Online Decision-Focused Learning under Semi-Bandit Feedback

Aabhash Dhakal
University of Würzburg, Germany
aabhash.dhakal@uni-wuerzburg.de

Tim Lachner
University of Würzburg, Germany
tim.lachner@uni-wuerzburg.de

Jayanta Mandi
Loughborough University, UK
j.mandi1@lboro.ac.uk

Marco Foschini
KU Leuven, Belgium
marco.foschini@kuleuven.be

Christoph M. Flath
University of Würzburg, Germany
christoph.flath@uni-wuerzburg.de

Abstract

Decision-focused learning (DFL) improves on the Predict-then-optimize (PtO) pipeline by training predictive models to minimize decision regret rather than prediction loss. Most DFL methods assume offline full-feedback data with complete action cost vectors. Many real-world decision systems instead operate in an online setting: in each round, the agent acts using past feedback and observes costs only for the selected components, yielding semi-bandit feedback. Bandit methods optimize regret in this regime, but train models with predictive updates rather than decision-focused updates. In this paper, we formalize online *semi-bandit DFL*: decision-focused learning for online optimization problems with semi-bandit feedback. The most straightforward extension would impute unobserved components with the model's own predictions and compute the decision-focused gradient. But this creates a self-confirming *imputation degeneracy*: imputation errors can flip counterfactual decisions that define the decision-focused gradient, biasing the representation update and reinforcing errors in later imputations. To address this degeneracy, we propose BAYESIANSPO, which combines an analytically updated Bayesian last-layer head with masked decision-focused learning. The Bayesian head supplies both (i) uncertainty to balance exploration and exploitation and (ii) stable imputations for unobserved components, while masking unobserved component gradients prevents those imputations from directly corrupting the representation update. Across knapsack, shortest-path, portfolio benchmarks, and an energy scheduling task, BAYESIANSPO achieves the lowest mean cumulative regret on all six instances and the improvement is statistically significant on five of them.

1 Introduction

Consider a dispatcher routing packages through a road network. Each delivery includes contextual information such as time of day, weather, and current traffic. A predictive model uses this context to estimate road travel times, and an optimizer then computes the shortest path using the predicted values. This is an example of a *predict-then-optimize* (PtO) pipeline [8], where a context is mapped to predicted action costs that are fed to a combinatorial optimizer (Figure 1). The goal is not to predict

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

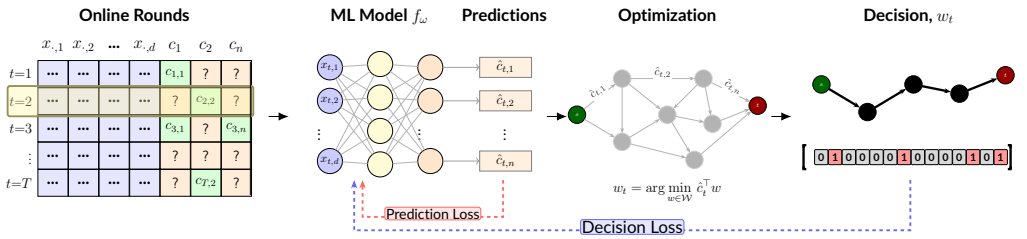


Figure 1: Online PtO pipeline. Each round, the learner observes a context, predicts arm costs, solves a downstream optimization problem, and receives semi-bandit feedback.

every road's travel time accurately; it is to recover the optimal solution. *Decision-focused learning* (DFL) improves PtO by training the predictor with an end-to-end objective that accounts for solution quality rather than prediction error [13, 21], and has become a widely adopted approach for PtO.

Standard DFL, however, relies on two strong assumptions: *offline* learning from a fixed historical dataset, and *full feedback*, where ground-truth costs are observed for all actions, including costs of the actions that were not taken. Real-world decision systems can violate both these assumptions. In the package delivery example, each round brings a new request with a new context. The dispatcher predicts travel times using the current model, chooses roads to traverse as part of a route, observes realized travel times only on traversed roads, and then updates the model using this feedback for future rounds. Such problems are modeled as *online contextual combinatorial bandits with semi-bandit feedback*.

Recent DFL work relaxes these assumptions one at a time: Capitaine et al. [9] study online full-feedback DFL, and Hu et al. [14, 15] study offline partial-feedback DFL. We are the first to study the missing quadrant, *online semi-bandit DFL* (Table 1). Our approach is suitable for applications where decisions are made sequentially and feedback is observed only for the selected arms (following bandit literature, we use *arm* for the elements that make up an action, and use terms *arm* and *component* interchangeably).

Online semi-bandit DFL combines three challenges. (i) *Exploration* requires navigating the trade-off between decision quality in the current round and collecting information useful for future rounds. (ii) *Imputation* is needed for completing unobserved costs before forming a decision-focused gradient. (iii) *Learning* requires training the predictive model with the decision-focused gradient. In this paper, we identify imputation as the main structural obstacle. The learner may fill missing costs using its own predictions, but these completed costs are then passed through the optimizer to form the decision-focused gradient. Poor imputations can change the counterfactual decisions that define the gradient, bias the representation update, and reinforce prediction errors that produced them. We call this self-reinforcing failure mode *imputation degeneracy*.

To tackle these challenges, our proposed approach, BAYESIANSPO, decouples cost prediction from decision-focused representation learning. We use Bayesian last layer learning with per-arm heads. These heads, updated analytically from observed costs, supply stable imputations and posterior uncertainty. A masked SPO+ update sets gradients on unobserved arms to zero, limiting *imputation degeneracy* by preventing imputed values from entering the representation directly, and concentrating the decision-focused signals on observed arms. Our contributions are:

(1) We identify the *imputation degeneracy* as a major obstacle for online semi-bandit DFL, and formalize the mismatch between decision-focused gradients and semi-bandit observations. (2) We propose BAYESIANSPO, which combines analytically updated per-arm posteriors with masked gradient updates, decoupling imputation from decision-focused representation learning. (3) Across all six knapsack, shortest-path, portfolio and energy-scheduling instances, BAYESIANSPO yields the lowest cumulative regret, with the reduction being statistically significant in five cases. (4) We provide a supporting analysis that bounds regret for UCB exploration and separately bounds the loss of detached masked SPO+ representation learning, with explicit terms for local imputation errors and residual effects.

2 Related Work and Positioning

The DFL landscape spans two axes: the setting and the feedback structure.

Table 1: The DFL landscape. Prior works address: *learning*, *exploration*, and *imputation*. The online $\times$ semi-bandit cell is the first setting where all three must be handled simultaneously.

	Offline	**Online**
Full feedback	Elmachtoub and Grigas [13]	Capitaine et al. [9]
Semi-bandit	Hu et al. [14, 15]	**This work**

Offline, full feedback. This is the mature DFL setting, where learners train on complete cost vectors by propagating gradients through an optimizer. It includes SPO+ [13], extensions to integer programs [20], and blackbox-solver surrogates [26], building on earlier optimization-layer work [12, 28].

Online, full feedback. In the online setting, the learner updates its predictor after each round. Capitaine et al. [9] regularize the downstream decision problem and develop decision-focused FTPL and OGD algorithms with expected static and dynamic regret guarantees. Related work on data sufficiency for optimization shows that useful observations are those spanning decision-changing directions [6]. These works, however, assume access to the complete cost vector every round.

Offline, semi-bandit feedback. The semi-bandit setting adds *imputation*, i.e., estimating the costs of actions that were not observed. Hu et al. [14] study *offline* policy learning with bandit feedback and extend the framework to semi-bandit feedback in [15]. However, their approach is not directly applicable to our setting, as it relies on pre-collected logged data and does not prescribe an active exploration rule for sequential decision-making.

Online, semi-bandit feedback (our setting). This setting combines learning, exploration, and imputation. Neither neighboring cell in Table 1 gives a direct comparator: offline semi-bandit estimators require pre-collected logged data, while online full-feedback DFL requires a complete cost vector in each round. We therefore compare against online semi-bandit methods with prediction-trained models. Combinatorial semi-bandits define the observation model [10, 11, 17]; UCB [5] and its linear contextual variants [19, 1] provide optimism-based exploration. NeuralUCB [29] and CN-UCB [16] extend this to neural features.

Our architecture builds on the Neural-Linear Bandit approach of Riquelme et al. [23] and related Bayesian last-layer methods of Weber et al. [27], which combine shared learned features with per-arm Bayesian linear heads and analytical posterior updates. While these works train representations for predictive accuracy, we adapt this architecture to decision-focused representation learning under semi-bandit feedback. We show that in this setting, a straightforward combination of exploration, cost imputation, and decision-focused learning can lead to imputation degeneracy. Our work combines posterior completion, gradient masking, and analytical posterior updates from observed costs. Together, these choices make online semi-bandit DFL viable by preserving the decision-focused learning signal while containing imputation degeneracy.

3 Problem Setup and Imputation Degeneracy

Contextual Optimization. In standard contextual optimization problems with linear objective functions, the true coefficient vector $c \in \mathbb{R}^n$ is unknown at decision time, but a context vector $x \in \mathbb{R}^d$ is observed and used to predict it, where n is the number of arms and d is the context dimension. Let $\hat{c} = f_\omega(x) \in \mathbb{R}^n$ be the cost vector predicted by a model with parameters ω. The downstream optimizer computes $w^\star(\hat{c}) = \operatorname{argmin}_{w \in \mathcal{W}} \hat{c}^\top w$, where $\mathcal{W}$ is the feasible set. We focus on binary feasible sets, $\mathcal{W} \subseteq \{0,1\}^n$, where semi-bandit feedback is naturally componentwise, with $w_i = 1$ indicating that arm i is selected. For clarity, we present the formalism for cost minimization.

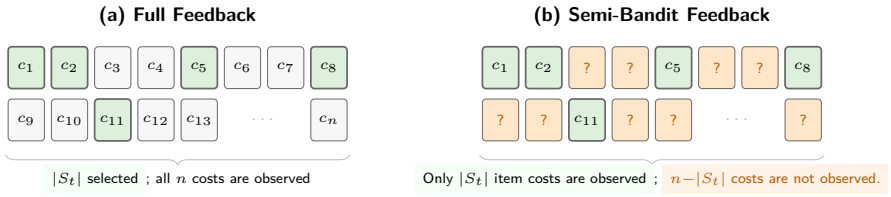


Figure 2: Full feedback vs. semi-bandit feedback on a knapsack instance with n items. The learner selects $|S_t|$ items (green). (a) Under full feedback, the complete cost vector, $(c_1, \ldots, c_n)$ is revealed. (b) Under semi-bandit feedback, only selected items' costs are revealed; the remaining $n - |S_t|$ entries are not observed. This leads to a challenge for *implementing SPO+, whose gradient requires the full cost vector.*

Decision-focused learning. For such contextual optimization problems, decision-focused learning (DFL) [21] trains the predictive model so that $w^\star(\hat{c})$ minimizes the downstream objective, $c^\top w^\star(\hat{c})$. Prior works [20, 13] report that DFL outperforms prediction-focused cost learning. Gradient-based DFL requires backpropagating through the optimizer, which is generally not possible for combinatorial optimization problems. Surrogate loss functions are therefore used instead. One of the most widely used DFL surrogates is the SPO+ loss [13]:

$$\ell_{\text{SPO+}}(\hat{c}, c) = \max_{w \in \mathcal{W}} (c - 2\hat{c})^\top w - (c - 2\hat{c})^\top w^\star(c) \tag{1}$$

This is the optimization gap of $w^\star(c)$ under the adjusted costs $2\hat{c} - c$. Its subgradient with respect to $\hat{c}$ is $\nabla_{\hat{c}} \ell_{\text{SPO+}}(\hat{c}, c) = 2\big(w^\star(c) - w^\star(2\hat{c} - c)\big)$.

Online contextual optimization under semi-bandit feedback. We study online contextual optimization over T rounds. In each round $t \in [T]$, the learner observes x_t and predicts costs, $\hat{c}_t = f_{\omega_t}(x_t)$. It then selects a feasible decision by using these predictions and an exploration rule. Without exploration, the greedy choice is

$$w_t \in \operatorname*{argmin}_{w \in \mathcal{W}} \hat{c}_t^\top w. \tag{2}$$

The decision-maker then incurs cost $c_t^\top w_t$. Under *semi-bandit feedback*, after choosing w_t it observes only selected-arm costs, $\{c_{t,i} : i \in S_t\}, S_t := \text{supp}(w_t)$; unselected costs remain missing. For example, in a shortest path problem, only traversed-edge travel times are observed, not edges belonging to alternative routes. In experiments, we report simulator-evaluated realized cumulative regret, $\mathcal{R}_{\text{real}}(T) = \sum_{t=1}^{T} [c_t^\top w_t - \min_{w \in \mathcal{W}} c_t^\top w]$, using the full realized cost vector only for offline evaluation. This is not observable in a deployed semi-bandit system. Because the realized oracle can exploit noise in c_t, for theoretical analysis we use *pseudo-regret*, measured against noiseless conditional mean costs, as defined in Section 5.

The imputation degeneracy. Standard DFL techniques such as SPO+ need a complete cost vector because the gradient is defined through the two solver calls $w^\star(c_t)$ and $w^\star(2\hat{c}_t - c_t)$. Under semi-bandit feedback, only $c_{t,i}$ for $i \in S_t$ is observed, so the learner must form a completed vector $\bar{c}_t$, with observed entries copied from c_t and missing entries *imputed*. Because this completed vector enters the solver calls that define the SPO+ gradient, imputation errors can flip the counterfactual decisions and bias resulting decision signal. The feature extractor is then updated using this biased decision signal, reinforcing the same errors in later imputations. We refer to this phenomenon as *imputation degeneracy*. Appendix B.10 provides a detailed example of how this occurs.

4 BayesianSPO: Bayesian Last-Layer Learning with Decision-Focused Representations

BAYESIANSPO responds to the imputation degeneracy with three design choices. First, per-arm Bayesian linear heads are updated analytically from observed costs, decoupling cost prediction from decision-focused representation learning. The heads provide predictions and uncertainty estimates

required for UCB exploration. Second, a shared feature extractor is trained using the decision-focused loss with completed cost vectors. Third, *masking* zeros out gradient entries for unobserved arms, blocking their direct gradient contributions from propagating through learning. While completed entries $\bar{c}_i$ for $i \notin S_t$ still go through the solver, their gradient entries are not allowed to train the representation.

Shared representation. We follow the Neural-Linear Bandit approach by combining shared learned features with per-arm Bayesian linear heads [23, 27]. While standard Bayesian last-layer models train the representation for prediction accuracy, we train it with masked SPO+ for decision quality. Specifically, a linear feature extractor $h = \phi_\psi(x) = W_1^\top x + b_1 \in \mathbb{R}^q$ maps $x \in \mathbb{R}^d$ to q latent features via parameters $\psi = \{W_1, b_1\}, W_1 \in \mathbb{R}^{d\times q}, b_1 \in \mathbb{R}^q$. Although the resulting cost predictor remains affine, the feature map is learned jointly across arms. Observations of selected arms therefore update features used by every Bayesian head. We train this representation for decision quality by learning a shared prior across arms and adapting the input directions emphasized by the features, which the Bayesian heads use for prediction and uncertainty estimation. We discuss nonlinear extractors in Appendix D.8.

Bayesian heads. Each component has one Bayesian linear regression head. Before selecting an action in round t, each head summarizes its past observations as

$$A_i = \lambda I_q + \sum_{s<t\,:\, i\in S_s} h_s h_s^\top, \qquad v_i = \sum_{s<t\,:\, i\in S_s} c_{s,i} h_s, \tag{3}$$

where $\lambda > 0$ is the prior precision and h_s is the feature vector computed when round s was observed. I_q denotes the $q \times q$ identity matrix and $c_{s,i}$ is the observed cost of component i in that round. Matrix A_i summarizes the observed feature directions, while v_i accumulates feature vectors weighted by their observed costs. These statistics give the posterior mean, predicted cost, and posterior standard deviation

$$\hat{\theta}_i = A_i^{-1} v_i, \qquad \hat{c}_i = h^\top \hat{\theta}_i, \qquad \sigma_i = \sqrt{h^\top A_i^{-1} h}. \tag{4}$$

Here $\hat{\theta}_i$ is the coefficient estimate for an arm/component i, $\hat{c}_i$ is its predicted cost, and σ_i measures the prediction uncertainty used for exploration.

Exploration. We use UCB in the usual optimism-under-uncertainty sense. Posterior means are shifted by $\beta\sigma_i$ in the direction favorable to the objective. For cost minimization, action selection uses the lower confidence bound $w_t \in \operatorname{argmin}_{w\in\mathcal{W}} \sum_i w_i(\hat{c}_i - \beta\sigma_i)$. For reward maximization, the corresponding optimistic rule is $\operatorname{argmax}_{w\in\mathcal{W}} \sum_i w_i(\hat{c}_i + \beta\sigma_i)$. We treat UCB as the default exploration rule and tune only its scale β. Figure 3 summarizes the per-round protocol. See Appendix E.5 for alternative exploration strategies.

Training with imputed costs. After observing $\{c_{t,i}\}_{i\in S_t}$, the learner adds $h_t h_t^\top$ to A_i and $c_{t,i} h_t$ to v_i for each observed arm. It then recomputes those arms' coefficient estimates and predictions. In the training expressions below, $\hat{\theta}_i$ and $\hat{c}_i$ denote these updated values; unobserved-arm values are unchanged. The learner constructs $M \geq 1$ completed cost vectors by combining true costs on observed arms with posterior-based completions, or *imputations* on the rest:

$$\bar{c}_i^{(m)} = \begin{cases} c_{t,i}, & i \in S_t, \\ \hat{c}_i, & i \notin S_t,\ m = 1, \\ \hat{c}_i + \sigma_i \xi_i^{(m)},\ \xi_i^{(m)} \sim \mathcal{N}(0,1), & i \notin S_t,\ 2 \leq m \leq M. \end{cases} \tag{5}$$

Training uses the loss $\ell_{\mathrm{SPO}+}/n$, normalized by the total number of arms, with completed costs held fixed when computing gradients. For each completed vector, its subgradient with respect to the predicted costs is $g^{(m)} = 2\big(w^\star(\bar{c}^{(m)}) - w^\star(2\hat{c} - \bar{c}^{(m)})\big)/n$, which we average over completions as $g = \frac{1}{M}\sum_{m=1}^M g^{(m)}$.

The per-arm coefficients $\hat{\theta}_i$ carry the resulting decision-focused gradient into the shared representation. $M = 1$ uses posterior-mean completion; $M > 1$ averages that completion with $M - 1$ posterior-sampled completions. Updating the Bayesian heads before computing $g^{(m)}$ ensures that representation

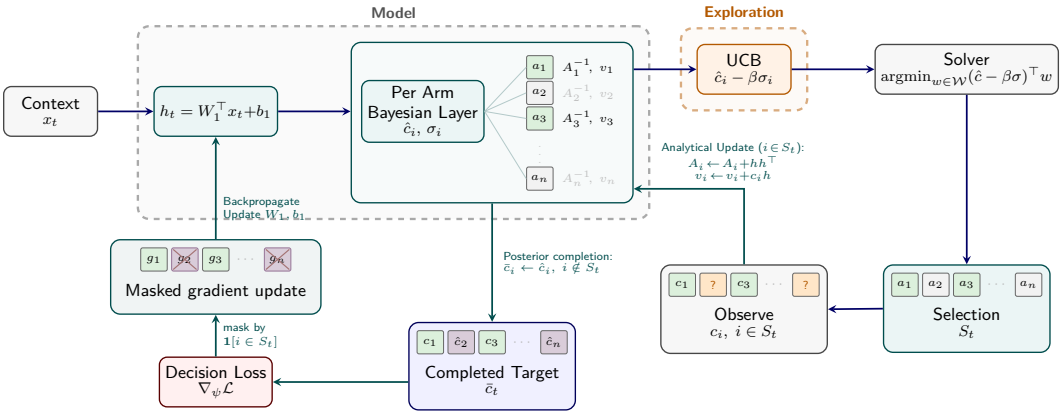


Figure 3: BAYESIANSPO. Context x_t passes through a shared feature extractor to per-arm Bayesian posteriors, which provide predictions, uncertainties, and imputations. Posterior-mean completion ($M = 1$) is illustrated. SPO+ gradient is computed on imputed targets, masked to observed arms, and backpropagated through the frozen $\hat{\theta}_i$ to update only the shared representation parameters. Mini-boxes illustrate: observed , missing , imputed .

learning uses predictions informed by the current round's observations. Because this pulls observed-arm predictions toward realized costs $c_{t,i}$, the two SPO+ solver calls can agree, zeroing out that completion's gradient contribution. An alternative is to evaluate the SPO+ subgradient at the pre-update predictions while still backpropagating through the updated heads.

Masking to contain degeneracy. We mask unobserved arms, setting $g_i \leftarrow 0$ for $i \notin S_t$ before backpropagating through the frozen heads $\hat{\theta}_i$, updating only the shared representation ψ via $\nabla_{W_1} = x_t\big(\sum_{i\in S_t} g_i\hat{\theta}_i\big)^\top$ and $\nabla_{b_1} = \sum_{i\in S_t} g_i\hat{\theta}_i$. Although the solver still uses imputed costs to form the counterfactual decisions, masking stops unobserved arms from directly driving the representation update. Meanwhile, the Bayesian heads supplying future imputations update analytically from observed costs alone, so SPO+ affects future posteriors only through changes to the feature map (ablated in Appendices E.2–E.3).

5 Regret and Local Imputation Control

We analyze the three mechanisms of BAYESIANSPO: UCB exploration, masked SPO+ learning, and the effect of imputation error. Recall that the feature map, ϕ_ψ converts a context x into a latent representation h, with ψ denoting its weights and bias. First, we analyze regret and errors in imputed costs while holding the learner's feature map fixed across rounds. All assumptions and full proofs can be found in Appendix B.

Separately, we study representation learning against a fixed reference $\phi_{\bar{\psi}}$, while the learner updates its representation ϕ_{ψ_t} online in each round. Here, ψ_t denotes the learner's parameters at round t, whereas $\bar{\psi}$ remains fixed across rounds. Evaluating both representations using the learner's Bayesian heads and completed costs, standard online gradient descent bounds the learner's additional cumulative masked SPO+ loss relative to this reference.

Consider posterior-mean completion ($M = 1$). The completed target vector $\bar{c}_t$ contains observed costs $c_{t,i}$ on selected arms, and imputed costs on missing arms. Let $\hat{\theta}^+_{t,i}$ denote the learner's Bayesian head after its update from the costs observed in round t. We construct the reference prediction vector p_t as

$$p_{t,i} = \begin{cases} \phi_{\bar{\psi}}(x_t)^\top \hat{\theta}^+_{t,i}, & i \in S_t, \\ \bar{c}_{t,i}, & i \notin S_t. \end{cases}$$

Within each round, holding $\hat{\theta}^+_{t,i}$ and $\bar{c}_t$ fixed matches the masked update, while the reference loss also reflects any mismatch between $\phi_{\bar{\psi}}$ and the learner's heads. Since p_t and $\bar{c}_t$ agree on missing arms by

construction, the reference's loss is bounded by its prediction errors on observed arms:

$$\ell_{\mathrm{SPO}+}(p_t, \bar{c}_t) \leq 2 \sum_{i \in S_t} |p_{t,i} - c_{t,i}|.$$

Combined with online gradient descent (OGD), this bounds cumulative loss by the reference's prediction errors plus an online-learning term (Appendix B.5). To further isolate the effect of imputation within the reference's loss, we examine how missing costs change the reference's loss through the two SPO+ decisions. Our theoretical step bounds imputation errors on uncertain missing arms that can replace selected arms at a similar or lower predicted cost.

Setup. Write $\tilde{c}_t := \mathbb{E}[c_t \mid x_t]$ for the conditional mean cost vector, distinct from the realized costs c_t and completed costs $\bar{c}_t$. We assume $\tilde{c}_{t,i} = \theta_i^{*\top} \phi^*(x_t)$, where ϕ^* is a shared latent feature map. The realized cost is $c_{t,i} = \tilde{c}_{t,i} + \epsilon_{t,i}$. Conditional on the history, current context, and selected action, the noise $\epsilon_{t,i}$ is independent, zero-mean, and σ-sub-Gaussian.

Let n be the number of arms, $k := \max_{w \in \mathcal{W}} \|w\|_0$ the largest number an action can select, and q the feature dimension. Let L bound the norm of each feature vector and $\lambda > 0$ denote the prior precision. We write $\alpha_{\lambda,L} = \max\{1, L^2/\lambda\}$, a factor accounting for the size of the features relative to the prior precision in the uncertainty bounds.

Let $\hat{c}_{t,i}$ and $\sigma_{t,i}$ denote the prediction and posterior standard deviation of the arm i before observing the costs of the round. Choose a confidence radius $\beta_T > 0$ such that $|\hat{c}_{t,i} - \tilde{c}_{t,i}| \leq \beta_T \sigma_{t,i}$ holds simultaneously for all arms and rounds with high probability. We apply the standard linear-UCB confidence bound to each arm and combine the failure probabilities [1]. For fixed problem parameters and confidence level, this gives $\beta_T = O(\sqrt{\log(1+T)})$ (Appendix B.2).

Standard decision-regret bound. With the learner's feature map held fixed, define pseudo-regret $\tilde{\mathcal{R}}(T) = \sum_{t=1}^{T} [\tilde{c}_t^\top w_t - \min_{w \in \mathcal{W}} \tilde{c}_t^\top w]$ using conditional mean costs. With $\beta = \beta_T$, linear UCB [1] gives, with high probability,

$$\tilde{\mathcal{R}}(T) \leq O\left(\beta_T \sqrt{\alpha_{\lambda,L} k n T q \log\left(1 + \frac{kTL^2}{nq\lambda}\right)}\right).$$

For fixed problem parameters, its T-dependence is worst-case optimal up to logarithmic factors. In Appendix B.9, we construct Gaussian instances whose optimal actions differ but whose observed costs are hard to distinguish, forcing $\Omega(\sqrt{T})$ expected regret.

Contested arms. For a tolerance $\tau_t \geq 0$, we call a missing arm *contested* if it can replace a selected arm without violating the constraints and if its predicted cost is at most τ_t higher than the selected arm's. This includes alternatives predicted to be cheaper. Assign each contested arm to one selected arm that it can replace under these conditions. Let κ denote the maximum number of contested arms assigned to any single selected arm in a round ($\kappa = 0$, if none are contested).

From observed uncertainty to imputation error. Standard semi-bandit analysis bounds cumulative uncertainty on selected arms, whereas SPO+ depends on unobserved costs through the combinatorial solver. Our key insight is to use UCB optimality across feasible swaps to show that any high-width contested missing arm j has at most twice the uncertainty of the selected arm r it can replace ($\sigma_{t,j} \leq 2\sigma_{t,r}$).

For two interchangeable arms with equal predicted costs, UCB can leave one unselected only if the selected arm is at least as uncertain. Fix an uncertainty threshold $\varepsilon > 0$. With UCB scale $\beta > 0$ and $\tau_t \leq \beta\varepsilon/2$, every contested missing arm j with $\sigma_{t,j} \geq \varepsilon$ can replace a selected arm r satisfying $\sigma_{t,j} \leq 2\sigma_{t,r}$. Using the confidence bound on prediction error, we bound arm j's imputation error by $4\beta_T \sigma_{t,r}^2/\varepsilon$, where $\sigma_{t,r}^2$ is the posterior variance for the selected arm r. Summing these bounds uses $V_T := \sum_{t=1}^{T} \sum_{i \in S_t} \sigma_{t,i}^2$, the total pre-observation variance on selected arms. Posterior updates bound this sum logarithmically in T.

SPO+ evaluates p_t using the completed costs $\bar{c}_t$ and adjusted costs $2p_t - \bar{c}_t$. Let H_t contain contested missing arms with $\sigma_{t,i} \geq \varepsilon$ that are selected by exactly one of these two decisions. We bound their

cumulative absolute imputation error relative to $\tilde{c}_t$, normalized by n. The bound scales linearly in $\kappa q \beta_T$, up to logarithmic factors.

Theorem 1 (Local imputation control)**.** *Under this setup, use UCB with any $\beta > 0$, posterior-mean completion, and thresholds $\varepsilon > 0$ and $0 \le \tau_t \le \beta\varepsilon/2$. Whenever the confidence bounds hold for every arm and round,*

$$\frac{1}{n}\sum_{t=1}^{T}\sum_{i\in H_t}|\hat{c}_{t,i}-\tilde{c}_{t,i}| \le \frac{8\alpha_{\lambda,L}\beta_T\kappa q}{\varepsilon}\log\left(1+\frac{kTL^2}{nq\lambda}\right).$$

For fixed ε, confidence level, and parameters, this local imputation contribution grows only polylogarithmically with T. The proof sums the per-arm bounds above, accounting for multiplicity κ, and applies the logarithmic bound on V_T (Appendix B.6). The same argument extends to bounded exchanges that remove and add at most m arms, with κ replaced by $m\kappa_m$ (Appendix B.4). On grid shortest paths, where one-edge swap is not feasible, this covers high-width missing edges in two-for-two detours within tolerance τ_t. Combining Theorem 1 with SPO+ convexity [13] and projected OGD [30] yields the complete training-loss bound in Appendix B.7, which also accounts for the reference loss with missing costs replaced by their conditional means, changes in the target-optimal decision, and residual imputation effects below ε or outside the contested set. Crucially, observations of selected arms control this local contribution without requiring regular observation of every arm.

6 Experiments and Results

We use the experiments to answer two questions. First, on online DFL benchmarks, does BAYESIANSPO reduce regret relative to prediction-focused baselines? Second, does BAYESIANSPO learn decision-focused representations under semi-bandit feedback?

