CORRUPTION-ROBUST SPARSE LINEAR CONTEXTUAL BANDITS WITH KNAPSACK CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study sparse linear contextual bandits with knapsack constraints under joint reward and consumption corruption. Consumption corruption creates a challenge beyond corrupted rewards: it affects not only statistical estimates, but also the recorded budget, resource prices, and stopping decisions that govern future allocation. We develop Robust Optimistic Primal–Dual (ROPD), an estimator-modular framework that combines corruption-aware confidence widths with online resource prices and a budget-safety rule. With concrete sparse implementation, ROPD achieves regret against a clean population-LP benchmark of $\widetilde{O}(T^{2/3} + \Gamma T^{1/3})$ under forced exploration and population-design coverage, and $\widetilde{O}(\sqrt{T} + \Gamma)$ under on-policy realized-design coverage, for a supplied valid corruption bound Γ under the stated proportional-budget scaling and fixed model/design parameters. When the corruption level is unknown, Shared-Grid adapts confidence radii around common point estimates fitted to a single realized history, incurring explicit initialization and master-comparison costs; its sharper on-policy guarantee additionally requires recommendation coverage. Both methods preserve observed budgets on every realization and bound clean resource violation by cumulative consumption corruption. These results connect corruption-robust sparse estimation with resource accounting, pricing, and stopping in high-dimensional online allocation.

1 INTRODUCTION

Contextual bandits with knapsack constraints (CBwK) couple learning with resource allocation: after observing a context, a learner chooses an action whose reward must be learned while its consumption depletes finite budgets (Badanidiyuru et al., 2018; Agrawal & Devanur, 2016; Agrawal et al., 2016). In high-dimensional applications, only a small subset of features may govern reward and consumption, motivating sparse action-specific models (Bastani & Bayati, 2020; Hao et al., 2020; Ma et al., 2024).

Corruption of consumption feedback creates a difficulty beyond corrupted rewards. A corrupted reward changes what the learner infers about an action, whereas corrupted consumption also changes the state used to allocate future resources—consumption estimate, recorded budget, resource prices, and potentially stopping time. For example, an advertisement that is truly expensive but repeatedly recorded as inexpensive can be favored while the observed budget account still appears feasible.

A robust estimator alone therefore does not resolve the allocation problem. The analysis must control how corrupted observations affect later predictions, relate observed consumption to the latent clean resource use, and account for reward lost when corrupted resource information changes stopping. In the sparse setting, these guarantees also depend on whether the contexts actually collected for each action are informative. When the corruption level is unknown, candidates calibrated to different levels may recommend different actions even though feedback is observed only for the actions that are actually played.

We address these issues with Robust Optimistic Primal–Dual (ROPD), which combines corruption-aware prediction widths with online resource prices and a budget-safety rule. The allocation analysis is estimator-modular: it uses only pointwise confidence and cumulative uncertainty along the recommended actions. Norm-constrained Lasso provides one concrete sparse implementation. We study both forced exploration and on-policy learning, and for unknown corruption introduce Shared-Grid, which compares confidence calibrations around common point estimates fitted to one realized history.

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

Table 1: Guarantee regimes and leading terms when corruption is absent.

Corruption information	Sampling	Coverage	Corruption-free term
Bound $\Gamma \geq C_Z$	Forced	Population design	$\widetilde{O}(T^{2/3})$
Bound $\Gamma \geq C_Z$	On-policy	Realized design	$\widetilde{O}(\sqrt{T})$
Not supplied	Forced	Population design	$\widetilde{O}(T^{2/3}) + \mathrm{Reg}_{\mathrm{mst}}(T)$
Not supplied	On-policy	Realized design and recommendation coverage	$\widetilde{O}(\sqrt{T}) + \mathrm{Reg}_{\mathrm{mst}}(T)$

Orders assume proportional budgets, fixed model and coverage parameters, dominated initialization and burn-in, and negligible failure costs. $\mathrm{Reg}_{\mathrm{mst}}(T)$ is the master-comparison cost; Corollaries 1–2 give the full bounds.

1.1 Main Contributions

1. Joint statistical and resource effects of corruption. We measure joint reward–consumption corruption through the priced budget C_Z and separate its two roles: contaminated samples affect future prediction, while corrupted consumption directly changes resource accounting, price updates, and stopping. This decomposition yields regret guarantees together with exact observed-budget feasibility and an explicit bound on clean resource violation.

2. Estimator-modular robust allocation with two data routes. Given a valid bound $\Gamma \geq C_Z$, ROPD combines optimistic predictions with online resource prices. Its allocation theorem requires only pointwise confidence and cumulative recommendation-width control; norm-constrained Lasso is one concrete implementation. Forced exploration under population design gives the $\widetilde{O}(T^{2/3})$ corruption-free regime, while realized on-policy design gives $\widetilde{O}(\sqrt{T})$ under the conditions summarized in Table 1.

