When Is Low Fidelity Worth Buying? Exact Certificates and Empirical Safeguards for Multi-Fidelity Bayesian Optimization

Anonymous authors
Paper under double-blind review

Abstract

Multi-fidelity Bayesian optimization can reduce the cost of expensive high-fidelity evaluations, but low-fidelity information is valuable only when its benefit justifies the additional evaluation cost. Existing approaches often model low-fidelity sources or select among fidelities without explicitly accounting for this procurement decision. We introduce a cost-aware framework that treats each high-fidelity query and its optional low-fidelity purchase as a single prepaid macro-round. The framework combines an exact finite-family certificate for reliable procurement with an empirical safeguard that probes low-fidelity usefulness before activation and falls back to high-fidelity-only optimization when the evidence is insufficient. We prove expected and high-probability bounds on cost-adjusted simple regret, a finite-horizon comparison bound against matched high-fidelity-only procurement, and an exact bounded-cost fallback property. Experiments show that the method actively exploits informative low-fidelity sources, rejects unsupported sources, and achieves a substantial improvement on a standard multi-fidelity benchmark while preserving the declared cost accounting. These results demonstrate how certification and empirical safeguards can turn low-fidelity evaluation into a controlled procurement option.

1 Introduction

In engineering, scientific computing, and AutoML, high-fidelity (HF) evaluations are accurate but expensive, while low-fidelity (LF) evaluations are cheaper and may be biased. LF can guide the search, but it is useful only when the improvement it enables justifies its cost. We study when an optimizer should buy an LF evaluation.

For example, an optimizer may buy a cheap evaluation, use it to choose the next accurate evaluation, and find that the cheap result did not help. Its price has still been paid. If the relation between the two sources was estimated poorly, the cheap result may also point the search in the wrong direction. We therefore ask two questions: when is a cheap evaluation worth buying, and how much can we spend checking it before we decide?

Prior work studies how to combine cheap and accurate observations, how to choose among information sources, and how evaluation prices should affect stopping (Kennedy & O'Hagan, 2000; Le Gratiet & Garnier, 2014; Perdikaris et al., 2017; Cutajar et al., 2019; Xing et al., 2021; Mikkola et al., 2023; Foumani & Bostanabad, 2023; Andrés-Thió et al., 2024; Candelieri et al., 2025; Lee et al., 2020; Xie et al., 2024; 2026; Wilson, 2024; Ishibashi et al., 2023; Hu et al., 2025). Our focus is the cost of buying a cheap observation before the accurate query it is intended to guide. Counting only the accurate query makes this decision look cheaper than it is.

Our approach has two parts. Exact CGS is a decision rule for purchasing LF within a search round. It counts both the LF and HF costs and, under an explicit model-coverage assumption, provides a formal guarantee relative to HF-only search. Safe-LF is a separate empirical safeguard: it compares paired LF and HF evaluations, limits the screening cost, and falls back to HF-only search when the evidence is weak. Passing this empirical screen does not establish the model-coverage assumption required by Exact CGS.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The paper makes four contributions:

1. We formalize cost-adjusted simple regret (CASR), which charges each LF evaluation together with the HF evaluation it is intended to guide.
2. We derive exact expected and high-probability CASR guarantees for Exact CGS under finite-family model coverage, and compare it with a matched HF-only policy.
3. We give a pointwise accounting identity for the cost of Safe-LF screening, including the cost of rejected LF sources, and state explicitly that empirical screening does not certify model coverage.
4. We evaluate both methods on controlled problems and benchmarks. Results show when LF helps and when safeguards reject it; current real-data results demonstrate bounded rejection costs, but not consistent performance gains.

2 Problem Setup

We consider a finite set of candidate points and seek to minimize an unknown objective f. An accurate evaluation returns $f(x)$ and costs $c_H > 0$. A cheap evaluation costs $c_L \geq 0$, but its value may differ from $f(x)$. Before round t, the algorithm knows its previous observations, denoted by $\mathcal{H}_t$, and its best accurate result, denoted by Y_{t-1}. If it evaluates point x_t accurately, the improvement is

$$\Delta_t = (Y_{t-1} - f(x_t))_+.$$

Let $I_t = 1$ mean that the cheap evaluation is also purchased in this round; $I_t = 0$ means it is not. The round's cost, the total cost through round τ, and the final score are

$$K_t = c_H + I_t c_L, \qquad C_\tau = C_0 + \sum_{t \leq \tau} K_t, \qquad \mathrm{CASR}(\tau) = Y_\tau - f^* + C_\tau.$$

Here C_0 is the known initialization cost, including any cheap evaluations run before the search, and Y_0 is the best accurate result after initialization. The final score adds the remaining error $Y_\tau - f^*$ to all costs paid. R_0 is the score that would result from stopping immediately.

We use expected improvement (EI), the predicted amount by which a new accurate evaluation may beat the best result so far. Gaussian-process models are a common way to estimate it (Mockus, 1989; Rasmussen & Williams, 2006; Jones et al., 1998; Srinivas et al., 2010; Brochu et al., 2010). Other work models the relation between cheap and accurate evaluations or uses staged budgets (Kennedy & O'Hagan, 2000; Le Gratiet & Garnier, 2014; Perdikaris et al., 2017; Cutajar et al., 2019; Xing et al., 2021; Xiong et al., 2013; Li et al., 2017; Falkner et al., 2018; Klein et al., 2017). Here we study the decision made after EI has been estimated: does its expected benefit justify the full cost of the round?