Benchmarks. We consider three combinatorial structures: **Knapsack** at n=32 and n=100; **Shortest path** (SP) on 8×8 and 20×20 grids (n=112 and 760 edges); and **Portfolio optimization** with n=32 assets under a cardinality-k=5 constraint. All benchmarks come from PyEPO [25]; we replace PyEPO's continuous portfolio with a cardinality-constrained binary version such that semi-bandit feedback is well-defined (selected assets are observed, unselected ones are not; Appendix C). We convert each offline PtO instance into a T-round online problem: each round the environment draws $x_t \sim \mathcal{N}(0, I_5)$ and hidden coefficients c_t; the agent observes only x_t, predicts coefficients, selects $w_t \in \mathcal{W}$, and then observes only the selected coefficients $\{c_{t,i} : i \in \mathrm{supp}(w_t)\}$.

Additionally, we adapt the benchmark of Mandi et al. [20] to online learning using historical Irish SEMO (Single Electricity Market Operator) electricity prices. Each round is one day, with eight calendar and system predictors for each of 48 half-hour slots (384 inputs). These include forecast wind, system load, ex-ante electricity price, and CO_2 intensity. The learner predicts the cost of starting each task at each allowed time. A solver schedules 20 tasks on five machines to minimize total predicted cost, subject to task windows and machine capacities. During replay, we reveal only the 20 selected costs to simulate semi-bandit feedback and use the full vector to evaluate regret. We use the first 394 days for preprocessing and tuning, then evaluate online on the remaining 395 days.

Methods and protocol. We compare BAYESIANSPO against linear bandits (LinUCB, Thompson sampling); neural bandits (ε-Greedy, NeuralUCB, BLL-ONLINE); and BAYESIANMSE, which uses the same last layer head, exploration and imputation as BAYESIANSPO, but trains the shared representation with masked MSE instead of masked SPO+ (Appendix D.1). We report additional design ablations that vary decoupling, masking, surrogate choice, exploration, and posterior completion (Appendix E). For all instances in Table 2, all tunable methods, including M for BAYESIANSPO, use 50 Optuna trials on 3 tuning seeds and disjoint evaluation seeds. We run T=10,000 rounds over 10 seeds and report the realized cumulative regret $\mathcal{R}_{\mathrm{real}}(T)$, which captures both early exploration and late policy quality [18]. Significance uses a one-sided Wilcoxon signed-rank test at $p < 0.05$.

Results. Table 2 shows the main comparison. The two Bayesian last-layer methods, BAYESIANSPO and BAYESIANMSE achieve lower cumulative regret than other baselines. Their shared architecture combines analytical heads, UCB exploration, and posterior imputation with a linear feature map. The main difference between the two methods is the decision-focused adaptation of the shared

prior toward input directions that matter for the optimization problem via the representation loss. BAYESIANSPO has the lowest mean cumulative regret on all six benchmarks and is significantly better on five. The largest gain is on portfolio; knapsack and SP 20×20 show smaller gains, and SP 8×8 is essentially tied.

BAYESIANSPO also has the lowest mean regret among the nonlinear Bayesian last-layer methods (Appendix D.8), and pseudo-regret favors the same method as realized regret in all 50 seed-benchmark pairs (Appendix D.9). On the energy scheduling problem (Figure 4), BAYESIANSPO reaches the lowest final regret of 92.81 million compared to 101.14 for BAYESIANMSE, a statistically significant 8.2% reduction (one-sided paired Wilcoxon $p = 0.0031$).

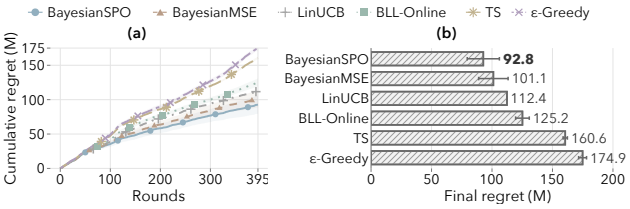


Figure 4: Energy scheduling task. (a) Cumulative regret over 395 days. (b) Final regret. Means shown in millions (cost units). Bands/Error bars show ±1 s.d.

Table 2: Cumulative regret after $10,000$ rounds (mean ± s.d. over 10 seeds). Best in **bold**, * marks significance over BAYESIANMSE. Knapsack and SP values in K.

Method	Knapsack 32	Knapsack 100	SP 8×8	SP 20×20	Portfolio
EpsilonGreedy	149.6 ± 25.7	436.0 ± 20.3	49.8 ± 3.1	173.7 ± 6.3	782.3 ± 60.4
LinUCB	35.4 ± 5.0	99.4 ± 8.1	30.0 ± 1.2	92.4 ± 1.5	357.2 ± 27.9
TS	38.6 ± 4.1	99.1 ± 9.8	29.6 ± 1.2	91.3 ± 1.6	391.0 ± 61.3
NeuralUCB	75.5 ± 6.3	224.1 ± 6.8	13.6 ± 1.7	62.2 ± 3.1	731.0 ± 43.0
BLL-Online	19.9 ± 2.4	66.9 ± 7.2	8.6 ± 0.9	51.4 ± 2.4	341.6 ± 41.9
BayesianMSE	18.5 ± 1.5	54.6 ± 4.8	6.5 ± 1.1	26.2 ± 2.6	101.5 ± 9.8
BayesianSPO	$\mathbf{17.3}^* \pm \mathbf{1.7}$	$\mathbf{49.6}^* \pm \mathbf{4.2}$	$\mathbf{6.4 \pm 1.0}$	$\mathbf{22.9}^* \pm \mathbf{2.2}$	$\mathbf{68.6}^* \pm \mathbf{3.4}$

Decision-relevant diagnostic. To separate aggregate performance from the learned representation, we use a four-arm task whose oracle selects one arm from $\{0, 1\}$ and one arm from $\{2, 3\}$. Only two input coordinates (x_0, x_1) determine the oracle action; the remaining (x_2, x_3, x_4) shift costs without changing the decision. Under full feedback, DFL observes all four costs each round and can learn both latent decision margins. Under semi-bandit feedback, only the two selected arms are observed, so many pairwise comparisons are missing. The task tests whether the learned representation tracks decision-relevant variation rather than nuisance cost variation. Figure 5 shows how each method allocates its two-dimensional representation capacity on the controlled four-arm diagnostic. In this run, under full feedback, BAYESIANSPO recovers the oracle geometry. Under semi-bandit feedback, BAYESIANMSE allocates a larger share of input-row norm to nuisance features (x_2, x_3, x_4) that reduce cost prediction error without affecting the decision. In contrast, BAYESIANSPO allocates more capacity to decision-relevant features. Appendix C.4 examines this comparison across ten training seeds and four representation widths.

7 Conclusion

We identify *imputation degeneracy* as a major obstacle to DFL under semi-bandit feedback and address it by decoupling prediction from decision-focused representation learning. Empirically, BAYESIANSPO achieves the lowest mean cumulative regret on all six instances. Under semi-bandit feedback, our controlled diagnostic shows that it places more relative weight on decision-relevant inputs.

Limitations. (i) Theory: Our regret bounds assume a fixed feature map and linear conditional means in those coordinates, while the learner updates its representation online. The synthetic benchmarks use nonlinear conditional means. Moreover, the training loss analysis holds the Bayesian heads and completed costs fixed within each representation update, and the global-coverage refinement

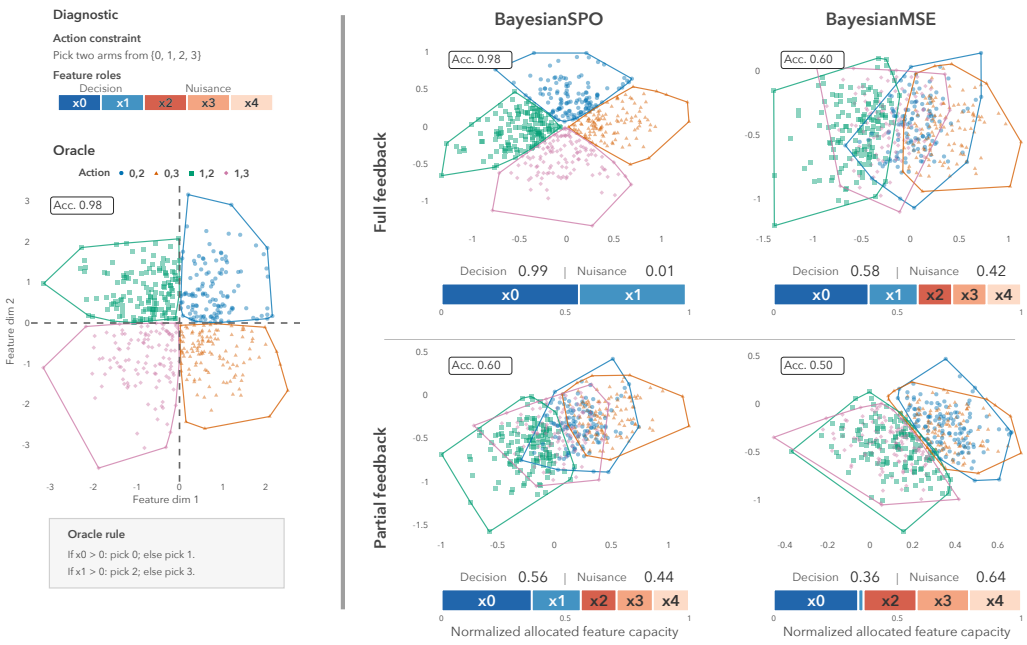


Figure 5: Decision-relevant representations. Left column shows oracle geometry; grid shows learned 2D representations, grouped by oracle action. Bars show input-row ℓ_2 norms. Under full feedback, BAYESIANSPO recovers geometry while BAYESIANMSE allocates capacity to nuisance features. Under partial feedback in this run, masked BAYESIANSPO assigns 56% of the total row norm to decision inputs, versus 36% for BAYESIANMSE.

requires each arm to be observed in a positive fraction of the rounds when its uncertainty is high. (ii) Method: Exploration uses standard UCB scores $\hat{c}_i - \beta\sigma_i$ to balance predicted cost against uncertainty. However, uncertainty in individual costs need not translate into uncertainty about which feasible decision is optimal, and the exploration rule does not explicitly prioritize observations by their usefulness for the masked SPO+ update. (iii) Computation: BAYESIANSPO requires additional solver calls and, when $M > 1$, for each posterior completion. In online PtO, however, the solver is already called every round to act, making these incremental additions rather than the epoch-scale solver burden of offline DFL.

Future work. Online semi-bandit DFL opens a broader agenda around recovering counterfactual decision signal without relying on self-imputation. Future work may look at optimization geometry to guide exploration, action selection or identify counterfactual values needed for a reliable gradient. Beyond our strictly online setting, replaying past observations through the current representation and recomputing the Bayesian heads could help keep posterior estimates aligned with changing features.

Acknowledgments and Disclosure of Funding

A.D. acknowledges support from the project "KI-Regio—Artificial Intelligence for Regional Value Chains", co-funded by the European Union through the European Regional Development Fund (ERDF) and by the Free State of Bavaria.
M.F. acknowledges funding from the Flemish Government under "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" (`https://www.flandersairesearch.be/en`).

The authors declare no competing interests.

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, volume 24, 2011.

[2] Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International Conference on Machine Learning*, pages 127–135, 2013.

[3] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019.

[4] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008. doi: 10.1007/s10994-007-5040-8.

[5] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002.

[6] Omar Bennouna, Amine Bennouna, Saurabh Amin, and Asuman Ozdaglar. What data enables optimal decisions? An exact characterization for linear optimization. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[7] Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. Learning with differentiable perturbed optimizers. In *Advances in Neural Information Processing Systems*, volume 33, pages 9508–9519, 2020.

[8] Dimitris Bertsimas and Nathan Kallus. From predictive to prescriptive analytics. *Management Science*, 66 (3):1025–1044, 2020.

[9] Aymeric Capitaine, Maxime Haddouche, Eric Moulines, Michael I. Jordan, Etienne Boursier, and Alain Durmus. Online decision-focused learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=FJhtHBphCt`.

[10] Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pages 151–159, 2013.

[11] Richard Combes, Mohammad Sadegh Talebi Mazraeh Shahi, Alexandre Proutiere, and Marc Lelarge. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

[12] Priya Donti, Brandon Amos, and J Zico Kolter. Task-based end-to-end model learning in stochastic optimization. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[13] Adam N Elmachtoub and Paul Grigas. Smart "Predict, then Optimize". *Management Science*, 68(1):9–26, 2022.

[14] Yichun Hu, Nathan Kallus, Xiaojie Mao, and Yanchen Wu. Contextual linear optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL `https://papers.neurips.cc/paper_files/paper/2024/hash/2f7001684feb23ca6569352eda963f39-Abstract-Conference.html`.

[15] Yichun Hu, Nathan Kallus, Xiaojie Mao, and Yanchen Wu. Contextual linear optimization with partial feedback. arXiv preprint arXiv:2405.16564v3, 2025. URL `https://arxiv.org/abs/2405.16564v3`. Version 3, November 10, 2025.

[16] Taehyun Hwang, Kyuwook Chai, and Min-hwan Oh. Combinatorial neural bandits. In *International Conference on Machine Learning*, pages 14203–14236, 2023.

[17] Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543, 2015.

[18] Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[19] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pages 661–670, 2010.

[20] Jayanta Mandi, Emir Demirović, Peter J Stuckey, and Tias Guns. Smart predict-and-optimize for hard combinatorial optimization problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 1603–1610, 2020. doi: 10.1609/aaai.v34i02.5521.

[21] Jayanta Mandi, James Kotary, Senne Berden, Maxime Mulamba, Victor Bucarey, Tias Guns, and Ferdinando Fioretto. Decision-focused learning: Foundations, state of the art, benchmark and future opportunities. *Journal of Artificial Intelligence Research*, 80:1623–1701, August 2024. doi: 10.1613/jair.1.15320.

[22] James Oxley. *Matroid Theory*. Oxford University Press, 2nd edition, 2011.

[23] Carlos Riquelme, George Tucker, and Jasper Snoek. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. In *International Conference on Learning Representations*, 2018.

[24] Daniel Russo and Benjamin Van Roy. Learning to optimize via information-directed sampling. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[25] Bo Tang and Elias B. Khalil. PyEPO: A PyTorch-based end-to-end predict-then-optimize library for linear and integer programming. *Mathematical Programming Computation*, 16(3):297–335, 2024.

[26] Marin Vlastelica, Anselm Paulus, Vít Musil, Georg Martius, and Michal Rolínek. Differentiation of blackbox combinatorial solvers. In *International Conference on Learning Representations*, 2020.

[27] Noah Weber, Janez Starc, Arpit Mittal, Roi Blanco, and Lluís Màrquez. Optimizing over a Bayesian Last Layer. In *Third Workshop on Bayesian Deep Learning, NeurIPS*, 2018. URL `https://bayesiandeeplearning.org/2018/papers/117.pdf`.

[28] Bryan Wilder, Bistra Dilkina, and Milind Tambe. Melding the data-decisions pipeline: Decision-focused learning for combinatorial optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1658–1665, 2019.

[29] Dongruo Zhou, Lihong Li, and Quanquan Gu. Neural contextual bandits with UCB-based exploration. In *Proceedings of the 37th International Conference on Machine Learning*, pages 11492–11502, 2020.

[30] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML)*, pages 928–936, 2003.

Appendix Contents

A BayesianSPO algorithm **14**
B Proofs **16**
B.1 Setup, notation, and assumptions 16
B.2 UCB exploration regret 18
B.3 SPO+ flip-set geometry 20
B.4 Decision-relevant UCB coverage via local exchanges 20
B.5 Detached masked-SPO+ comparator regret 22
B.6 Local imputation sensitivity and comparator residuals 23
B.7 Local and global training-loss bounds 25
B.8 Extensions and specializations 28
B.9 Lower bounds 28
B.10 Imputation degeneracy in a two-path problem 31
B.11 Structural separation: exact cardinality vs. shortest path 32
C Benchmark problems, data generation, and counterfactual structure **35**
C.1 Knapsack 35
C.2 Shortest path 35
C.3 Portfolio optimization 36
C.4 Decision-relevant representations task 36
C.5 Summary and instance hyperparameters 37
C.6 Connection to the formal structural separation 38
D Implementation Details and Extended Results **40**
D.1 Baselines 40
D.2 Reproducibility. 40
D.3 Hyperparameter Tuning Details 41
D.4 Regret trajectories 42
D.5 Regret-vs-horizon scaling 42
D.6 Compute and Wall-Clock Time 44
D.7 Where the gains come from 44
D.8 Nonlinear feature extractors 46
D.9 Pseudo-regret and realized regret 47
D.10 Energy scheduling with historical prices 48
E Core Design Ablations **49**
E.1 Architecture: Bayesian vs. deterministic head 49
E.2 Coupled vs. decoupled head update 50
E.3 Masking Ablations 51
E.4 Learning Objective 52
E.5 Exploration Design 52
E.6 Imputation Design 53
F Sensitivity Sweeps **57**
F.1 Capacity-constrained feedback-density sweep 57
F.2 β (UCB scale) 58
F.3 λ (prior precision) 58
F.4 DGP noise width w_0 58

A BayesianSPO algorithm

Algorithm 1 uses an explicit objective-sign convention. Let $s_{\mathrm{obj}} = -1$ for cost minimization, with $w^\star(c) = \mathrm{argmin}_{w \in \mathcal{W}} c^\top w$, and $s_{\mathrm{obj}} = +1$ for reward maximization, with $w^\star(c) = \mathrm{argmax}_{w \in \mathcal{W}} c^\top w$.

In each round $t = 1, \ldots, T$, the agent receives a context $x_t \in \mathbb{R}^d$, selects a feasible action $w_t \in \mathcal{W} \subseteq \{0,1\}^n$ over n arms, and observes per-arm coefficients $\{c_{t,i} : i \in S_t\}$ on the selected support $S_t = \mathrm{supp}(w_t)$. The feature extractor maps x_t to $h_t = \phi_\psi(x_t) \in \mathbb{R}^q$ via parameters ψ, where q is the feature dimension. Each arm $i \in [n]$ has an independent Bayesian linear head with sufficient statistics (A_i, v_i), posterior mean $\hat{\theta}_i = A_i^{-1} v_i$, predicted coefficient $\hat{c}_{t,i} = h_t^\top \hat{\theta}_i$, and posterior standard deviation $\sigma_{t,i} = \sqrt{h_t^\top A_i^{-1} h_t}$.

Let ℓ_t denote the SPO+ loss normalized by n and averaged over M completions, with heads, completed targets, and missing prediction coordinates held fixed during the representation update. The pseudocode shows a plain masked gradient step with learning-rate schedule η_t; the experiments use Adam with global gradient clipping (Appendix D.3). The completion count $M \geq 1$ is tuned per dataset: $M = 1$ uses deterministic posterior-mean completion, while $M > 1$ averages that mean completion with $M - 1$ posterior draws.

Algorithm 1 BAYESIANSPO with masked SPO+ representation updates and posterior imputation

Require: Feasible set $\mathcal{W}$, solver $w^\star$ for the chosen objective sense, objective sign $s_{\mathrm{obj}} \in \{-1, +1\}$, horizon T, prior precision λ, exploration scale β, completions $M \geq 1$, learning-rate schedule η_t.

Initialize feature parameters ψ.
for each arm $i \in [n]$ **do**
  $A_i \leftarrow \lambda I_q, \quad A_i^{-1} \leftarrow \lambda^{-1} I_q, \quad v_i \leftarrow 0.$
end for
for $t = 1, \ldots, T$ **do**
  Observe context x_t and compute $h_t = \phi_\psi(x_t)$.
  for each arm $i \in [n]$ **do**
    $\hat{\theta}_i \leftarrow A_i^{-1} v_i.$
    $\hat{c}_{t,i} \leftarrow h_t^\top \hat{\theta}_i.$
    $\sigma_{t,i} \leftarrow \sqrt{h_t^\top A_i^{-1} h_t}.$
  end for
  Select $w_t \leftarrow \mathrm{argmin}_{w \in \mathcal{W}} \sum_i w_i(\hat{c}_{t,i} - \beta \sigma_{t,i})$ for cost minimization; use $\mathrm{argmax}_{w \in \mathcal{W}} \sum_i w_i(\hat{c}_{t,i} + \beta \sigma_{t,i})$ for reward maximization.
  Observe $\{c_{t,i} : i \in S_t\}$ with $S_t = \mathrm{supp}(w_t) = \{i \in [n] : w_{t,i} = 1\}$.
  for each observed arm $i \in S_t$ **do**
    $A_i^{-1} \leftarrow A_i^{-1} - \dfrac{A_i^{-1} h_t h_t^\top A_i^{-1}}{1 + h_t^\top A_i^{-1} h_t}.$
    $A_i \leftarrow A_i + h_t h_t^\top.$
    $v_i \leftarrow v_i + c_{t,i} h_t.$
  end for
  For each $i \in S_t$, recompute $\hat{\theta}_i \leftarrow A_i^{-1} v_i$ and overwrite $\hat{c}_{t,i} \leftarrow h_t^\top \hat{\theta}_i$ for training; unobserved-arm values are unchanged.
  Initialize $\nabla_{h_t} \ell_t \leftarrow 0$.
  for $m = 1, \ldots, M$ **do**
    Construct completed cost vector $\bar{c}_t^{(m)} \in \mathbb{R}^n$ by setting, for each $i \in [n]$,

$$\bar{c}_{t,i}^{(m)} \leftarrow \begin{cases} c_{t,i}, & i \in S_t, \\ \hat{c}_{t,i}, & i \notin S_t,\ m = 1, \\ \hat{c}_{t,i} + \sigma_{t,i} \xi_i^{(m)},\ \xi_i^{(m)} \sim \mathcal{N}(0,1), & i \notin S_t,\ 2 \leq m \leq M. \end{cases}$$

    Compute normalized SPO+ arm-gradient: $g_t^{(m)} \leftarrow \frac{2 s_{\mathrm{obj}}}{n} \Big(w^\star(2\hat{c}_t - \bar{c}_t^{(m)}) - w^\star(\bar{c}_t^{(m)}) \Big).$
    Mask and accumulate: $\nabla_{h_t} \ell_t \leftarrow \nabla_{h_t} \ell_t + \sum_i g_{t,i}^{(m)} \mathbf{1}\{i \in S_t\}\, \hat{\theta}_i.$
  end for
  Update feature extractor: $\nabla_{h_t} \ell_t \leftarrow M^{-1} \nabla_{h_t} \ell_t; \quad \psi \leftarrow \psi - \eta_t \nabla_\psi \phi_\psi(x_t)^\top \nabla_{h_t} \ell_t.$
end for

Algorithm 1 reuses mutable prediction variables. Definition 3 (Appendix B.5) uses $\bar{f}_t(W_t)$ for the post-observation, masked training predictions, keeping $\hat{c}_t$ for the pre-action predictions.

Posterior statistics under online representation updates. For the regret analysis, we hold the feature map fixed across rounds. For the representation-learning analysis, we hold the Bayesian heads and completed costs fixed during each representation update. The implementation updates the feature map online without recomputing past feature vectors, so the stored per-arm statistics (A_i, v_i) combine features computed by different versions of the map. Replaying past observations through the current feature map and rebuilding these statistics would remove this stale-coordinate approximation at additional computational cost. The architecture ablation in Appendix E.1 evaluates the decoupled Bayesian head empirically.

Only observed costs enter the posterior sufficient statistics (A_i, v_i). During the SPO+ step, the current posterior coefficients are fixed linear maps, and backpropagation updates only the feature-extractor parameters ψ. With a linear extractor $h_t = W_1^\top x_t + b_1$, the last step is $\nabla_{W_1} = x_t(\nabla_{h_t}\ell_t)^\top$ and $\nabla_{b_1} = \nabla_{h_t}\ell_t$. For $M > 1$, we average the posterior-mean completion gradient with $M - 1$ posterior-sampled completion gradients. Including the posterior-mean gradient generally makes this a biased estimate of the expected gradient under posterior sampling.

B Proofs

The analysis has two branches. For fixed realizable features, standard linear-UCB confidence and potential arguments bound decision pseudo-regret (Section B.2) [1]. For representation learning, we hold each round's post-observation heads, completed costs, and missing prediction coordinates fixed. SPO+ convexity [13, Proposition 3] and projected OGD [30, Theorem 1] then bound regret relative to a fixed comparison representation (Section B.5). These are separate guarantees for fixed-feature decisions and detached training loss.

Our argument connects the comparison representation's loss to imputation error. Feasible exchanges can be used to relate uncertainty on contested missing arms to uncertainty on observed arms (Section B.4). A target-perturbation bound identifies how those missing costs affect the SPO+ loss (Section B.6). We combine the local bound with OGD to obtain the full training-loss bound in Section B.7, then give a refinement under additional observation coverage in Section B.7.1. Extensions for posterior-completion averaging and Gaussian contexts follow in Section B.8.

Section B.9 gives information-theoretic lower bounds on decision pseudo-regret. Section B.10 gives a two-path counterexample for posterior-mean imputation. Section B.11 develops the exact-cardinality and matroid results and compares their feasible exchanges with those of grid paths.

B.1 Setup, notation, and assumptions

We restate the round protocol from Section 3 and list the assumptions used in the proofs.

Let $\mathcal{F}_{t-1}$ be the history through round $t-1$, generated by past contexts, actions, observed costs, and the learner's past random choices. Throughout, $\lambda > 0$, $\mathcal{W} \subseteq \{0,1\}^n$ is nonempty, and $k = \max_{w \in \mathcal{W}} \|w\|_0$. The case $k = 0$ has a single zero action and zero decision regret and SPO+ loss, positive-gap specializations concern $k \geq 1$.

Round protocol. In each round $t \in [T]$:

1. The environment reveals context $x_t \in \mathbb{R}^d$.
2. The model computes latent representation $h_t = \phi_{\psi_t}(x_t) = W_{1,t}^\top x_t + b_{1,t} \in \mathbb{R}^q$ from $\mathcal{F}_{t-1}$-measurable parameters $\psi_t = \{W_{1,t}, b_{1,t}\}$.
3. Per-arm Bayesian linear regression provides predictions and uncertainties. Superscripts $-$ and $+$ distinguish coefficients before and after observing the round's selected costs:

$$A_{t,i} = \lambda I_q + \sum_{s<t,\, i \in S_s} h_s h_s^\top,$$
$$v_{t,i} = \sum_{s<t,\, i \in S_s} c_{s,i}\, h_s, \qquad \hat{\theta}^-_{t,i} = A_{t,i}^{-1} v_{t,i},$$
$$\hat{c}_{t,i} = h_t^\top \hat{\theta}^-_{t,i}, \qquad \sigma_{t,i} = \sqrt{h_t^\top A_{t,i}^{-1} h_t}.$$

   These sums use observations from rounds $s < t$, available before action selection. The current round is added after its selected costs are revealed.
4. UCB action selection (minimization convention): $w_t = \operatorname{argmin}_{w \in \mathcal{W}} \sum_i w_i(\hat{c}_{t,i} - \beta\sigma_{t,i})$.
5. The environment reveals $c_{t,i}$ only for the selected arms,

$$S_t := \operatorname{supp}(w_t) = \{i \in [n] : w_{t,i} = 1\}, \qquad |S_t| \leq k := \max_{w \in \mathcal{W}} \|w\|_0.$$

   Here k is the largest number of arms any feasible action can select.
6. Update the sufficient statistics and posterior coefficients:

$$A^+_{t,i} = A_{t,i} + \mathbf{1}\{i \in S_t\} h_t h_t^\top, \qquad v^+_{t,i} = v_{t,i} + \mathbf{1}\{i \in S_t\} c_{t,i} h_t, \qquad \hat{\theta}^+_{t,i} = (A^+_{t,i})^{-1} v^+_{t,i}.$$

   Freeze $\hat{\theta}^+_{t,i}$ during the representation-gradient step on the detached masked SPO+ loss (Definition 3), which updates ψ_t to ψ_{t+1}. For $i \notin S_t$, $\hat{\theta}^+_{t,i} = \hat{\theta}^-_{t,i}$.

Cost model. In the proof, each realized cost is decomposed into a predictable mean and a random fluctuation:

$$c_{t,i} = \theta_i^{*\top} \phi^*(x_t) + \epsilon_{t,i}.$$

The fluctuation $\epsilon_{t,i}$ has conditional mean zero given x_t. Under the fixed-feature abstraction, the learner's coordinates identify the latent feature system, and we write this mean/noiseless cost as

$$\tilde{c}_{t,i} := \mathbb{E}[c_{t,i} \mid x_t] = \theta_i^{*\top} \phi^*(x_t).$$

For the fixed-feature analysis, we identify $h_t = \phi^*(x_t)$, so $\tilde{c}_{t,i} = h_t^\top \theta_i^*$ and Assumption 4 bounds the coefficient norm in these same coordinates. A fixed invertible change to features Rh_t changes the coefficients to $R^{-\top}\theta_i^*$; the bound S must then apply to those transformed coefficients. Outside this abstraction, the environment mean depends on ϕ^*.

Pseudo-regret (primary theoretical object).

$$\tilde{\mathcal{R}}(T) = \sum_{t=1}^{T} \big[\tilde{c}_t^\top w_t - \min_{w \in \mathcal{W}} \tilde{c}_t^\top w\big].$$

This compares the learner to the action that is optimal for the conditional mean costs. It avoids the realized-oracle noise floor in $\mathcal{R}_{\mathrm{real}}(T)$, where $\min_w c_t^\top w$ can exploit favorable noise realizations. The theorem statements below are therefore pseudo-regret statements.

The realized benchmark regret $\mathcal{R}_{\mathrm{real}}(T)$ is used for empirical comparison in simulation, where the full realized cost vector is available after the fact. In general, $\mathcal{R}_{\mathrm{real}}(T) - \tilde{\mathcal{R}}(T)$ is not a zero-mean martingale term. The realized oracle $\mathrm{argmin}_w c_t^\top w$ can select actions with favorable noise, creating a positive noise floor even for a learner that acts optimally for $\tilde{c}_t$. We therefore keep the theorem in pseudo-regret and use realized regret only for empirical comparison.