3. Adaptation to unknown corruption on shared data. Shared-Grid compares corruption calibrations around common point estimates fitted to the single realized history, rather than analyzing candidates on counterfactual estimator histories. Its guarantee retains initialization and master-comparison costs explicitly; the sharper on-policy result additionally requires coverage of candidate recommendations by realized plays.

Relation to closest work. Clean high-dimensional CBwK already combines sparse estimation with primal–dual resource control (Ma et al., 2024), while corruption-robust linear and contextual bandits primarily address corrupted reward feedback without knapsack state (He et al., 2022; Liu et al., 2024). Thus the new difficulty here is not sparse estimation or dual pricing in isolation: corrupted consumption simultaneously affects prediction, resource-price path, recorded budget, and stopping. We analyze these effects against a clean population-LP benchmark and, when the corruption level is unknown, adapt on a single realized estimator history. Appendix A gives the broader comparison.

2 Model, Feedback, and Benchmark

2.1 Interaction and Feedback

Over T rounds, the learner chooses from K non-null actions $\mathcal{A} = [K] = \{1, \ldots, K\}$, and a null action with zero reward and consumption. Write $\mathcal{A}_0 = \mathcal{A} \cup \{\texttt{null}\}$. Each context x_t lies in $\mathbb{R}^d$, and the m resource budgets are $B = (B_1, \ldots, B_m)^\top$ with $B_i > 0$ for every $i \in [m]$.

Clean potential outcomes are the rewards and consumptions each action would generate before corruption. Let $\mathcal{F}_{t-1}$ contain the full past, including earlier clean potential outcomes and corruption; the learner uses only its observed subhistory. The budget-safety test determines whether the round is active before x_t is drawn. On an active round, the learner observes x_t; the environment generates clean potential outcomes and fixes each action's corruption; then the learner selects $a_t \in \mathcal{A}_0$ using fresh randomness conditionally independent of those outcomes and corruption. Only the selected feedback is observed. Both the action and whether the resulting observation is retained for estimation are decided before the round's reward and consumption feedback is observed. Appendix B specifies the full protocol.

For unknown coefficients $\mu_a^\star, w_{a,i}^\star$ and $a \in \mathcal{A}$, clean conditional means are

$$r^0(a,x) = \langle \mu_a^\star, x \rangle, \quad b_i^0(a,x) = \langle w_{a,i}^\star, x \rangle, \quad i \in [m]. \tag{1}$$

Write $b^0 = (b_1^0, \ldots, b_m^0)^\top$; the null action has zero means.

Assumption 1 (Sparse linear contextual model)**.** Given $\mathcal{F}_{t-1}$, x_t has distribution $\mathcal{D}$ and satisfies $\|x_t\|_2 \le 1$. For every action and resource, $r_t^{\mathrm{cl}}(a) = r^0(a,x_t) + \eta_t^r(a)$ and $b_{t,i}^{\mathrm{cl}}(a) = b_i^0(a,x_t) + \eta_{t,i}^b(a)$, where $\mathbb{E}[\eta_t^r(a) \mid \mathcal{F}_{t-1}, x_t] = 0$ and $\mathbb{E}[\eta_{t,i}^b(a) \mid \mathcal{F}_{t-1}, x_t] = 0$. The reward noise η_t^r and consumption noise $\eta_{t,i}^b$ are conditionally sub-Gaussian with parameters at most σ_r and σ_b, respectively. Clean rewards and every clean consumption coordinate lie in $[0,1]$. For every a, i, the support size $\|\cdot\|_0$ and Euclidean norm satisfy $\|\mu_a^\star\|_0, \|w_{a,i}^\star\|_0 \le s_0$ and $\|\mu_a^\star\|_2, \|w_{a,i}^\star\|_2 \le R$, where s_0 is true sparsity and $R > 0$ is the parameter-norm bound.

Observed feedback is $r_t(a) = \mathrm{clip}(r_t^{\mathrm{cl}}(a) + \xi_t^r(a), 0, 1)$ and $b_{t,i}(a) = \mathrm{clip}(b_{t,i}^{\mathrm{cl}}(a) + \xi_{t,i}^b(a), 0, 1)$, where ξ^r, ξ^b are raw adversarial perturbations and $\mathrm{clip}(z,0,1) = \min\{1, \max\{0, z\}\}$. Writing b_t, b_t^{cl} and c_t^b for the corresponding resource vectors. The post-clipping deviations are

$$c_t^r(a) = r_t(a) - r_t^{\mathrm{cl}}(a), \quad c_{t,i}^b(a) = b_{t,i}(a) - b_{t,i}^{\mathrm{cl}}(a). \tag{2}$$

