

Subdata Selection: A Unified Framework for Optimal Selection and Statistical Efficiency Assessment

Min Yang¹ Wei Zheng² John Stufken³ Ming-Chung Chang⁴ Ting Tian⁵
Xueqin Wang⁶

¹University of Illinois at Chicago ²University of Tennessee, Knoxville

³George Mason University ⁴Academia Sinica

⁵Sun Yat-sen University ⁶University of Science and Technology of China

NeurIPS 2026

Why select subdata?

- **Full data:** N observations $(y_i, \mathbf{x}_i)$ with a $p \times 1$ feature vector $\mathbf{x}_i$.
Subdata: n of them, with $n \ll N$.
- Two settings where we must work with subdata:
 - N is too large for the available computing budget;
 - labels y_i are expensive (active learning): choose which $\mathbf{x}_i$ to label.
- Selection uses only the features $\mathbf{x}_i$. Labels are not needed at the selection stage.
- **Goal:** keep the n points that retain the most information about the model parameters θ .
- Other approaches exist (coresets, support points). We focus on the **information-based** approach for parametric models.

The information-based formulation

- Parametric model with Fisher information $I_{\theta}(\mathbf{x}_i)$ for point i . With indicators $\delta_i \in \{0, 1\}$, $\sum_i \delta_i = n$, the subdata carries

$$I_{\theta}(\delta) = \sum_{i=1}^N \delta_i I_{\theta}(\mathbf{x}_i).$$

- $\text{Var}(\hat{\theta}) \propto I_{\theta}^{-1}$, so choose a convex criterion Φ and solve

$$\delta^{\text{opt}} = \underset{\delta}{\text{argmin}} \Phi(I_{\theta}(\delta)).$$

- With $\binom{N}{n}$ candidate subsets, exact search is hopeless even for modest N and n .

Criteria used

- D : $\Phi = \log \det I_{\theta}^{-1}$, the volume of the confidence ellipsoid
- A : $\Phi = \text{Tr } I_{\theta}^{-1}$, the total variance $E\|\hat{\theta} - \theta\|^2$
- Same for a submatrix when only some parameters matter; results extend to Φ_p criteria

Theorem 1

D -optimal exact subdata selection is **NP-hard**.

Two open questions

Q1. Assessment

How efficient is a given subdata S' relative to the (unknown) optimal subdata S_{opt} ?
Until now, methods could only be compared with each other, not with the optimum.

Q2. Selection

Which n points should be selected to reach high efficiency, for a general parametric model, arbitrary N and n , and several optimality criteria?

This work answers both questions within one framework
built on the theory of optimal approximate designs.

Key idea: relax to bounded approximate designs

- An *approximate design* (Kiefer, 1974) puts weights on the data points:
 $\xi = \{(\mathbf{x}_i, w_i)\}_{i=1}^N$, $w_i \geq 0$, $\sum_i w_i = 1$, with $\Phi(\xi) = \Phi(\sum_i w_i I_{\theta}(\mathbf{x}_i))$.
- Subdata S of size n is the design with $w_i = 1/n$ on S and $w_i = 0$ elsewhere.
- **Bounded relaxation:** cap every weight at $1/n$, so each point is used at most once:

$$\xi^* = \underset{\xi}{\operatorname{argmin}} \Phi(\xi) \quad \text{subject to} \quad 0 \leq w_i \leq \frac{1}{n}, \quad \sum_{i=1}^N w_i = 1.$$

A continuous convex problem instead of a search over $\binom{N}{n}$ subsets.

- **Rounding:** ξ^* has at least n points with positive weight, and nearly n of them sit at $1/n$. The subdata S^* consists of the n points with the largest weights.

$\mathcal{X}_1$: $w_i = 0$
not selected

$\mathcal{X}_2$: $0 < w_i < 1/n$
fractional

$\mathcal{X}_3$: $w_i = 1/n$
selected

A yardstick for any subdata selection method

Define the efficiency of subdata S as $\text{Eff}(S) = \Phi(S_{\text{opt}})/\Phi(S)$: subdata of size n with efficiency e is as informative as optimal subdata of size en . Since the relaxed problem contains every subset of size n , $\Phi(\xi^*) \leq \Phi(S_{\text{opt}}) \leq \Phi(S^*)$, and hence for subdata S from *any* method

$$\frac{\Phi(\xi^*)}{\Phi(S)} \leq \text{Eff}(S) \leq \frac{\Phi(S^*)}{\Phi(S)} = \frac{\Phi(\xi^*)}{\Phi(S)} \bigg/ \frac{\Phi(\xi^*)}{\Phi(S^*)}.$$

- The width of the bracket is governed by $\Phi(\xi^*)/\Phi(S^*)$. In all our experiments it is at least **99.99%** for large N and n , so the lower and upper bounds essentially coincide.
- This gives, for the first time, a rigorous benchmark for the statistical efficiency of subdata from *any* method, model-based or not.
- Applying the benchmark shows that many existing methods operate well below their theoretical potential.