Confidence assumptions. The fixed-feature confidence results use bounded features, realizability, conditional sub-Gaussian noise, and bounded true coefficients.

Assumption 1 (Bounded features)**.** *$\|h_t\| \le L$ for all t.*

Assumption 2 (Realizability)**.** *$\exists\, W_1^*, b_1^*$ with $\phi^*(x) \equiv W_1^{*\top} x + b_1^*$.*

Assumption 3 (Sub-Gaussian noise)**.** *Let $\mathcal{H}_t$ denote the information available before the round's costs are revealed, comprising the history $\mathcal{F}_{t-1}$, current context x_t, and selected arms S_t. The features h_t and selected arms S_t are known given this information. For every $i \in [n]$ and $a \in \mathbb{R}$,*

$$\mathbb{E}[\epsilon_{t,i} \mid \mathcal{H}_t] = 0, \qquad \mathbb{E}[\exp(a\epsilon_{t,i}) \mid \mathcal{H}_t] \le \exp(a^2\sigma^2/2).$$

Conditioning on the history provides the martingale-noise condition required by the confidence bound.

Assumption 4 (Bounded parameters)**.** *$\|\theta_i^*\| \le S$ for all i.*

Representation-update assumptions. The OGD results use bounded contexts and frozen posterior coefficients, together with a bounded projected domain.

Assumption 5 (Bounded augmented contexts)**.** *For $x_t^+ = (x_t^\top, 1)^\top$, $\|x_t^+\| \le B_x$ for all t.*

Assumption 6 (Bounded detached posterior coefficients)**.** *During the detached representation-gradient step, the frozen post-observation coefficients satisfy $\|\hat{\theta}_{t,i}^+\| \le B_\theta$. The constant B_θ may be read either as an explicit projection/regularity assumption or as a horizon-dependent high-probability radius (Remark 3).*

Assumption 7 (Bounded feature-update domain)**.** *For the OCO statement, append the bias as the final row of the feature matrix:*

$$W = \begin{bmatrix} W_1 \\ b_1^\top \end{bmatrix} \in \mathbb{R}^{(d+1)\times q}, \qquad h_t = W_t^\top x_t^+.$$

The iterates lie in a compact convex set $\mathcal{P}_T$ with Frobenius diameter D_W, meaning $\|W - W'\|_F \le D_W$ for all $W, W' \in \mathcal{P}_T$. The analyzed update is projected OGD. The experiments use unprojected Adam updates with gradient clipping. The bound-domain projection is part of the surrogate analysis.

Additional assumptions (used only where stated):

Assumption 8 (Observation coverage)**.** *At the horizon T, there exist $\varepsilon > 0$ and $p_{\min} \in (0, 1]$ such that, for every arm i,*

$$\sum_{t=1}^{T} \mathbf{1}\{i \in S_t,\ \sigma_{t,i} \geq \varepsilon\} \ \geq\ p_{\min} \sum_{t=1}^{T} \mathbf{1}\{\sigma_{t,i} \geq \varepsilon\}.$$

We call an arm i high-width *in round t if its posterior standard deviation $\sigma_{t,i}$ is at least ε. Among these rounds, each arm must be observed at least a fraction $p_{\min}$ of the time. This fraction is measured over the full horizon T, so initial rounds without observations are allowed.*

Remark 1 (Role of global coverage)**.** *The main theorem bounds errors on contested missing arms through the uncertainty of selected arms they can feasibly replace. Assumption 8 is an optional stronger condition for the global refinement in Corollary 3.*

Assumption 9 (Uniform context conditioning)**.** *For a specified tolerance $\varepsilon > 0$, there exists $T_{\mathrm{cond}} \leq T$ such that for $t \geq T_{\mathrm{cond}}$ and every arm i, $\sigma_{t,i}^2 \leq \varepsilon^2$ uniformly over h_t in the reachable feature set. From T_{cond} onward, every arm's uncertainty stays within the tolerance for all reachable contexts. We use this condition to prove eventual target-optimizer agreement (Corollary 4).*

B.2 UCB exploration regret

We isolate exploration cost by treating ψ as fixed throughout this subsection.

Lemma 1 (Per-arm self-normalized confidence bound)**.** *Under Assumptions 1–4 with a fixed feature map realizing the conditional means, for any $\delta \in (0, 1)$, there is an event $\mathcal{E}_T$ of probability $\geq 1 - \delta$, on which, for all $t \leq T$ and $i \in [n]$,*

$$\|\hat{\theta}_{t,i}^{-} - \theta_i^*\|_{A_{t,i}} \leq \beta_T, \qquad \|\hat{\theta}_{t,i}^{+} - \theta_i^*\|_{A_{t,i}^{+}} \leq \beta_T,$$

where

$$\beta_T = \sigma\sqrt{2\log\tfrac{n}{\delta} + q\log\!\big(1 + \tfrac{TL^2}{q\lambda}\big)} + \sqrt{\lambda}\, S.$$

Here σ is the sub-Gaussian scale of observation noise, while $\sigma_{t,i}$ is the posterior standard deviation. In particular, $|\hat{c}_{t,i} - \tilde{c}_{t,i}| \leq \beta_T \sigma_{t,i}$ for the pre-action prediction.

Proof. For each arm, apply the time-uniform self-normalized inequality of Abbasi-Yadkori et al. [1], Theorem 1, to masked features $\mathbf{1}\{i \in S_s\}h_s$ and observed masked noises $\mathbf{1}\{i \in S_s\}\epsilon_{s,i}$. The interleaved pre-cost and post-cost filtration makes the designs predictable and the masked noises adapted, with the conditional bound from Assumption 3. At any prefix, write $V = \lambda I + \sum_{s:i\in S_s} h_s h_s^\top$ and $Z = \sum_{s:i\in S_s} h_s \epsilon_{s,i}$. Ridge regression gives

$$\|\hat{\theta} - \theta_i^*\|_V = \|Z - \lambda\theta_i^*\|_{V^{-1}} \leq \|Z\|_{V^{-1}} + \sqrt{\lambda}\, S.$$

The self-normalized bound controls $\|Z\|_{V^{-1}}$ by $\sigma\sqrt{2\log(n/\delta) + \log(\det V/\det(\lambda I))}$. Both the pre-action and post-observation statistics are prefixes with at most T observations, so the same event covers both, including $A_{T,i}^{+}$. The determinant AM–GM inequality bounds the log ratio by $q\log(1 + TL^2/(q\lambda))$. A union bound over the n arms gives the time-uniform event $\mathcal{E}_T$. Finally, Cauchy–Schwarz gives $|h_t^\top(\hat{\theta}_{t,i}^{-} - \theta_i^*)| \leq \sigma_{t,i}\|\hat{\theta}_{t,i}^{-} - \theta_i^*\|_{A_{t,i}}$. □

Exploration-radius convention. The decision-regret bounds use $\beta \geq \beta_T$ and $\beta > 0$. The confidence event $\mathcal{E}_T$ itself holds for any nonanticipating action rule. For posterior-mean completion, it bounds imputation errors by $\beta_T\sigma_{t,i}$, independently of the UCB scale. The local imputation theorem requires only $\beta > 0$ and $\tau_t \leq \beta\varepsilon/2$ for the exchange argument. For $M > 1$, use the completion radius γ_T of Remark 10 instead of β_T.

Lemma 2 (Per-round pseudo-regret)**.** *For $\beta \geq \beta_T$ with $\beta > 0$, on the confidence event of Lemma 1,*

$$\tilde{c}_t^\top w_t - \tilde{c}_t^\top w^\star(\tilde{c}_t) \ \leq\ 2\beta \sum_{i\in S_t} \sigma_{t,i}.$$

Proof. The confidence event gives $\tilde{c}_{t,i} \in [\hat{c}_{t,i} - \beta\sigma_{t,i},\ \hat{c}_{t,i} + \beta\sigma_{t,i}]$. By UCB optimism, $\sum_i w_{t,i}(\hat{c}_{t,i} - \beta\sigma_{t,i}) \le \sum_i w^\star(\tilde{c}_t)_i(\hat{c}_{t,i} - \beta\sigma_{t,i})$. Combining,

$$\tilde{c}_t^\top w_t - \tilde{c}_t^\top w^\star(\tilde{c}_t) \le 2\beta \sum_{i \in S_t} \sigma_{t,i}. \qquad \square$$

Lemma 3 (Observed variance budget)**.** *Under Assumption 1 and the covariance updates in the round protocol, let $\alpha_{\lambda,L} := \max\{1, L^2/\lambda\}$. With $A_{T+1,i}$ including all observations through round T,*

$$V_T := \sum_{t=1}^{T} \sum_{i \in S_t} \sigma_{t,i}^2 \le 2\alpha_{\lambda,L} \sum_{i=1}^{n} \log \frac{\det A_{T+1,i}}{\det(\lambda I_q)} \le 2\alpha_{\lambda,L} nq \log\left(1 + \frac{kTL^2}{nq\lambda}\right).$$

Proof. Since $\sigma_{t,i}^2 \le L^2/\lambda$, we have $\sigma_{t,i}^2 \le \alpha_{\lambda,L} \min\{1, \sigma_{t,i}^2\}$. The standard elliptical-potential lemma bound [1, Lemma 11] therefore gives $\sum_{t: i \in S_t} \sigma_{t,i}^2 \le 2\alpha_{\lambda,L} \log(\det A_{T+1,i}/\det(\lambda I_q))$ for each arm. Write $N_i = \sum_t \mathbf{1}\{i \in S_t\}$. The trace bound and AM–GM give $\log(\det A_{T+1,i}/\det(\lambda I_q)) \le q \log(1 + N_i L^2/(q\lambda))$. Summing over arms and using concavity of log with $\sum_i N_i \le kT$ proves the aggregate bound. $\square$

Remark 2 (Using the observed variance budget)**.** *The budget V_T sums posterior variances on observed arms. Feasible exchanges relate this budget to contested missing arms (Section B.4). The global bound extends control across all arms under observation coverage (Assumption 8 and Theorem 5).*

Theorem 2 (Fixed-feature UCB, standard consequence)**.** *Under Assumptions 1–4 with features fixed realizing $\tilde{c}_{t,i} = h_t^\top \theta_i^*$ and UCB radius $\beta = \beta_T > 0$, with probability $\ge 1 - \delta$,*

$$\tilde{\mathcal{R}}(T) \ \le\ O\left(\beta_T \sqrt{\alpha_{\lambda,L} knT\, q \log\bigl(1 + kTL^2/(nq\lambda)\bigr)}\right).$$

Proof. By Lemma 2, $\tilde{\mathcal{R}}(T) \le 2\beta \sum_t \sum_{i \in S_t} \sigma_{t,i}$. Cauchy–Schwarz with $|S_t| \le k$ gives

$$\sum_t \sum_{i \in S_t} \sigma_{t,i} \le \sqrt{kTV_T}$$

Substituting the aggregate bound on V_T from Lemma 3 and the full confidence radius $\beta = \beta_T$ from Lemma 1, including the regularization-bias term $\sqrt{\lambda}S$, yields the displayed bound. $\square$

Expected regret and horizon dependence. To compare with expected lower bounds, let $U_T(\delta)$ denote the upper bound in Theorem 2 at confidence $1 - \delta$. Under the fixed-feature assumptions $\|h_t\| \le L$ and $\|\theta_i^*\| \le S$, we have $|\tilde{c}_{t,i}| \le LS$ and $0 \le \tilde{\mathcal{R}}(T) \le 2kLST$. Splitting the expectation over the confidence event and its complement gives

$$\mathbb{E}[\tilde{\mathcal{R}}(T)] \le U_T(\delta) + 2kLST\,\delta.$$

For $T \ge 2$, taking $\delta = 1/T$ in the confidence radius keeps $\beta_T = O(\sqrt{\log T})$ and yields an expected $\widetilde{O}(\sqrt{T})$ upper bound with all other problem parameters fixed. The Gaussian instances in Appendix B.9 force $\Omega(\sqrt{T})$ expected regret, so the horizon dependence is worst-case optimal up to logarithmic factors.

Remark 3 (Posterior coefficient scale)**.** *The regularized initialization and observation updates in the round protocol give*

$$A_{t,i}^+ = \lambda I_q + \sum_{s \le t,\, i \in S_s} h_s h_s^\top \succeq \lambda I_q,$$

because each outer product $h_s h_s^\top$ is positive semidefinite. Under the fixed-feature assumptions of Lemma 1, its confidence event therefore gives

$$\sqrt{\lambda}\,\|\hat{\theta}_{t,i}^+ - \theta_i^*\| \le \|\hat{\theta}_{t,i}^+ - \theta_i^*\|_{A_{t,i}^+} \le \beta_T.$$

The triangle inequality then yields $\|\hat{\theta}_{t,i}^+\| \le S + \beta_T/\sqrt{\lambda}$. Thus Assumption 6 holds with $B_\theta(T) = S + \beta_T/\sqrt{\lambda}$ on that event, including the final post-observation update at $t = T$.

B.3 SPO+ flip-set geometry

Definition 1 (SPO+ gradient)**.** *Given predictions $\hat{c}$, target c, and objective sign $s_{\mathrm{obj}} \in \{-1, +1\}$ (-1 for cost minimization, $+1$ for reward maximization), set*

$$w^{\mathrm{spo}} = w^\star(2\hat{c} - c), \quad w^{\mathrm{opt}} = w^\star(c), \quad g_i = \frac{2s_{\mathrm{obj}}}{n}\big(w_i^{\mathrm{spo}} - w_i^{\mathrm{opt}}\big) \in \tfrac{2s_{\mathrm{obj}}}{n} \cdot \{-1, 0, +1\}.$$

The factor $1/n$ normalizes the SPO+ subgradient by the number of arms. At round t, evaluate this formula with a single completion $\bar{c}_t$ and post-observation predictions $h_t^\top \hat{\theta}_{t,i}^+$, before masking. The flip set is $D_t = \{i : g_{t,i} \neq 0\}$. When $M > 1$, the flip-set statements apply separately to each completion before gradient averaging.

Lemma 4 (Flip-set cardinality)**.** *For $\mathcal{W} \subseteq \{0,1\}^n$ with $\|w\|_0 \leq k$ for all $w \in \mathcal{W}$, $|D_t| \leq 2k$.*

Proof. $D_t = \mathrm{supp}(w^{\mathrm{spo}}) \triangle \mathrm{supp}(w^{\mathrm{opt}})$ is the symmetric difference of two sets of size at most k, hence at most $2k$ elements. □

Lemma 5 (Decision stability)**.** *Let $u = w^\star(c)$ under the objective sign in Definition 1, and define the signed decision gap $\Delta(c) := \min_{w \in \mathcal{W}\setminus\{u\}} s_{\mathrm{obj}} c^\top (u - w)$, with $\min \varnothing := +\infty$. For any perturbation v and alternative $w \in \mathcal{W} \setminus \{u\}$,*

$$s_{\mathrm{obj}}(c + v)^\top (u - w) \geq s_{\mathrm{obj}} c^\top (u - w) - \|v\|_\infty \|u - w\|_1.$$

If the right-hand side is positive for every alternative, then $w^\star(c + v) = u$. For binary actions with support size at most $k \geq 1$, it suffices that $2k\|v\|_\infty < \Delta(c)$.

Proof. Expand the left-hand side and use $|v^\top(u - w)| \leq \|v\|_\infty \|u - w\|_1$. Positivity makes u strictly better than every alternative. The binary specialization follows from $\|u - w\|_1 \leq 2k$. □

SPO+ flip collapse. Taking $v = 2(p - c)$ gives $w^\star(2p - c) = w^\star(c)$ and a zero SPO+ subgradient whenever $\|p - c\|_\infty < \Delta(c)/(4k)$. Thus the generic collapse constant is $\chi_{\mathcal{W}} = 4k$.

B.4 Decision-relevant UCB coverage via local exchanges

A feasible exchange relates a missing arm's uncertainty to the uncertainty of a selected arm. Squaring and summing these relations bounds uncertainty on contested missing arms by the observed variance budget V_T.

Definition 2 (Local exchanges and contested arms)**.** *Let $e_i \in \mathbb{R}^n$ denote the ith standard coordinate vector. In the minimization convention, a pair (r, j) is a feasible local exchange if $r \in S_t$, $j \notin S_t$, and $w_t - e_r + e_j \in \mathcal{W}$. Its estimated exchange gap is the predicted cost difference $\hat{c}_{t,j} - \hat{c}_{t,r}$.*

For a tolerance $\tau_t \geq 0$, let $\mathcal{C}_t(\tau_t)$ be the set of missing arms $j \notin S_t$ that admit a feasible exchange (r, j) with $\hat{c}_{t,j} - \hat{c}_{t,r} \leq \tau_t$. These are the contested arms defined in Section 5, the set is independent of the current SPO+ gradient and the comparator. Only its intersection with the comparator flip set contributes to the loss bound below. For each such j, choose one selected arm $r_t(j)$ satisfying these exchange and cost conditions, and define

$$\kappa = \max_{t, r \in S_t} |\{j \in \mathcal{C}_t(\tau_t) : r_t(j) = r\}|.$$

Set $\kappa = 0$ if no arms are contested in any round.

With only an exact-cardinality constraint, any selected item can be exchanged for any missing item. On the directed grids considered here, changing routes requires exchanging multiple edges, so no one-edge exchange is feasible (Section B.11).

Lemma 6 (Local UCB exchange charge)**.** *Let $\beta > 0$. If (r, j) is a feasible local exchange and w_t is the UCB action, then*

$$\beta(\sigma_{t,j} - \sigma_{t,r}) \leq \hat{c}_{t,j} - \hat{c}_{t,r}.$$

Consequently, if $\varepsilon > 0$, if $\hat{c}_{t,j} - \hat{c}_{t,r} \leq \tau_t \leq \beta\varepsilon/2$ and $\sigma_{t,j} \geq \varepsilon$, then $\sigma_{t,j} \leq 2\sigma_{t,r}$ and $\sigma_{t,r} \geq \varepsilon/2$. If this tolerance condition holds in every round, then

$$\sum_{t=1}^{T} \sum_{\substack{j \in \mathcal{C}_t(\tau_t) \\ \sigma_{t,j} \geq \varepsilon}} \sigma_{t,j}^2 \leq 4\kappa V_T.$$

Proof. Compare the UCB action w_t with the feasible exchange $w_t - e_r + e_j$. Optimality and cancellation of unchanged arms give

$$\hat{c}_{t,r} - \beta\sigma_{t,r} \le \hat{c}_{t,j} - \beta\sigma_{t,j},$$

which rearranges to the displayed inequality. Under the threshold conditions,

$$\sigma_{t,j} \le \sigma_{t,r} + \varepsilon/2 \le \sigma_{t,r} + \sigma_{t,j}/2,$$

so $\sigma_{t,j} \le 2\sigma_{t,r}$. Each selected arm occurs as $r_t(j)$ for at most κ missing arms in a round. Squaring and summing therefore bounds the missing-arm sum by $4\kappa \sum_t \sum_{r\in S_t} \sigma_{t,r}^2 = 4\kappa V_T$. □

Count of high-width contested arms. Under the same threshold conditions, let $\Gamma_T(\varepsilon,\tau) := \sum_{t=1}^{T} \sum_{j\in\mathcal{C}_t(\tau_t)} \mathbf{1}\{\sigma_{t,j} \ge \varepsilon\}$ count qualifying arm-round occurrences, including repeats. Each contributes at least ε^2 to the squared-width sum. With Assumption 1, Lemmas 6 and 3 give, deterministically,

$$\Gamma_T(\varepsilon,\tau) \le \frac{4\kappa V_T}{\varepsilon^2} \le \frac{8\alpha_{\lambda,L}\kappa nq}{\varepsilon^2} \log\left(1 + \frac{kTL^2}{nq\lambda}\right). \tag{6}$$

Remark 4 (Interpreting the multiplicity κ)**.** *The multiplicity depends on the realized trajectory and the choices of $r_t(j)$. Since j ranges only over missing arms, $\kappa \le n$, or $\kappa \le n-k$ when every action has exactly k selected arms. The latter value is attainable in the full exact-cardinality family when many replacements are tied near the boundary. Any uniform bound on the number of contested replacements available to a selected arm also bounds κ.*

Bounded-size exchanges. Fix an integer $m \ge 1$. A feasible bounded exchange consists of sets $R \subseteq S_t$ and $J \subseteq [n] \setminus S_t$ with $1 \le |R|, |J| \le m$ such that

$$w_t - \sum_{r\in R} e_r + \sum_{j\in J} e_j \in \mathcal{W}.$$

Its aggregate predicted gap is $\widehat{\Delta}_t(R,J) := \sum_{j\in J} \hat{c}_{t,j} - \sum_{r\in R} \hat{c}_{t,r}$. Let $\mathcal{C}_t^{(m)}(\tau_t)$ contain the missing arms j belonging to at least one such addition set J with $\widehat{\Delta}_t(R,J) \le \tau_t$. For each $j \in \mathcal{C}_t^{(m)}(\tau_t)$, choose one qualifying exchange $(R_t(j), J_t(j))$ containing j, and define

$$\kappa_m := \max_{t,r\in S_t} \left|\{j \in \mathcal{C}_t^{(m)}(\tau_t) : r \in R_t(j)\}\right|,$$

with $\kappa_m = 0$ if every contested set is empty. This counts how many missing arms use the same selected arm in their chosen removal sets, including when several missing arms use the same exchange.

Lemma 7 (Bounded-exchange UCB charge)**.** *Let w_t be the UCB action with $\beta > 0$. Every feasible bounded exchange obeys*

$$\beta\left(\sum_{j\in J} \sigma_{t,j} - \sum_{r\in R} \sigma_{t,r}\right) \le \widehat{\Delta}_t(R,J).$$

If $\varepsilon > 0$ and $0 \le \tau_t \le \beta\varepsilon/2$ in every round, then

$$\sum_{t=1}^{T} \sum_{\substack{j\in\mathcal{C}_t^{(m)}(\tau_t)\\ \sigma_{t,j}\ge\varepsilon}} \sigma_{t,j}^2 \le 4m\kappa_m V_T.$$

Consequently, under posterior-mean completion and the confidence event of Lemma 1, any subsets $H_t^{(m)} \subseteq \mathcal{C}_t^{(m)}(\tau_t) \cap \{j : \sigma_{t,j} \ge \varepsilon\}$ satisfy

$$\frac{1}{n}\sum_{t=1}^{T} \sum_{j\in H_t^{(m)}} |\hat{c}_{t,j} - \tilde{c}_{t,j}| \le \frac{4\beta_T m\kappa_m}{n\varepsilon} V_T. \tag{7}$$

Proof. Compare w_t with the exchanged action under the objective $\hat{c}_t - \beta\sigma_t$. Cancelling unchanged coordinates gives the first inequality. For each contested j with $\sigma_{t,j} \geq \varepsilon$, its chosen exchange therefore satisfies

$$\sigma_{t,j} \leq \sum_{\ell \in J_t(j)} \sigma_{t,\ell} \leq \sum_{r \in R_t(j)} \sigma_{t,r} + \varepsilon/2 \leq \sum_{r \in R_t(j)} \sigma_{t,r} + \sigma_{t,j}/2.$$

Thus $\sigma_{t,j} \leq 2\sum_{r \in R_t(j)} \sigma_{t,r}$. Cauchy–Schwarz and $|R_t(j)| \leq m$ give

$$\sigma_{t,j}^2 \leq 4m \sum_{r \in R_t(j)} \sigma_{t,r}^2.$$

Each selected arm occurs in at most κ_m chosen removal sets when counted once per contested missing arm. Summing proves the variance bound. On the confidence event, $|\hat{c}_{t,j} - \tilde{c}_{t,j}| \leq \beta_T \sigma_{t,j} \leq \beta_T \sigma_{t,j}^2/\varepsilon$ for $j \in H_t^{(m)}$, which proves (7). □

Under Assumption 1, substituting the observed variance budget from Lemma 3 gives the logarithmic bound of Theorem 1 with κ replaced by $m\kappa_m$. For $m = 1$, the same witness choices recover Lemma 6.

Example 1 (High-width arms outside the contested set)**.** *In top-1 selection with two arms, either arm can replace the other. A sufficiently high predicted cost for arm 2 can exceed its UCB uncertainty bonus, meaning UCB may never select it even while its width is large. Such an arm can lie outside the contested set.*

B.5 Detached masked-SPO+ comparator regret

Definition 3 (Detached masked SPO+ loss)**.** *At round t, hold the post-observation heads, completed costs, and missing prediction values fixed while varying W. For the posterior-mean completion analyzed in Theorem 1, reuse the completed cost vector from Section 4:*

$$\bar{c}_{t,i} = \begin{cases} c_{t,i}, & i \in S_t, \\ \hat{c}_{t,i}, & i \notin S_t, \end{cases}$$

Using the augmented context x_t^+, define the masked prediction map by

$$\bar{f}_{t,i}(W) = \begin{cases} (x_t^+)^\top W \hat{\theta}_{t,i}^+, & i \in S_t \\ \hat{c}_{t,i}, & i \notin S_t \end{cases}$$

Here W_t is the feature parameter before the representation step. Missing heads are unchanged by the observation, so their frozen predictions equal the pre-action $\hat{c}_{t,i}$ used in the confidence and exchange arguments. Holding these quantities fixed is precisely StopGrad; only selected prediction coordinates depend on W. The detached masked loss is

$$\bar{L}_t(W) = \frac{1}{n} \ell_{\mathrm{SPO}+}(\bar{f}_{t,W}, \bar{c}_t).$$

The factor $1/n$ matches the per-arm gradient normalization in Definition 1. The loss computation uses the completed vector from the current forward pass, with missing predictions held constant for backpropagation. The masked SPO+ update is the exact subgradient of this masked-prediction loss in W. For $M > 1$, condition on the mean-anchored completed vectors and replace $\bar{L}_t$ by the normalized average $\frac{1}{nM}\sum_{m=1}^{M} \ell_{\mathrm{SPO}+}(\bar{f}_{t,W}, \bar{c}_t^{(m)})$, keeping every missing prediction coordinate detached. The OCO gradient bound below is unchanged. The width-based imputation bounds require the stated completion error inequality for every completion being averaged.

Projected OGD updates the representation using this loss:

$$W_{t+1} = \Pi_{\mathcal{P}_T}\left(W_t - \eta_t \nabla_W \bar{L}_t(W_t)\right),$$

where $\Pi_{\mathcal{P}_T}$ is Frobenius projection onto the domain in Assumption 7, and η_t is the step size.

Lemma 8 (Loss bound from observed prediction errors)**.** *For posterior-mean completion ($M = 1$), let $p_t = \bar{f}_t(W)$ be the masked predictions from Definition 3. For every W,*

$$0 \leq \ell_{\mathrm{SPO}+}(p_t, \bar{c}_t) \leq 2 \sum_{i \in S_t} |p_{t,i} - c_{t,i}|.$$

Proof. Write $u = w^\star(\bar{c}_t)$ and $a = w^\star(2p_t - \bar{c}_t)$. In the minimization convention,

$$\ell_{\mathrm{SPO}+}(p_t, \bar{c}_t) = (2p_t - \bar{c}_t)^\top (u - a) = \bar{c}_t^\top (u - a) + 2(p_t - \bar{c}_t)^\top (u - a).$$

Optimality of a makes the loss nonnegative, and optimality of u gives $\bar{c}_t^\top (u - a) \le 0$. On missing arms, $p_{t,i} = \bar{c}_{t,i}$; on observed arms, $\bar{c}_{t,i} = c_{t,i}$. Since actions are binary, $|u_i - a_i| \le 1$, which gives the upper bound. □

Theorem 3 (Detached masked OGD, standard consequence)**.** *Under Assumptions 5, 6, and 7, let $G := 2kB_\theta B_x/n$. The projected detached update with step-size schedule $\eta_t = \eta_0/\sqrt{t+1}$, $\eta_0 > 0$, satisfies, for any fixed comparator $\bar{W} \in \mathcal{P}_T$,*

$$\sum_{t=1}^{T} \bar{L}_t(W_t) - \sum_{t=1}^{T} \bar{L}_t(\bar{W}) \;\le\; O\!\left(\left(\frac{D_W^2}{\eta_0} + \eta_0 G^2\right)\sqrt{T}\right).$$

The comparator $\bar{W}$ ranges over the domain $\mathcal{P}_T$. The augmented realizing matrix formed from W_1^*, b_1^* is one possible choice when it belongs to this domain. For $D_W, G > 0$, choosing $\eta_0 = D_W/G$ gives $O(D_W G\sqrt{T})$. When $G = 0$, take a zero representation step; the losses are constant on the convex domain and comparator regret is zero. If $D_W = 0$, the domain is a singleton and the same conclusion holds.