2.1 Assumptions and guarantee

The guarantees require the following conditions. The history $\mathcal{H}_t$ includes all observations and random choices made before round t. **Assumption 1 (conservative improvement estimate).** At every candidate point, the true expected improvement is at least the model's estimate minus a known error allowance η_t:

$$\mathbb{E}[(Y_{t-1} - f(x))_+ \mid \mathcal{H}_t] \geq \widehat{\mathrm{EI}}_t(x) - \eta_t, \qquad \eta_t \geq 0.$$

Assumption 2 (decide before observing the result). The choice to buy the cheap evaluation, the error allowance, and the rule for selecting the accurate evaluation use only information already available. Thus the round's cost is known before the algorithm decides whether to continue.

Assumption 3 (no repeated accurate evaluations). There are finitely many candidates, accurate evaluations are exact, and the model assigns no further improvement to a point already evaluated. Each continuing round therefore uses a new point.

The first two conditions support the cost guarantees; the third gives a finite upper bound on the number of rounds. The practical screening method described later does not automatically satisfy Assumption 1. Its performance must therefore be judged from experiments rather than inferred from the theorem.

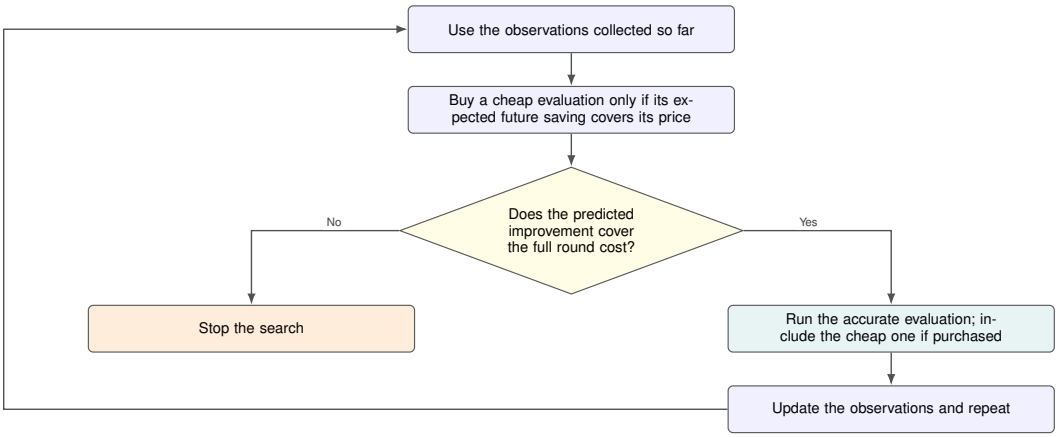


Figure 1: The theoretical rule. It buys a cheap evaluation only when its expected future saving covers its price, and runs the current round only when the predicted improvement covers the full cost. The practical screening method is a separate empirical procedure described in the experiments.

3 Counting the Cost of Each Round

The total score changes by the round's cost minus the improvement found in that round:

Lemma 3.1 (How each round changes the total score)**.** *For any policy and any finite stopping time* τ*,*

$$\mathrm{CASR}(\tau) = R_0 + \sum_{t=1}^{\tau} (K_t - \Delta_t).$$

Buying a cheap evaluation does not directly improve the best accurate result; it only adds to the cost and may help choose future points. The rule below continues only when the estimated improvement is large enough to cover the whole round, including any cheap evaluation. Under Assumption 1, this makes the expected increase in the total score nonpositive:

$$\mathbb{E}[\Delta_t \mid \mathcal{H}_t] \geq \widehat{\mathrm{EI}}_t(x_t) - \eta_t, \qquad \eta_t \geq 0, \tag{1}$$

The allowance η_t accounts for possible overestimation. The continuation rule is therefore

$$\max_{x \in \mathcal{X}} \widehat{\mathrm{EI}}_t(x) \geq K_t + \eta_t. \tag{2}$$

If the rule fails, the algorithm stops. Otherwise, it evaluates the candidate with the largest conservative improvement estimate and buys the cheap evaluation only if the separate purchase decision calls for it.

4 When Is a Cheap Evaluation Worth Buying?

4.1 A conservative estimate of improvement

Suppose we have a finite set Θ of possible prediction models. For each candidate, we calculate the expected improvement under every model and use the smallest value:

$$\widehat{\mathrm{EI}}_t(x) = \min_{\theta \in \Theta} \mathbb{E}_{P_{\theta,t}}[(Y_{t-1} - f(x))_+ \mid \mathcal{H}_t].$$

If the set contains the true model, this estimate cannot exceed the true expected improvement:

$$\mathbb{E}[(Y_{t-1} - f(x))_+ \mid \mathcal{H}_t] \geq \widehat{\mathrm{EI}}_t(x). \tag{3}$$

If the set does not contain the true model, but we know that the estimate's error is at most $\delta_t(x)$,

$$0 \leq \mathbb{E}[(Y_{t-1} - f(x))_+ \mid \mathcal{H}_t] - \widehat{\mathrm{EI}}_t(x) \leq \delta_t(x),$$

then setting η_t to at least the largest such error gives the conservative bound required by Assumption 1. This is a formal result only when model coverage or a valid error bound is available.