2.2 Population-LP Benchmark and Effective Corruption

The clean population linear program (LP) assigns static, context-dependent probabilities $y(a \mid x)$ to non-null actions, with remaining mass on `null`:

$$\begin{aligned} V^{\mathrm{UB}} := \max_{y} \quad & T\,\mathbb{E}_x\left[\sum_{a\in\mathcal{A}} r^0(a,x) y(a \mid x)\right] \\ \text{s.t.} \quad & T\,\mathbb{E}_x\left[\sum_{a\in\mathcal{A}} b_i^0(a,x) y(a \mid x)\right] \le B_i, \quad i \in [m], \\ & y(a \mid x) \ge 0, \quad \sum_{a\in\mathcal{A}} y(a \mid x) \le 1, \quad \forall x. \end{aligned} \tag{3}$$

Fix a measurable optimizer $y^\star$ before interaction. Define clean expected regret by

$$\mathrm{Reg}(T) := V^{\mathrm{UB}} - \mathbb{E}\left[\sum_{t=1}^T r^0(a_t, x_t)\right], \tag{4}$$

where the expectation includes contexts, clean noise, learner randomness, and randomized corruption. The learner does not know or solve the LP. For $Z > 0$ and normalized resource prices $\lambda \in \Lambda := \{\lambda \in \mathbb{R}_+^m : \|\lambda\|_1 \le 1\}$, define the clean Lagrangian score $F^0(a,x;\lambda) = r^0(a,x) - Z\lambda^\top b^0(a,x)$.

Assumption 2 (Lagrangian scale)**.** Choose a positive scale $Z > 0$ satisfying $Z \ge V^{\mathrm{UB}}/B_{\min}$, where $B_{\min} = \min_i B_i$.

The choice $Z = T/B_{\min}$ is computable from the horizon and budgets and satisfies Assumption 2 because $V^{\mathrm{UB}} \le T$. Thus, in the stopping argument, depletion of any resource carries enough dual value to offset the remaining benchmark reward. Under proportional budgets, $B_{\min} = \Theta(T)$ and $B_{\max} := \max_i B_i = O(T)$, hence $Z = O(1)$; our formal bounds nevertheless keep the dependence on these quantities explicit.

For every $\lambda \in \Lambda$, the score change caused by one corrupted observation satisfies

$$\left|c_t^r(a) - Z\lambda^\top c_t^b(a)\right| \le |c_t^r(a)| + Z\|c_t^b(a)\|_\infty. \tag{5}$$

This motivates the effective corruption budget

$$C_Z := \sum_{t=1}^T \max_{a\in\mathcal{A}} \left\{|c_t^r(a)| + Z\|c_t^b(a)\|_\infty\right\}. \tag{6}$$

Known corruption means that the learner receives a deterministic input Γ satisfying $C_Z \le \Gamma$ on every admissible realization; unknown corruption means that no such input is supplied.

2.3 Observed and Clean Budget Feasibility

For an operating budget $0 < B^{\mathrm{op}} \le B$, define cumulative observed use $U_t = \sum_{\tau<t} b_\tau(a_\tau)$. A non-null action is allowed only if

$$U_{t,i} \le B_i^{\mathrm{op}} - 1, \quad \forall i \in [m]. \tag{7}$$

Otherwise the learner selects `null` from that round onward. Let $C_b(T) := \sum_{t=1}^{T} \|c_t^b(a_t)\|_\infty$ be the cumulative consumption corruption along the played path and let $[z]_+ := \max\{z, 0\}$. Define the observed and clean realized budget violations by $\mathrm{Viol}_{\mathrm{obs}}(T) := \max_{i\in[m]}[\sum_{t=1}^{T} b_{t,i}(a_t) - B_i]_+$ and $\mathrm{Viol}_{\mathrm{cl}}(T) := \max_{i\in[m]}[\sum_{t=1}^{T} b^{\mathrm{cl}}_{t,i}(a_t) - B_i]_+$.

With $B^{\mathrm{op}} = B$, the safety rule gives exact observed feasibility, $\mathrm{Viol}_{\mathrm{obs}}(T) = 0$, on every realization. If $C_Z \le \Gamma$, it also gives $\mathrm{Viol}_{\mathrm{cl}}(T) \le C_b(T) \le \Gamma/Z$; Proposition 2 in Appendix B proves these claims and gives the buffered version for strict clean feasibility. Appendix B also collects the feedback, budget, and coverage notation.

3 Estimation and Optimistic Primal–Dual Learning

The proposed method separates estimation from allocation. An estimation module supplies action-specific reward and consumption predictions with confidence widths, while the allocation rule uses these outputs to choose actions and update resource prices. The allocation guarantees depend only on the estimator interface in Section 3.2, not on a particular regression method. We use norm-constrained Lasso as one concrete implementation. This section defines that estimator and states its prediction guarantee; the regularization, design calibration, proof, and certified optimization details are deferred to Appendix E.