Proof. The detached prediction map is affine in W, so SPO+ convexity makes $\bar{L}_t$ convex [13, Proposition 3]. At any W, evaluate g_t at $\bar{f}_t(W)$ using the frozen post-observation heads. The chain rule for the augmented matrix W, including its bias row, gives

$$\nabla_W \bar{L}_t(W) = x_t^+ \Big(\sum_{i \in S_t} g_{t,i} \hat{\theta}_{t,i}^+\Big)^\top,$$

$$\|\nabla_W \bar{L}_t(W)\|_F \le \sum_{i \in S_t} |g_{t,i}|\, \|\hat{\theta}_{t,i}^+\|\, \|x_t^+\| \le \frac{2kB_\theta B_x}{n} = G.$$

Here $|S_t| \le k$ and $|g_{t,i}| \le 2/n$ by Definition 1; the triangle inequality gives the same bound for $M > 1$ averaged completions. The masked update is therefore a valid subgradient step with norm at most G. Apply Zinkevich [30, Theorem 1] on the projected domain of diameter D_W. Balancing its two terms with $\eta_0 = D_W/G$ gives the stated specialization. □

We compare representations using the same Bayesian heads and missing-arm imputations. The reference loss also reflects any mismatch between its feature coordinates and the learner's heads. For posterior-mean completion ($M = 1$), combining the balanced OGD bound with Lemma 8 gives, with $p_t = \bar{f}_t(\bar{W})$,

$$\sum_{t=1}^{T} \bar{L}_t(W_t) \le O(D_W G\sqrt{T}) + \frac{2}{n} \sum_{t=1}^{T} \sum_{i \in S_t} |p_{t,i} - c_{t,i}|.$$

Remark 5 (Horizon dependence of the gradient bound)**.** *Taking $B_\theta(T) = S + \beta_T/\sqrt{\lambda}$ from Remark 3 makes $G = 2kB_\theta(T)B_x/n$ horizon-dependent. Both the general OGD bound and its balanced $O(D_W G\sqrt{T})$ specialization retain that dependence.*

B.6 Local imputation sensitivity and comparator residuals

The observed-error bound controls cumulative training loss. To isolate the effect of imputation on the reference's loss, we compare posterior-mean completion with a target that uses conditional means on missing arms. The difference has two parts: imputation errors on coordinates where the SPO+ decisions disagree, and changes in the target-optimal decision.

Lemma 9 (SPO+ target perturbation with optimizer switches)**.** *Write $\mathcal{L}(p, c) = \ell_{\mathrm{SPO}+}(p, c)/n$ in the minimization convention, with a fixed deterministic optimizer tie rule. For any prediction p, reference target c^0, and completed target $\bar{c} = c^0 + e$, set*

$$u = w^\star(\bar{c}), \qquad u^0 = w^\star(c^0), \qquad a = w^\star(2p - \bar{c}).$$

Then

$$\mathcal{L}(p, \bar{c}) \le \mathcal{L}(p, c^0) + \frac{e^\top (a - u)}{n} + \frac{(2p - c^0)^\top (u - u^0)}{n}.$$

Proof. Let $a^0 = w^\star(2p - c^0)$. Expanding the two losses gives

$$n\big[\mathcal{L}(p, \bar{c}) - \mathcal{L}(p, c^0)\big] = (c^0 - 2p)^\top (a - a^0) + e^\top (a - u) + (2p - c^0)^\top (u - u^0).$$

The first term is nonpositive because a^0 maximizes $(c^0 - 2p)^\top w$. Dropping it proves the inequality. The argument uses optimizer optimality and permits ties. □

A stable-target example. For one round, let c^0 keep observed costs and use conditional means on missing arms, while $\bar{c}$ uses posterior-mean completion. In the notation of Lemma 9, suppose $u = u^0$ and j is the only missing arm in $\mathrm{supp}(a - u)$. Let j have width $\sigma_j \geq \varepsilon > 0$ and suppose it can replace a selected arm r at a predicted cost increase of at most $\tau \leq \beta\varepsilon/2$, where $\beta > 0$ is the UCB scale. On the confidence event, the perturbation bound and Lemma 6 give

$$\mathcal{L}(p, \bar{c}) - \mathcal{L}(p, c^0) \leq \frac{|e_j|}{n} \leq \frac{\beta_T \sigma_j}{n} \leq \frac{4\beta_T \sigma_r^2}{n\varepsilon}.$$

The loss increase is bounded by observed-arm variance. Summing over rounds uses V_T; the full bound below also allows several missing disagreements and changes in the target-optimal decision.

Definition 4 (Reference target and comparator decisions)**.** *Fix a comparison representation $\bar{W} \in \mathcal{P}_T$ across all rounds and set $p_t = \bar{f}_t(\bar{W})$. Its predictions use the masked map in Definition 3. Form the reference target by retaining observed costs and replacing missing costs with their conditional means:*

$$c^0_{t,i} = \begin{cases} c_{t,i}, & i \in S_t, \\ \tilde{c}_{t,i}, & i \notin S_t, \end{cases} \qquad e_t = \bar{c}_t - c^0_t.$$

Thus e_t is zero on observed coordinates. With posterior-mean completion, the confidence event gives $|e_{t,i}| \leq \beta_T \sigma_{t,i}$ on missing coordinates. The completed target, reference target, and adjusted costs give the decisions

$$u_t = w^\star(\bar{c}_t), \qquad u^0_t = w^\star(c^0_t), \qquad a_t = w^\star(2p_t - \bar{c}_t).$$

The comparator's disagreement set is $D^*_t = \mathrm{supp}(a_t - u_t)$. It records the arms selected by exactly one of the two SPO+ decisions at the reference prediction p_t.

Which missing arms contribute? Only missing arms in D^*_t contribute to $e_t^\top(a_t - u_t)$. For $\varepsilon > 0$ and $\tau_t \geq 0$, divide them according to whether they are contested and whether their uncertainty exceeds the threshold:

$$H_t = D^*_t \cap \mathcal{C}_t(\tau_t) \cap \{i : \sigma_{t,i} \geq \varepsilon\},$$
$$J_t = D^*_t \cap \mathcal{C}_t(\tau_t) \cap \{i : \sigma_{t,i} < \varepsilon\},$$
$$K_t = (D^*_t \setminus S_t) \setminus \mathcal{C}_t(\tau_t).$$

Thus H_t contains contested missing arms with high uncertainty, J_t contains the remaining contested arms, and K_t contains missing disagreements outside the contested set. These sets partition $D^*_t \setminus S_t$. The sets and accumulated terms below depend on $\bar{W}$ and, where applicable, on ε and the tolerances τ_t; we suppress these arguments except in the comparator loss and switch term.

Accumulating the contributions. Because actions are binary, $|a_{t,i} - u_{t,i}| = 1$ on the disagreement set. Each missing coordinate contributes at most $|e_{t,i}|/n$. We will bound the sum over H_t through observed-arm uncertainty. Write the other two sums as

$$\mathcal{J}_T = \frac{1}{n} \sum_{t=1}^{T} \sum_{i \in J_t} |e_{t,i}|, \qquad \mathcal{K}_T = \frac{1}{n} \sum_{t=1}^{T} \sum_{i \in K_t} |e_{t,i}|.$$

Using $[a]_+ = \max\{a, 0\}$, write the reference loss and the contribution from changing the target-optimal decision as

$$\mathcal{A}_T(\bar{W}) = \sum_{t=1}^{T} \mathcal{L}(p_t, c^0_t),$$

$$\mathcal{S}_T(\bar{W}) = \frac{1}{n} \sum_{t=1}^{T} \big[(2p_t - c^0_t)^\top (u_t - u^0_t)\big]_+.$$

On the confidence event, each low-width error is at most $\beta_T\varepsilon$, so $\mathcal{J}_T \leq (\beta_T\varepsilon/n) \sum_t |J_t|$.

Theorem 4 (High-width local imputation sensitivity)**.** *Under Assumption 1, use posterior-mean completion and UCB with any $\beta > 0$, fix $\varepsilon > 0$ and suppose $0 \le \tau_t \le \beta\varepsilon/2$ for all t. On the confidence event of Lemma 1,*

$$\frac{1}{n}\sum_{t=1}^{T}\sum_{i\in H_t}|e_{t,i}| \le \frac{4\beta_T\kappa}{n\varepsilon}V_T \le \frac{8\alpha_{\lambda,L}\beta_T\kappa q}{\varepsilon}\log\left(1+\frac{kTL^2}{nq\lambda}\right).$$

Proof. For every $j \in H_t$, confidence, $\sigma_{t,j} \ge \varepsilon$, and Lemma 6 give

$$|e_{t,j}| \le \beta_T\sigma_{t,j} \le \frac{\beta_T}{\varepsilon}\sigma_{t,j}^2 \le \frac{4\beta_T}{\varepsilon}\sigma_{t,r_t(j)}^2.$$

Each selected arm occurs as $r_t(j)$ for at most κ missing arms per round. Sum over H_t and all rounds, then divide by n, to obtain $4\beta_T\kappa V_T/(n\varepsilon)$. Substituting the observed variance budget from Lemma 3 proves the logarithmic bound. □

Dependence on the realized trajectory. Under the same conditions, Cauchy–Schwarz applied to $\sum_t\sum_{j\in H_t}\sigma_{t,j}^2 \le 4\kappa V_T$ also gives

$$\frac{1}{n}\sum_{t=1}^{T}\sum_{i\in H_t}|e_{t,i}| \le \frac{2\beta_T}{n}\sqrt{\kappa V_T\sum_{t=1}^{T}|H_t|} \le \frac{2\beta_T}{n}\sqrt{\kappa\Gamma_T(\varepsilon,\tau)V_T}. \tag{8}$$

Here we use $|e_{t,j}| \le \beta_T\sigma_{t,j}$ and $\sum_t|H_t| \le \Gamma_T(\varepsilon,\tau)$.

Remark 6 (Choosing the uncertainty threshold)**.** *The threshold ε separates the contested contribution bounded by the theorem from $\mathcal{J}_T$. To keep low-width imputation errors below the decision-gap scale, suppose the conditional-mean decision gaps are at least $\Delta_{\min} > 0$ and choose*

$$\varepsilon = \frac{\Delta_{\min}}{4k\beta_T}, \qquad 0 \le \tau_t \le \frac{\beta\Delta_{\min}}{8k\beta_T}.$$

Substitution gives a local contribution of $O(\alpha_{\lambda,L}\beta_T^2\kappa kq\log(1+kTL^2/(nq\lambda))/\Delta_{\min})$. The local proof applies for any positive ε.

Remark 7 (Interpreting the comparator terms)**.** *The reference loss $\mathcal{A}_T$ includes observed-target noise and the comparison representation's prediction error through the frozen estimated heads. It can remain positive even when the true feature map is realizable.*

For posterior-mean completion ($M = 1$), the switch contribution is bounded by the reference's prediction errors on observed arms whose inclusion changes when imputed costs are replaced by conditional means:

$$\mathcal{S}_T(\bar{W}) \le \frac{2}{n}\sum_{t=1}^{T}\sum_{i\in S_t}|p_{t,i}-c_{t,i}|\,|u_{t,i}-u_{t,i}^0|.$$

To see this, expand $2p_t - c_t^0 = 2(p_t-\bar{c}_t) + (2\bar{c}_t - c_t^0)$. Optimality gives

$$\bar{c}_t^\top(u_t-u_t^0) \le 0, \qquad (c_t^0)^\top(u_t-u_t^0) \ge 0.$$

Thus $(2\bar{c}_t - c_t^0)^\top(u_t-u_t^0) \le 0$, while $p_t - \bar{c}_t$ vanishes on missing arms. Taking positive parts and bounding the remaining inner product by absolute values proves the inequality. Changes confined to missing arms therefore contribute zero to this term.

The local rate bounds the sum over H_t; the full training-loss bound retains $\mathcal{A}_T$, $\mathcal{S}_T$, $\mathcal{J}_T$, and $\mathcal{K}_T$.

B.7 Local and global training-loss bounds

Proof of Theorem 1. For posterior-mean completion, $e_{t,i} = \hat{c}_{t,i} - \tilde{c}_{t,i}$ on every missing arm. Apply Theorem 4. □

Corollary 1 (Full detached masked-SPO+ loss bound)**.** *Assume the confidence event of Lemma 1, the feature-norm bound L, and UCB with $\beta > 0$. For the detached masked loss $\bar{L}_t = \ell_{\mathrm{SPO}+}/n$, fix a comparator $\bar{W}$ in the projected feature domain and $\varepsilon > 0$. Under the bounded-domain and gradient assumptions, projected OGD with $\eta_t = D_W/(G\sqrt{t+1})$ (zero step if $G = 0$), posterior-mean completion, and $0 \le \tau_t \le \beta\varepsilon/2$ obeys, on the confidence event,*

$$\begin{aligned}\sum_{t=1}^{T} \bar{L}_t(W_t) \le O(D_W G\sqrt{T}) \\ + \frac{4\beta_T \kappa V_T}{n\varepsilon} \\ + \mathcal{J}_T + \mathcal{K}_T + \mathcal{A}_T(\bar{W}) + \mathcal{S}_T(\bar{W}).\end{aligned}$$

The local term can be replaced by the sharper trajectory-sensitive bound (8)*. The remaining terms are the reference loss ($\mathcal{A}_T$), changes in the target-optimal decision ($\mathcal{S}_T$), and low-width or nonlocal imputation effects ($\mathcal{J}_T$ and $\mathcal{K}_T$), defined in Section B.6. Remark 7 bounds the switch term using prediction errors on observed arms.*

Proof. Apply Lemma 9 at $(p_t, c_t^0, \bar{c}_t)$. Since e_t vanishes on observed arms and $a_t - u_t$ outside D_t^*, partition the remaining coordinates into H_t, J_t, K_t. Taking absolute values of coordinate errors and the positive part of the switch term gives

$$\sum_t \bar{L}_t(\bar{W}) \le \mathcal{A}_T(\bar{W}) + \mathcal{S}_T(\bar{W}) + \frac{1}{n}\sum_t \sum_{i \in H_t} |e_{t,i}| + \mathcal{J}_T + \mathcal{K}_T.$$

Bound the sum over H_t by Theorem 4 or (8), then add the OGD bound from Theorem 3. □

Bounded-exchange loss bound. Replace $\mathcal{C}_t(\tau_t)$ by $\mathcal{C}_t^{(m)}(\tau_t)$ in the partition of Section B.6 to define $H_t^{(m)}, J_t^{(m)}, K_t^{(m)}$ and their corresponding residual sums $\mathcal{J}_T^{(m)}, \mathcal{K}_T^{(m)}$. Under the same assumptions, the proof of Corollary 1 then applies with local term $4\beta_T m \kappa_m V_T/(n\varepsilon)$ by Lemma 7. The reference loss and switch term are unchanged. Missing disagreements outside the enlarged contested set remain in $\mathcal{K}_T^{(m)}$.

B.7.1 Global high-width sensitivity under coverage

Observation coverage assumption lets us bound high-width missing contributions across all arms. For $\varepsilon > 0$, define the high-width rounds and their count by

$$\mathcal{B}_\varepsilon = \{t \le T : \max_i \sigma_{t,i} \ge \varepsilon\}, \qquad B_T(\varepsilon) = |\mathcal{B}_\varepsilon|.$$

Split the missing-coordinate errors on the comparator disagreement set D_t^* into

$$\mathcal{I}_T^{\mathrm{high}} = \frac{1}{n}\sum_t \sum_{\substack{i \in D_t^* \setminus S_t \\ \sigma_{t,i} \ge \varepsilon}} |e_{t,i}|,$$

$$\mathcal{I}_T^{\mathrm{low}} = \frac{1}{n}\sum_t \sum_{\substack{i \in D_t^* \setminus S_t \\ \sigma_{t,i} < \varepsilon}} |e_{t,i}|.$$

Theorem 5 (Global coverage and imputation sensitivity)**.** *Under Assumptions 1 and 8 at threshold $\varepsilon > 0$,*

$$B_T(\varepsilon) \le \frac{V_T}{p_{\min}\varepsilon^2} \le \frac{2\alpha_{\lambda,L} nq \log(1 + kTL^2/(nq\lambda))}{p_{\min}\varepsilon^2}. \tag{9}$$

For posterior-mean completion, on the confidence event of Lemma 1, also

$$\mathcal{I}_T^{\mathrm{high}} \le \frac{\beta_T L}{n\sqrt{\lambda}} \min\left\{ \frac{(1-p_{\min})V_T}{p_{\min}\varepsilon^2},\ \min\{2k, n\} B_T(\varepsilon) \right\},$$

$$\mathcal{I}_T^{\mathrm{low}} \le \frac{\beta_T \varepsilon}{n} \sum_t |\{i \in D_t^* \setminus S_t : \sigma_{t,i} < \varepsilon\}|.$$

Proof. Each observed high-width occurrence contributes at least ε^2 to V_T. Coverage bounds each arm's total high-width count by $1/p_{\min}$ times its observed count, and its missing count by $(1 - p_{\min})/p_{\min}$ times that count. Summing over arms and bounding the union count gives the bound on $B_T(\varepsilon)$ and a total missing count of at most $(1 - p_{\min})V_T/(p_{\min}\varepsilon^2)$. Lemma 3 gives the logarithmic bound. Alternatively, at most $\min\{2k, n\}$ missing disagreements occur per round in $\mathcal{B}_\varepsilon$. Multiply either missing-count bound by $|e_{t,i}| \le \beta_T \sigma_{t,i} \le \beta_T L/\sqrt{\lambda}$ and divide by n. The low-width bound follows from $|e_{t,i}| \le \beta_T \varepsilon$. □

Corollary 2 (Target agreement)**.** *If c_t^0 has a unique optimizer and $2k\|e_t\|_\infty < \Delta(c_t^0)$, then $u_t = u_t^0$, so this round contributes zero to $\mathcal{S}_T$. For noiseless observed targets, $c_t^0 = \tilde{c}_t$. On the confidence event with posterior-mean completion and $\beta_T > 0$, if these targets have gaps at least $\Delta_{\min} > 0$, choosing $\varepsilon = \Delta_{\min}/(4k\beta_T)$ gives $u_t = u_t^0$ outside $\mathcal{B}_\varepsilon$.*

Proof. Apply Lemma 5 with $c = c_t^0$ and $v = e_t$. Outside $\mathcal{B}_\varepsilon$, confidence gives $2k\|e_t\|_\infty < 2k\beta_T\varepsilon = \Delta_{\min}/2$, proving the specialization. □

Remark 8 (High-width rounds and target agreement)**.** *The quantity $B_T(\varepsilon)$ counts high-width rounds wherever they occur in the sequence. Eventual agreement from a particular time T_{cond} instead uses the uniform conditioning in Assumption 9. Agreement sets the switch contribution to zero on those rounds. The gap specialization above uses noiseless targets because c_t^0 otherwise retains observed noise, which can change its optimizer and decision gap.*

Corollary 3 (Global-coverage refinement)**.** *Under Assumption 1, suppose each arm is observed in at least a $p_{\min}$ fraction of its rounds with width at least $\varepsilon > 0$. Use posterior-mean completion and suppose the confidence event holds. For the projected update under Assumptions 5, 6, and 7, take $\eta_t = D_W/(G\sqrt{t+1})$ (zero step when $G = 0$). Every fixed comparator $\bar{W} \in \mathcal{P}_T$ then satisfies*

$$\sum_t \bar{L}_t(W_t) \le O(D_W G\sqrt{T}) + \mathcal{A}_T(\bar{W}) + \mathcal{S}_T(\bar{W}) + \frac{\beta_T L(1 - p_{\min})V_T}{n\sqrt{\lambda}\, p_{\min}\varepsilon^2} + \mathcal{I}_T^{\mathrm{low}}.$$

Here $\mathcal{I}_T^{\mathrm{low}}$ covers missing disagreements below the width threshold. The high-width term covers all remaining missing disagreements and can be replaced by the bound using $B_T(\varepsilon)$ in Theorem 5 when smaller. For noiseless reference targets with gap at least $\Delta_{\min} > 0$, $\beta_T > 0$, and $\varepsilon = \Delta_{\min}/(4k\beta_T)$, the target optimizer is unchanged outside $\mathcal{B}_\varepsilon$, so the switch contribution is zero on those rounds.

Proof. Summing Lemma 9 bounds the comparator loss by $\mathcal{A}_T(\bar{W})$, $\mathcal{S}_T(\bar{W})$, $\mathcal{I}_T^{\mathrm{high}}$, and $\mathcal{I}_T^{\mathrm{low}}$. Apply Theorem 5 to the high-width term and add the OGD bound from Theorem 3. Corollary 2 gives the stated target agreement. □

Corollary 4 (Eventual target agreement under explicit conditioning)**.** *Use posterior-mean completion with confidence radius $\beta_T > 0$. Suppose observed targets are noiseless, the conditional-mean decision gaps are at least $\Delta_{\min} > 0$, and the confidence event holds. If Assumption 9 holds with $\varepsilon = \Delta_{\min}/(4k\beta_T)$, then $u_t = u_t^0$ for $t \ge T_{\mathrm{cond}}$, so the switch contribution is zero on those rounds.*

Proof. Conditioning and confidence imply $2k\|e_t\|_\infty \le 2k\beta_T\varepsilon = \Delta_{\min}/2 < \Delta_{\min}$. Apply Lemma 5 with $c = c_t^0$ and $v = e_t$. □

Remark 9 (Rates of the controlled contribution)**.** *For fixed feature bound L, confidence level δ, positive gap $\Delta_{\min}$, and remaining problem parameters, take $\varepsilon = \Delta_{\min}/(4k\beta_T)$ with $\beta_T > 0$. If coverage at these thresholds holds with $p_{\min}$ bounded below independently of T, then $B_T(\varepsilon) = O(\beta_T^2 \log T)$ and $\mathcal{I}_T^{\mathrm{high}} = O(\beta_T^3 \log T)$. Since $\beta_T = O(\sqrt{\log T})$, this high-width contribution grows polylogarithmically with T.*

B.8 Extensions and specializations

Remark 10 (Posterior-completion averaging in the theory)**.** *For $M > 1$, average the detached losses over the mean-anchored completions. The OGD bound is unchanged. Apply the target-perturbation lemma to each completion and average its reference loss, switch term, and coordinate-error contributions, using its own disagreement set. For the local bound, keep $\beta > 0$ for UCB and $0 \le \tau_t \le \beta\varepsilon/2$. If a completion-error radius $\gamma_T \ge 0$ satisfies $|\bar{c}_{t,i}^{(m)} - \tilde{c}_{t,i}| \le \gamma_T \sigma_{t,i}$ simultaneously for every missing entry and completion, the same charging proof bounds each local contribution, and hence their average, by $4\gamma_T \kappa V_T/(n\varepsilon)$. The global high- and low-width error bounds likewise use γ_T in place of β_T.*

For the Gaussian sampler in Algorithm 1, fix a sampling failure probability $\delta_s \in (0,1)$. Under the fixed-feature confidence assumptions, the completion-error bound holds with probability at least $1 - \delta - \delta_s$ for

$$\gamma_T = \beta_T + \sqrt{2\log\left(\frac{2nT(M-1)}{\delta_s}\right)}.$$

Apply the Gaussian tail bound $\Pr(|\xi| > a) \le 2e^{-a^2/2}$ to at most $nT(M-1)$ draws and intersect their simultaneous event with $\mathcal{E}_T$. For $M = 1$, take $\gamma_T = \beta_T$. UCB retains its own radius; its decision-regret guarantee still requires $\beta \ge \beta_T$. Target-agreement thresholds use $\gamma_T > 0$, giving $\varepsilon = \Delta_{\min}/(4k\gamma_T)$ under the same gap and noiseless-target assumptions. The observed-error bound in Lemma 8 remains specific to $M = 1$.

Remark 11 (Combining confidence and OGD)**.** *The detached OGD argument applies to any sequence of frozen heads and targets satisfying its norm bounds. To control the imputation contribution, we also use confidence calibrated in fixed coordinates, together with UCB exchanges for the local bound or observation coverage for the global bound.*

Remark 12 (Gaussian-context specialization)**.** *For $x_t \sim \mathcal{N}(0, I_d)$, fix a context failure probability $\zeta \in (0,1)$ and set*

$$u = \log(T/\zeta), \qquad B_{x,T} = \sqrt{1 + d + 2\sqrt{du} + 2u}.$$

The chi-squared tail bound $\Pr(\|x_t\|^2 > d + 2\sqrt{du} + 2u) \le e^{-u}$ and a union bound give $\max_{t \le T} \|x_t^+\| \le B_{x,T}$ with probability at least $1 - \zeta$. Let C_T be a deterministic bound on the operator norm of the augmented feature matrix. For projected iterates, one may take $C_T = \sup_{W \in \mathcal{P}_T} \|W\|_{\mathrm{op}}$; for a fixed map, use its norm. On the context event, $\|h_t\| \le L_T := C_T B_{x,T}$.

For the fixed-feature confidence bounds, apply the determinant-form self-normalized inequality under the original filtration, as in Lemma 1, then intersect its event with the context event. This intersection has probability at least $1 - \delta - \zeta$. On it, the feature envelope bounds the determinants and supplies the substitutions

$$L = L_T, \qquad \alpha_{\lambda,L} = \max\{1, L_T^2/\lambda\}, \qquad \beta_T = \beta_T(L_T), \qquad G = \frac{2kB_\theta(T)B_{x,T}}{n},$$

where $\beta_T(L_T)$ is the radius in Lemma 1 with L_T in place of L. In fixed coordinates, $B_\theta(T) = S + \beta_T(L_T)/\sqrt{\lambda}$. The observed variance budget satisfies $V_T \le 2\alpha_{\lambda,L} nq \log(1 + kTL_T^2/(nq\lambda))$. For $T \ge 2$, taking $\zeta = 1/T$ gives $B_{x,T} = O(\sqrt{d + \log T})$ and probability at least $1 - \delta - 1/T$. The confidence statements still require fixed realizable feature coordinates.

B.9 Lower bounds

The lower bounds use instances with different optimal actions but similar observation laws. The paired constructions here and in Appendix B.11 share the testing argument below.

Lemma 10 (KL divergence under semi-bandit feedback)**.** *Consider two instances ν, ν' with the same deterministic context sequence and feasible set. Costs are independent across arms and rounds, with Gaussian means $\mu_{t,i}, \mu'_{t,i}$ and common variance $\sigma^2 > 0$. For any common nonanticipating policy, the laws of its action-observation histories through round h satisfy*

$$\mathrm{KL}(\mathbb{P}_\nu^h \,\|\, \mathbb{P}_{\nu'}^h) = \frac{1}{2\sigma^2}\mathbb{E}_\nu\left[\sum_{t=1}^{h}\sum_{i \in S_t}(\mu_{t,i} - \mu'_{t,i})^2\right].$$

For constant means, this is $\sum_i (\mu_i - \mu'_i)^2 \mathbb{E}_\nu[N_i(h)]/(2\sigma^2)$, where $N_i(h) = \sum_{t \le h} \mathbf{1}\{i \in S_t\}$.

Proof. We adapt the divergence decomposition of Lattimore and Szepesvári [18, Lemma 15.1] to semi-bandit observations. At a fixed history the policy uses the same action kernel under both instances, contributing zero conditional KL. Given the action, the observed costs have product Gaussian laws, whose KL is the sum of the selected-arm squared mean differences divided by $2\sigma^2$. Summing and taking expectation under ν proves the identity. □

Two-instance testing. For probability laws P, Q on the same history space and any event E, Pinsker's and Bretagnolle–Huber's inequalities [18, Eq. (14.12) and Theorem 14.2] give

$$P(E) + Q(E^c) \geq \max\left\{1 - \sqrt{\frac{\mathrm{KL}(P\|Q)}{2}},\ \frac{1}{2}e^{-\mathrm{KL}(P\|Q)}\right\}. \tag{10}$$

Under a common nonanticipating policy, suppose the one-round regrets on instances ν^+, ν^- satisfy $r_t^+ \geq \Delta \mathbf{1}_{E_t}$ and $r_t^- \geq \Delta \mathbf{1}_{E_t^c}$, where $\Delta, \sigma > 0$ and E_t depends on the history and current action. If pre-action KL is at most $2(t-1)\Delta^2/\sigma^2$, then $H = \min\{T, 1 + \lfloor \sigma^2/(4\Delta^2) \rfloor\}$ keeps it at most $1/2$ through round H. Appending the action preserves KL, so (10) gives error-probability sum at least $1/2$. Summing and using nonnegative regret on later rounds gives

$$\max_{\nu \in \{\nu^+, \nu^-\}} \mathbb{E}_\nu[\tilde{\mathcal{R}}(T)] \geq \frac{\Delta H}{4} \geq \frac{1}{16}\min\left\{\Delta T, \frac{\sigma^2}{\Delta}\right\}, \tag{11}$$

since $H \geq \frac{1}{4}\min\{T, \sigma^2/\Delta^2\}$.

Theorem 6 (Exploration lower bound)**.** *Fix $L, S, \sigma > 0$ and integers $1 \leq q \leq T$ and $1 \leq k \leq n/2$. For every nonanticipating algorithm there is a contextual combinatorial semi-bandit instance with n arms, exactly k selected arms per round, fixed q-dimensional feature map with $\|h_t\| \leq L$, coefficients $\|\theta_i^*\| \leq S$, and independent Gaussian noise of variance σ^2, such that*

$$\mathbb{E}[\tilde{\mathcal{R}}(T)] \ \geq \frac{1}{16\sqrt{2}}\min\left\{\frac{LSkT}{\sqrt{q}},\ \sigma\sqrt{qknT}\right\}.$$

Writing $a = \lfloor n/k \rfloor$, the rate $\Omega(\sigma\sqrt{qknT})$ holds when $T \geq \sigma^2 q^2 a/(4L^2S^2)$.

Proof. Partition the arms into k groups G_g, each of size $a_g \in \{a, a+1\}$, and require one selection per group. Set

$$H = \lfloor T/q \rfloor, \qquad \Delta_{\exp} = \min\left\{\frac{LS}{\sqrt{q}},\ \frac{\sigma}{2}\sqrt{\frac{qa}{T}}\right\}.$$

Choose context dimension $d = q$, and let $e_\ell \in \mathbb{R}^q$ be the ℓ-th standard coordinate vector. Use q deterministic context blocks of H rounds each. In block ℓ, take $x_t = e_\ell$ and $h_t = Lx_t$. On the remaining $T - qH$ rounds take $x_t = h_t = 0$. This is a fixed linear feature map with bounded augmented contexts. Independently for each group–block pair (g, ℓ), choose a special arm $V_{g,\ell}$ uniformly in G_g. For $i \in G_g$ set

$$\theta^*_{i,\ell} = -\frac{\Delta_{\exp}}{L}\mathbf{1}\{i = V_{g,\ell}\}.$$

Thus $\|\theta_i^*\| \leq \Delta_{\exp}\sqrt{q}/L \leq S$. In each group and block, the special arm has mean cost $-\Delta_{\exp}$ and all others have mean zero. All noises are independent $\mathcal{N}(0, \sigma^2)$.