This gives the theoretical rule: when the true model is covered, compare a conservative estimate of improvement with the full round cost. The practical screen used in our experiments is different. It uses a limited set of paired cheap and accurate evaluations, then applies correlation and other checks to decide whether to use the cheap source. These checks are evidence about usefulness, not proof that the model-coverage condition holds.

Theorem 4.1 (Expected and high-probability total-score bounds)**.** *Assume Conditions 1 and 2, $c_H > 0$, $c_L \geq 0$, and the rule in equation 2. Let C_0 be the initialization cost and let τ_B be the last round executed before the budget is reached or the rule says to stop. Then*

$$\mathbb{E}[\mathrm{CASR}(\tau_B)] \leq \mathbb{E}[R_0].$$

If f has range at most D, so that $0 \leq \Delta_t \leq D$, let $\nu_t = \mathrm{Var}(\Delta_t \mid \mathcal{H}_t)$, $V_B = \sum_{t \leq \tau_B} \nu_t$, and $q = c_H + c_L + D$. Then, for any $v, \varepsilon > 0$,

$$\mathbb{P}\{\mathrm{CASR}(\tau_B) - R_0 \geq \varepsilon,\ V_B \leq v\} \leq \exp\left(-\frac{\varepsilon^2}{2(v + q\varepsilon/3)}\right).$$

This theorem says that, under the conservative-estimate condition, following the rule does not increase the expected score above the score at initialization. The high-probability bound additionally limits the chance of a large increase. These are guarantees for the mathematical rule under its assumptions; they do not automatically apply to the empirical screening procedure.

4.2 Deciding whether the cheap result is worth its price

The previous rule decides whether to run the current round. A separate question is whether to pay for the cheap evaluation. To answer it, compare the expected future score with and without that evaluation, keeping the current accurate point fixed. Let $V_{t,h}$ denote the lowest expected future score change over at most h rounds, given the current observations and remaining budget. Let $Q^{(0)}$ and $Q^{(1)}$ denote that score change when the cheap evaluation is not bought or is bought in the current round. An action that exceeds the remaining budget is not allowed. For a feasible choice $i \in \{0, 1\}$,

$$Q^{(i)}_{t,h,b_t} = ic_L + \mathbb{E}\Big[c_H - \Delta_t + V_{t+1,h-1}(\mathcal{H}^{(i)}_{t+1}, b_t - K_t(i)) \mid \mathcal{H}_t, I_t = i\Big]. \tag{4}$$

Define $d_{t,h,b_t} = Q^{(0)}_{t,h,b_t} - Q^{(1)}_{t,h,b_t}$. A positive value means that buying the cheap evaluation gives a lower expected future score. The cheap result cannot improve the current accurate result; it can only help choose later points. Therefore buy it only when its expected reduction in future score is at least its price.

Proposition 4.2 (When buying the cheap evaluation is worthwhile)**.** *Let S_{t,h,b_t} be the expected reduction in future score caused by seeing the cheap result, with the current accurate point held fixed. Then the net benefit of buying it is $S_{t,h,b_t} - c_L$. Buying is best when this quantity is nonnegative. If an estimate of S has error at most ϵ_t, buying only when the estimate exceeds $c_L + \epsilon_t$ limits the loss from a mistaken decision to at most $2\epsilon_t$ in that round.*

The expected future reduction can be estimated by simulating possible outcomes. More simulations reduce random estimation error, but they do not remove errors caused by a wrong prediction model or by searching only a limited set of candidate points. The next results state a comparison guarantee when these errors are bounded. This guarantee concerns the mathematical purchase rule, not the correlation screen used in the practical experiments.

Theorem 4.3 (Comparison with accurate-only search)**.** *Consider at most H buying decisions. Both policies use the same rule, based on the available history, to choose the current accurate point. One policy may buy the cheap evaluation; the accurate-only policy never does. At round t, let $h_t \in \{1, \ldots, H\}$ be the number of decisions left and let $b_t = B - C_{t-1}$ be the remaining budget. Let $\widehat{d}_t$ estimate d_{t,h_t,b_t}, and suppose*

$$\mathbb{P}\{|\widehat{d}_t - d_{t,h_t,b_t}| > \varepsilon_t \mid \mathcal{H}_t\} \leq \alpha_t.$$

Write $\mathrm{CASR}_{\mathrm{Acc}}$ *for the score of the accurate-only policy. Assume possible future scores lie in a range of width at most* L_B*; under the stated bounded-objective and budget convention, one may take* $L_B = D + B$*. If the buying rule acts only when* $\widehat{d}_t \geq \varepsilon_t$ *and the purchase fits the remaining budget, then*

$$\mathbb{E}[\mathrm{CASR}_{\mathrm{Buy}}] \leq \mathbb{E}[\mathrm{CASR}_{\mathrm{Acc}}] + 2\sum_{t=1}^{H} \varepsilon_t + L_B \sum_{t=1}^{H} \alpha_t. \tag{5}$$

This result compares only the decision to buy the cheap evaluation. The two policies must start from the same points, use the same budget, and choose accurate evaluation points by the same rule. Under those conditions, the additional expected score is bounded by the purchase-estimation error and the chance that its error bound fails. The result does not compare different rules for choosing accurate points.