3.1 Sparse Estimation and Prediction Confidence

Fix an action a and one response channel: reward with parameter $\theta^\star = \mu_a^\star$, or resource i with $\theta^\star = w_{a,i}^\star$. Let $\mathcal{S}_a^{\mathrm{est}}(t) = \{\tau < t : a_\tau = a \text{ and the sample is retained}\}$ and $n = N_a^{\mathrm{est}}(t) := |\mathcal{S}_a^{\mathrm{est}}(t)|$. The matrix $X \in \mathbb{R}^{n\times d}$ contains the retained contexts as rows, and $y \in \mathbb{R}^n$ contains the observed responses. For analysis, write $y = X\theta^\star + \eta + \zeta$, where η is clean noise and ζ is post-clipping corruption. Retention is decided before feedback; all channels for an action use the same contexts.

With regularization $\rho_n > 0$, one implementation is norm-constrained Lasso (Tibshirani, 1996):

$$\widehat{\theta} \in \operatorname*{arg\,min}_{\|\theta\|_2 \le R} \left\{ \frac{\|y - X\theta\|_2^2}{2n} + \rho_n \|\theta\|_1 \right\}, \qquad n \ge 1. \tag{8}$$

The penalty promotes sparse estimates without imposing a support constraint; set $\widehat{\theta} = 0$ when $n = 0$. For the following Proposition 1, $s \ge \max\{1, s_0\}$ is the calibration sparsity, n_0 is the minimum retained-sample threshold, $\kappa > 0$ is the retained-design restricted-strong-convexity (RSC) curvature, and $\delta_{\mathrm{est}} \in (0, 1)$ is the estimation failure level. Appendix E.2 gives the calibration of ρ_n, κ, n_0 and Appendix E.8 gives the certified optimization tolerance. The fitted coefficients and regularization do not depend on the corruption candidate Γ.

Proposition 1 (Prediction confidence for the concrete estimator)**.** *Under Assumption 1 and the calibration specified in Appendix E.2, there is a noise event of probability at least $1 - \delta_{\mathrm{est}}$, simultaneous over actions, response channels, and obtained sample prefixes. On this event, whenever a retained prefix $n \ge n_0$ satisfies the calibrated RSC condition, any feasible fit meeting the certified optimization tolerance of Appendix E.8 satisfies*

$$\sup_{\|x\|_2 \le 1} |\langle \widehat{\theta} - \theta^\star, x\rangle| \le \min\left\{ 2R,\ \frac{7\rho_n\sqrt{s}}{\kappa} + \frac{4C_n}{\kappa n} \right\}, \qquad C_n := \|\zeta\|_1. \tag{9}$$

For an exact minimizer, the coefficient 7 can be replaced by 6.

The two nontrivial terms have a simple interpretation: because the regularization level ρ_n decreases at the usual $n^{-1/2}$ statistical rate, $7\rho_n\sqrt{s}/\kappa$ is the clean sparse-estimation error with $\sqrt{s}$ reflecting sparsity and $1/\kappa$ the conditioning of the retained design. Retained corruption contributes separately through C_n/n, while the cap $2R$ follows directly from the norm constraint. The complete proof, including the RSC argument and the treatment of large corruption, is given in Appendix E.5.

3.2 Coverage and the Estimator Interface

Write $\widehat{\mu}_{a,t-1}$ and $\widehat{w}_{a,i,t-1}$ for the reward and resource fits. Their clipped predictions are $\widehat{r}_{a,t} = \mathrm{clip}(\langle \widehat{\mu}_{a,t-1}, x_t \rangle, 0, 1)$ and $\widehat{b}_{a,i,t} = \mathrm{clip}(\langle \widehat{w}_{a,i,t-1}, x_t \rangle, 0, 1)$. More generally, $\beta^{r,\Gamma}_{a,t}$ and $\beta^{b,\Gamma}_{a,i,t}$ denote nonnegative absolute-error radii for the reward and resource-i predictions of action a at round t, under corruption candidate Γ. They are fixed from the observed history and x_t before the fresh action draw. A candidate Γ is a hypothesized upper bound on C_Z; candidates share point estimates and differ only in their radii.