Fix a cell (g, ℓ) and all other special-arm labels. Let Q_i be the full trajectory law when this cell's special arm is i, and Q_0 the reference law setting all means in this cell to zero while leaving the other cells unchanged. Let N_i count selections of i in this block. By Lemma 10,

$$\mathrm{KL}(Q_0 \,\|\, Q_i) = \frac{\Delta_{\exp}^2}{2\sigma^2}\mathbb{E}_0[N_i].$$

Although observations in other cells can affect the policy, their conditional laws at a fixed history agree and contribute no KL. Since $0 \leq N_i \leq H$, Pinsker's inequality gives

$$\mathbb{E}_i[N_i] \leq \mathbb{E}_0[N_i] + \frac{H\Delta_{\exp}}{2\sigma}\sqrt{\mathbb{E}_0[N_i]}.$$

Under Q_0, $\sum_{i\in G_g} N_i = H$. Averaging over the uniform special-arm label and applying Jensen's inequality therefore yields

$$\frac{1}{a_g}\sum_{i\in G_g}\mathbb{E}_i[N_i] \le \frac{H}{a_g} + \frac{H\Delta_{\exp}}{2\sigma}\sqrt{\frac{H}{a_g}} \le \frac{3H}{4},$$

using $a_g \ge a \ge 2$, $qH \le T$, and the definition of $\Delta_{\exp}$. The expected regret in this cell is at least $\Delta_{\exp}H/4$ for every fixed assignment of the other labels. Average those labels under the same product prior and sum over all kq cells. Since $H = \lfloor T/q \rfloor$ and $T \ge q$, we have $qH \ge T/2$. The prior-averaged regret is therefore at least

$$\begin{aligned}\frac{kqH\Delta_{\exp}}{4} &\ge \frac{kT\Delta_{\exp}}{8}\\ &= \min\left\{\frac{LSkT}{8\sqrt{q}},\ \frac{\sigma\sqrt{qak^2T}}{16}\right\}\\ &\ge \frac{1}{16\sqrt{2}}\min\left\{\frac{LSkT}{\sqrt{q}},\ \sigma\sqrt{qknT}\right\}.\end{aligned}$$

The last inequality uses $a = \lfloor n/k \rfloor \ge n/(2k)$. At least one instance in the finite prior support attains this lower bound. Under the stated horizon condition, $\Delta_{\exp} = (\sigma/2)\sqrt{qa/T}$, giving the untruncated rate.

□

Theorem 7 (Shortest-path feedback lower bound)**.** *Fix $L, S, \sigma > 0$ and $0 < \Delta \le LS/2$. There are two four-arm shortest-path instances ν^+, ν^- with two edges per path, fixed features of norm at most L, parameter norms at most S, and independent Gaussian noise of variance σ^2, such that every nonanticipating semi-bandit algorithm and every integer $T \ge 1$ satisfy*

$$\max_{\nu\in\{\nu^+,\nu^-\}}\mathbb{E}_\nu[\tilde{\mathcal{R}}(T)] \ge \frac{1}{16}\min\left\{\Delta T, \frac{\sigma^2}{\Delta}\right\}.$$

Proof. Let the feasible paths be $a = (1,1,0,0)$ and $b = (0,0,1,1)$, with mean costs

$$\mu^+ = (\Delta, 0, 0, 2\Delta), \qquad \mu^- = (\Delta, 0, 0, 0).$$

Path a is optimal under ν^+ and b under ν^-, with gap Δ in both. A constant scalar feature $h = L$ and coefficients $\theta_i^* = \mu_i/L$ realize the means within the stated bounds. Let E_t denote selection of path b. The one-round mean regret is $\Delta\mathbf{1}_{E_t}$ under ν^+ and $\Delta\mathbf{1}_{E_t^c}$ under ν^-. Only edge 4 differs. Writing $N_b(t-1)$ for the number of bottom-path selections before round t, Lemma 10 gives

$$\mathrm{KL}(\mathbb{P}_{\nu^+}^{t-1} \,\|\, \mathbb{P}_{\nu^-}^{t-1}) = \frac{2\Delta^2}{\sigma^2}\mathbb{E}_{\nu^+}[N_b(t-1)] \le \frac{2(t-1)\Delta^2}{\sigma^2}.$$

Apply the shared cutoff bound (11). □

Theorem 8 (Two-arm minimax decision lower bound)**.** *Fix $L, S, \sigma > 0$. For every integer horizon $T \ge 1$ satisfying $\sigma/\sqrt{T} \le 2LS$ and every nonanticipating algorithm, there exists a two-arm Gaussian instance ν_T with fixed context, feature norm at most L, parameter norms at most S, and gap $\Delta_T = \sigma/\sqrt{T}$ such that, even with full feedback,*

$$\mathbb{E}_{\nu_T}[\tilde{\mathcal{R}}(T)] \ge \frac{\sigma\sqrt{T}}{8\mathrm{e}}.$$

Proof. Let the two cost means be $(\Delta/2, -\Delta/2)$ under ν^+ and $(-\Delta/2, \Delta/2)$ under ν^-. Costs are independent across arms and rounds, with Gaussian noise of variance σ^2. The context is constant, one arm is selected before each observation, and full feedback reveals both costs. The per-round KL is Δ^2/σ^2, so $\mathrm{KL}(\mathbb{P}_{\nu^+}^T \,\|\, \mathbb{P}_{\nu^-}^T) = T\Delta^2/\sigma^2$. Let A be the event that arm 1 is selected more than $T/2$ times. Regret is at least $\Delta T/2$ on A under ν^+ and on A^c under ν^-. The exponential bound in (10), multiplied by $\Delta T/4$, gives

$$\max\left\{\mathbb{E}_{\nu^+}[\tilde{\mathcal{R}}],\ \mathbb{E}_{\nu^-}[\tilde{\mathcal{R}}]\right\} \ge \tfrac{\Delta T}{8}\exp(-T\Delta^2/\sigma^2).$$

Choosing $\Delta = \Delta_T = \sigma/\sqrt{T}$ proves the stated bound on at least one of the two instances. These means are realized by a constant scalar feature $h = L$ and coefficients $\pm\Delta_T/(2L)$, whose magnitudes are at most S by the horizon condition. A constant context and affine feature bias give this realization. The hard-instance gap depends on T; this is a minimax lower bound over a class allowing such gaps. Semi-bandit feedback is only less informative. □

Theorem 9 (Combined lower bound, max form)**.** *Fix $L, S, \sigma > 0$, $1 \le q \le T$, $1 \le k \le n/2$, and $0 < \Delta_{\min} \le LS/2$, with $\sigma/\sqrt{T} \le 2LS$. Let $\mathfrak{I}$ contain every special-arm assignment in the exploration construction and both instances of each two-instance construction, with their respective dimensions and action sets. Every nonanticipating algorithm satisfies*

$$\sup_{\nu\in\mathfrak{I}} \mathbb{E}_\nu[\tilde{\mathcal{R}}(T)] \ge \Omega\Big(\max\big\{\, \min\{LSkT/\sqrt{q},\ \sigma\sqrt{qknT}\},\ \min\{\Delta_{\min}T,\ \sigma^2/\Delta_{\min}\},\ \sigma\sqrt{T}\big\}\Big).$$

Here n, k, q describe the exploration construction and $\Delta_{\min}$ specifies the shortest-path gap. The union $\mathfrak{I}$ retains each construction's own gaps, dimensions, and action sets.

Proof. Theorems 6, 7, 8 bound the supremum over $\mathfrak{I}$ on possibly different instances, giving their maximum. The shortest-path term saturates at order $\sigma^2/\Delta_{\min}$ once $T \ge \sigma^2/\Delta_{\min}^2$. □

B.10 Imputation degeneracy in a two-path problem

Consider the diamond-graph in Figure 6, with top path $a = (1, 1, 0, 0)$ and bottom path $b = (0, 0, 1, 1)$, two parallel two-edge paths between s and t. The selected top edges each have observed cost 1. Conditional on these observations, suppose the two missing costs C_3, C_4 are independent uniform variables on $[0, 2]$, and let the predictor be $\hat{c} = (1, 1, 1, 1)$. The posterior mean is the unique squared-error-optimal completion $\bar{c}^{\text{mean}} = (1, 1, 1, 1)$. Both paths then have completed cost 2.

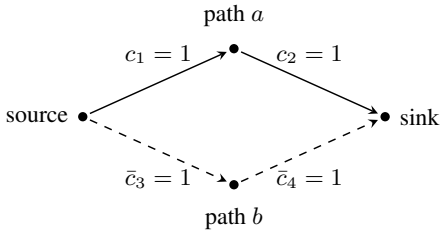


Figure 6: Observed costs on path a (solid) and posterior-mean imputations on path b (dashed). Both completed paths cost 2.

For a cost target c, the normalized minimization SPO+ subgradient is

$$g(\hat{c}, c) = \frac{2}{4}\big[w^\star(c) - w^\star(2\hat{c} - c)\big].$$

At $\bar{c}^{\text{mean}} = \hat{c}$, the two solver inputs coincide. A fixed deterministic tie rule therefore gives $g(\hat{c}, \bar{c}^{\text{mean}}) = 0$. In contrast, take true costs $c = (1, 1, \frac{1}{4}, \frac{1}{4})$, within the stated support. Then $w^\star(c) = b$ and $w^\star(2\hat{c} - c) = a$, so

$$g(\hat{c}, c) = \tfrac{1}{2}(b - a) \ne 0.$$

These optimizer choices hold throughout the positive-probability region where the true bottom-path cost is less than 2. Thus squared-error-optimal completion can suppress a nonzero full-information SPO+ update.

For the completion $\bar{c}^{\text{opp}} = (1, 1, \frac{3}{2}, \frac{3}{2})$, the target optimizer is a and the shifted optimizer is b. Both solver optima are unique, and the resulting subgradient $\frac{1}{2}(a - b)$ is the opposite of $g(\hat{c}, c)$. Preserving both true-cost optimizer outputs is sufficient for gradient agreement.

Writing $C = C_3 + C_4$ and $c = (1, 1, C_3, C_4)$, the realized full-information gradient is

$$g(\hat{c}, c) = \begin{cases} \frac{1}{2}(b - a), & C < 2, \\ \frac{1}{2}(a - b), & C > 2. \end{cases}$$

The conditional distribution of C is symmetric about 2, and $\Pr(C = 2) = 0$. Hence

$$\mathbb{E}[g(\hat{c}, c) \mid \text{observed costs}] = 0 = g(\hat{c}, \bar{c}^{\text{mean}}).$$

The example shows that completion can suppress or reverse a realized full-information gradient. It does not establish conditional expected-gradient bias for posterior-mean completion in this symmetric case. This samplewise distinction motivates the discriminative-imputation direction in Appendix E.6.

An update can reinforce the missing-cost error. Use the same two paths, now with fixed noiseless costs $c = (-\frac{1}{2}, -\frac{1}{2}, -1, -1)$. Path a costs -1 and path b costs -2, so choosing a incurs one unit of regret. Let z be a shared scalar latent variable, with predictions $\hat{c}_i(z) = \hat{\theta}_i z$ and fixed head coefficients $\hat{\theta} = (\frac{1}{2}, \frac{1}{2}, 1, 1)$. For $z > 0$, the predicted path costs are z and $2z$, so the greedy learner chooses a and observes only its two costs.

Posterior-mean completion fills the two missing entries with z. The completed and adjusted costs are

$$\bar{c} = (-\tfrac{1}{2}, -\tfrac{1}{2}, z, z), \qquad 2\hat{c} - \bar{c} = (z + \tfrac{1}{2}, z + \tfrac{1}{2}, z, z).$$

Their path totals are $(-1, 2z)$ and $(2z + 1, 2z)$, respectively. The first solver therefore chooses a and the second chooses b, with no ties. The normalized SPO+ gradient is $g = \frac{1}{2}(a - b)$.

Let $\rho \in [0, 1]$ be the fraction of each missing-edge gradient retained before backpropagation. Full masking uses $\rho = 0$; no masking uses $\rho = 1$. Holding the head coefficients and completed target fixed during differentiation, the scalar gradient passed to z is

$$g_z(\rho) = \underbrace{2\left(\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right)}_{\text{two observed edges}} + \rho \underbrace{2\left(-\tfrac{1}{2}\right)(1)}_{\text{two missing edges}} = \tfrac{1}{2} - \rho.$$

Gradient descent subtracts this quantity. Masking gives $g_z = \frac{1}{2}$ and decreases z, bringing predictions toward the true costs, which are attained at $z = -1$. With the true costs available, the SPO+ gradient is $\frac{1}{2}(b - a)$ and its scalar gradient is also $\frac{1}{2}$. At $\rho = \frac{1}{2}$ the update stops. Without masking, $g_z = -\frac{1}{2}$, so z increases. This makes the unobserved optimal path look still more expensive: its absolute prediction error is $|2z - (-2)| = 2z + 2$. As long as $z > 0$, the learner keeps choosing a and repeating this error-reinforcing update.

A sequential run with updating Bayesian heads. We also run the agents for 50 rounds on this noiseless problem, updating the heads from observed costs before each representation step. The context is constant at zero, so the learned feature is its scalar bias z. We initialize $z_0 = 0.43$ and each head at the coefficient above, with $A_i = 10$ and $v_i = 10\hat{\theta}_i$. All agents use UCB with $\beta = 0.2$ and Adam with learning rate $0.05/\sqrt{t}$ at round t (seed 0). SPO+ uses one posterior-mean completion. The full-feedback control observes all four costs for both head and representation updates. The other agents observe only the chosen path.

Masked BAYESIANSPO and the full-feedback control first select the optimal path in round 29, each incurring cumulative regret 28. The unmasked variant never selects it within 50 rounds and incurs regret 50; its final absolute path-b prediction error is 3.597, compared with 1.445 under masking. BAYESIANMSE first selects the optimal path in round 30 and incurs regret 29. Its missing-edge squared-error gradient is zero because the imputed target equals the prediction.

The same initial gradient-direction contrast holds at $z_0 = 0.33$ and $z_0 = 0.53$. At $z_0 = 0.53$, masking decreases z but does not reach the optimal path within 50 rounds. The example shows how missing-edge gradients can reinforce an imputation error. Masking blocks this direct contribution; imputed costs still determine the two optimizer outputs used to form the gradient.

B.11 Structural separation: exact cardinality vs. shortest path

The local imputation bound requires a feasible exchange of one selected item for one missing item. Theorem 4 bounds each high-width contested imputation error by the squared uncertainty

of a selected arm that the missing arm can replace. The full exact-cardinality family admits every selected–unselected exchange, whereas simple directed grid paths admit no one-item exchanges. This distinction determines which missing-coordinate contributions the local theorem controls.

Definition 5 (One-item exchange accessibility)**.** *Relative to a selected action w_t, a missing item j is **swap-accessible** if $w_t - e_r + e_j \in \mathcal{W}$ for some selected item r. Otherwise, any feasible alternative containing j requires a different kind of move and is outside this one-item exchange argument.*

For the full exact-cardinality family $\mathcal{W} = \{w \in \{0,1\}^n : \mathbf{1}^\top w = k\}$, the exchange of selected arm r for missing arm j changes the conditional-mean objective by $\tilde{c}_{t,j} - \tilde{c}_{t,r}$.

For a simple directed grid, let A_{flow} be its node–edge incidence matrix and b_{flow} its source–sink demand vector. Feasible paths satisfy $A_{\text{flow}} w = b_{\text{flow}}$. If $w' = w - e_r + e_j$ were also feasible, then the incidence columns would satisfy $(A_{\text{flow}})_r = (A_{\text{flow}})_j$, which is impossible for distinct grid edges. A detour replacing a right–up segment with up–right changes only four coordinates while sharing the rest of the path. This is a two-for-two exchange, covered by Lemma 7 with $m = 2$ when its aggregate predicted gap meets the tolerance. Missing disagreement edges without a qualifying bounded exchange remain in the nonlocal residual of the extended bound.

Corollary 5 (Refined flip-set collapse for exact cardinality)**.** *Let $\mathcal{W} = \{w \in \{0,1\}^n : \mathbf{1}^\top w = k\}$ with $1 \le k < n$. For any prediction vector p and target cost vector c, if the target optimum has decision gap $\Delta(c) \ge \Delta_{\min} > 0$, the SPO+ subgradient at (p, c) is zero whenever*

$$\|p - c\|_\infty < \Delta_{\min}/4,$$

a factor-k improvement over the general $\Delta_{\min}/(4k)$ obtained from Lemma 5. Thus $\chi_{\mathcal{W}} = 4$ for this family. For an at-most-k constraint under arbitrary signed costs or a weighted knapsack, the general $\Delta_{\min}/(4k)$ threshold applies.

Proof. Let $u = w^\star(c)$. An alternative replacing r selected items has $\|w - u\|_1 = 2r$ and signed cost gap at least $r\Delta_{\min}$, since pairing removed and added items gives r feasible single-item exchanges, each with gap at least $\Delta_{\min}$. Apply Lemma 5 with $v = 2(p - c)$. Its right-hand side is at least $r(\Delta_{\min} - 4\|p - c\|_\infty) > 0$. □

Corollary 6 (Cardinality exceptional-round budget)**.** *For the full exact-cardinality family of Corollary 5, fix $\Delta_{\min} > 0$ and use the confidence radius $\beta_T > 0$. Under Assumption 1 and global coverage at threshold $\varepsilon = \Delta_{\min}/(4\beta_T)$, the count bound* (9) *gives*

$$B_T\left(\frac{\Delta_{\min}}{4\beta_T}\right) \le O\Big(\tfrac{\alpha_{\lambda,L} n \beta_T^2 q \log(1 + kTL^2/(nq\lambda))}{p_{\min} \Delta_{\min}^2}\Big).$$

For target-optimizer stability, also use posterior-mean completion and assume the confidence event, noiseless observed targets, and conditional-mean decision gaps at least $\Delta_{\min}$. The same stability argument with $v = e_t$ gives target agreement outside $\mathcal{B}_\varepsilon$, where $2\|e_t\|_\infty < 2\beta_T\varepsilon = \Delta_{\min}/2$. The generic threshold $\Delta_{\min}/(4k\beta_T)$ instead gives the count bound $O(\alpha_{\lambda,L} n k^2 q \log(1 + kTL^2/(nq\lambda))\, \beta_T^2/(p_{\min}\Delta_{\min}^2))$.

Theorem 10 (Exact-cardinality feedback lower bound)**.** *Fix integers $1 \le k < n$, constants $L, S, \sigma > 0$, and $0 < \Delta \le LS$. For $\mathcal{W} = \{w \in \{0,1\}^n : \mathbf{1}^\top w = k\}$, there are two fixed-context instances ν^+, ν^- with feature norm at most L, parameter norms at most S, and independent Gaussian noise of variance σ^2, such that every nonanticipating semi-bandit algorithm and every integer $T \ge 1$ satisfy*

$$\max_{\nu \in \{\nu^+, \nu^-\}} \mathbb{E}_\nu[\tilde{\mathcal{R}}(T)] \ge \frac{1}{16} \min\left\{\Delta T, \frac{\sigma^2}{\Delta}\right\}.$$

Proof. Give items $1, \dots, k-1$ mean cost zero and items $k+2, \dots, n$ mean cost Δ. For the pair $a = k$, $b = k+1$, set

$$(\mu_a^+, \mu_b^+) = (0, \Delta), \qquad (\mu_a^-, \mu_b^-) = (\Delta, 0).$$

Each instance has exactly k zero-mean items, so its unique optimal action has cost zero. The feature $h = L$ and coefficients $\theta_i^* = \mu_i/L$ give the required bounded realization. Let $E_t = \{b \in S_t\}$. Under ν^+, selecting b incurs mean regret at least Δ. Under ν^-, omitting b incurs mean regret at least Δ, since only $k-1$ other items have zero mean and the action must select k items. Hence

$$r_t^+ \ge \Delta \mathbf{1}_{E_t}, \qquad r_t^- \ge \Delta \mathbf{1}_{E_t^c}$$

for every feasible action, including actions selecting both or neither contested item. Only a, b differ between the two instances. Lemma 10 gives

$$\mathrm{KL}(\mathbb{P}_{\nu^+}^{t-1} \,\|\, \mathbb{P}_{\nu^-}^{t-1}) = \frac{\Delta^2}{2\sigma^2}\mathbb{E}_{\nu^+}[N_a(t-1) + N_b(t-1)] \le \frac{(t-1)\Delta^2}{\sigma^2}.$$

This is no larger than $2(t-1)\Delta^2/\sigma^2$, so (11) proves the result. □

This construction isolates one contested pair. A round can reveal zero, one, or both contested costs, depending on the selected action. In the two-path construction, the differing edge is observed only on the bottom path.

Lemma 11 (Matroid exchanges and coordinate retention)**.** *Let $\mathcal{W}$ be the set of binary incidence vectors of the bases of a matroid on $[n]$, with $\|w\|_0 = k$ for all $w \in \mathcal{W}$. For the current-update actions in Definition 1, partition the flip set $D_t = D_t^- \cup D_t^+$ by*

$$D_t^- = \mathrm{supp}(w^{\mathrm{opt}}) \setminus \mathrm{supp}(w^{\mathrm{spo}}), \qquad D_t^+ = \mathrm{supp}(w^{\mathrm{spo}}) \setminus \mathrm{supp}(w^{\mathrm{opt}}).$$

Then:

(a) $|D_t^-| = |D_t^+|$.

(b) (Basis exchange.) For every $a \in D_t^-$, there exists $b \in D_t^+$ such that $w^{\mathrm{opt}} - e_a + e_b \in \mathcal{W}$.

(c) (Conditional.) If $D_t \neq \varnothing$ and $D_t^- \subseteq S_t$, then $|D_t \cap S_t|/|D_t| \ge 1/2$.

Proof. (a) $|D_t^-| = k - |w^{\mathrm{opt}} \cap w^{\mathrm{spo}}| = |D_t^+|$. (b) Matroid basis exchange axiom; see Oxley [22]. (c) Under the condition, $|D_t \cap S_t| \ge |D_t^-| = |D_t|/2$. Parts (a) and (c) hold for any two equal-cardinality actions. □

Remark 13 (UCB inclusion under exact cardinality)**.** *For the full exact cardinality family in the minimization convention, UCB selects the k items with smallest adjusted costs $u_{t,i} = \hat{c}_{t,i} - \beta\sigma_{t,i}$, up to tie-breaking. A large predicted cost can offset the optimism bonus even when posterior width is large. The count bound* (6) *controls high-width contested alternatives through the uncertainty of selected arms they can replace. Global coverage remains an additional assumption, and Appendix C reports corresponding empirical diagnostics.*

Remark 14 (General matroid coverage)**.** *For matroid families ($\mathcal{W}$ = bases of a matroid), the exchange axiom identifies feasible local alternatives. Whether UCB selects them also depends on predicted means and uncertainty bonuses. Establishing Assumption 8 requires additional conditions on adjusted scores or a separate analysis of observation frequencies.*

Portfolio with a risk constraint. The portfolio family in Appendix C.3 is $\mathcal{W} = \{w \in \{0,1\}^n : \mathbf{1}^\top w = k,\ w^\top \Sigma w \le \gamma\bar{\sigma}^2 k\}$. Here Σ is the return-noise covariance, $\bar{\sigma}^2 = \mathrm{tr}(\Sigma)/n$ is its average marginal variance, and γ is the risk-budget multiplier. The risk constraint can exclude exchanges required by the matroid basis-exchange axiom. A sufficient condition for the refined collapse result is that the unique optimum over the full exact-cardinality family is risk-feasible and has gap at least $\Delta_{\min}$ within that unrestricted family. Corollary 5 then preserves that same optimum after the SPO+ shift, and it remains feasible for the constrained problem. For other instances, the generic bound applies, while Theorem 1 controls feasible contested one-item exchanges and Lemma 7 extends this to bounded exchanges.

The cardinality lower-bound construction transfers to a risk-constrained class if the risk constraint is nonbinding on every exact-k action, for example $\Sigma = \sigma^2 I$ with a risk budget at least $\sigma^2 k$. Applying this transfer to the empirical active-risk benchmark requires verifying the nonbinding-risk condition.

C Benchmark problems, data generation, and counterfactual structure

This appendix specifies the optimization problems, the data-generating process, and the per-benchmark counterfactual structure that determines how much decision-relevant signal survives masking. All three benchmarks come from the PYEPO [25] suite for PtO and DFL. We adapt these offline benchmarks to the online semi-bandit setting by drawing a fresh context per round and revealing objective coefficients only on the arms in the selected support. In this appendix, c_t denotes the generic coefficient vector passed to the solver: costs for shortest path, item values for knapsack, and portfolio returns (written as r_t in Appendix C.3); the solver objective sign is set accordingly.

Knapsack and shortest path are already discrete in PyEPO and carry over directly. For portfolio, we replace PyEPO's continuous asset weights ($w \in [0,1]^n$, $\mathbf{1}^\top w = 1$) with cardinality-constrained binary selection ($w \in \{0,1\}^n$, $\mathbf{1}^\top w = k$). Each action then selects exactly k assets: returns are observed on $\mathrm{supp}(w_t)$ and remain missing elsewhere. The factor-model DGP for returns is taken unchanged from PyEPO.

At each round t, we draw a fresh five-dimensional context $x_t \sim \mathcal{N}(0, I_5)$ across all benchmarks. For knapsack and shortest path, the per-arm mean objective coefficient follows the PyEPO polynomial feature map

$$\mu_{t,i} = \frac{(Q_i^\top x_t/\sqrt{d} + 3)^{\deg} + 1}{3.5^{\deg}}, \tag{12}$$

where $Q \in \{0,1\}^{n \times d}$ has i.i.d. $\mathrm{Bernoulli}(0.5)$ entries, sampled once at instance initialization and held fixed across all rounds and seeds. The realized coefficient is $c_{t,i} = \mu_{t,i} \cdot (1 + U_{t,i})$ with multiplicative noise $U_{t,i} \sim \mathrm{Uniform}(-w_0, w_0)$, $w_0 = 0.25$, sampled i.i.d. per arm and per round; degrees are $\deg = 4$ for knapsack and $\deg = 6$ for shortest path. Conditional on x_t, the additive fluctuation $\epsilon_{t,i} = \mu_{t,i} U_{t,i}$ is bounded and sub-Gaussian with scale $|\mu_{t,i}| w_0$. This Gaussian-context DGP permits unbounded conditional means. Applying the uniform noise-scale assumption (Assumption 3) would require an additional bound on those means. Portfolio uses a factor-model DGP (Appendix C.3). Seeds are split: tuning uses $\{0,1,2\}$, evaluation uses $\{100, 101, \ldots, 109\}$, with DGP, agent initialization, and environment rollout driven by separate streams from each seed.

C.1 Knapsack

Formulation. The 0/1 knapsack problem at round t is

$$\max_{w \in \{0,1\}^n} \; c_t^\top w \quad \text{s.t.} \quad a^\top w \le B, \tag{13}$$

where $c_t \in \mathbb{R}_+^n$ is the per-item value vector (common polynomial DGP, $\deg = 4$), $a \in [1,5]^n$ are item weights drawn once at instance initialization from $\mathrm{Uniform}(1,5)$, and B is the capacity. We solve with Gurobi. Instances: $(n, B) = (32, 20)$ and $(100, 50)$. Their observation rates depend on the selected-set sizes, Table 7 reports empirical rates for tuned BAYESIANSPO.

Counterfactual structure. Knapsack allows for item swaps when the resulting set satisfies the capacity constraint. For these exchanges, the SPO+ gradient on selected items can therefore carry information about whether each item should remain in the solution, while the unselected side of the swap is handled through imputation. Portfolio adds pairwise risk coupling to this selected-item decision signal.

C.2 Shortest path

Formulation. On an $N \times N$ grid, the shortest-path problem has the flow formulation

$$\min_{w \in \{0,1\}^n} \; c_t^\top w \quad \text{s.t.} \quad A_{\text{flow}}\, w = b_{\text{flow}}, \tag{14}$$

where $A_{\text{flow}} \in \{-1, 0, +1\}^{|V| \times n}$ is the node-edge incidence matrix of the directed grid (only right and up moves), b_{flow} is the source–sink demand vector with $(b_{\text{flow}})_v = +1$ at the source, -1 at the sink, and 0 otherwise, and c_t are edge costs from the common polynomial DGP with $\deg = 6$. The arm set is the $n = 2N(N-1)$ directed edges, and a feasible path uses exactly $2(N-1)$ of them. We solve this problem exactly using a custom NumPy dynamic-programming algorithm that computes minimum costs in topological order and recovers the path by backtracking. Instances: $N = 8$ ($n = 112$, observation rate $\sim 12\%$) and $N = 20$ ($n = 760$, observation rate $\sim 5\%$).

Counterfactual structure. Monotone grid paths can differ across most of their edges, but can also differ by a short detour. Exposing a missing edge requires a multi-edge path change. Lemma 7 controls high-width missing edges in bounded detours whose aggregate predicted gap satisfies the contestability condition. Edges without such detours remain outside the local bound. Empirically, SP 20×20 improves, while SP 8×8 is a narrow endpoint where BAYESIANMSE is already near competitive.