Proposition 4.4 (Cost of checking and rejecting the cheap source)**.** *Suppose the policy sets a maximum of* P *checks. During these checks, it chooses points using only the accurate-evaluation history; cheap results do not change later choices. Compare it with accurate-only search using the same initialization, candidate set, choice rule, and random seed. If the policy rejects the cheap source after* $p \leq P$ *checks and never buys it for later rounds, both policies choose the same accurate points and*

$$\mathrm{CASR}_{\mathrm{check}} - \mathrm{CASR}_{\mathrm{Acc}} = pc_L \leq Pc_L.$$

If the policy skips all checks whenever $Pc_L > \rho B$*, with* $0 < \rho \leq 1$*, then* $p = 0$ *and the two trajectories, best observed accurate values, and total scores are identical for every run.*

Proof. The policies have identical accurate-evaluation histories before the first check, so they choose the same point. Since the cheap result does not affect later choices during checking, they continue to choose identical accurate points. The only extra cost is the p cheap checks, proving the first claim. If the planned checks exceed the allowed budget fraction, no check is made and the two runs are identical. □

Theorem 4.5 (Error bound for simulated buying decisions)**.** *Suppose that searching a finite candidate set adds at most* ξ_t *error to the estimated buying benefit. Suppose also that, at each future step* s*, the prediction model differs from the true outcome distribution by at most* $\kappa_{t,s}$*, and simulation adds at most* ϵ_t^{MC} *error. Then, when all these bounds hold together, the error in the buying decision is at most*

$$\varepsilon_t = \xi_t + G \sum_{s=t}^{t+H-1} (t + H - s)\kappa_{t,s} + \epsilon_t^{\mathrm{MC}}, \tag{6}$$

where $G = D + c_H + c_L$*. If these bounds fail with probability at most* α_t*, substituting this error into equation 5 gives the corresponding comparison with accurate-only search.*

4.3 The mathematical decision rule

Algorithm 1 summarizes the theoretical rule. It first decides whether the cheap evaluation is expected to save enough in future rounds to cover its price. It then checks whether the expected improvement justifies the full current cost. This rule has the guarantee proved above only when the model assumptions hold. The empirical screening procedure tested later is a separate method and does not inherit that guarantee.

In the experiments, the screening method does not use a verified set of models to certify the improvement estimate. It runs a limited number of cheap and accurate evaluations at points chosen by the accurate-only rule. It then uses the observed agreement, along with additional checks on the real-data tasks, to decide whether to use cheap evaluations in later rounds. If the checks reject the cheap source, the later accurate-evaluation choices follow the accurate-only rule. We report this method as an empirical test of the buying idea, not as an implementation of the formal guarantee.

5 Experiments

The experiments answer three questions. Can screening use cheap evaluations when they are informative? Does it reject weak or uninformative evaluations without changing the later accurate-only

Algorithm 1 Theoretical rule for buying a cheap evaluation

Initialize with n_{init} accurate evaluations and any cheap evaluations run in advance; record their cost as C_0, and set $t \leftarrow 1$.
while $t \leq T$ and $C_{t-1} < B$ **do**
    Set $b_t \leftarrow B - C_{t-1}$; estimate the conservative improvement and the future value of the cheap result; buy it only when the estimated future saving covers its price and the round fits the budget; set $K_t \leftarrow c_H + I_t c_L$.
    if $\max_{x \in \mathcal{X}} \widehat{\mathrm{EI}}_t(x) < K_t + \eta_t$ or $C_{t-1} + K_t > B$ **then**
        stop.
    else
        Evaluate the point with the largest conservative improvement estimate; also evaluate it cheaply if $I_t = 1$.
        Update the best accurate result and the model; set $C_t \leftarrow C_{t-1} + K_t$ and $t \leftarrow t + 1$.
    end if
end while

Table 1: Screening on a finite synthetic grid (100 paired seeds; mean $\pm$95% confidence interval; lower score is better).

Cheap-source setting	Accurate-only score	Screened score	Paired score change	Accepted runs	Later cheap evaluations/run	Rejected runs
Strong	.1097±.0201	.0940±.0173	$-.0157 \pm .0102$	90%	6.08	7%
Weak	.1097±.0201	.1103±.0201	$+.0006 \pm .0000$	0%	0	90%
Null	.1097±.0201	.1103±.0201	$+.0006 \pm .0000$	0%	0	90%

search? Does buying cheap evaluations improve results on standard and real-data benchmarks? We begin with controlled problems, then report a standard benchmark and real-data screens. The results support the first two claims more clearly than the third.

5.1 Experimental setup

The controlled study uses 100 paired random seeds, a budget equal to 20 accurate evaluations, and at most six checks. The LCBench and BSCW screens use 10 paired seeds and at most eight checks. Each method is charged for both accurate and cheap evaluations. We report the change in final score relative to accurate-only search. A separate comparison tests published methods for choosing evaluation types on the same finite candidate sets. We also include the standard `mf2` benchmark functions (van Rijn, 2024).

5.2 Controlled tests: when does screening help?

We first test the screening method on a finite grid with synthetic objective functions. Each check evaluates the same point cheaply and accurately; the point is chosen before the cheap result is seen. This lets us test whether the screen can tell a useful cheap source from a weak or uninformative one, while keeping the accurate-only comparison matched.