The interface has two requirements. First, let $\delta_{\mathrm{cov}} \in (0,1)$ denote the failure level for the design/coverage event. On one joint event of failure at most $\delta_{\mathrm{est}} + \delta_{\mathrm{cov}}$, require simultaneously for every valid $\Gamma \geq C_Z$,

$$|\widehat{r}_{a,t} - r^0(a, x_t)| \leq \beta^{r,\Gamma}_{a,t}, \quad |\widehat{b}_{a,i,t} - b^0_i(a, x_t)| \leq \beta^{b,\Gamma}_{a,i,t} \quad (a \in \mathcal{A},\ i \in [m],\ t \in [T]). \tag{10}$$

For the Lasso implementation, Proposition 1 and $C_Z \leq \Gamma$ give the concrete radii in Appendix E.4; unit radii are used before n_0. Set predictions and radii to zero for `null`. The combined score radius $\beta^{Z,\Gamma}$ puts reward and resource uncertainty on the clean-score scale; since $\|\lambda\|_1 \leq 1$, the largest resource radius suffices. Clean scores lie in $[-Z, 1]$, whose length is $h := 1 + Z$:

$$\beta^{Z,\Gamma}_{a,t} := \beta^{r,\Gamma}_{a,t} + Z \max_i \beta^{b,\Gamma}_{a,i,t}, \quad h := 1 + Z. \tag{11}$$

Second, let $I_t \in \{0, 1\}$ indicate whether the pre-action safety test (7) passes; it remains one on an active round even if `null` is subsequently chosen. For the recommendation $\widehat{a}_t$ used in the optimistic comparison, require on the same event

$$2 \sum_{t=1}^{T} I_t \beta^{Z,\Gamma}_{\widehat{a}_t,t} \leq \mathcal{R}^0_{\mathrm{est}} + \psi_{\mathrm{est}}(\Gamma). \tag{12}$$

Here $\mathcal{R}^0_{\mathrm{est}}$ captures cumulative statistical uncertainty in the clean case and $\psi_{\mathrm{est}}(\Gamma)$ captures the additional width caused by retained corruption; both are analysis quantities rather than algorithm inputs. The optimistic comparison charges uncertainty only to the action recommended by the learner, so no separate cumulative-width control is needed along the population-LP policy. Shared-Grid requires the corresponding bound uniformly over valid candidates' recommendations on one shared history; see (24). Any estimator satisfying (10) and (12) can replace the concrete module.

Design conditions for the concrete rates. The modular allocation theorem does not require a particular design assumption; it only requires the confidence and cumulative-width interface above. To obtain the concrete Lasso rates, we use two sufficient design conditions. Restricted-curvature conditions are standard in sparse estimation (Negahban et al., 2012; Raskutti et al., 2010), while contextual-bandit analyses must additionally account for the action-specific covariates generated by adaptive sampling (Bastani & Bayati, 2020). The exact calibrations and their proofs are deferred to Appendix E.3.

Assumption 3 (Population design for the concrete estimator)**.** For a generic context $x \sim \mathcal{D}$, known constants $\kappa_x, L_x > 0$ satisfy

$$\mathbb{E}[xx^\top] \succeq \kappa_x I_d, \quad \mathbb{E} \exp\left(\frac{\langle u, x \rangle^2}{L_x^2 \|u\|_2^2} \right) \leq 2 \quad (u \neq 0). \tag{13}$$

Here κ_x offers the lower bound for population covariance curvature and L_x controls the directional sub-Gaussian scale.

On-policy learning (OP) retains every non-null observation, so its confidence depends on the action-specific designs generated by the algorithm itself.

Condition 4 (On-policy retained-design coverage)**.** For supplied design parameters $\kappa_{\mathrm{op}}, \upsilon_{\mathrm{op}} > 0$ and threshold n_0, the action-specific realized retained designs satisfy the calibrated RSC event defined in Appendix E.3, simultaneously after n_0 samples per action, with probability at least $1 - \delta_{\mathrm{cov}}$. Here κ_{op} is the RSC curvature and υ_{op} controls its ℓ_1 tolerance.

Unlike Assumption 3, this is a condition on the algorithm's actual adaptive, corrupted, and potentially stopped trajectory; population richness alone does not imply it. It is sufficient for the concrete

Lasso implementation to satisfy the estimator interface, rather than an additional requirement of the general allocation theorem. Shared-Grid OP additionally uses the separate recommendation-coverage condition in Section 5.

Rate constants. Let $\ell := \log\big(16eK(m+1)dT/\min\{\delta_{\text{est}}, \delta_{\text{cov}}\}\big)$ be the concentration logarithm, $\kappa_{\text{FE}} := \kappa_x/2$ the FE retained-design curvature, and let $H_T := 1 + \log(1+T)$ denote the harmonic factor. Let A_{FE} and A_{op} be the corresponding corruption-free $n^{-1/2}$ score-width coefficients; κ_{op} is from Condition 4. Their full parameter dependence is given in Appendix E.3; the A's are not universal constants.