Equivalence theorem for bounded designs (Theorem 2)

Let $F_\Phi(\xi; \mathbf{x}_i)$ be the directional derivative of Φ at ξ toward the point mass at $\mathbf{x}_i$ (Φ convex and linearly differentiable, which covers A , D and Φ_p).

Characterization of the optimal bounded design

ξ^* minimizes Φ over all bounded designs ($0 \leq w_i \leq 1/n$, $\sum_i w_i = 1$) if and only if $\mathcal{X}$ can be partitioned as $\mathcal{X}_1 \cup \mathcal{X}_2 \cup \mathcal{X}_3$ with

- (a) $w_i^* = 0$ on $\mathcal{X}_1$, $0 < w_i^* < 1/n$ on $\mathcal{X}_2$, $w_i^* = 1/n$ on $\mathcal{X}_3$;
- (b) $F_\Phi(\xi^*; \mathbf{x}_i) = s$ is constant on $\mathcal{X}_2$;
- (c) $\max_{\mathbf{x}_i \in \mathcal{X}_3} F_\Phi(\xi^*; \mathbf{x}_i) \leq s \leq \min_{\mathbf{x}_i \in \mathcal{X}_1} F_\Phi(\xi^*; \mathbf{x}_i)$ (if $\mathcal{X}_2 = \emptyset$, compare the max and the min directly).

- Reading: at the optimum every fully selected point is at least as valuable (smaller F_Φ) as every unselected point, so no exchange can lower Φ .
- Discrete-space version of Sahm and Schwabe (2001). A violation tells the algorithm exactly which points to move.

The main algorithm: IBOSS OBD

Preparation (subdata only)

1. Start from IBOSS subdata S of size n (Wang, Yang and Stufken, 2019).
2. Compute $F_{\Phi}(\xi_S; \mathbf{x}_i)$ for all i . Swap the $\lfloor n/p_1 \rfloor$ worst points in S with the $\lfloor n/p_1 \rfloor$ best points outside S . Repeat p_1 times $\Rightarrow$ **IBOSS+**.
3. One-point exchanges ($\mathbf{x}_{1*}$ in, $\mathbf{x}_{3*}$ out), n times $\Rightarrow$ **IBOSS++**.
4. Set $\mathcal{X}_3^{(0)} = S$, $\mathcal{X}_1^{(0)} = \mathcal{X} \setminus S$, $\mathcal{X}_2^{(0)} = \emptyset$.

- Whenever the conditions fail, the iteration strictly decreases Φ . Final subdata: the n largest weights of $\xi^* \Rightarrow$ **IBOSS OBD**. The same run yields $\Phi(\xi^*)$, the yardstick for every other method.

Iteration (fractional weights)

1. If $\mathcal{X}_2^{(t)} \neq \emptyset$, optimize its weights in $(0, 1/n)$ by Newton's method; move points that reach 0 or $1/n$ to $\mathcal{X}_1$ or $\mathcal{X}_3$.
2. If the equivalence conditions hold (within ϵ), stop and output ξ^* .
3. Otherwise move $\mathbf{x}_{1*} = \operatorname{argmin}_{\mathcal{X}_1^{(t)}} F_{\Phi}$ and $\mathbf{x}_{3*} = \operatorname{argmax}_{\mathcal{X}_3^{(t)}} F_{\Phi}$ into $\mathcal{X}_2^{(t+1)}$; set $t \leftarrow t + 1$.

Theoretical guarantees

- **Convergence (Theorem 3).** Under a mild boundedness condition on Φ along pairwise weight transfers (Assumption 1, verified empirically in all our examples), $\Phi(\xi_{\mathcal{P}(t)}) \rightarrow \min_{\xi \in \Xi} \Phi(\xi)$ as $t \rightarrow \infty$.
- **Verifiable stopping rule (Theorem 4).** Relaxing the equivalence conditions by ϵ guarantees $\Phi(\xi_\epsilon) - \Phi(\xi^*) \leq \epsilon$, which sidesteps floating-point issues in checking exact optimality.
- **Polynomial time (Theorem 5).** With the ϵ stopping rule the algorithm is polynomial, with worst-case cost $O(nN^2p_1^2/\epsilon)$. With $\epsilon = 10^{-5}$ only a few iterations are needed in practice.
- **Cheap variant.** IBOSS+ costs only $O(Np^3)$ and is already highly efficient, a claim that can be made only because the benchmark exists.
- **Scope.** General parametric models, arbitrary N and n , A -, D - and Φ_p -optimality, full or partial parameter vectors.