With 100 paired seeds, the screen accepts the strongly informative source in 90% of runs and lowers the average total score. It rejects the weak and uninformative sources; the difference from accurate-only search is the cost of the checks. This supports the screening behavior in these controlled settings, but does not show that correlation alone guarantees future benefit.

5.3 Standard benchmark: the value of informative cheap data

The standard `mf2` suite compares accurate-only search with methods that always use the cheap observations or stop using them after detecting a problem. This comparison shows whether cheap data can help on established function pairs; it does not test the paired-check screening rule. Across these four functions, always using cheap data gives a lower score, and the gated method also improves on accurate-only search.

Table 2: Results on four standard multi-fidelity functions (20 seeds; lower total score is better).

Function	Dim.	B (accurate only)	C (always cheap)	R (uses cheap when reliable)
Borehole	8	39.5	**28.2**	33.6
Currin	2	4.20	**1.59**	1.73
Park91a	4	6.49	**2.82**	3.60
Park91b	4	0.92	**0.51**	0.68

Table 3: LCBench regret under one shared protocol (10 seeds; mean ±95% confidence interval; lower is better). The screening method and accurate-only baseline use paired runs. The other methods use cheap evaluations throughout or stop using them after an error check. The final column reports the correlation between cheap and accurate values.

Dataset	Accurate-only	AR1 model	Gated AR1	Residual GP	Screen	Correlation
Fashion-MNIST	1.619±1.227	2.092±1.189	1.099±.875	2.594±1.247	1.619±1.227	.920
adult	2.405±.209	2.725±.354	2.460±.227	2.733±.324	2.405±.209	.736
higgs	1.726±1.902	2.877±1.328	1.425±.997	3.308±1.504	1.726±1.902	.817
jasmine	.439±.069	.575±.097	.605±.133	.666±.148	.439±.069	.824
vehicle	1.383±.935	3.085±2.623	3.511±2.132	3.245±1.462	1.383±.935	.905
volkert	5.057±3.560	6.706±3.089	5.662±2.822	9.689±3.165	5.057±3.560	.879

Paired-check result on Currin. We also ran the screening method on Currin with the same candidate set and budget for both policies. Across 100 paired seeds, its mean score was 1.468 ± 0.183, compared with 2.720 ± 0.241 for accurate-only search; the mean paired difference was -1.252 ± 0.239. The screen accepted the cheap source in every run, but acceptance only permits later purchases: 62% of runs made at least one later cheap purchase under the budget and buying rule. This is evidence that the practical screen can use a helpful source on this benchmark. It is one benchmark result, not evidence of a general guarantee.

5.4 Real-data tests: screening controls cost, gains remain unclear

We tested the screening method on LCBench, a set of machine-learning configurations, and the Benchmark Supercritical Wing (BSCW), an engineering design problem. We used the same candidate points, starting evaluations, budget, and random seeds when comparing methods. On LCBench, the cheap result is the validation error after 10% of training and the accurate result is the final error. The screening method checks whether its planned eight evaluations fit within 5% of the budget; if not, it runs accurate-only search without checking. When checking fits, additional spatial and agreement checks decide whether to use cheap evaluations later.

The LCBench results do not show a gain over accurate-only search for the screening method. On BSCW, all methods reach zero observed regret in this small screen; the proposed method makes eight cheap checks and no later cheap purchases. These results show how the method limits spending when evidence is insufficient. They do not establish improved performance on real data.

We also compare with published methods that choose evaluations by estimating how much each observation may improve later decisions. We adapted these rules to the same finite candidate sets and total-cost accounting. This is a separate baseline comparison; it does not test the screening method. Table 5 reports the results using the same starting points and random seeds.

The table summarizes how many checks were allowed, how the method decided whether to use cheap evaluations, and what it did when they were rejected.

Table 4: BSCW results under the same protocol (10 seeds). All methods reach zero observed regret; the table also reports evaluation counts and total cost.

Method	Regret	Total cost	Accurate evaluations	Cheap evaluations
Accurate-only	0.000	20.000	20.0	0.0
AR1 model	0.000	19.760	19.0	18.9
Gated AR1	0.000	19.760	19.0	18.9
Residual GP	0.000	19.735	19.0	18.7
Screen	0.000	19.400	19.0	8.0

Table 5: Comparison with two published evaluation-selection methods on finite candidate sets (10 seeds; mean $\pm$95% confidence interval; lower error is better). Both use the same budget, starting points, and candidate set. H/L is the mean number of accurate/cheap evaluations.

Dataset	MF-KG regret	rMF-KG regret	MF-KG H/L	rMF-KG H/L
Fashion-MNIST	1.717$\pm$1.216	5.176$\pm$1.802	16.3/9.2	3.0/42.0
adult	2.759$\pm$.345	3.870$\pm$.681	16.6/8.4	3.0/42.0
higgs	3.163$\pm$1.085	8.136$\pm$2.121	18.0/5.0	3.0/42.0
jasmine	.469$\pm$.156	2.572$\pm$1.251	18.0/5.0	3.0/42.0
vehicle	5.532$\pm$1.654	23.883$\pm$8.616	17.9/5.2	3.0/42.0
volkert	9.350$\pm$2.960	16.351$\pm$3.615	18.0/5.0	3.0/42.0
BSCW	.0036$\pm$.0019	.0042$\pm$.0018	12.1/19.5	9.2/26.8

Table 6: Summary of the screening rules used in the experiments.

