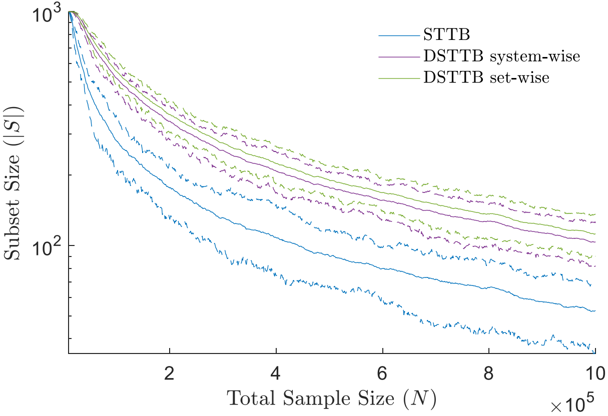
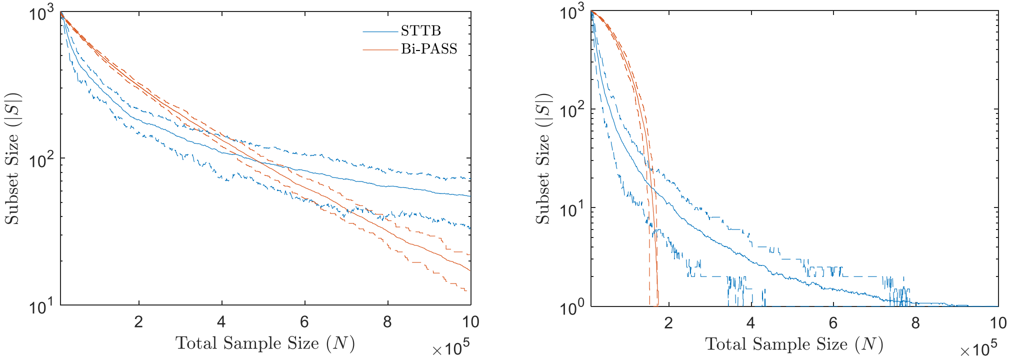


Figure: Mean and 10th/90th percentiles (dashed) of returned subset size for STTB and DSTTB on a problem instance with 1000 systems with expected performances from the Beta($a = 4$, $b = 2$) distribution.

Comparing STTB and bi-PASS

bi-PASS performs better on instances where most systems are clearly inferior.



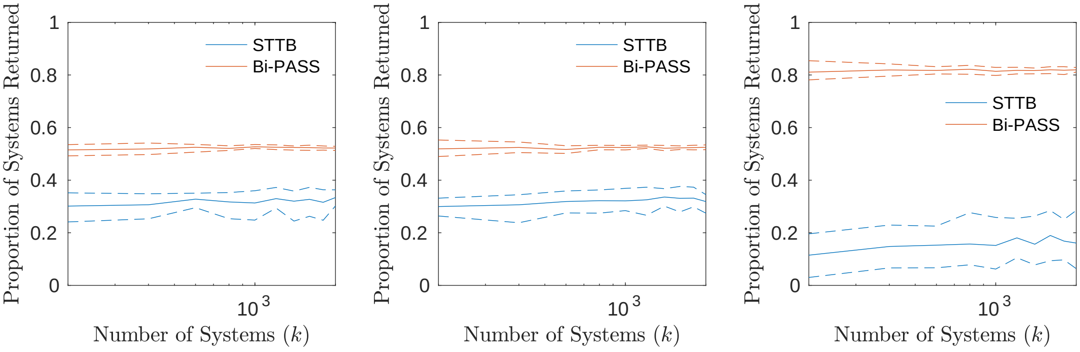
(a) Beta($a = 4$, $b = 2$)

(b) Beta($a = 2$, $b = 4$)

Figure: Mean and 10th/90th percentiles (dashed) of subset size for STTB and bi-PASS on two fixed problem instances when varying the total sample size.

Comparing STTB and bi-PASS

Let budget grow with the number of systems: $N = 40k$.



(a) Beta($a = 4$, $b = 2$) (b) Uniform(0, 1) (c) Beta($a = 2$, $b = 4$)

Figure: Mean and 10th/90th percentiles (dashed) of the proportion of systems returned by STTB and bi-PASS when run with a total sample size of $N = 40k$ on three problem instances.

Section 4: Screening in Parallel Computing Environments

(Pairing DSTTB and STTB for efficient screening)

Motivation for Parallel Computing

Parallel computing is an accessible tool for large R&S problems, i.e., $k > 10{,}000$.

Can parallelize existing selection procedures or develop new schemes for comparing systems in parallel environments (Hunter and Nelson, 2017).