3.3 Optimistic Primal–Dual Rule

At prices $\lambda_t \in \Lambda$, first form the plug-in estimate $\widehat{F}_{a,t}$ of the clean score $F^0(a, x_t; \lambda_t)$; candidate Γ then recommends the action with the largest estimate plus its score-confidence width:

$$\widehat{F}_{a,t}(\lambda_t) := \widehat{r}_{a,t} - Z\sum_{i=1}^{m} \lambda_{t,i}\widehat{b}_{a,i,t}, \tag{14}$$

$$\widehat{a}_t^{\Gamma} \in \arg\max_{a\in\mathcal{A}_0}\{\widehat{F}_{a,t}(\lambda_t) + \beta_{a,t}^{Z,\Gamma}\}, \tag{15}$$

with zero score and radius for `null` and the fixed order $\texttt{null}, 1, \ldots, K$ for ties. $\widehat{a}_t^{\Gamma}$ is the optimistic recommendation for candidate Γ, computed on rounds that pass the pre-action safety test (7). The actually played action a_t may differ because of forced exploration or the Shared-Grid master; if the safety test fails, the learner plays `null` from that round onward.

Embed λ_t in a probability vector $q_t \in \mathbb{R}_+^{m+1}$: $\lambda_t = q_{t,1:m}$ and $q_{t,m+1} = 1 - \|\lambda_t\|_1$, where the last coordinate is slack. For stepsize $\eta_q > 0$, set $\widetilde{g}_t = (b_t(a_t) - B^{\text{op}}/T, 0)$, i.e., observed excess consumption with a zero slack coordinate, and update

$$q_{t+1,j} \propto q_{t,j}e^{\eta_q \widetilde{g}_{t,j}}, \quad \lambda_{t+1} = q_{t+1,1:m}. \tag{16}$$

Normalize q_{t+1} to sum to one. The update increases the relative weight of resources with larger excess use. We denote its cumulative price-learning penalty by

$$\mathcal{D}_T := Z\left[\frac{\log(m+1)}{\eta_q} + \eta_q\left(1 + \frac{B_{\max}}{T}\right)^2 T\right]. \tag{17}$$

Choosing $\eta_q \asymp \frac{\sqrt{\log(m+1)/T}}{1+B_{\max}/T}$ gives $\mathcal{D}_T = O\left(Z\left(1+\frac{B_{\max}}{T}\right)\sqrt{T\log(m+1)}\right)$.

4 Robust Learning with Known Corruption

Given $\Gamma \geq C_Z$, ROPD uses the same point estimate with corruption-aware radii. Write the optimistic score as

$$U_{a,t}^{\Gamma} := \widehat{F}_{a,t}(\lambda_t) + \beta_{a,t}^{Z,\Gamma}, \tag{18}$$

with $U_{\texttt{null},t}^{\Gamma} = 0$. Here $U_{a,t}^{\Gamma}$ is a scalar optimistic score for action a, distinct from the cumulative observed-consumption vector U_t.

Forced mode (FE) retains only independent exploration samples, using ϵ_t for the probability of a uniform forced action at round t; on-policy mode (OP) retains every non-null sample and sets $\epsilon_t = 0$. Appendix C gives the complete procedure.

Guarantees. Set $\delta_{\text{tot}} := \delta_{\text{est}} + \delta_{\text{cov}}$, combining estimator and coverage failure probabilities.

Theorem 1 (Known corruption: regret and resource control)**.** *Suppose Assumptions 1 and 2 hold and the estimator/coverage pair satisfies the interface in (10)–(12) for a deterministic supplied bound Γ with $C_Z \leq \Gamma$ on every admissible realization. Algorithm 1 with $B^{\text{op}} = B$ satisfies the following expected-regret bound:*

$$\text{Reg}(T) \leq O\left(\mathcal{R}_{\text{est}}^0 + h\sum_{t=1}^{T}\epsilon_t + \mathcal{D}_T + \Gamma + \psi_{\text{est}}(\Gamma) + Z\right) + h\delta_{\text{tot}}T. \tag{19}$$

Moreover, on every realization, $\text{Viol}_{\text{obs}}(T) = 0$ *and* $\text{Viol}_{\text{cl}}(T) \leq C_b(T) \leq C_Z/Z \leq \Gamma/Z$.

Algorithm 1 ROPD with a Known Corruption Bound

Require: Valid bound Γ; mode FE or OP; exploration rates $\epsilon_t \in [0,1]$ (zero in OP); $0 < B^{\mathrm{op}} \le B$, Z, and η_q.
Estimation module: predictions and radii satisfying (10) and (12), with the module's required inputs.
Initialize $U_1 = 0$, $q_{1,j} = 1/(m+1)$, $\lambda_1 = q_{1,1:m}$, and all retained action-specific histories as empty.
for $t = 1, \dots, T$ **do**
  if $U_{t,i} > B_i^{\mathrm{op}} - 1$ for some i **then**
    Play `null` on rounds $t, \dots, T$ and terminate.
  end if
  Observe x_t; obtain predictions and radii from retained data and compute $U_{a,t}^{\Gamma}$ for all actions.
  Set $\widehat{a}_t \in \arg\max_{a \in \mathcal{A}_0} U_{a,t}^{\Gamma}$ with the fixed tie rule; initialize $a_t = \widehat{a}_t$.
  In mode FE, with probability ϵ_t, override $\widehat{a}_t$ by a uniform draw a_t from $\mathcal{A}$ and mark the round as forced.
  Play a_t and observe $r_t(a_t)$ and $b_t(a_t)$.
  In mode FE, retain the sample iff the round was forced; in mode OP, retain every non-null sample.
  Set $U_{t+1} = U_t + b_t(a_t)$ and update q_{t+1} and λ_{t+1} by (16).
end for