Part	Rule	Result
Checks	At most 6 in controlled tests; 8 on real data	fixed by the experiment
Correlation check	Use cheap data only above a conservative 0.80 threshold	rejects weak/null cases
After rejection	Continue choosing accurate evaluations only	100% of weak/null runs

6 Conclusion

We studied when a cheap evaluation is worth buying before an accurate one. Our theoretical result gives a cost rule when the true prediction rule is among the models considered: continue only when a conservative estimate of improvement covers the full price of the round. Our experiments also studied a practical screen that uses a limited number of paired evaluations before deciding whether to use the cheap source. On controlled problems, it used strongly informative cheap data and rejected weak or uninformative data after bounded checks. It improved the result on the Currin benchmark. On the current real-data screens, it did not show a consistent improvement over accurate-only search; on BSCW all methods reached zero observed regret, and the screen made checks but no later cheap purchases. Thus the experiments support the value of screening in controlled settings and its ability to limit spending, but they do not show that the practical screen inherits the theoretical guarantee or reliably improves real-data results. A key next step is to find checks whose outcomes predict future savings more directly than correlation alone.

AI use statement

We used generative AI tools for English translation and language editing, restructuring of the LaTeX source, and formatting of figure source code. We did not use generative AI tools to run experiments, generate experimental data, or make final scientific claims. All AI-assisted text, equations, citations, and artifacts were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

Ethics statement

This work uses mathematical analysis and computational experiments on synthetic and public benchmark data. It does not involve human subjects, private personal data, or deployment of a high-risk decision system. We are not aware of additional ethical concerns beyond the standard requirements for documenting data provenance and computational procedures.

Reproducibility statement

The paper specifies the objective, evaluation costs, decision rules, and reported measures. The appendix gives the proofs, experiment settings, and definitions of the comparison methods. Methods use matched starting points, candidate sets, and random seeds where stated. Additional implementation and data-processing details accompany the supplementary material.

References

N. Andrés-Thió et al. Characterising harmful data sources when constructing multi-fidelity surrogate models. *Artificial Intelligence*, 336:104207, 2024.

Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

A. Candelieri, A. Ponti, and F. Archetti. *Multiple information source Bayesian optimization*. Springer, 2025.

Kurt Cutajar, Martin Pullin, Andreas Damianou, Neil D. Lawrence, and Javier Gonzalez. Deep gaussian processes for multi-fidelity modeling. *arXiv preprint arXiv:1903.07320*, 2019.

Stefan Falkner, Aaron Klein, and Frank Hutter. Bohb: Robust and efficient hyperparameter optimization at scale. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1437–1446, 2018.

Z. Z. Foumani and R. Bostanabad. Multi-fidelity cost-aware bayesian optimization. *Computer Methods in Applied Mechanics and Engineering*, 2023.

Shuang Hu, Hong Wang, Zhen Dai, Bryan Kian Hsiang Low, and See-Kiong Ng. Adjusted expected improvement for cumulative regret minimization in noisy bayesian optimization. *Journal of Machine Learning Research*, 26(46):1–33, 2025.

Hayato Ishibashi, Masayuki Karasuyama, Ichiro Takeuchi, and Hideitsu Hino. A stopping criterion for bayesian optimization by the gap of expected minimum simple regrets. *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 206:6463–6497, 2023.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

Aaron Klein, Stefan Falkner, Jost Tobias Springenberg, and Frank Hutter. Fast bayesian optimization of machine learning hyperparameters on large datasets. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 528–536, 2017.

L. Le Gratiet and J. Garnier. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 4(5), 2014.

E. H. Lee, V. Perrone, C. Archambeau, and M. Seeger. Cost-aware bayesian optimization. *arXiv preprint arXiv:2003.10870*, 2020.

Lisha Li, Kevin Jamieson, Giulia DeSalvo, Afshin Rostamizadeh, and Ameet Talwalkar. Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2017.

P. Mikkola, J. Martinsson, A. Durand, et al. Multi-fidelity bayesian optimization with unreliable information sources. In *Proceedings of AISTATS*, 2023.

Jonas Mockus. *Bayesian Approach to Global Optimization*. Kluwer Academic Publishers, 1989.

P. Perdikaris, M. Raissi, A. Damianou, N. D. Lawrence, and G. E. Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A*, 473(2198), 2017.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. *International Conference on Machine Learning*, 2010.

Sarah van Rijn. mf2: Multifidelity test functions. `https://github.com/sjvrijn/mf2`, 2024.

James T. Wilson. Stopping bayesian optimization with probabilistic regret bounds. *Advances in Neural Information Processing Systems*, 37, 2024.

Q. Xie, R. Astudillo, P. I. Frazier, Z. Scully, and A. Terenin. Cost-aware bayesian optimization via the pandora's box gittins index. In *Advances in Neural Information Processing Systems*, 2024.

Q. Xie, L. Cai, A. Terenin, P. I. Frazier, and Z. Scully. Cost-aware stopping for bayesian optimization. In *Proceedings of the International Conference on Machine Learning*, 2026.

Weiwei Xing, A. A. Shah, Peng Wang, Shandian Zhe, Qing Fu, and Robert M. Kirby. Residual gaussian process: A tractable nonparametric bayesian emulator for multi-fidelity simulations. *Applied Mathematical Modelling*, 97:36–56, 2021.