Issues that can undermine statistically validity and efficiency:

- **Sequences of the inputs and outputs are different on multiple processors.**
- Observations with shorter simulation time are available earlier $\Rightarrow$ the sequence of output observations may not be independent.
- In the master-worker framework, the master processor can become a bottleneck if overwhelmed by communication tasks.

Divide-and-Conquer Scheme

Certain tasks are performed on worker processors before being performed again on master processor with combined results (Ni et al., 2017).

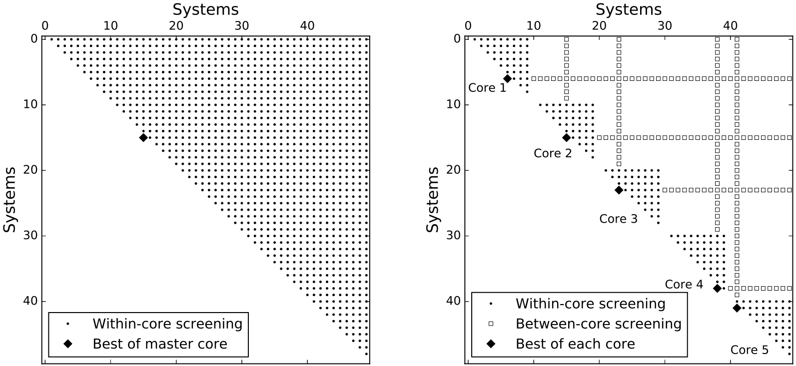


Figure: Each dot represents a pair of systems to be screened. In the left panel, the master performs all-pairs screening. In the right panel, each worker core receives 10 systems, performs all pairs screening among those 10 systems, and also screens against "killer" systems, i.e., those having the highest sample mean on other workers.

PASS Scheme

Full pairwise comparisons necessitate extensive communication across systems.

In PASS, the master processor maintains and updates an adaptive standard based on results returned by workers.

Results from individual systems are compared to the standard.

Decoupling Systems

Zhong et al. (2022) lower the time complexity of KN procedure and Paulson's procedure from $O(k^2 \log k)$ to $O(k^2)$ and $O(k \log k)$ to $O(k)$ respectively, by replacing the term $\hat{\sigma}_{ij}^2$ with $\hat{\sigma}_i^2 + \hat{\sigma}_j^2$. Makes the procedure more conservative.

By applying a similar decoupling on STTB, DSTTB lowers the time complexity from $O(k^2)$ to $O(k)$.

$$\mathcal{S}^{\mathrm{STTB}} = \left\{ i : \widehat{\mu}_i + \delta \geq \widehat{\mu}_j - t_{\beta,\nu}\sqrt{\frac{\widehat{\sigma}_i^2}{n} + \frac{\widehat{\sigma}_j^2}{n}} \text{ for all } j \neq i \right\}$$

versus

$$\mathcal{S}^{\mathrm{DSTTB}} = \left\{ i : \widehat{\mu}_i + t_{\beta,\nu}\sqrt{\frac{\widehat{\sigma}_i^2}{n}} + \delta \geq \widehat{\mu}_j - t_{\beta,\nu}\sqrt{\frac{\widehat{\sigma}_j^2}{n}} \text{ for all } j \neq i \right\}.$$

Prescreening

Prescreening

Initially screen all systems, followed by a second round of screening *without additional samples*.

Can lower the overall running time without compromising screening power.

Proposition

For any $\boldsymbol{\mu} \in \mathbb{R}^k$, $\alpha \in (0, 1)$ and $\delta \geq 0$, when DSSTB is set up to deliver the system-wise PGS guarantee, $\mathbb{P}(\mathcal{S}^{\mathrm{STTB}} \subseteq \mathcal{S}^{\mathrm{DSTTB}}) = 1$.

Prescreening Experiments

Compare the effects of prescreening for different combination of STTB and DSSTB.

All combinations conclude with STTB $\Rightarrow$ ultimately return the same subset.

- Instance with expected performances following Beta($\alpha = 2, \beta = 4$).
- $k = 100{,}000$ and $n_i = 20$ for $i = 1, 2, \ldots, k$.
- Time for simulation is not counted.
- The experiments were run on a single processor, so communication costs between different processors are not counted.

Prescreening Experiments

Table: Mean and 10th/90th percentiles of wall-clock time (in seconds) spent screening by combinations of the STTB and DSTTB procedures given different numbers of processors (p) on a problem with $k = 100{,}000$ systems. DSTTB + STTB indicates that STTB is used to screen the systems returned by DSTTB.