Proof overview. In (19), $\mathcal{R}_{\mathrm{est}}^0$ is clean estimation uncertainty, $h\sum_t \epsilon_t$ forced exploration, $\mathcal{D}_T$ price learning, Γ direct corruption/resource transfer, $\psi_{\mathrm{est}}(\Gamma)$ corruption-induced prediction width, and Z the stopping boundary. Confidence controls the clean-score gap, the price update controls observed resource imbalance, and $ZB_i \ge V^{\mathrm{UB}}$ offsets benchmark reward lost after stopping. Appendix F gives the full allocation argument.

Corollary 1 (Known corruption: concrete rates)**.** *Under Assumptions 1–2, run Algorithm 1 with $B^{\mathrm{op}} = B$, a deterministic valid bound $\Gamma \ge C_Z$, and the estimator and certified precision of Proposition 1. For a universal C, with the rate constants defined in Section 3.2:*

(i) ***Forced exploration.*** *Under the population-design Assumption 3 (calibration in Appendix E.3), use (43) and a constant forced-exploration probability $\epsilon_t \equiv \epsilon$ chosen by (44). Then*

$$\mathrm{Reg}(T) \le C\left[h^{1/3}A_{\mathrm{FE}}^{2/3}K^{1/3}T^{2/3} + h\sqrt{K(n_0+\ell)T} + \left(1 + \frac{KH_T}{\kappa_{\mathrm{FE}}\epsilon}\right)\Gamma + \mathcal{D}_T + Z\right] + h\delta_{\mathrm{tot}}T. \tag{20}$$

(ii) ***On-policy learning.*** *Under the retained-design Condition 4, with $\epsilon_t = 0$,*

$$\mathrm{Reg}(T) \le C\left[A_{\mathrm{op}}\sqrt{KT} + hKn_0 + \left(1 + \frac{KH_T}{\kappa_{\mathrm{op}}}\right)\Gamma + \mathcal{D}_T + Z\right] + h\delta_{\mathrm{tot}}T. \tag{21}$$

The $A_{\mathrm{FE}}/A_{\mathrm{op}}$ and n_0 terms capture clean learning and burn-in, while the terms proportional to Γ quantify the additional effect of corruption under the corresponding design conditioning. The terms $\mathcal{D}_T + Z$ arise from dual-price learning and stopping, and $h\delta_{\mathrm{tot}}T$ is the failure-event cost. With proportional budgets, fixed remaining parameters, dominated burn-in, the stepsize following (17), and $\delta_{\mathrm{tot}} \le T^{-2}$, the FE and OP bounds simplify to $\widetilde{O}(T^{2/3} + \Gamma T^{1/3})$ and $\widetilde{O}(\sqrt{T} + \Gamma)$, respectively. Here $\widetilde{O}$ suppresses polylogarithmic factors in d, K, m, T and inverse failure levels. Appendix C gives the full FE analysis.

Comparison of rates. Reward-corrupted linear bandits without resource constraints achieve near-optimal or minimax corruption dependence (He et al., 2022; Liu et al., 2024), while clean sparse CBwK exhibits $T^{2/3}$-type learning costs and, under stronger coverage, $\sqrt{T}$ behavior (Ma et al., 2024). Our results extend these CBwK regimes to joint reward and consumption corruption, giving $\widetilde{O}(T^{2/3} + \Gamma T^{1/3})$ under FE and $\widetilde{O}(\sqrt{T} + \Gamma)$ under OP, together with observed-budget feasibility and control of clean resource violation. The rates are not directly comparable across settings and no matching lower bound establishes their optimality here.

5 Adaptation to Unknown Corruption

Grid and algorithm. Because each post-clipping reward deviation is at most 1 and each consumption deviation has ℓ_∞ norm at most 1, $C_Z \le C_{\max,Z} := T(1+Z)$. Use the dyadic grid

$$\mathcal{G} := \{0\} \cup \{2^j : j = 0, \dots, \lceil \log_2 C_{\max,Z} \rceil\}. \tag{22}$$

The analysis uses the smallest valid point

$$\Gamma^\star := \min\{\Gamma \in \mathcal{G} : \Gamma \geq C_Z\}. \tag{23}$$

It is zero when $C_Z = 0$; otherwise $C_Z \leq \Gamma^\star \leq \max\{1, 2C_Z\}$. Here $\Gamma^\star$ is analysis-only: the learner neither observes C_Z nor selects this point. Every candidate recommends by (15) using the same fitted coefficients. A *master* combines these recommendations. We use EXP4.P (Beygelzimer et al., 2011) with $M := |\mathcal{G}| + 1$ experts: the grid candidates and one uniform expert on $\mathcal{A}_0$. Its comparison failure level is δ_{master}. Before master-controlled play, use a prescribed safe cyclic initialization of at most $t_{\text{init}} \leq T$ non-null rounds, where t_{init} is a nonnegative multiple of K. Retain these observations and update prices and accounting, but keep master weights fixed. Let p_t^{M} denote the master's smoothed played-action distribution on $\mathcal{A}_0$; Appendix D gives its update.