Simulation design

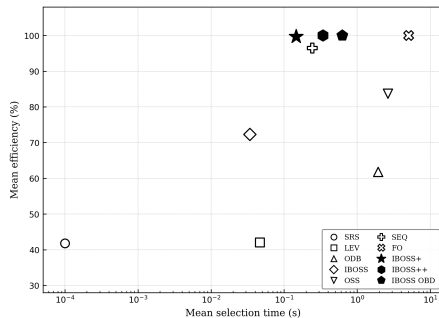
Three scenarios

- (i) Linear first-order model, $p = 10$:
 D -opt for all parameters; A -opt for the first five slopes.
- (ii) Full second-order logistic regression, $p = 3$:
 D -opt for main effects; A -opt for all parameters.
- (iii) Clusterwise linear regression, $p = 10$, $G = 10$ clusters:
 D - and A -opt for all parameters. No closed-form information matrix; a 110-dimensional surrogate is used (Liu, Stufken and Yang, 2026).

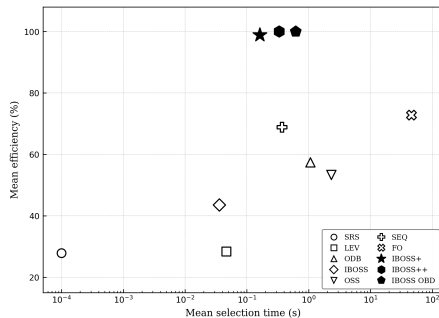
Setup

- $\mathbf{x}_i \sim N_p(\mathbf{1}_p, \Sigma_p)$, unit variances, correlation 0.5
- $N = 100,000$; $n = 1,000$ ($n = 10,000$ in (iii))
- 100 replications; $\epsilon = 10^{-5}$; Julia on a laptop
- Reported efficiency = $\Phi(\xi^*)/\Phi(S)$, the lower bound; the upper bound is within 0.01%
- Competitors: SRS, ODB, LEV, IBOSS, OSS, SEQ, FO (not all apply to every scenario)

Results, Scenario (i): linear model, $p = 10$



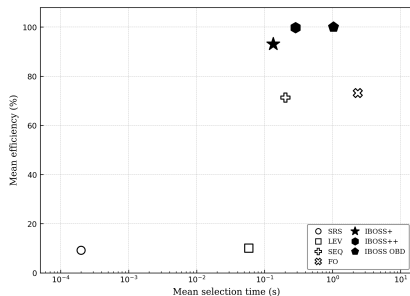
D-opt, all parameters



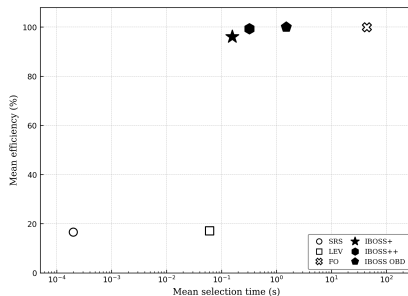
A-opt, first five slope parameters

- **IBOSS+, IBOSS++ and IBOSS OBD** (filled symbols) are at or near 100% under both criteria.
- SEQ and FO compete under *D*-optimality for all parameters, but fall to about 70% when only five slopes matter; FO then needs 46 s versus 0.6 s for IBOSS OBD.

Results, Scenario (ii): second-order logistic regression, $p = 3$



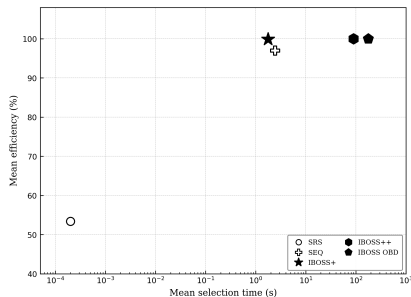
D-opt, main-effect parameters



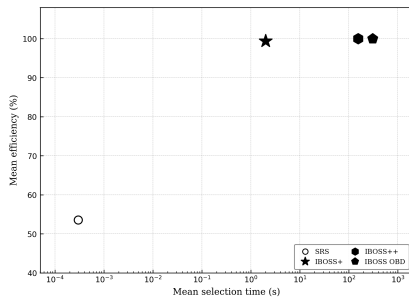
A-opt, all parameters

- A stress test: most points carry almost no information. SRS and LEV keep only 9 to 17% of it; IBOSS++ and IBOSS OBD are at or near 100%, and IBOSS+ reaches 93 to 96% in about 0.15 s.
- FO matches IBOSS OBD only under *A*-optimality, at 44 s versus 1.5 s. SEQ is excluded there because its realized subdata sizes fall far below the target.