C.3 Portfolio optimization

Formulation. The cardinality-constrained mean-variance portfolio at round t is

$$\max_{w\in\{0,1\}^n} \; r_t^\top w \quad \text{s.t.} \quad \mathbf{1}^\top w = k,\; w^\top \Sigma w \le \gamma \bar{\sigma}^2 k, \tag{15}$$

where $r_t \in \mathbb{R}^n$ is the realized return vector at round t, Σ is the conditional covariance of returns given x_t, $k = 5$ is the cardinality, $\gamma = 1.0$ is the risk-budget multiplier, and $\bar{\sigma}^2 = \frac{1}{n}\mathrm{tr}(\Sigma)$. Within each instance, Σ is fixed across contexts and rounds. We solve with Gurobi as a mixed-integer quadratically constrained program. Instance: $n = 32$, observation rate $\sim 16\%$. Returns follow the factor model $r_{t,i} = \left(0.05\, Q_i^\top x_t/\sqrt{d}+\beta_0^{1/\deg}\right)^{\deg}+(L_{\mathrm{fac}} f_t)_i+\zeta_{t,i}$, whose first term is the conditional mean. Here the signal scale is 0.05, the baseline return is $\beta_0 = 0.10$, $\deg = 4$, $Q \in \{0,1\}^{n\times d}$ is Bernoulli(0.5) (sampled once), factor loadings $(L_{\mathrm{fac}})_{ij} \sim \mathrm{Uniform}(-2.5{\cdot}10^{-3}w_0, 2.5{\cdot}10^{-3}w_0)$ are sampled once at noise width $w_0 = 0.25$, common factors $f_t \sim \mathcal{N}(0, I_d)$ are drawn independently of x_t each round, and idiosyncratic return noise is $\zeta_{t,i} \sim \mathcal{N}(0, (10^{-2}w_0)^2)$. The risk constraint uses $\Sigma = L_{\mathrm{fac}} L_{\mathrm{fac}}^\top + (10^{-2}w_0)^2 I_n$.

Counterfactual structure. Portfolio shares knapsack's selected-set swap structure only for swaps that remain feasible under the risk constraint. Other competitive portfolios can require multi-asset changes even when each portfolio has risk slack. The local exchange argument covers feasible swaps, and the generic bound retains nonlocal contributions. The refined collapse result applies under the stronger unrestricted-optimum condition in Appendix B.11. Portfolio's $\sim 32\%$ gain over BAYESIANMSE is a separate empirical result.

C.4 Decision-relevant representations task

In addition to the PyEPO-derived benchmarks, we use one controlled task to compare the learned representations. The decision-relevant representations task is a controlled maximization problem with four unit-weight arms. Each feasible action selects exactly two of the four arms. For context $x \in \mathbb{R}^5$, the noiseless arm values are

$$v_0(x) = 3 + \frac{x_0 + 2x_2 + 2x_3 + 2x_4}{\sqrt{5}}, \qquad v_1(x) = 3 + \frac{-x_0 + 2x_2 + 2x_3 + 2x_4}{\sqrt{5}}, \tag{16}$$

$$v_2(x) = 3 + \frac{x_1 + 2x_2 + 2x_3 + 2x_4}{\sqrt{5}}, \qquad v_3(x) = 3 + \frac{-x_1 + 2x_2 + 2x_3 + 2x_4}{\sqrt{5}}. \tag{17}$$

The figure experiments use noise width zero. More generally, the implementation can multiply each value by independent $\mathrm{Uniform}(1 - w_0, 1 + w_0)$ noise.

In the noiseless construction, each pair has one value above the common offset and one below it (apart from probability-zero ties). The two largest values therefore contain one arm from each pair. Since $v_0 - v_1 = 2x_0/\sqrt{5}$ and $v_2 - v_3 = 2x_1/\sqrt{5}$, the oracle selects arm 0 instead of 1 iff $x_0 > 0$, and arm 2 instead of 3 iff $x_1 > 0$. The coordinates x_2, x_3, x_4 enter all four arms with the same coefficient, so they create a large common value shift but do not change the oracle decision. Under semi-bandit feedback, arm dependent errors in estimating this shared component can affect both action selection and the imputations used for representation learning.

Each representation in Figure 5 is trained once on a single seed with the same $T = 10{,}000$ training rounds. The task uses 500 held-out contexts, hidden dimension 2, noise width 0, $\beta = 1.0$, $\lambda = 0.1$, learning rate 0.03, and $M = 8$ posterior completions for BAYESIANSPO. All learned agents use the same Bayesian last-layer posterior and UCB action rule; only the feature-map objective and feedback mask differ.

The probe accuracy in Figure 5 and Table 3 is measured by five-fold logistic-regression accuracy for predicting the oracle action, using 400 contexts to fit the probe and 100 to test it in each stratified fold. The probe uses fixed L2 regularization (`C=1`) without feature standardization. While the probe is refitted, the representation stays fixed across folds. The bars show each input row's norm in the learned matrix $W_1 \in \mathbb{R}^{5\times 2}$, normalized by the sum over all rows:

$$\frac{\|W_{1,j:}\|_2}{\sum_{k=0}^{4} \|W_{1,k:}\|_2}.$$

The reported decision share sums this quantity over $j \in \{0, 1\}$, and the nuisance share sums it over $j \in \{2, 3, 4\}$. Under our i.i.d. initialization, an untrained extractor has an expected decision share of 40% at every width, since two of the five inputs are decision-relevant.

With a two-dimensional representation, fitting the common cost variation competes with representing both decision coordinates. Under full feedback, adding the same value to all arms shifts every two-arm objective equally, leaving the SPO+ loss unchanged. MSE rewards accurate prediction of this common component. Figure 5 illustrates the resulting input weights and feature geometry in one run. Under both feedback settings in this run, BAYESIANSPO assigns a larger share of input-row norm to decision coordinates than BAYESIANMSE.

With an affine extractor and $q \geq d + 1$, a fixed full-row-rank feature map induces a shared regularizer on the per-arm affine coefficients, as in multi-task feature learning [4]; Figure 5 uses $q < d + 1$, where the map must also allocate limited capacity.

Effect of representation width. We repeat the diagnostic at widths $q \in \{2, 3, 6, 16\}$ with ten training seeds (200–209), using the same $10{,}000$-round data stream and 500 held-out contexts. Within each width and seed, both methods and feedback settings share the same initial feature weights. We retain the native initialization rule and the hyperparameters above, without retuning across widths. Table 3 reports variability across these training seeds; each probe uses all q learned features.

Table 3: Representation diagnostics across widths (mean $\pm$ standard deviation over ten training seeds). Decision share is the percentage of total input-row norm assigned to x_0, x_1. Acc. is five-fold probe accuracy (%) using all q learned features.

		BAYESIANSPO		BAYESIANMSE	
Feedback	Width q	Decision share	Acc.	Decision share	Acc.
Full	2	$\mathbf{98.8 \pm 0.5}$	$\mathbf{96.6 \pm 1.2}$	61.3 ± 11.2	58.0 ± 1.9
	3	64.9 ± 8.0	$\mathbf{95.8 \pm 0.8}$	$\mathbf{68.6 \pm 8.6}$	93.2 ± 2.6
	6	$\mathbf{55.9 \pm 5.1}$	$\mathbf{95.1 \pm 0.6}$	42.7 ± 5.9	87.7 ± 3.7
	16	$\mathbf{52.6 \pm 2.0}$	$\mathbf{95.8 \pm 0.3}$	41.0 ± 3.8	93.8 ± 0.8
Semi-bandit	2	$\mathbf{56.6 \pm 10.0}$	$\mathbf{55.9 \pm 3.8}$	44.9 ± 8.4	53.7 ± 3.1
	3	$\mathbf{67.5 \pm 6.1}$	$\mathbf{94.5 \pm 2.7}$	58.0 ± 9.0	90.2 ± 6.1
	6	$\mathbf{48.9 \pm 4.4}$	$\mathbf{92.5 \pm 1.4}$	41.8 ± 6.2	87.5 ± 3.6
	16	$\mathbf{42.8 \pm 4.8}$	$\mathbf{94.5 \pm 1.1}$	39.6 ± 3.0	94.1 ± 0.7

Under full feedback, BAYESIANSPO's decision share falls from 98.8% at $q = 2$ to 52.6% at $q = 16$, while probe accuracy stays between 95.1% and 96.6%. Under semi-bandit feedback, its decision share peaks at $q = 3$ and declines at larger widths. Its mean probe accuracy is 55.9% at $q = 2$ and ranges from 92.5% to 94.5% at $q \geq 3$. Decision share measures relative input-weight magnitude. At $q = 6$ and $q = 16$, all learned matrices W_1 have rank 5, so the features preserve every input coordinate. Probe accuracy at these widths measures how readily logistic regression, with the same regularization at every width, learns the oracle actions from the resulting feature geometry.

C.5 Summary and instance hyperparameters

Table 4 summarizes the counterfactual structure of each benchmark family; Table 5 lists the per-instance hyperparameters used in the main evaluation.

Table 4: Feasible one-item exchanges across benchmark families. Capacity and risk constraints can rule out individual swaps. Theorem 1 controls feasible contested one-item exchanges; Lemma 7 extends this to bounded exchanges, including two-for-two grid detours.

Family	**One item exchanges?**	**Coupling Constraint**	**Local argument**
Shortest path	No	Flow conservation	Bounded detours only
Knapsack	Conditional (capacity)	Shared capacity	Feasible swaps only
Portfolio	Conditional (risk)	Risk constraint	Feasible swaps only

Table 5: Benchmark instance hyperparameters. d: context dimension. $\deg$: polynomial degree in the DGP. w_0: noise width.

Benchmark	n	d	$\deg$	w_0	**Other**
Knapsack 32	32	5	4	0.25	capacity $B = 20$, weights $\sim U[1, 5]$
Knapsack 100	100	5	4	0.25	capacity $B = 50$, weights $\sim U[1, 5]$
SP 8×8	112	5	6	0.25	grid 8×8, 14 edges/path
SP 20×20	760	5	6	0.25	grid 20×20, 38 edges/path
Portfolio	32	5	4	0.25	$k = 5$, $\gamma = 1.0$, signal scale $= 0.05$, $\beta_0 = 0.10$

C.6 Connection to the formal structural separation

Appendix B.11 separates feasible one-item exchanges from multi-edge path changes. The local theorem controls high-width contested arms. Lemma 7 extends this control to bounded multi-edge changes, such as two-for-two detours on grid paths. The full exact-cardinality family also has the refined collapse constant $\chi_{\mathcal{W}} = 4$; under the additional global coverage assumption this improves the exceptional-round upper bound, as summarized in Table 6. These results compare exchange structure and exceptional-round counts. Comparing regret between benchmark families requires a separate argument.

Weighted knapsack and risk-constrained portfolio can admit useful one-item swaps, but their constraints may rule others out; the exact-cardinality specialization requires the additional conditions stated in Appendix B.11. The ratio k/n gives the fraction of edges observed per path. A per-edge observation-rate guarantee requires separate coverage conditions.

Table 6: Exceptional-round counts under global coverage. These upper bounds control $B_T(\Delta_{\min}/(\chi_{\mathcal{W}}\beta_T))$, the number of high-width rounds throughout the horizon. The eventual-agreement result uses the separate conditioning assumption, and total loss is bounded by the full loss decomposition. The full exact-cardinality family uses $\chi_{\mathcal{W}} = 4$; the generic binary case uses $4k$.

Family	**Exceptional-round upper bound**
Exact-cardinality selection	$O(\alpha_{\lambda,L} nq \log T\, \beta_T^2/(p_{\min}\Delta_{\min}^2))$
Portfolio*	generic binary rate unless the cardinality specialization applies
Shortest path (grid)	$O(\alpha_{\lambda,L} nk^2 q \log T\, \beta_T^2/(p_{\min}\Delta_{\min}^2))$

*A sufficient condition is that the unrestricted exact-k optimum is risk-feasible and has the stated gap within the unrestricted family. Otherwise use the generic threshold. Local bounds still apply to exchanges with contested arms, with remaining terms explicit.

Counterfactual gradient signal (structural intuition). For a nonempty SPO+ flip set D_t, the fraction of changed coordinates that are observed is $|D_t \cap S_t|/|D_t|$. If the two solver outputs have equal cardinality and their removed coordinates D_t^- all belong to S_t, then this fraction is at least $1/2$ by the counting argument in Lemma 11. This conditional counting statement also applies to equal-length grid paths. If the observed action differs from both solver outputs, the condition can fail in either family. The retained-coordinate count measures support coverage. The gradient's norm and direction also depend on the values propagated through those coordinates.

Empirical coverage diagnostic. The global observation-coverage assumption (Assumption 8) is stronger than the local coverage used in Theorem 1. Our diagnostic computes, for each arm i, the fraction of high-uncertainty rounds in which $i \in S_t$, then takes the minimum over arms. We record

pre-action posterior widths and accumulate per-arm eligible and selected-eligible counts. These counts measure observation coverage.

Table 7 reports the diagnostic for tuned BAYESIANSPO over $T = 10{,}000$ rounds and ten evaluation seeds, defining high uncertainty as posterior width above the within-run median pooled over all pre-action arm-round widths. A rarely selected component can drive the worst-arm statistic near zero. Portfolio and knapsack have higher median-arm coverage (0.073–0.113) than shortest path, especially SP 20×20 (median 0.008, worst-arm 0). The 75th-percentile threshold gives the same qualitative ordering.

Table 7: Empirical observation-coverage profile for high-uncertainty components. High uncertainty means posterior width at or above the median of all pre-action arm-round widths within a run. Values are averaged over ten seeds; $p_{\min}$ is the worst arm, while p_{10} and p_{50} are the 10th percentile and median over arms.

Benchmark	**Obs. rate**	$p_{\min}$	p_{10}	p_{50}
Portfolio	0.156	0.0539	0.0603	0.0726
Knapsack 32	0.273	0.0155	0.0414	0.1128
Knapsack 100	0.235	0.0029	0.0194	0.0813
SP 8×8	0.125	0.0008	0.0039	0.0396
SP 20×20	0.050	0.0000	0.0009	0.0084

D Implementation Details and Extended Results

This appendix reports implementation details and extended diagnostics supporting Table 2: baseline definitions (Appendix D.1), reproducibility (Appendix D.2), hyperparameter tuning details (Appendix D.3), full regret trajectories (Appendix D.4), regret-vs-horizon scaling (Appendix D.5), wall-clock timing (Appendix D.6), and saved-trajectory MAE diagnostics (Appendix D.7), nonlinear feature extractors (Appendix D.8), pseudo-regret and realized regret (Appendix D.9), and energy scheduling with historical prices (Appendix D.10). Core design ablations, including the PFY-vs.-SPO+ representation-objective comparison, are grouped separately in Appendix E.

D.1 Baselines

Our six comparators span three families: *classical* bandits that work directly on the raw context, *neural* bandits that learn shared representations, and BAYESIANMSE, which shares the Bayesian last-layer head, UCB rule, feature dimension, and masking schedule with BAYESIANSPO but trains the shared representation with MSE. The grouping separates generic bandit baselines from the matched representation-loss comparison. Hyperparameter search spaces for all baselines are listed in Table 8.

Classical bandit baselines. *LinUCB* [19] keeps per-arm sufficient statistics on the raw context and selects via Upper Confidence Bound. Our *Thompson-sampling* baseline adapts linear contextual Thompson sampling [2] to combinatorial actions: it maintains a Bayesian linear regression per arm on the raw context and passes posterior samples to the optimizer.

Neural bandit baselines. *ε-Greedy* exploits the best predicted action with probability $1 - \varepsilon_t$ and otherwise explores, using either constant or linearly decaying ε_t. Its predictor is a ReLU hidden layer followed by a deterministic linear head, trained jointly on observed-arm prediction MSE without uncertainty estimates. *NeuralUCB* [29] adds an NTK-derived covariance to a shared neural feature extractor and selects via UCB; it tests whether non-Bayesian deep exploration matches the Bayesian last-layer pattern when both have access to learned representations. *BLL-Online* [27] is the closest Bayesian last-layer baseline from the bandit literature: a shared feature extractor feeds a per-arm Bayesian linear head, action selection uses UCB, and the shared representation is trained by a marginalized prediction objective. In our implementation, its feature update retains a posterior-covariance term in addition to prediction residuals. This comparison evaluates a posterior-aware prediction update against masked SPO+ representation learning. BAYESIANMSE provides the matched comparison of training losses.

Matched MSE variant. BAYESIANMSE uses the same analytical Bayesian last-layer head, tuned UCB rule, posterior-mean target construction, feature dimension, and masking schedule as BAYESIANSPO. It replaces the masked SPO+ representation update with masked MSE. We say masked MSE because masking is functionally inert under MSE: the unobserved-arm residual is zero by construction, so the per-arm gradient is zero before any explicit masking is applied. Appendix E.3's weight-response curve confirms this empirically. The comparison with BAYESIANSPO therefore changes the representation-training objective while keeping the posterior and masking machinery fixed. The same shared neural extractor and per-arm Bayesian regression machinery (Appendix A) is reused.

Implementation. Methods with learned representations use a single affine feature layer, except ε-Greedy with its ReLU hidden layer and NeuralUCB with its deeper feature stack. LinUCB, Thompson sampling, BAYESIANSPO, and BAYESIANMSE use per-arm linear regressors. LinUCB and Thompson sampling operate on the raw context, while BAYESIANSPO and BAYESIANMSE use a shared feature map. The evaluation environment uses Python 3.11.13 with `jax` 0.8.2 and `jaxlib` 0.8.2. Per-method search spaces, sampler, pruning, and seed protocol are documented in Appendix D.3. Baseline implementations are released alongside the paper.

D.2 Reproducibility.

The code repository (`https://github.com/Aabhash/OnlineSemiBanditDFL`) contains the training and evaluation code, benchmark-generation code, configuration files, and dependency specifications (`pyproject.toml` and lock file). The code is self-contained for the reported experiments and

does not require specialized hardware. Knapsack and portfolio use external Gurobi solver through `gurobipy` as the optimization oracle. Shortest path uses an exact dynamic-programming oracle implemented in NumPy for the directed acyclic grid (Appendix C.2). Tuning uses $T_{\text{tune}} = 2{,}000$ rounds $\times$ 3 seeds, evaluation uses $T_{\text{eval}} = 10{,}000$ rounds $\times$ 10 seeds (disjoint from tuning). Optuna search spaces are listed in Appendix D.3. All experiments ran on a single Apple M3 laptop (8-core CPU, 16 GB RAM, macOS); total wall-clock times across final-evaluation seeds are reported in Appendix D.6, and total compute for Table 2 including hyperparameter tuning was approximately 40 CPU-hours.

D.3 Hyperparameter Tuning Details

D.3.1 Tuning Protocol

All hyperparameters are selected with Optuna [3] using the Tree-structured Parzen Estimator (TPE) sampler. For each reported agent–dataset pair with tunable parameters, we run 50 Optuna trials. Each non-pruned trial evaluates a candidate configuration for $T_{\text{tune}} = 2{,}000$ rounds on tuning seeds $\{0, 1, 2\}$ and reports the mean final cumulative regret. The objective is *minimized* (lower regret is better) regardless of whether the underlying environment is a minimization or maximization problem.

We use Optuna's percentile pruner to discard unpromising configurations before all three seeds are run. For each completed seed in a trial, we report the regret accumulated over the final 75% of the trajectory (rounds 501–$2{,}000$). After the first five startup trials, a configuration may be pruned if this mean late-stage regret is worse than the 75th percentile of completed trials at the same seed step. This reduces compute while avoiding early-round pruning of exploration-heavy configurations.

Tuning seeds $\{0, 1, 2\}$ and evaluation seeds $\{100, 101, \ldots, 109\}$ are strictly disjoint to prevent overfitting to specific random instantiations. Evaluation uses 10 seeds with a longer horizon of $T_{\text{eval}} = 10{,}000$ rounds to assess statistical significance.

Since BAYESIANMSE is the best amongst the baselines, BAYESIANSPO is compared with BAYESIANMSE on paired final cumulative regret values from the same ten evaluation seeds, using the one-sided Wilcoxon signed-rank alternative $\mathcal{R}_{\text{real}}^{\text{BAYESIANSPO}} < \mathcal{R}_{\text{real}}^{\text{BAYESIANMSE}}$ at $p < 0.05$.

All neural representation-learning agents with a tuned learning rate use the standard online-gradient-descent schedule $\eta_t = \eta_0/\sqrt{t+1}$ [30], with only the initial scale η_0 tuned. This is the canonical $O(1/\sqrt{t})$ rate for online convex optimization with bounded gradients. We tune η_0 on a three-point categorical grid to keep the online search stable and comparable across methods under the 50-trial, three-seed tuning budget. This fixed grid is a deliberate compromise for a protocol with few optimizer-scale parameters: it covers the stable operating range while avoiding a separate broad log-scale optimizer sweep. The implementation uses Adam with clipping; the projected-OGD analysis uses the same decay but additionally assumes projection and balances its bound with $\eta_0 = D_W/G$. Feature parameters are updated with Adam and global gradient clipping at norm 5.0; Bayesian last-layer statistics are updated analytically.

D.3.2 Search Spaces

Table 8 reports the hyperparameter search spaces for all agents. Continuous parameters sampled on a log scale are indicated with "(log)". Categorical parameters are sampled uniformly from the listed values.

For the Bayesian last-layer agents and matched ablations, the shared feature map is a single affine layer with no activation. With $d = 5$ and $q \in \{8, 16, 32\}$, this overparameterizes a linear feature subspace and shows that the gains can appear even in a simple linear setting. The Bayesian posterior regularization λ controls the prior precision of the per-arm Bayesian linear regression layer; larger values favor more conservative posterior updates. The exploration parameter β scales the posterior standard deviation in the UCB acquisition function.

In Table 8, `hidden_dim` is the feature or hidden width q, and β is the UCB scale for LinUCB and NeuralUCB. For Thompson sampling, `prior_var` is the coefficient-prior variance and `noise_var` is the assumed observation-noise variance. For BLL-Online, inverse `noise_var` scales the marginalized feature gradient. The PFY fields `pfy_n_samples`, `pfy_sigma`, and `pfy_noise` specify the

Table 8: Hyperparameter search spaces. All continuous ranges are sampled log-uniformly unless noted otherwise.

Agent	Parameter	Range / Choices	Scale
ε-Greedy	η_0	$\{3\times10^{-4},\ 10^{-3},\ 3\times10^{-3}\}$	categorical
	ε	$[0.01,\ 0.3]$	log
	decay	{`true`, `false`}	categorical
	hidden_dim	$\{8, 16, 32\}$	categorical
LinUCB	β	$[0.1,\ 5.0]$	log
	λ	$[0.01,\ 10.0]$	log
NeuralUCB	η_0	$\{3\times10^{-4},\ 10^{-3},\ 3\times10^{-3}\}$	categorical
	β	$[0.1,\ 5.0]$	log
	λ	$[0.01,\ 10.0]$	log
	batch_size	$\{4, 8, 16\}$	categorical
	hidden_dim	$\{8, 16, 32\}$	categorical
Thompson Sampling	prior_var	$[0.1,\ 10.0]$	log
	noise_var	$[0.01,\ 1.0]$	log
BayesianMSE	η_0	$\{3\times10^{-4},\ 10^{-3},\ 3\times10^{-3}\}$	categorical
	β	$[0.1,\ 5.0]$	log
	hidden_dim	$\{8, 16, 32\}$	categorical
	λ	$[0.01,\ 10.0]$	log
BayesianSPO	η_0	$\{3\times10^{-4},\ 10^{-3},\ 3\times10^{-3}\}$	categorical
	β	$[0.1,\ 5.0]$	log
	hidden_dim	$\{8, 16, 32\}$	categorical
	λ	$[0.01,\ 10.0]$	log
	M (imputation samples)	$\{1, 2, 4\}$	categorical
BayesianPFY	η_0	$\{3\times10^{-4},\ 10^{-3},\ 3\times10^{-3}\}$	categorical
	β	$[0.1,\ 5.0]$	log
	hidden_dim	$\{8, 16, 32\}$	categorical
	λ	$[0.01,\ 10.0]$	log
	M (imputation samples)	$\{1, 2, 4\}$	categorical
	pfy_n_samples	$\{1, 2\}$	categorical
	pfy_sigma	$[0.1,\ 10.0]$	log
	pfy_noise	{`gumbel`, `normal`}	categorical
BLL-Online	η_0	$\{3\times10^{-4},\ 10^{-3},\ 3\times10^{-3}\}$	categorical
	β	$[0.1,\ 5.0]$	log
	hidden_dim	$\{8, 16, 32\}$	categorical
	λ	$[0.01,\ 10.0]$	log
	noise_var	$[0.1,\ 10.0]$	log

perturbation sample count, scale, and distribution. With decay enabled, $\varepsilon_t = \varepsilon \max\{0, 1 - t/T\}$ at the tth action ($t = 1, \ldots, T$), where T is the tuning or evaluation horizon; otherwise $\varepsilon_t = \varepsilon$.

D.4 Regret trajectories

Figure 7 shows the full cumulative-regret and smoothed per-round-regret trajectories for the four principal comparators: BAYESIANSPO, BAYESIANMSE, BLL-ONLINE, and LINUCB. These plots provide the time-resolved view behind Table 2, showing how the Bayesian last-layer methods separate from LINUCB over training and how the BAYESIANSPO–BAYESIANMSE gap develops across benchmarks.

D.5 Regret-vs-horizon scaling

Figure 8 plots BAYESIANSPO's cumulative regret over the evaluation horizon on log–log axes, using the saved final-evaluation trajectories. Each benchmark curve is normalized by its own final value at $T = 10{,}000$, so the figure shows the shape of regret accumulation rather than absolute endpoint performance. A $\sqrt{T}$ reference is shown for scale. Fitted log–log slopes over rounds $1{,}000$–$10{,}000$

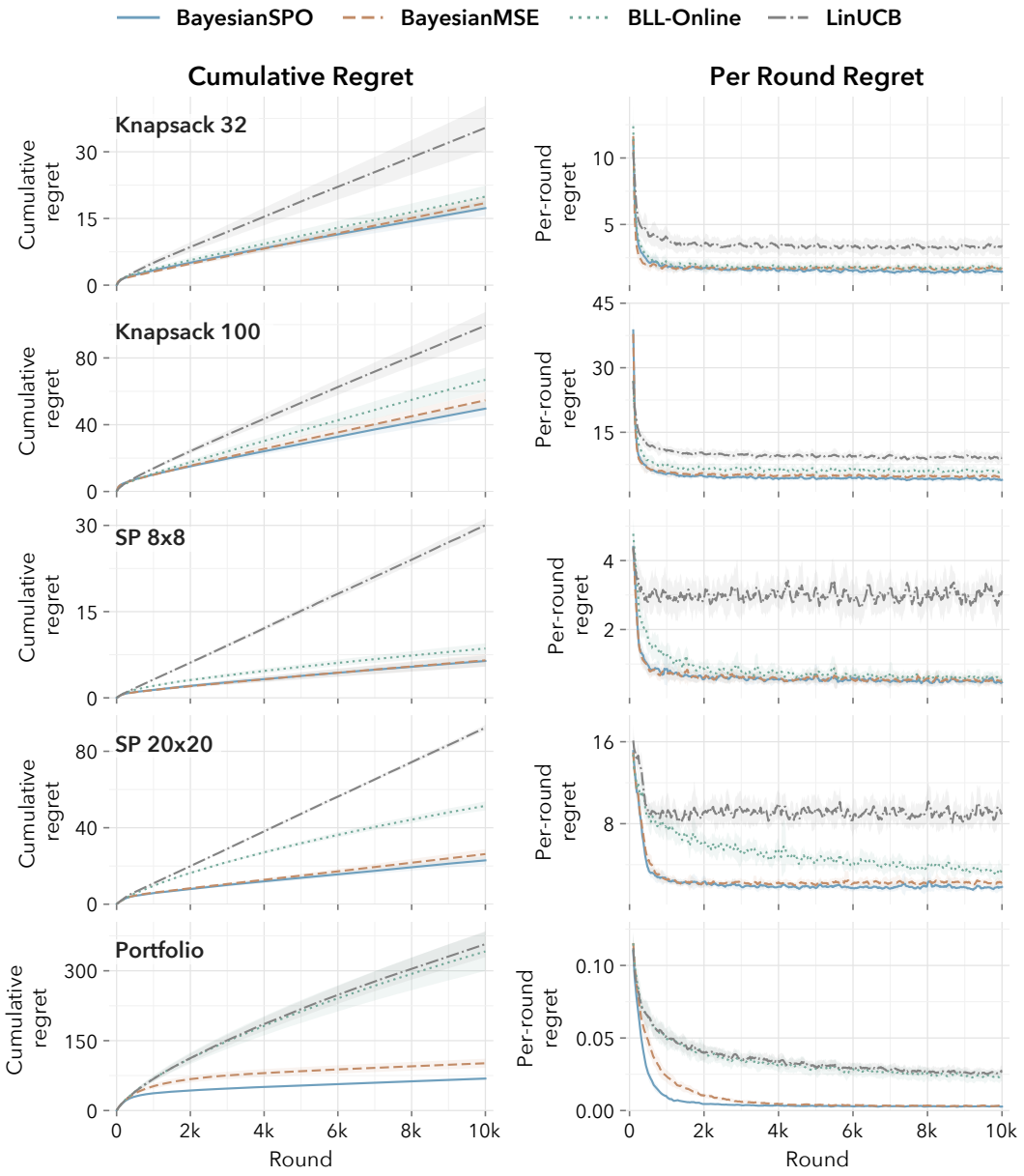


Figure 7: Regret trajectories across all five benchmarks. Each row is one benchmark; left panels show cumulative regret, and right panels show 100-round smoothed per-round regret. Cumulative panels use thousands for non-Portfolio benchmarks and raw units for Portfolio. Lines report means over ten evaluation seeds and bands show $\pm$ one s.d.

are 0.74 on Knapsack 32, 0.72 on Knapsack 100, 0.68 on SP 8×8, 0.62 on SP 20×20, and 0.27 on Portfolio. All fitted slopes are below one within this measured finite-horizon window.