S. Xiong, P. Z. G. Qian, and C. F. J. Wu. Sequential design and analysis of high-accuracy and low-accuracy computer codes. *Technometrics*, 55(1):37–46, 2013.

A Proofs

A.1 Proof of the total-score identity

The best accurate result is $Y_\tau = Y_0 - \sum_{t\le\tau}\Delta_t$, and the total cost is $C_\tau = C_0 + \sum_{t\le\tau} K_t$. Adding the remaining error and the cost gives Lemma 3.1.

A.2 Why the conservative estimate is valid

If the true prediction model is one of the models in Θ, then

$$\min_{\theta\in\Theta} \mathbb{E}_{P_{\theta,t}}[(Y_{t-1} - f(x))_+ \mid \mathcal{H}_t] \le \mathbb{E}_{P_{\theta^*,t}}[(Y_{t-1} - f(x))_+ \mid \mathcal{H}_t],$$

which proves equation 3. If the model only approximates the truth, the difference between the predicted and true outcome distributions can bound the improvement error. In particular, because the improvement changes by at most the change in the predicted objective value, a known upper bound on the 1-Wasserstein distance between those distributions can be included in η_t. A bound based on total variation also needs a known limit on the possible payoff; an unbounded Gaussian model does not provide that limit by itself. Adding more candidate models makes the smallest predicted improvement smaller, so the rule becomes more cautious without losing validity.

A.3 Proof of the budget guarantee

On a continuation round, maximization and the gate give

$$\widehat{\mathrm{EI}}_t(x_t) \ge K_t + \eta_t, \qquad \mathbb{E}[\Delta_t \mid \mathcal{H}_t] \ge K_t.$$

Let $W_t = K_t - \Delta_t$, and set $W_t = 0$ after stopping. Then $\mathbb{E}[W_t \mid \mathcal{H}_t] \le 0$. The accounting identity gives

$$\mathrm{CASR}(\tau_B) - R_0 = \sum_{t=1}^{\tau_B} W_t,$$

and summing the conditional inequalities over the finite budget gives $\mathbb{E}[\mathrm{CASR}(\tau_B)] \le \mathbb{E}[R_0]$.

For the probability bound, subtract each round's conditional mean to obtain a zero-mean fluctuation $X_t = W_t - \mathbb{E}[W_t \mid \mathcal{H}_t]$. If the range of f is at most D, each fluctuation has upper bound $q = c_H + c_L + D$, and its conditional variance is $\nu_t = \mathrm{Var}(\Delta_t \mid \mathcal{H}_t)$. Since each round costs at least c_H, the budget implies $\tau_B \le N_B := \lfloor (B - C_0)/c_H \rfloor$ when $B \ge C_0$. A standard concentration bound for bounded zero-mean fluctuations then gives, for $V_B = \sum_{t\le\tau_B} \nu_t \le v$,

$$\mathbb{P}\left\{ \sum_{t\le\tau_B} X_t \ge \varepsilon,\ V_B \le v \right\} \le \exp\!\left(-\frac{\varepsilon^2}{2(v + q\varepsilon/3)} \right),$$

Since $\nu_t \le D^2/4$, an unconditional version follows by substituting $v = N_B D^2/4$.

A.4 Why the search stops after finitely many rounds

After an exact accurate evaluation, the expected improvement at that point is zero. Since every new round costs at least $c_H > 0$, the method must choose a new point each time. It can therefore make at most $|\mathcal{X}| - n_{\mathrm{init}}$ further evaluations.

B Additional Experiment Details

B.1 Problems and comparison methods

Besides the controlled tests and benchmarks in the main paper, we evaluate additional synthetic functions, GP-generated objectives in dimensions four and six, and published comparison methods. These results are secondary checks on how the methods behave in different settings; they are not additional proof of the theoretical assumptions.

Table 7: LCBench results with a lower cheap-evaluation price (20 paired seeds; mean ±95% confidence interval). The score difference includes the cost of eight checks. The screen rejected later cheap evaluations in every run.

Dataset	Accurate-only regret	Screen regret	Paired score change	Later cheap evaluations/run	Decision
Fashion-MNIST	1.156±.725	1.325±.748	-.031±.313	0	Reject; accurate-only
adult	2.593±.224	2.609±.224	-.184±.032	0	Reject; accurate-only
higgs	1.472±1.195	1.894±1.233	+.222±.570	0	Reject; accurate-only
jasmine	.507±.092	.522±.090	-.185±.030	0	Reject; accurate-only
vehicle	2.447±1.151	2.447±1.151	-.200±.000	0	Reject; accurate-only
volkert	5.821±2.342	6.647±2.329	+.626±1.046	0	Reject; accurate-only

Table 8: Probe-budget sensitivity on LCBench (20 seeds per task; means over six tasks).

B	Accurate-only regret	Screen regret	Screen cost	Checks	Later cheap use	Rejected
8	5.877	5.877	8.00	0.0	0.0	100%
12	4.346	4.346	12.00	0.0	0.0	100%
16	3.190	3.426	15.80	8.0	0.0	100%
20	2.333	2.574	19.80	8.0	0.0	100%

B.2 Implementation details

We scale design variables to $[0, 1]^d$. Candidate sets, model-fitting restarts, starting evaluations, and stopping settings are fixed for each experiment. All reported costs include initialization and cheap evaluations. The simulated buying-rule diagnostic uses 64 simulations, 128 candidate points, and 1,500 random features. It records the final score, remaining error, number of cheap evaluations, and run time. Its simulation standard error measures random simulation noise; it does not measure errors caused by using the wrong prediction model.