Results, Scenario (iii): clusterwise linear regression, $p = 10$, $G = 10$



D -opt, all parameters



A -opt, all parameters

- A mixture-of-experts model handled through a 110-dimensional surrogate matrix, with $n = 10,000$. LEV and FO do not apply; SEQ applies only under D -optimality and reaches 97%.
- IBOSS+ reaches 99.4 to 99.9% in about 2 s. IBOSS++ and IBOSS OBD reach 100% but need minutes for the 110-dimensional weight optimization.

Mean statistical efficiency (%) across all settings

Setting	SRS	ODB	LEV	IBOSS	OSS	SEQ	FO	IBOSS+	IBOSS++	IBOSS OBD
(i) D , all	41.8	61.8	42.1	72.3	83.7	96.5	100.0	99.7	100.0	100.0
(i) A , five slopes	27.8	57.4	28.4	43.6	53.3	68.8	72.8	98.9	100.0	100.0
(ii) D , main effects	9.2	–	10.0	–	–	71.3	73.2	93.1	99.8	100.0
(ii) A , all	16.6	–	17.1	–	–	–	100.0	96.1	99.3	100.0
(iii) D , all	53.4	–	–	–	–	97.0	–	99.9	100.0	100.0
(iii) A , all	53.6	–	–	–	–	–	–	99.4	100.0	100.0

- Means over 100 replications with $N = 100,000$; a dash means the method does not apply or was excluded (see the paper).
- Efficiencies are measured against $\Phi(\xi^*)$. The lower bound for IBOSS OBD is 99.999%, so its efficiency is reported as 100.
- Mean selection time: IBOSS+ 0.13 to 2.0 s in every setting; IBOSS OBD 0.6 to 1.5 s in (i) and (ii), 176 and 306 s in (iii); FO 5 and 46 s in (i), 2.3 and 44 s in (ii).

Takeaways

Selection

- IBOSS OBD converges to the optimal bounded design in polynomial time.
- All three proposed methods are near-perfectly efficient for linear, logistic and clusterwise models.
- IBOSS+ is the practical choice: high efficiency at $O(Np^3)$ cost.

Where it matters

Costly data collection, labeling or analysis: clinical trials, environmental monitoring, large surveys, choosing which examples to label.

Assessment

- First tight bounds on the efficiency of subdata from *any* method.
- Many existing methods sit far below their potential, especially for partial-parameter criteria.

Open directions

- Small N , where the bracket can widen.
- Models without an explicit information matrix.
- Efficiency across a family of candidate models.
- Software package in preparation.

References and acknowledgments

- Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Annals of Statistics*, 2, 849–879.
- Sahm, M. and Schwabe, R. (2001). A note on optimal bounded designs. In *Optimum Design 2000*, 131–140.
- Welch, W. (1982). Algorithmic complexity: three NP-hard problems in computational statistics. *J. Statist. Comput. Simul.*, 15, 17–25.
- Wang, H., Yang, M. and Stufken, J. (2019). Information-based optimal subdata selection for big data linear regression. *JASA*, 114, 393–405.
- Liu, Y., Stufken, J. and Yang, M. (2026). Information-based optimal subdata selection for clusterwise linear regression. *Statistica Sinica*, 36, 715–740.
- Ahipaşaoğlu, S. D. (2021). A branch-and-bound algorithm for the exact optimal experimental design problem. *Statistics and Computing*, 31, 65.
- Pronzato, L. and Wang, H. (2021). Sequential online subsampling for thinning experimental designs. *J. Statist. Plann. Inference*, 212, 169–193.
- Full references for the competing methods (SRS, ODB, LEV, IBOSS, OSS, SEQ, FO) are in the paper.

Funding. Min Yang: NSF grant 2210546. John Stufken: NSF grant 2304767. Ming-Chung Chang: Academia Sinica grant AS-CDA-111-M05 and NSTC grants 113-2124-M-001-020, 114-2628-M-001-008-MY3 and 114-2124-M-001-008. The authors declare no competing interests. **Contact:** Min Yang, myang2@uic.edu