D.6 Compute and Wall-Clock Time

Figure 9 reports total wall-clock time across the ten final-evaluation seeds for each method and benchmark, including action selection and parameter updates. BAYESIANSPO is in the same computational regime as BAYESIANMSE and BLL-ONLINE, while LINUCB remains cheaper because it does not update learned representations. In deployment, the optimizer is called once per round to act; regret is only an offline evaluation metric because unselected costs are missing. Each posterior completion used in the SPO+ update therefore adds two training-time solver calls, for $w^\star(\bar{c}^{(m)})$ and $w^\star(2\hat{c} - \bar{c}^{(m)})$. These calls are the incremental training cost relative to PtO; the combinatorial bottleneck is the same as in solver-based DFL. Mandi et al. [20] address that repeated-solve bottleneck for hard combinatorial SPO through continuous relaxations, MIP warm starts, and LP-basis reuse. The tuned $M = 1$ cells (Knapsack 32, Knapsack 100, Portfolio) require two extra solver calls per round beyond action selection, and the tuned $M > 1$ cells (SP 8×8 at $M = 4$, SP 20×20 at $M = 2$) require $2M$ extra solver calls.

D.7 Where the gains come from

As a lightweight diagnostic of how the regret gains are expressed in the learned predictor, Table 9 reports mean MAE over the final 2000 rounds for the two closest Bayesian methods on contested and non-contested components, along with the relative method gap $\Delta_{\%} = 100(\mathrm{MAE}(\text{BAYESIANMSE}) - \mathrm{MAE}(\text{BAYESIANSPO}))/\mathrm{MAE}(\text{BAYESIANMSE})$. Positive $\Delta_{\%}$ means BAYESIANSPO has lower MAE.

We use the full realized coefficient vector to group components, independently of which components the learner selects. For knapsack, each component's score is the absolute reduced cost in the continuous knapsack relaxation. For shortest path, it is the excess cost of the cheapest source–sink path containing that edge over the optimal path cost. In both cases, components at or below the within-round 25th percentile of scores are marked proxy-contested, including ties, so the group can exceed 25% of components. Portfolio has no corresponding slack diagnostic and uses the middle 25% of components by realized-return rank; this rank fallback also applies when a slack routine raises an exception. Other components form the complement.

Across four of five benchmarks, relative MAE advantage of BAYESIANSPO is larger on contested components. The diagnostic approximates decision relevance using slack or rank scores and hence is a proxy, with component groups recomputed from realized costs each round. This offline proxy is distinct from the theorem's locally contested set. In Portfolio and SEMO, common-mode shifts add the same constant to every feasible decision and so leave the optimal selection unchanged. For fully observed cost vectors, SPO+ is unchanged by such shifts, whereas MSE penalizes prediction errors in them (Appendix C.4).

Table 9: Final-2000-round contested vs. non-contested component MAE, sampled every 25 rounds within the final 2000 rounds ($T = 10{,}000$, ten seeds). The $\Delta_{\%}$ columns report relative improvement of BAYESIANSPO over BAYESIANMSE; positive values favor BAYESIANSPO. Bold marks the larger relative method separation within each dataset.

Dataset	Contested components MAE			Non-contested components MAE		
	BayesianSPO	**BayesianMSE**	$\Delta_{\%}$	**BayesianSPO**	**BayesianMSE**	$\Delta_{\%}$
Knapsack 32	1.512	1.765	$+\mathbf{14.3}\%$	2.950	3.248	$+9.2\%$
Knapsack 100	1.590	1.851	$+\mathbf{14.1}\%$	3.229	3.267	$+1.2\%$
SP 8×8	0.251	0.301	$+\mathbf{16.4}\%$	0.579	0.602	$+3.9\%$
SP 20×20	0.309	0.349	$+\mathbf{11.5}\%$	0.591	0.578	-2.2%
Portfolio	0.0034	0.0038	$+11.4\%$	0.0083	0.0102	$+\mathbf{19.0}\%$

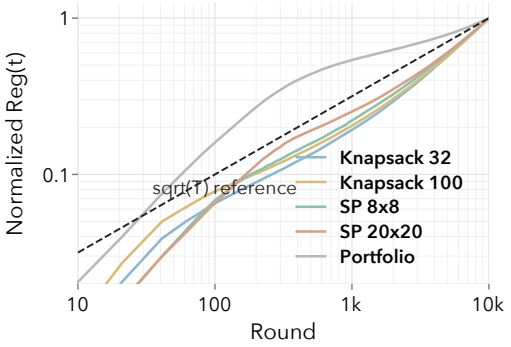


Figure 8: Regret-vs-horizon scaling for BAYESIANSPO. Curves show mean cumulative regret over ten evaluation seeds, normalized by each benchmark's regret at $T = 10{,}000$. The dashed line is a $\sqrt{T}$ reference normalized to the same endpoint.

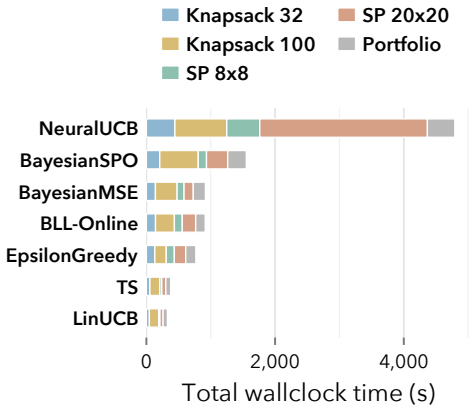


Figure 9: Total wall-clock time across evaluation seeds for the final-evaluation methods (T=10,000 rounds), stacked by benchmark.

D.8 Nonlinear feature extractors

The masked representation update in Appendix A applies to any differentiable feature extractor. We test nonlinear extractors by replacing the linear maps in BAYESIANSPO, BAYESIANMSE, and BLL-ONLINE with residual networks. A learned scalar gate controls each residual block's contribution. The gates start at zero, so each nonlinear extractor initially reduces to its linear map. We train the weights and gates online while keeping each method's training loss and analytical Bayesian head updates unchanged. Residual weights receive gradients once their gates move away from zero.

Evaluation protocol. We train the linear and nonlinear models from scratch on the same ten episodes (seeds 100–109), each with $T = 10{,}000$ rounds of semi-bandit feedback. We retain each method's tuned settings. Table 10 shows the results.

We retain the architectures selected in preliminary nonlinear experiments using development runs and validation. The Knapsack 100 and SP 8×8 architectures are also used for Knapsack 32 and SP 20×20, respectively. Within each benchmark, all three methods share the same residual-block design. Knapsack uses five ReLU blocks and a residual scale of 2. Shortest path uses three ReLU blocks and scale 1. Portfolio uses two blocks with squared activations, identity-initialized residual weights, and scale 0.1. The scale multiplies the hyperbolic tangent of each learned gate.

Table 10: Nonlinear extractors over ten seeds, $T = 10{,}000$. Regret is final cumulative realized regret (mean $\pm$ standard deviation); Knapsack and SP values are divided by 10^3, and Portfolio values are unscaled. Reduction is the percentage decrease from each method's fresh linear control. Positive reductions favor the nonlinear extractor. Bold marks the lowest mean regret in each row.

	BAYESIANSPO		BAYESIANMSE		BLL-ONLINE	
Benchmark	Regret	Reduction (%)	Regret	Reduction (%)	Regret	Reduction (%)
Knapsack 32	$\mathbf{16.55 \pm 1.97}$	4.5	17.90 ± 1.42	3.0	19.05 ± 1.64	4.4
Knapsack 100	$\mathbf{48.66 \pm 4.00}$	1.9	52.59 ± 4.83	3.7	69.13 ± 7.08	-3.3
SP 8×8	$\mathbf{5.69 \pm 0.93}$	11.3	6.04 ± 1.01	7.4	7.14 ± 0.49	16.9
SP 20×20	$\mathbf{20.88 \pm 2.58}$	9.0	21.07 ± 2.41	19.7	49.39 ± 1.51	3.9
Portfolio	$\mathbf{68.44 \pm 3.53}$	0.3	100.53 ± 8.90	0.9	356.31 ± 36.35	-4.3

BAYESIANSPO has the lowest mean regret among nonlinear methods on all five benchmarks (Table 10). Replacing the linear extractor with a nonlinear one lowers mean regret for both BAYESIANSPO and BAYESIANMSE on all five benchmarks. BLL-ONLINE improves on Knapsack 32 and both shortest-path instances, but has higher mean regret on Knapsack 100 and Portfolio.

Further work on nonlinear extractors. The Bayesian heads retain statistics computed from earlier features. As the extractor changes, those statistics remain tied to the old feature coordinates. Recomputing the heads from past observations or controlling feature drift are possible next steps. The current analysis fixes the feature coordinates, so extending it to changing representations remains future work, along with joint tuning of network depth, residual scale, and learning rate.

D.9 Pseudo-regret and realized regret

Pseudo-regret asks how well the learner acts relative to the best decision for the expected costs given the context, $\tilde{c}_t = \mathbb{E}[c_t \mid x_t]$. Realized regret compares the learner's cost with the lowest feasible cost after the realized vector c_t is known. Favorable noise can lower this hindsight minimum. A learner that always chooses an action optimal for $\tilde{c}_t$ therefore has zero pseudo-regret, but can still have positive expected realized regret.

Write $c_t = \tilde{c}_t + \epsilon_t$, where ϵ_t is the vector of cost noise. For a method a choosing actions w_t^a, the two metrics satisfy

$$\begin{aligned}\mathcal{R}^a_{\text{real}}(T) = \tilde{\mathcal{R}}^a(T) + \sum_{t=1}^{T} \epsilon_t^\top w_t^a \\ + \sum_{t=1}^{T} \left[\min_{w \in \mathcal{W}} \tilde{c}_t^\top w - \min_{w \in \mathcal{W}} c_t^\top w \right].\end{aligned} \tag{18}$$

The last sum is the same for every learner on an episode, so it cancels when comparing two methods. For the BAYESIANMSE minus BAYESIANSPO comparison, the remaining difference between metrics is

$$\sum_{t=1}^{T} \epsilon_t^\top \left(w_t^{\text{MSE}} - w_t^{\text{SPO}} \right).$$

Both methods choose their actions before the current cost noise is revealed. Under conditionally mean-zero noise, each summand has expectation zero given the shared past, the context, and both actions. The two metrics therefore give the same expected difference between the methods' regrets. Their individual regret values and percentage reductions can still differ.

Direct evaluation. We run both methods for $T = 10{,}000$ rounds with the submitted tuned settings and paired episode seeds 100–109. We compute both metrics on these fresh trajectories and report pseudo-regret in Table 11. Table 2 remains the main realized-regret comparison. The learner never sees the conditional mean costs used for evaluation. Both methods use the same reference cost, so it cancels when we subtract their regrets on a paired episode.

Table 11: Pseudo-regret from fresh evaluations over $T = 10{,}000$ rounds (mean $\pm$ standard deviation over ten paired seeds). Knapsack and SP values are divided by 10^3; Portfolio is unscaled.

Benchmark	BAYESIANSPO	BAYESIANMSE
Knapsack 32	$\mathbf{10.19 \pm 1.84}$	11.50 ± 1.57
Knapsack 100	$\mathbf{27.62 \pm 3.75}$	32.72 ± 4.46
SP 8×8	$\mathbf{5.59 \pm 1.19}$	5.75 ± 1.06
SP 20×20	$\mathbf{21.05 \pm 2.13}$	24.34 ± 2.56
Portfolio	$\mathbf{48.98 \pm 4.92}$	79.89 ± 8.32

BAYESIANSPO has lower mean pseudo-regret on all five benchmarks. Comparing realized regret and pseudo-regret on these same fresh trajectories, the two metrics favor the same method in all 50 seed–benchmark combinations. Within each benchmark, the Pearson correlation between the BAYESIANMSE minus BAYESIANSPO differences under the two metrics ranges from 0.9982 to 0.9998. Paired percentile-bootstrap 95% intervals for the percentage reduction exclude zero on four benchmarks and include zero on SP 8×8, using 20,000 resamples of the ten seeds.

D.10 Energy scheduling with historical prices

We use the energy-aware scheduling benchmark of Mandi et al. [20], with the released PredOpt `load3` instance and historical SEMO electricity prices from November 2011 to December 2013.[1] One round corresponds to one day. The context contains eight predictors for each of 48 half-hour slots, giving 384 inputs. The predictors are the holiday flag, weekday, week of year, month, wind-production forecast, ex-ante system load, ex-ante electricity price, and CO_2 intensity. The outcome is the vector of 48 realized system marginal prices (`SMPEP2`).

Decisions and feedback. The instance has 20 tasks and five machines. Task j has power demand P_j, duration d_j, earliest start r_j, and completion deadline ℓ_j, with times measured in half-hour slots. We consider every start that lets the task finish within its window, $s \in \{r_j, \dots, \ell_j - d_j\}$. This gives $\sum_{j=1}^{20}(\ell_j - d_j - r_j + 1) = 309$ possible task-start pairs. If $\pi_{t,h}$ is the realized price in slot h on day t, the cost of starting task j at slot s is

$$c_{t,(j,s)} = \tfrac{1}{2} P_j \sum_{h=s}^{s+d_j-1} \pi_{t,h}.$$

The factor $1/2$ accounts for half-hour slots. The model predicts these 309 costs directly. A mixed-integer program assigns each task a start time and a machine to minimize total predicted cost. Its 1,545 binary variables describe the possible assignments, subject to machine resource capacities in each slot. Each schedule selects 20 task-start pairs. During replay, the learner observes their realized costs, while the other 289 costs remain hidden. We use the complete cost vector reconstructed from historical prices only to evaluate regret.

Protocol. The data contain 789 complete days. We use the first 394 for preprocessing and hyperparameter selection, then keep those choices fixed and train the agents online on the remaining 395 days in chronological order. We standardize contexts using the first split and divide training costs by its median absolute nonzero component cost, $68{,}574$. Reported regret is converted back to the benchmark's original cost units. Each method uses its own Gurobi solver. All 30 learner seeds (0–29) see the same sequence of held-out days, so the reported variation measures the effect of learner randomness on that sequence.

Tuning selects width 8, prior precision $\lambda = 4$, and $\beta = 0$ for both BAYESIANSPO and BAYESIAN-MSE. Their learning rates are 0.003 and 0.001, respectively, and BAYESIANSPO uses one completion ($M = 1$). Tuning also selects a zero uncertainty bonus for BLL-ONLINE and LinUCB. These four methods therefore learn from partial feedback without an exploration bonus. Day-to-day variation in the observation inputs can expose the learner to different task-start costs through greedy action selection, alleviating the need for explicit exploration. The remaining methods use their tuned exploration settings.

Results. BAYESIANSPO has the lowest mean final regret among the six methods in Figure 4. It reaches 92.81 ± 13.29 million benchmark cost units, compared with 101.14 ± 12.17 for BAYESIAN-MSE (mean $\pm$ standard deviation), and has lower regret in 23 of the 30 paired runs. A one-sided paired Wilcoxon signed-rank test on final regret gives $p = 0.0031$ in favor of BAYESIANSPO. Comparing the two method means gives an 8.2% reduction. Computing the percentage reduction for each seed first and then averaging gives 7.24%, with a paired-bootstrap 95% interval of $[1.39\%, 12.60\%]$ from 100,000 resamples.

[1] Data and scheduling instances: `https://github.com/PredOpt/predopt-benchmarks/tree/main/Energy`.

E Core Design Ablations

This appendix groups the ablations around five parts of BAYESIANSPO: architecture, masking, learning, exploration, and imputation. These design choices determine where uncertainty enters, where gradients flow, and which objective trains the shared representation. Sensitivity sweeps over calibrated hyperparameters are separated in Appendix F; trajectory and prediction diagnostics are in Appendix D.

Ablation protocol. Unless stated otherwise, each ablation holds the main-run tuned hyperparameters from Section 6 and Appendix D.3 fixed, varies one knob, and reports final cumulative regret over the same $T = 10{,}000$ ten-seed protocol as Table 2. Core-design ablations use three representative instances (Knapsack 100, SP 20×20, and Portfolio); sensitivity sweeps and extended diagnostics retain all five benchmarks.

Sweeps center the knob around the tuned value where applicable. Deterministic head variants inherit the corresponding tuned Bayesian feature-extractor hyperparameters. These comparisons measure the effect of replacing the head at those settings. The coupled-head ablation is the exception: because joint head-and-representation training changes the optimization landscape, it includes the coarse $\eta_0/2, \eta_0, 2\eta_0$ learning-rate sweep described below.

Table 12 gives the roadmap for the core ablations.

Table 12: Roadmap of core design ablations. Rows are grouped by component; each line gives the ablation knob and where it is reported.

Component	Ablation	Description	Where
Architecture	Bayesian vs. non-Bayesian head	(A_i, v_i) posterior vs. deterministic θ_i; tests the uncertainty/completion scaffold.	App. E.1
	Decoupled vs. coupled head	Analytical BLR head update vs. joint SGD with ψ; tests prediction–DFL decoupling.	App. E.2
Masking	Full mask vs. unmasked endpoint	$w_{\text{miss}} = 0$ vs. $w_{\text{miss}} = 1$ on $i \notin S_t$ gradients; tests masking benefit.	App. E.3.1
	Unobserved-arm weight sweep	Sweep w_{miss} between full mask and full leakage; tests partial unobserved gradients.	App. E.3.2
Learning	SPO+ vs. MSE	$\ell_{\text{SPO}+}$ vs. $\|\hat{c} - \bar{c}\|_2^2$; isolates decision-focused representation learning.	Tab. 2; App. D.7
	SPO+ vs. PFY	SPO+ gradient vs. perturbed Fenchel–Young gradient; tests surrogate choice.	App. E.4
Exploration	UCB vs. greedy	Optimistic costs $\hat{c} - \beta\sigma$ vs. posterior means $\hat{c}$; tests the uncertainty bonus.	App. E.5.1
	UCB vs. Thompson sampling	Confidence bound vs. sampled $\tilde{\theta}_i \sim \mathcal{N}(\hat{\theta}_i, A_i^{-1})$; tests probability matching.	App. E.5.1
Imputation	Posterior mean vs. mean-anchored posterior completions	$M = 1 : \bar{c}_i = \hat{c}_i$; $M > 1$: average the mean completion with $M - 1$ posterior draws; tests uncertainty-aware completion.	App. E.6

BAYESIANSPO's head has two roles: the posterior supplies UCB uncertainty and cost imputations, and the analytical update keeps the head separate from masked SPO+ representation learning. Figure 10 summarizes both architecture checks.

E.1 Architecture: Bayesian vs. deterministic head

This ablation replaces the analytical Bayesian last layer with a deterministic linear head trained jointly with the feature map. The non-Bayesian head SPO variant keeps the masked SPO+ representation objective and inherits the corresponding tuned feature dimension and learning rate. Without posterior uncertainty, the deterministic head uses ε-greedy exploration with the fixed schedule $\varepsilon_t = 0.1 \max\{0, 1 - t/T\}$, which decays linearly over the evaluation horizon.

Removing the Bayesian head also removes posterior-directed exploration and posterior-mean imputations. This is still the relevant end-to-end replacement a practitioner would obtain by dropping the analytical head without adding a separate uncertainty system.

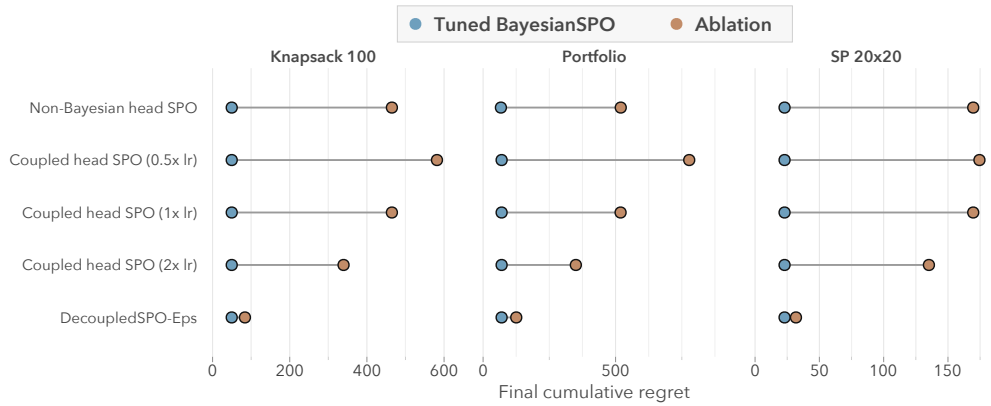


Figure 10: Architecture design ablations. Each row connects tuned BAYESIANSPO to an ablated variant within the same benchmark. The posterior row removes the Bayesian posterior head; the coupled-head rows replace the analytical head update with joint SPO+ training under a coarse learning-rate multiplier sweep. The DECOUPLEDSPO-EPS row retains the analytical head with ε-greedy exploration.

Figure 10 reports mean final cumulative regret across the three core-design instances and ten seeds. Replacing BAYESIANSPO with the non-Bayesian head SPO variant increases regret by about 7.5–$9.5\times$ across these instances. This row is the same algorithmic endpoint as the $1\times$ coupled-head variant below; the coupled sweep repeats it to show the learning-rate multiplier context.

E.2 Coupled vs. decoupled head update

Last-layer decoupling with a Bayesian head separates analytical posterior updates from masked-SPO+ updates of the features. An obvious alternative is to drop the analytical step and train head and features jointly under the same masked SPO+ loss. Each arm has a deterministic learnable parameter θ_i optimized together with ψ, replacing the Bayesian sufficient statistics (A_i, v_i) and posterior mean $\hat{\theta}_i = A_i^{-1} v_i$.

Three variants probe the comparison. DECOUPLEDSPO-UCB is the paper's main agent, with an analytical head and UCB exploration. DECOUPLEDSPO-EPS keeps the analytical head but uses ε-greedy on the same schedule as the coupled variant, isolating the head-update mechanism from the exploration rule. COUPLEDSPO-EPS replaces the Bayesian head with a deterministic linear head θ_i jointly trained with ψ via masked SPO+ gradients, dropping the analytical posterior update entirely. This deterministic variant uses ε-greedy exploration (Appendix E.1).

Coarse η_0 sweep on the coupled variant. Coupling trains both ψ and $\{\theta_i\}$, whereas the decoupled update sends masked SPO+ gradients only to ψ. We therefore evaluate COUPLEDSPO-EPS at $\eta_0/2$, η_0, and $2\eta_0$, using the same scheduled learning-rate form $\eta_t = \eta_0/\sqrt{t+1}$, and compare against the best of the three. The coupled architecture uses this three-point diagnostic budget, compared with 50 Optuna trials for the main method.

Figure 10's coupled-head rows compare each coupled variant against tuned BAYESIANSPO. With ε-greedy exploration, DECOUPLEDSPO-EPS is 1.4–$1.8\times$ worse than the UCB default, but the best coupled-head variant is still another 2.8–$4.3\times$ worse than DECOUPLEDSPO-EPS. The best coupled-head variant in this sweep remains about 5.0–$6.9\times$ worse than the final method across the core-design instances. Decoupling the predictive head from the decision-focused representation update reduces regret even under the same action-selection rule.

E.3 Masking Ablations

Masking zeroes the SPO+ gradient on unobserved arms before it reaches the representation update. The default is full masking, $g_i \leftarrow 0$ for $i \notin S_t$; the opposite endpoint lets imputed missing-arm gradients flow at full strength. These ablations ask whether the posterior-mean boundary pathology from Appendix E.6 matters at $T = 10{,}000$, holding tuned hyperparameters and UCB exploration fixed.

E.3.1 Masking Ablation

The endpoint ablation compares full masking ($w_{\text{miss}} = 0$) with full missing-arm leakage ($w_{\text{miss}} = 1$). The weight w_{miss} is applied only to unobserved-arm gradients during the representation update; observed-arm gradients are unchanged. Positive paired values below indicate higher regret than the full-mask run for the same method, benchmark, and seed.

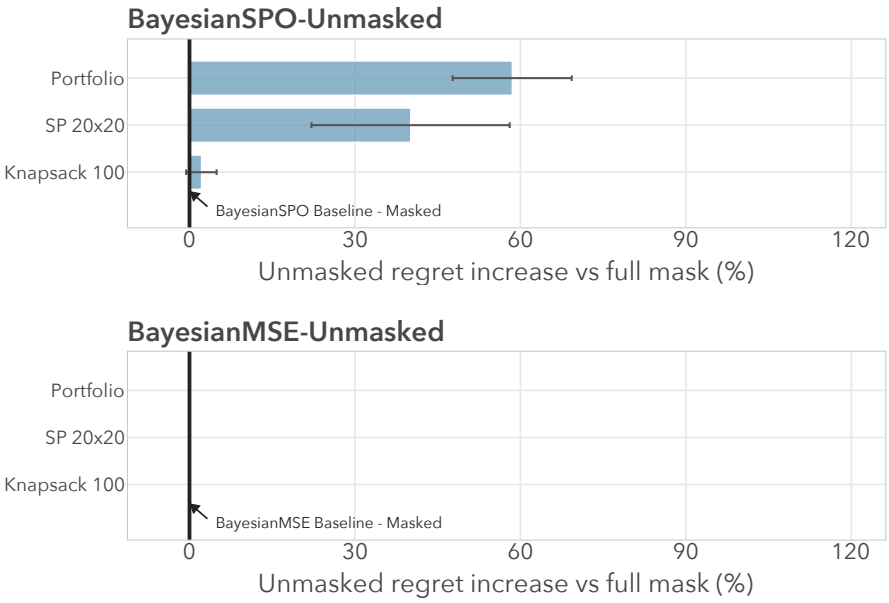


Figure 11: Endpoint view of the masking ablation. Horizontal bars report paired regret increase from setting $w_{\text{miss}} = 1$ relative to full masking ($w_{\text{miss}} = 0$). The BAYESIANMSE reference is plotted at zero because unobserved-arm residuals are zero by construction.

Figure 11 shows that full leakage increases BAYESIANSPO regret on every core-design instance: 2.2% on Knapsack 100, 40.1% on SP 20×20, and 58.5% on Portfolio, averaging 33.6%. The flat BAYESIANMSE reference is expected because missing entries are filled with the model's posterior mean, making the unobserved-arm MSE residual zero before masking. The endpoint test therefore isolates a decision-focused masking failure mode: for SPO+ gradients, full-strength imputed missing-arm updates can materially damage regret.

E.3.2 Unobserved-Arm Weight Ablation

The weight ablation turns the endpoint comparison into a sweep over $w_{\text{miss}} \in \{0, 0.05, 0.25, 0.5, 0.75, 1\}$. Intermediate values let some missing-arm gradient signal flow while damping imputation-induced leakage, so the sweep asks whether partial leakage is ever preferable to either endpoint.

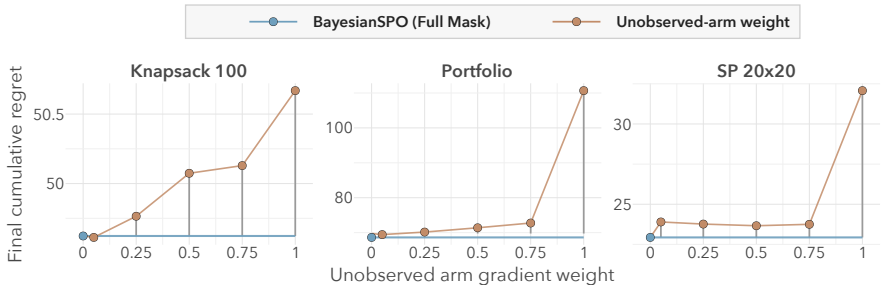


Figure 12: Matched regret view of the unobserved-arm weight ablation. The blue line marks the full-mask BAYESIANSPO baseline ($w_{\text{miss}} = 0$); orange points trace positive unobserved-arm gradient weights, and gray row segments show the same paired regret increase relative to full masking that defines the dose-response statistic. Values are mean final cumulative regret over ten paired seeds.

Figure 12 reports the paired regret change from full masking to each positive missing-arm weight. Among the displayed nonzero weights, the best mean paired changes are approximately 0.0% on Knapsack 100, $+3.6\%$ on SP 20×20, and -0.4% on Portfolio. Partial leakage therefore offers no consistent improvement over full masking. These results motivate studying imputation through its effect on the decision-focused gradient, as discussed in Appendix E.6.

E.4 Learning Objective

This subsection examines the choice of SPO+ as BAYESIANSPO's representation loss. We hold the per-arm Bayesian head, UCB exploration, and gradient masking fixed, and replace the SPO+ representation update with a perturbed Fenchel–Young (PFY) decision-focused surrogate [7]. Its gradient compares the average perturbed optimizer output with the target decision. Table 13 therefore asks whether the gain comes from using *any* decision-focused surrogate inside the Bayesian semi-bandit setup, or from the SPO+ surrogate specifically.

BAYESIANSPO obtains lower cumulative regret on all three core-design instances. The relative gap is smallest on SP 20×20, larger on Knapsack 100, and largest on Portfolio. The comparison uses a low online solver budget: PFY uses one or two perturbation samples per round, keeping solver calls in the same regime as the main method. Assessing their performance requires a separate tuning and evaluation budget.

Table 13: PFY vs. SPO+ surrogate comparison ($T = 10{,}000$, 10 seeds). Regret is mean $\pm$ standard deviation across seeds, in thousands for knapsack and shortest path and raw units for portfolio. Time is mean wall-clock seconds per evaluation seed. Regret and time gaps are $(\text{PFY} - \text{SPO})/\text{SPO}$, so positive values mean higher PFY regret or time.