On Workers	On Master	$p = 10$	$p = 20$	$p = 50$	$p = 100$
–	STTB	38.32 [35.71, 39.96]			
–	DSTTB + STTB	**19.62** [18.07, 21.21]			
STTB	STTB	**2.55** [2.16, 2.86]	**2.64** [2.57, 2.74]	**2.65** [2.56, 2.73]	**2.84** [2.73, 3.02]
DSTTB	STTB	3.97 [3.81, 4.19]	3.99 [3.84, 4.11]	4.14 [4.01, 4.25]	4.20 [4.12, 4.30]
DSTTB	DSTTB + STTB	5.02 [4.44, 5.55]	4.75 [4.09, 5.28]	4.97 [4.74, 5.35]	4.76 [4.39, 5.22]
DSTTB + STTB	DSTTB + STTB	3.26 [3.07, 3.44]	3.14 [2.94, 3.29]	3.24 [3.14, 3.35]	3.50 [3.31, 3.65]

Section 5: Pairing Screening and Selection Procedures

(How to split the guarantees, Experiments to find the balance of sample size and parameters between two stages)

Motivation for Pairing Screening and Selection Procedures

Goal: To deliver the $1-\alpha$ PGS guarantee with $\delta \geq 0$.

Approach: Run a screening procedure to eliminate non-competitive systems followed by a more sampling-intensive selection procedure on the survivors.

Screening + Selection has potential to reduce the *total sample size* required to deliver a fixed-confidence guarantee.

Example: NSGS procedure of Nelson et al. (2001), which pairs STTB and Rinott's procedure (Rinott, 1978).

- Rinott's procedure has two stages. Each system's second-stage sample size depends on the variance estimated from the first stage.

No Data Reuse

Screening and selection procedures are assumed to be run independently.

- **Advantage:** Selection procedure be run as if only the surviving systems existed.
- **Disadvantage:** Need to **abandon** the data obtained during screening when running the selection procedure.

Splitting α and δ:

- The desired confidence level $1-\alpha$ will be split, i.e., $1-\alpha=(1-\alpha_1)(1-\alpha_2)$.
- The desired tolerance δ will be split, i.e., $\delta=\delta_1+\delta_2$.

The 1st-stage screening procedure should deliver the $1-\alpha_1$ **PGI guarantee** with δ_1.
The 2nd-stage selection procedure should deliver the $1-\alpha_2$ **PGS guarantee** with δ_2.

Stage 1 (Screening)	Stage 2 (Selection)
$\alpha_1,\ \delta_1$	$\alpha_2,\ \delta_2$

Data Reuse

Reusing data from screening procedure (Stage 1) would seem to improve the efficiency of a combined procedure.

- Surviving systems are more likely to have higher estimated expected performance than their true expected performance, which might break the frequentist guarantee.

An example of data reuse is Nelson et al. (2001):

- STTB ($\delta_1 = 0$ and $1 - \alpha_1$) + Rinott's procedure ($\delta_2 = \delta$ and $1 - \alpha_2$), where $1 - \alpha = 1 - \alpha_1 - \alpha_2$.
- Requiring Rinott's constant to be based on the original number of system, resulting in a larger sample size for second stage.

Wilson (2001) shows that for NSGS procedure, a less conservative decomposition $1 - \alpha = (1 - \alpha_1)(1 - \alpha_2)$ can be achieved, as the best system surviving the first stage and the event of it being selected are positively correlated.

- Presents an opportunity for reusing data without issue for other combinations.

Numerical Experiments

How to choose the first stage sample size n_0 to reduce the total sample size?

Higher n_0 leads to smaller returned subset from Screening procedure, thus reducing the sample size needed by second stage.

Setup:

- $k = 100$ systems.
- $\mu_i = i,$ for $i = 1, 2, \ldots, k$; all systems share a common σ^2.
- STTB + Rinott's procedure, without reusing data.
- $\alpha_1 = \alpha_2 = 1 - \sqrt{0.95}$.
- $\delta_1 = 0$ and $\delta_2 = \delta = 5.5$, thus 5 out 100 systems are good.