B.3 How price and budget affect screening

We repeated the LCBench screen with a lower cheap-evaluation price of 0.1 and a total budget of 20, using the same candidate sets, starting points, stopping rule, and 20 paired seeds. At this price, all eight planned checks fit within the budget, so the experiment tests rejection after seeing evidence. The screen ran all eight checks and then rejected further cheap evaluations in all 120 runs. The reported score difference includes the cost of those checks; it is not evidence of a performance gain.

We also varied the total budget while keeping the cheap-evaluation price at 0.1. Table 8 reports the average over the six tasks. With budgets of 8 and 12, the checks do not fit, so the method skips them and runs accurate-only search. With budgets of 16 and 20, it performs eight checks but never accepts further cheap evaluations. This shows how budget limits control checking costs; it does not show that cheap evaluations improve these real-data results.

C Additional Comparison Proofs

C.1 Proof of the future-savings calculation

Equation equation 4 separates the price of the cheap evaluation from its effect on later decisions. For a fixed history and future decision rule, let $G_{t,m}^{(i)}$ be one simulated future score when choice i is made. If at most N_B rounds remain and the objective range is at most D, each simulation lies in an interval of length at most

$$A_B = N_B(D + c_H + c_L).$$

For M independent simulations, a standard concentration bound gives, with probability at least $1 - \alpha_t$,

$$J_t^{(0)} \geq \widehat{J}_t^{(0)} - w_t, \qquad J_t^{(1)} \leq \widehat{J}_t^{(1)} + w_t, \qquad w_t = A_B\sqrt{\frac{\log(4/\alpha_t)}{2M}}.$$

Thus we can form a conservative estimate of the future saving from buying the cheap evaluation. To make the simulation error at most ε, it is enough to use

$$M \geq \frac{A_B^2 \log(4/\alpha_t)}{2\varepsilon^2}.$$

Writing the expected future score reduction, with remaining budget b_t, as S_{t,h,b_t} gives

$$Q^{(0)}_{t,h,b_t} - Q^{(1)}_{t,h,b_t} = S_{t,h,b_t} - c_L.$$

Therefore buying is best exactly when the expected future saving covers the price c_L. If the estimated saving differs from S by at most ϵ_t, buying only when the estimate exceeds $c_L + \epsilon_t$ limits the loss from a mistaken decision to at most $2\epsilon_t$. This proves Proposition 4.2.

C.2 Proof of the comparison with accurate-only search

Both methods choose the current accurate point by the same rule. The choice not to buy the cheap evaluation is therefore exactly the accurate-only option. When every estimate satisfies $|\widehat{d}_t - d_{t,h_t,b_t}| \leq \varepsilon_t$, a purchase is made only when it has nonnegative true value. If the method declines a purchase that would have helped, its value is less than $2\varepsilon_t$. Thus the loss on such a round is at most $2\varepsilon_t$. If an estimate exceeds its error bound, the extra score is at most L_B. Adding these bounds over at most H rounds gives equation 5.

C.3 Proof of the simulation error bound

Suppose the finite candidate search adds error ξ_t, the model's future outcome distributions differ from the true ones by at most $\kappa_{t,s}$, and simulation adds error ϵ_t^{MC}. Let a single future cost contribution vary over an interval of length at most G. Replacing the true outcome distribution at future step s by the model's distribution then changes its expected cost by at most $G\kappa_{t,s}$. Adding these effects over at most H future steps gives the model error

$$G \sum_{s=t}^{t+H-1} (t + H - s)\kappa_{t,s}.$$

Adding candidate-pool optimization and Monte Carlo errors yields

$$|\widehat{d}_t - d_{t,h_t,b_t}| \leq \xi_t + G \sum_{s=t}^{t+H-1} (t + H - s)\kappa_{t,s} + \epsilon_t^{\mathrm{MC}} =: \varepsilon_t. \tag{7}$$

Substitution into Theorem 4.3 proves Theorem 4.5. If each simulation is independent and the possible total future score lies in a range of width at most GH, a standard concentration bound gives

$$\epsilon_t^{\mathrm{MC}} \leq 2GH\sqrt{\frac{\log(4/\alpha_t)}{2M}},$$

up to a smaller constant when paired rollouts are used. This explains why a standard-error estimate in the current GP implementation controls only the Monte Carlo term, not the model-mismatch term $\kappa_{t,s}$.

C.4 Finite search and candidate-grid error

For a finite candidate set, exact accurate evaluations, distinct starting points, and a model that remembers observed values, the expected improvement at an evaluated point is zero. Since a new round costs at least $c_H > 0$, each round must select a new point, so at most $|\mathcal{X}| - n_{\mathrm{init}}$ further evaluations are possible. If f changes by at most L times the distance between points and the candidate set comes within distance δ of every point in the full domain, restricting the search to that set loses at most $L\delta$ in objective value.

C.5 Why cheap evaluations alone cannot identify the best point

For two candidates a and b, consider two possible objectives that swap values 0 and Γ at these points but produce the same cheap results. A first accurate evaluation chosen from cheap results alone has expected error at least $\Gamma/2$ on one of the two objectives. This example shows why the method still needs accurate evaluations to guide the search.