Dataset	SPO regret	PFY regret	PFY gap	SPO time	PFY time	Time gap
Knapsack 100	$\mathbf{49.6 \pm 4.2}$	56.5 ± 6.7	$+13.9\%$	59.8	125.6	$+110.0\%$
SP 20×20	$\mathbf{22.9 \pm 2.2}$	24.9 ± 1.2	$+8.4\%$	32.9	24.8	-24.6%
Portfolio	$\mathbf{68.6 \pm 3.4}$	302.2 ± 32.8	$+340.4\%$	28.9	96.5	$+233.9\%$

PFY is faster on SP 20×20, but about $2.1\times$ slower on Knapsack 100 and $3.3\times$ slower on Portfolio.

E.5 Exploration Design

How masking, exploration, and imputation relate. The three missing-information axes act at different times. Masking controls which coordinates of the current SPO+ gradient reach the representation update. Exploration changes which arms become observable in future rounds. Imputation determines the completed target used by the current SPO+ solver calls before masking is applied.

E.5.1 Exploration strategies

The strategies compared in Figure 13 use the same learned representation and Bayesian last-layer posterior; they differ only in how the next feasible action is selected from that posterior.

Greedy. Select $w_t = w^\star(\hat{c}_t)$ using posterior means, with no uncertainty bonus or random exploration. This is the Greedy variant in Figure 13.

ε-Greedy. Select a random feasible action with probability ε_t; otherwise use the greedy decision. Exploration is undirected i.e. it does not preferentially observe uncertain or decision-relevant arms. This is the standard baseline for contextual bandits without Bayesian posteriors. The separate architecture and decoupling ablations (Appendices E.1 and E.2) use linear decay, $\varepsilon_t = 0.1 \max\{0, 1 - t/T\}$ for $t = 1, \ldots, T$. The main ε-Greedy baseline tunes the initial probability and whether to use this linear decay form or a constant probability (Appendix D.3).

UCB (Upper/Lower Confidence Bound). Adds a tuned exploration bonus $\beta\sigma_i$ to predictions, where $\sigma_i = \sqrt{h^\top A_i^{-1} h}$ is the posterior standard deviation. Uncertain arms receive larger bonuses, so UCB naturally explores arms with high posterior variance. For minimization problems, we use LCB (lower confidence bound): $\hat{c}_i - \beta\sigma_i$.

Thompson sampling. Draw independent $z_i \sim \mathcal{N}(0, 1)$ for each arm and solve the combinatorial problem on the sampled coefficients $c_i^{\mathrm{TS}} = \hat{c}_i + s_{\mathrm{TS}}\sigma_i z_i$. Raw TS uses $s_{\mathrm{TS}} = 1$, equivalent to drawing $\tilde{\theta}_i \sim \mathcal{N}(\hat{\theta}_i, A_i^{-1})$ and evaluating it at the current feature vector. The scale sweep varies s_{TS} while retaining the same posterior and representation update.

Figure 13 reports the matched exploration ablation for BAYESIANSPO and BAYESIANMSE on the three core-design instances. It varies only the action-selection rule while holding the Bayesian last-layer architecture fixed; the head-architecture ablation is separate (Figure 10). Values are paired cumulative-regret ratios relative to each method's tuned UCB default on a log scale; the solid vertical line marks UCB $= 1$. Greedy action selection is near UCB on SP 20×20 but much worse on the capacity-constrained benchmarks, roughly doubling regret on Knapsack 100 and reaching $5.35\times$ UCB for BAYESIANSPO on Portfolio ($3.46\times$ for BAYESIANMSE). Raw Thompson sampling is also worse than UCB across the board and is especially unstable on Portfolio ($16.28\times$ UCB for BAYESIANSPO, $11.19\times$ for BAYESIANMSE).

The TS scale sweep uses $s_{\mathrm{TS}}/\beta \in \{0.125, 0.25, 0.5, 1, 2\}$ on Knapsack 100 and SP 20×20, and $\{0.0625, 0.125, 0.25\}$ on Portfolio, around the tuned UCB β.

Decision-guided exploration as future work. One possible direction is to select observations by whether resolving that uncertainty could change the optimizer's decision. Near-boundary or locally exchangeable components give one possible acquisition target, for example by boosting UCB on decision-relevant arms or approximating information-directed sampling [24] over optimal decisions. We leave these rules to future work because they introduce an additional exploration objective and tuning problem beyond the posterior UCB mechanism studied here.

E.6 Imputation Design

When computing the SPO+ gradient, unobserved arms ($i \notin S_t$) must be assigned some completed target value $\bar{c}_i$. The choice of imputation determines what the model learns from each round. The posterior mean is the natural prediction-focused completion, and BAYESIANSPO treats it as the $M = 1$ case of a more general posterior-completion average. Figure 14 visualizes the four cells of the imputation $\times$ masking design space.

Model imputation (posterior mean). Set $\bar{c}_i = \hat{c}_i$ for unobserved arms. This is the MSE-optimal imputation given the posterior. It can still lie on decision boundaries where the solver is ambiguous. Masking removes direct gradient contributions from unobserved arms; imputed costs still affect both solver outputs.

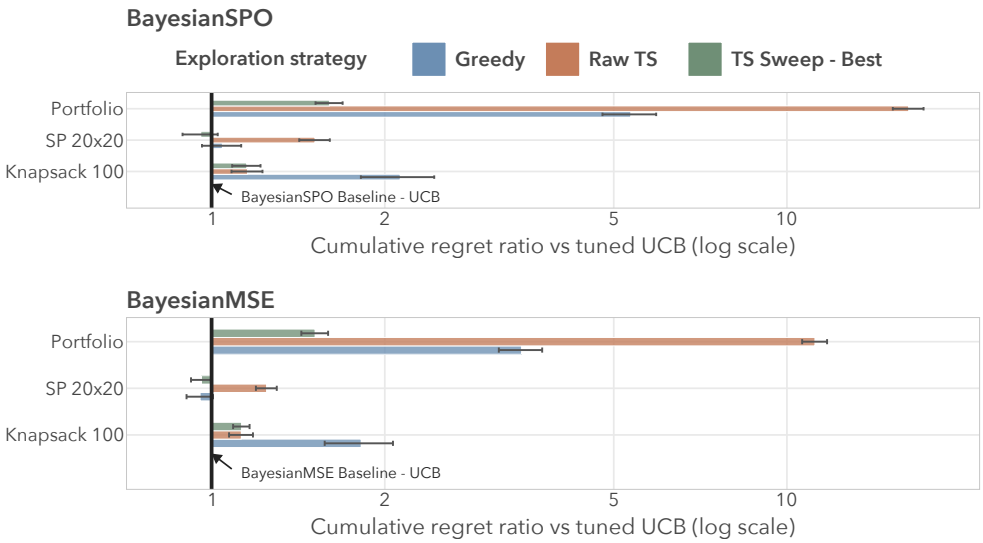


Figure 13: Exploration-strategy ablation using tuned BAYESIANSPO and BAYESIANMSE hyperparameters ($T = 10{,}000$, ten seeds). Greedy uses posterior means without random exploration. Each row compares alternatives against the tuned UCB default for the same representation loss and benchmark; the solid vertical line marks UCB $= 1$ on a log scale, and lower is better.

Posterior-completion averaging (M completions). For each round, observed arms use their true costs in every completion, $\bar{c}_i^{(m)} = c_{t,i}$ for $i \in S_t$. For missing arms, the first completion uses the posterior mean, $\bar{c}_i^{(1)} = \hat{c}_i$, and completions $m = 2, \ldots, M$ use posterior draws, $\bar{c}_i^{(m)} = \hat{c}_i + \sigma_i \xi_i^{(m)}$, with $\xi_i^{(m)} \sim \mathcal{N}(0,1)$. The masked SPO+ gradients are averaged, $g = \frac{1}{M} \sum_m g^{(m)}$.

The first completion is deterministic and retains weight $1/M$ in the gradient average. Conditional on the round's frozen quantities, let $g(C)$ be the masked gradient from a random posterior completion C. The mean-anchored average has expectation $M^{-1} g(\bar{c}^{(1)}) + (1 - M^{-1})\mathbb{E}[g(C)]$; an ordinary i.i.d. Monte Carlo average would sample all M completions. At $M = 1$ the procedure uses only posterior-mean completion; at $M > 1$ it adds posterior draws to average across optimizer boundaries. Each completion requires two training solver calls, giving $2M$ calls beyond action selection (Appendix D.6). Masking is retained throughout.

Empirical M sweep. We compare posterior-completion counts on each main benchmark. Table 14 reports the aggregate M selected by the final tuning summaries used for Table 2; Figure 15 reports the ten-seed means.

Table 14 shows the empirical tuning pattern: $M = 1$ on knapsack and portfolio, $M = 4$ on SP 8×8, and $M = 2$ on SP 20×20. The shortest-path coverage diagnostic in Appendix C is also sparser. Figure 14 illustrates sensitivity to point completions near optimizer boundaries; the sweep tests the usefulness of averaging directly.

Table 14: Aggregate tuner-selected posterior-completion count M for BAYESIANSPO.

Benchmark	Structural class	Selected M
Knapsack 32	Capacity-feasible swaps	1
Knapsack 100	Capacity-feasible swaps	1
Portfolio	Risk-feasible swaps	1
SP 8×8	Path-coupled	4
SP 20×20	Path-coupled	2

Increasing M slightly worsens regret on knapsack. For Portfolio, the displayed mean is lowest at $M = 1$ (68.63), compared with 70.04–70.45 for $M \in \{4, 8, 12\}$, with overlapping confidence intervals. SP 8×8 has its lowest mean regret at $M = 8$–12. On SP 20×20, mean regret ranges from 22.93K to 23.63K across the sweep, with the lowest mean at the tuned $M = 2$. The 95% confidence intervals overlap across these settings.

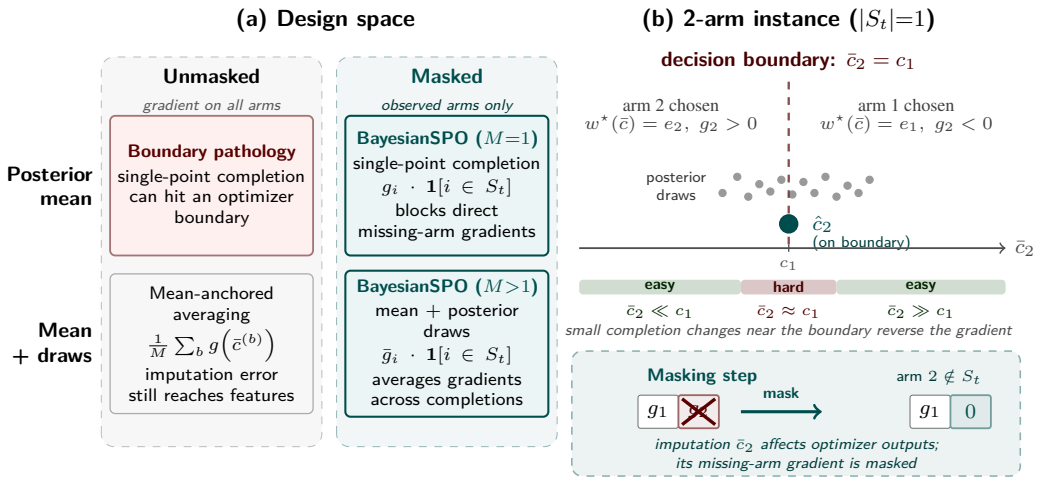


Figure 14: **The imputation × masking design space. (a)** The right column is BAYESIANSPO with the imputation hyperparameter M tunable: M=1 (top-right) is deterministic posterior-mean completion, and M>1 (bottom-right) averages the posterior-mean completion with $M-1$ posterior draws. The left column shows unmasked alternatives, where gradients from imputed missing-arm entries reach the representation update. The tuner selects M=1 on knapsack and portfolio and M>1 on shortest path. Panel (b) illustrates a possible reason for preferring multiple completions, not a measured boundary-distance classification of those benchmarks. **(b)** A minimal 2-arm instance with $\mathcal{W} = \{e_1, e_2\}$, arm 1 observed, and fixed tied predictions $\hat{c} = (c_1, c_1)$. The $\bar{c}_2$ axis varies the missing cost while $\bar{c}_1 = c_1$. On either side of the boundary $\bar{c}_2 = c_1$, the two SPO+ solver outputs are opposite, giving the displayed signs of g_2. At the boundary they coincide and a consistent tie rule gives zero gradient. The colored regions indicate sensitivity to completion changes: away from the boundary the gradient is locally constant; crossing it reverses the gradient. Posterior draws (grey cloud) illustrate this variation, not guaranteed recovery of the true-cost gradient. The *masking step* zeroes the unobserved-arm entry after computing the SPO+ gradient. Imputation can still alter optimizer outputs and observed-arm gradients, but its direct missing-arm gradient is blocked.

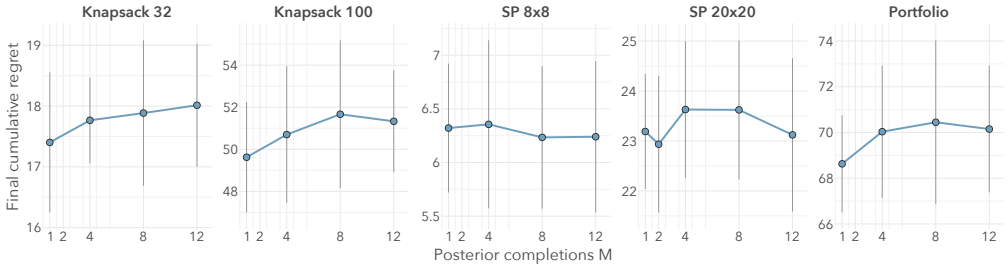


Figure 15: Posterior-completion count sweep for BAYESIANSPO. Each facet holds all other tuned hyperparameters fixed and varies only M, reporting mean final cumulative regret with 95% confidence intervals over ten seeds. Knapsack and SP regret is in thousands; Portfolio is unscaled.

Why posterior-mean completion can be a poor gradient target. Posterior-mean completion minimizes missing-cost MSE, but the SPO+ update depends on optimizer outputs induced by the completed vector. Appendix B.10 gives a diamond-graph example where the mean completion ties the observed and unobserved paths. At the specified predictor, both solver inputs coincide and a consistent tie rule yields a zero subgradient despite a nonzero realized full-information update. Off-boundary completions can also reverse that update. Preserving both true-cost optimizer outputs is sufficient for gradient agreement.

The optimality-cone view makes the design tension explicit. For a decision w, let $\Lambda(w) = \{c : w \in \operatorname{argmin}_{a \in \mathcal{W}} c^\top a\}$. Prediction-focused imputation asks for a likely missing cost vector; decision-

focused imputation would ask for a completion that separates plausible candidate solutions, preferably away from the boundaries of these cones. This echoes the "relevant extreme directions" of Bennouna et al. [6]: the informative directions are those along which costs cross between optimality cones.

Discriminative imputation as future work. The conservative masking strategy does not exhaust the imputation design space. A more ambitious direction is *discriminative imputation*: choose unobserved-arm targets that separate candidate solutions, rather than the most likely missing costs. One version may prefer completions away from optimality-cone boundaries, while another may train imputed values to approximate the SPO+ gradient itself. The per-dataset M pattern is a cheap first step in this direction, but learning non-trivial decision-focused imputations remains open.

F Sensitivity Sweeps

These sweeps probe the stability of the BAYESIANSPO versus BAYESIANMSE comparison around the tuned operating point. Each holds the tuned hyperparameters of Appendix D.3 fixed and varies a single knob, following the protocol introduced at the start of Appendix E. The capacity sweep below varies feedback density in the capacity-constrained benchmarks and reports the empirical number of selected components; the remaining three (β, λ, and DGP noise width w_0) are confirmatory robustness checks.

Table 15: Roadmap of sensitivity sweeps. Each sweep varies one knob around the tuned operating point while holding the remaining configuration fixed.

Component	Sweep	Description	Where
Feedback density	Capacity / cardinality	Knapsack capacity B and portfolio cardinality k change the observed-arm count $\lvert S_t \rvert$.	App. F.1
Exploration	UCB scale	Tests sensitivity to UCB optimism scale β.	App. F.2
Posterior	Prior precision	Tests posterior calibration through prior precision λ.	App. F.3
Noise variation in data	DGP noise width w_0	Tests robustness to the DGP noise width w_0 (implemented as `noise_width`).	App. F.4

F.1 Capacity-constrained feedback-density sweep

We sweep the weighted knapsack benchmarks at $n = 32$ and $n = 100$ by multiplying each benchmark capacity by $\{0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4\}$, and sweep the cardinality-constrained portfolio benchmark over $k \in \{2, 3, 4, 5, 6, 8, 10\}$. For weighted knapsack, the selected-set size depends on item weights, contexts, and the learned policy as well as capacity. Figure 16 therefore uses the empirical number of selected components per round on the horizontal axis. The sweep tests how much observed decision signal masked BAYESIANSPO needs before the decision-focused update outperforms prediction-focused training.

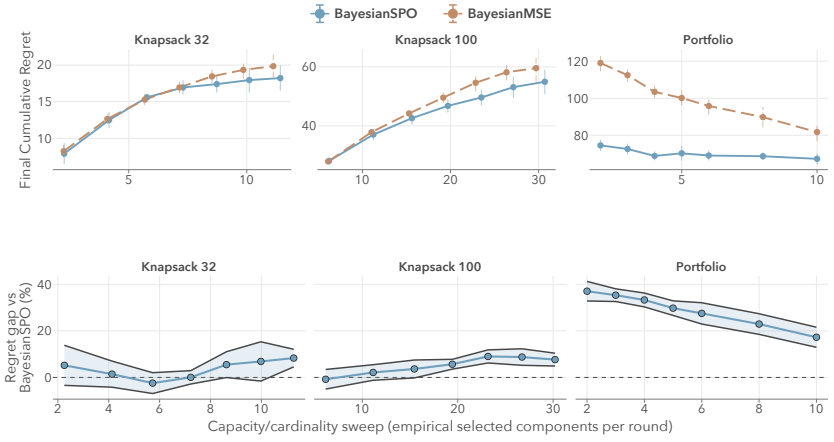


Figure 16: Capacity-constrained feedback-density sweep on weighted knapsack and cardinality-constrained portfolio. The top panel shows final cumulative regret for the matched Bayesian last-layer losses. The bottom panel averages the paired per-seed percentage gaps $100(R_{\mathrm{MSE}} - R_{\mathrm{SPO}})/R_{\mathrm{MSE}}$, where R is final cumulative regret. We vary capacity or cardinality. The horizontal axis reports mean selected components per round, averaged over seeds for each method in the top panel and over both methods and paired seeds in the bottom panel.

F.2 β (UCB scale)

We sweep $\beta \in \{0.1, 0.3, 1.0, 3.0, 5.0\}$ on every benchmark, matching the tuning range for the UCB scale. Figure 17 reports the resulting final cumulative regret directly. The shortest-path benchmarks have flatter sensitivity curves than Portfolio, with sensitivity especially apparent on SP 20×20 at large β. Portfolio is U-shaped, and knapsack improves over the displayed range as β increases. These dataset-dependent responses support per-benchmark calibration of the UCB scale.

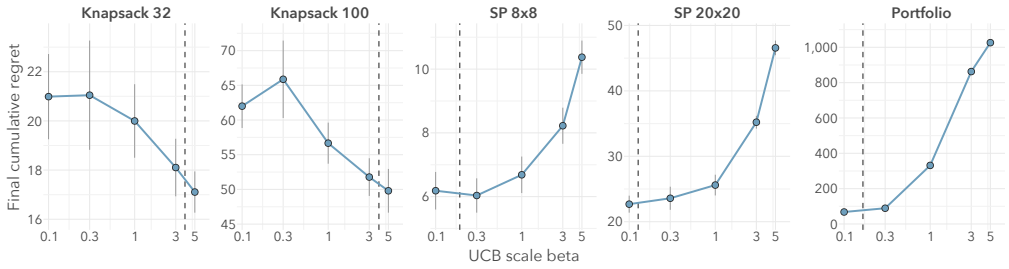


Figure 17: Sensitivity to the UCB scale. Each panel reports final cumulative regret over ten seeds for the same fixed β grid. Dashed vertical lines mark the nearest grid point to the actual tuned value used for the main BAYESIANSPO runs.

F.3 λ (prior precision)

We sweep $\lambda \in \{0.01, 0.1, 1.0, 10.0\}$. Figure 18 reports final cumulative regret directly. The capacity-constrained benchmarks prefer weak prior precision, especially Portfolio where large λ prevents the Bayesian head from adapting and increases regret. The shortest-path benchmarks are flatter and favor moderate precision. These results support choosing the prior precision separately for each dataset.

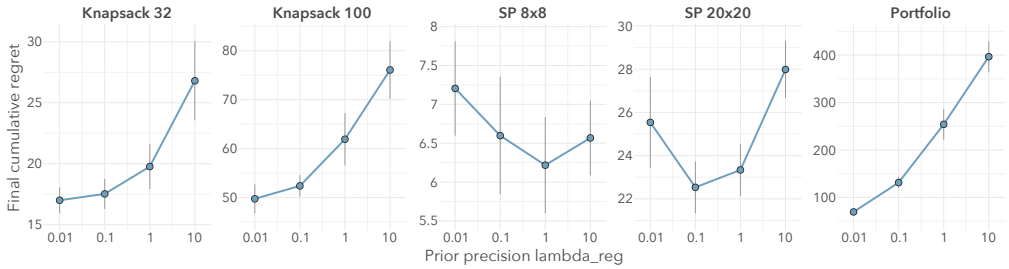


Figure 18: Sensitivity to Bayesian last-layer prior precision. Each panel reports final cumulative regret over ten seeds for the displayed λ values. Dashed vertical lines mark the nearest grid point to the actual value used for the main BAYESIANSPO runs.

F.4 DGP noise width w_0

We sweep the DGP noise width w_0 across $\{0, 0.25, 0.5, 1.0\}$. Knapsack and shortest path use multiplicative uniform noise; portfolio uses additive factor and idiosyncratic noise (Appendix C.3). In portfolio, both Σ and the risk budget scale with w_0^2, so positive noise widths leave the mathematical risk-feasible set unchanged for a fixed seed. At $w_0 = 0$, $\Sigma = 0$ and every exact-cardinality selection satisfies the risk constraint. The zero-noise portfolio point therefore also changes the constraint geometry. On SP 8×8, the paired regret gap varies with noise width.

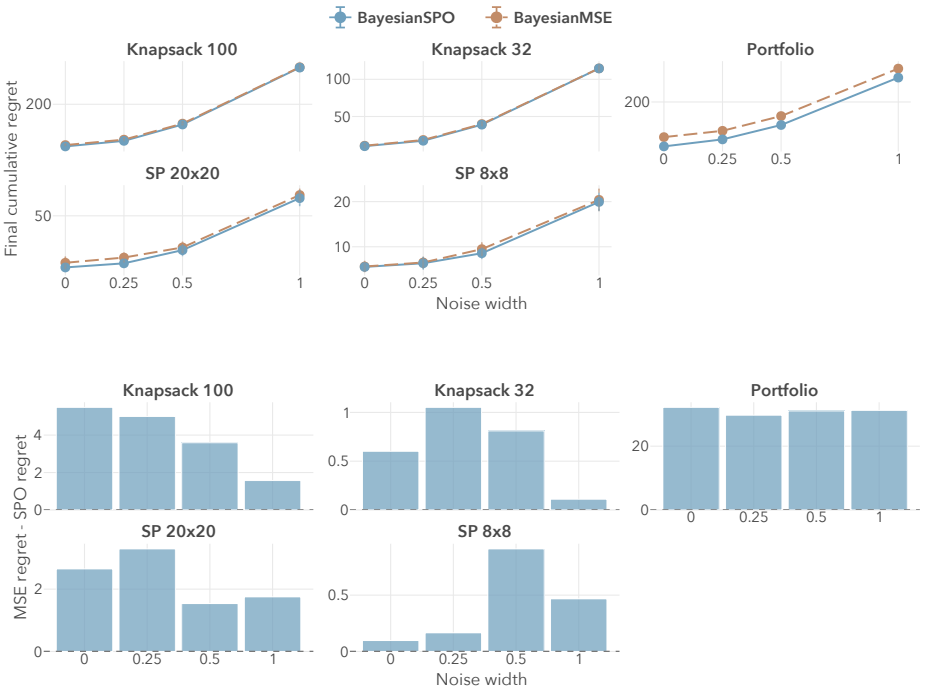


Figure 19: Sensitivity to DGP noise width w_0. The top panel reports final cumulative regret by method and benchmark; the bottom panel reports the regret gap $\mathcal{R}(\text{BAYESIANMSE}) - \mathcal{R}(\text{BAYESIANSPO})$, so positive values favor BAYESIANSPO.

Broader impact. This work is methodological: it improves learning for online combinatorial decision systems under partial feedback, with potential benefits in settings such as routing, allocation, recommendation, and inventory where better exploration can reduce cumulative cost while learning. The same adaptivity can also amplify harms if deployed with biased objectives, incomplete feedback, or constraints that omit fairness, safety, or service-quality requirements. Deployments should therefore pair the algorithm with domain-specific constraints, monitoring of exploration behavior, and audits of who benefits or is burdened by the learned decisions.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The abstract and Section 1 state the paper's scope as online semi-bandit decision-focused learning, identify imputation degeneracy, introduce BAYESIANSPO, and summarize the theory and empirical claims. The limitations of the theory and experiments are discussed in Section 5, Section 7, and the appendices.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: Section 7 discusses theoretical scope, benchmark misspecification, and solver-call overhead. Section 6 and Appendix C specify the benchmark scope and binary portfolio adaptation. Appendix B.7 states the additional observation-coverage condition used only by the global refinement (Corollary 3).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

Justification: Section 5 states the modelling assumptions and refers to the formal setup in Appendix B.1. Appendix B gives the supporting lemmas, full proofs, and lower-bound/counterexample material. Corollaries 1 and 3 in Appendix B.7 state the full loss bounds and the global refinement's additional coverage condition.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes].

   Justification: Section 6 describes benchmarks, methods, tuning, seeds, horizon, and metrics. Appendices C, D.1, D.3, E, and D provide benchmark formulations, data-generation details, baseline definitions, hyperparameter ranges, and extended results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions, model checkpoints, hosted models, or other means appropriate to the work.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes].

   Justification: For review, an anonymized repository was provided at `https://anonymous.4open.science/r/OnlineBanditDFL/`. The repository has now been made public at `https://github.com/Aabhash/OnlineSemiBanditDFL`. It contains training and evaluation code, benchmark-generation code, configuration files, and dependency specifications. The code is self-contained for the reported experiments and does not require specialized hardware. Knapsack and portfolio use the external Gurobi solver through `gurobipy` as the optimization oracle.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution.
- The instructions should contain the exact command and environment needed to run to reproduce the results.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes].

   Justification: Section 6 provides the protocol, including tuning trials, tuning and evaluation seeds, horizon, and evaluation metric. Appendix D.3 lists hyperparameter spaces, Appendix D.1 defines baselines, and Appendix C specifies benchmark details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes].

   Justification: Table 2 reports means and standard deviations over ten evaluation seeds, with one-sided Wilcoxon signed-rank tests against the best baseline. Additional appendix figures report confidence intervals or paired summaries for ablations and sweeps. Figure 5 is a single-seed illustration (Appendix C.4), with cross-validation measuring probe performance for that fixed representation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated.
   - The method for calculating the error bars should be explained.
   - The assumptions made should be given.
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it.
   - For asymmetric distributions, the authors should be careful not to show invalid symmetric error bars.
   - If error bars are reported in tables or plots, the authors should explain how they were calculated and reference the corresponding figures or tables.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes].

Justification: Appendix D.2 states how experiments were conducted. Appendix D.6 reports wall-clock timing for the final evaluation methods.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes].

Justification: The work is algorithmic; it does not involve human subjects, private data, scraped data, or high-risk model release. We have reviewed the NeurIPS Code of Ethics and are not aware of any deviation.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes].

Justification: The broader-impact paragraph near the end of the appendix discusses positive uses in routing, allocation, recommendation, and inventory, and notes risks from biased objectives, incomplete feedback, and missing fairness, safety, or service-quality constraints.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include malicious or unintended uses, fairness considerations, privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments.
- The authors should consider possible harms when the technology is used as intended, when it gives incorrect results, and following misuse.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: The paper does not release a high-risk pretrained model, scraped dataset, image generator, language model, or similarly sensitive asset. The released materials are code and generated benchmark configurations for optimization experiments.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards.
- Datasets that have been scraped from the Internet could pose safety risks.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes].

Justification: The code repository includes a GPL-3.0-or-later license for our released code. The paper cites PyEPO for the benchmark formulations used to generate synthetic instances, and the repository documents Gurobi/`gurobipy` as an external optimization-solver dependency subject to its own license. For energy scheduling, we credit the benchmark of Mandi et al. [20] and the PredOpt release [21], which supplies the `load3` instance and historical Irish SEMO electricity data (`https://github.com/PredOpt/predopt-benchmarks/tree/main/Energy`). The PredOpt repository uses the MIT License (`https://github.com/PredOpt/predopt-benchmarks/blob/main/LICENSE`), which requires retaining its copyright and permission notices; these are reproduced in `data/Energy/README.md`. The electricity observations originate from SEMO; a separate license for these observations is not specified in the cited PredOpt release.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license should be included for each asset.
- For scraped data from a particular source, the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes].

Justification: The new assets are the BAYESIANSPO implementation, training/evaluation scripts, generated benchmark configurations, and result-generation utilities. The paper and appendix document the algorithm, benchmark generation, hyperparameter tuning, evaluation protocol, and compute requirements; the public repository provides the corresponding code and dependency files.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable).

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing, user studies, or other research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing, user studies, or other human-subjects research, so IRB review is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research.
    - We recognize that the procedures for this may vary significantly between institutions and locations.
    - For initial submissions, do not include any information that would break anonymity.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A].

    Justification: LLMs were used as assistance tools for manuscript drafting and editing, reference discovery, theory/proof-exposition checks, and data-processing/plotting code assistance.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.