Numerical Experiments

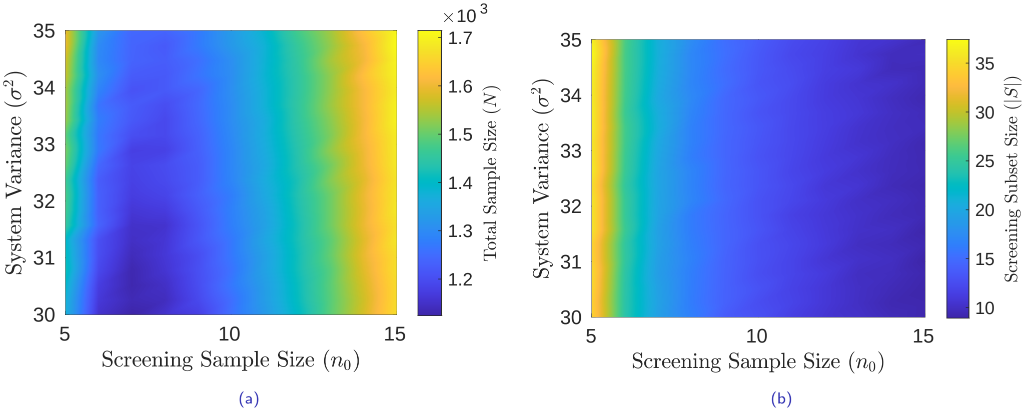


Figure: Heat maps of (a) the average total sample size and (b) the average size of the subset returned by STTB with respect to the sampling variance σ^2 and the initial sample size n_0 for a combined procedure of STTB (with $\delta = 0$) paired with Rinott's selection procedure on a problem instance with $k = 100$ systems.

Numerical Experiments

How to split the α and δ to reduce the total sample size?

Lower δ_1 and α_1 lead to a larger returned subset, meanwhile lower δ_2 and α_2 lead to a larger sample size for second stage.

Setup:

- $k = 100$ systems.
- $n_0 = 10$.
- $\mu_i = i,$ for $i = 1, 2, \ldots, k$; all systems share a common $\sigma^2 = 30$.
- STTB + Rinott's procedure, without reusing data.

Numerical Experiments

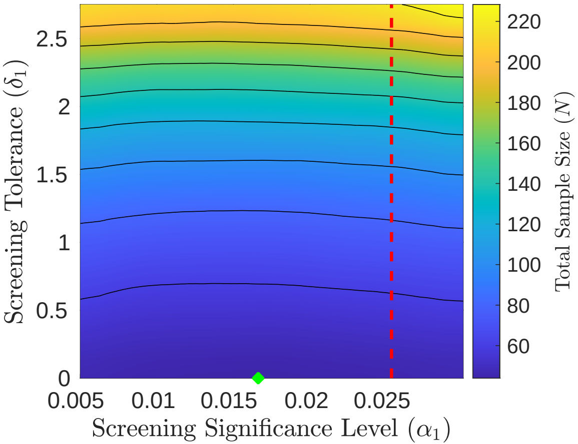


Figure: Heat map of the average total sample size for a combined procedure of STTB (with δ_1 and $1-\alpha_1$) and Rinott's selection procedure (with δ_2 and $1-\alpha_2$) on a problem instance with $k = 100$ systems. The red line represents evenly splitting α_1 and α_2; the green diamond indicates the optimal α_1 when $\delta_1 = 0$.

Conclusion

Takeaway

Highlighted a few screening procedures, discussing their differences in designs, guarantees, and empirical performance under different regimes.

Future Research

- Other procedures that sample fully sequentially in parallel computing.
- New screening frameworks include Bayesian and bootstrapping methods.
- Studying the rate at which sequential screening procedures eliminate non-competitive systems.

Bisection Parallel Adaptive Survivor Selection (bi-PASS)

Algorithm 1: Bisection Parallel Adaptive Survivor Selection (bi-PASS)

$\mathcal{S} = \{1, 2, \ldots, k\}$
for $i \in \mathcal{S}$ **do**
  $n_i = n_0$ and simulate each system n_0 times
Compute the initial estimated standard $\widehat{\mu}^* = |\mathcal{S}|^{-1} \sum_{i \in \mathcal{S}} \widehat{\mu}_i$ and total sample size $N = \sum_{i \in \mathcal{S}} n_i$
while $|\mathcal{S}| > 1 \text{ and } N < N_{max}$ **do**
  for $i \in \mathcal{S}$ **do**
    if $(\widehat{\mu}_i - \widehat{\mu}^*) n_i / \widehat{\sigma}_i^2 \leq -g(n_i / \widehat{\sigma}_i^2)$ **then**
      $\mathcal{S} = \mathcal{S} \setminus \{i\}$ and $\widehat{\mu}^* = |\mathcal{S}|^{-1} \sum_{i \in \mathcal{S}} \widehat{\mu}_i$
    else
      Simulate System i and update n_i, $\widehat{\mu}_i$, $\widehat{\sigma}_i^2$, N, and $\widehat{\mu}^*$
return $\mathcal{S}$