Algorithm 2 Shared-Grid with Unknown Corruption

Require: Grid $\mathcal{G}$; mode FE or OP; rates $\epsilon_t \in [0, 1/2]$ (zero in OP); $0 < B^{\text{op}} \leq B$, Z, η_q, t_{init}, and EXP4.P.
Estimation module: shared predictions and radii satisfying (10) and (24), with the module's required inputs.
Initialize $U_1 = 0$, $q_{1,j} = 1/(m+1)$, $\lambda_1 = q_{1,1:m}$, one shared collection of action-specific histories, and the EXP4.P master.
Run prescribed safe initialization for t_{init} rounds and update the shared histories, budget account, and prices.
for $t = t_{\text{init}} + 1, \ldots, T$ **do**
  if $U_{t,i} > B_i^{\text{op}} - 1$ for some i **then**
    Play `null` on rounds $t, \ldots, T$ and terminate.
  end if
  Observe x_t and obtain shared predictions; compute $\widehat{F}_{a,t}(\lambda_t)$ for all actions.
  For every $\Gamma \in \mathcal{G}$, compute its radii and recommendation $\widehat{a}_t^\Gamma$ by (15).
  Let the master combine the candidates and the uniform-advice expert into the smoothed distribution p_t^{M}.
  In mode FE, draw a forced-exploration coin with probability ϵ_t. If it succeeds, sample uniformly from $\mathcal{A}$; otherwise, draw $a_t \sim p_t^{\text{M}}$. In mode OP, always draw $a_t \sim p_t^{\text{M}}$.
  Play a_t and observe $r_t(a_t)$ and $b_t(a_t)$.
  On a non-forced round, normalize the observed score as $[r_t(a_t) - Z\lambda_t^\top b_t(a_t) + Z]/(1+Z)$ and update the master by importance weighting with $p_t^{\text{M}}(a_t)$; otherwise keep its weights fixed.
  In mode FE, retain the sample iff the round was forced; in mode OP, retain every non-null sample.
  Set $U_{t+1} = U_t + b_t(a_t)$ and update q_{t+1} and λ_{t+1} by (16).
end for

Uniform validity. The bound $\mathcal{R}_{\text{sh}}^0$ is the shared-history analogue of $\mathcal{R}_{\text{est}}^0$: it controls corruption-free cumulative recommendation widths, while $\psi_{\text{est}}(\Gamma)$ is the retained-corruption cost. Appendix G gives its form. One event must cover every valid grid point:

$$2 \sum_{t=t_{\text{init}}+1}^{T} I_t \beta_{\widehat{a}_t^\Gamma, t}^{Z,\Gamma} \leq \mathcal{R}_{\text{sh}}^0 + \psi_{\text{est}}(\Gamma). \tag{24}$$

Unlike known-corruption OP, a candidate can recommend an action that the master does not play. Bounding these recommendation widths therefore requires control of recommendation frequency relative to realized plays, rather than an additional property of Lasso or other estimation methods.

Let $N_a(t) := \sum_{\tau<t} \mathbf{1}\{a_\tau = a\}$ be the number of times non-null action a has actually been played before round t. Let δ_{cand} be the failure probability for covered recommendations, whose candidate-specific counts are $N_a^{\text{rec},\Gamma}(t) := \sum_{\tau=t_{\text{init}}+1}^{t-1} I_\tau \mathbf{1}\{\widehat{a}_\tau^\Gamma = a\}$.

Condition 5 (Covered recommendations)**.** For a recommendation-to-play ratio $\chi \geq 1$, additive count slack $n_{\text{cand}} \geq 0$, and failure level $\delta_{\text{cand}} \in (0, 1)$,

$$\mathbb{P}\left(N_a^{\text{rec},\Gamma}(t) \leq n_{\text{cand}} + \chi N_a(t) \text{ for all } \Gamma \in \mathcal{G},\, a \in \mathcal{A}, \text{ and } t \leq T+1\right) \geq 1 - \delta_{\text{cand}}. \tag{25}$$

Thus χ is the recommendation-to-play ratio and n_{cand} is an additive count slack; together they ensure that actions frequently recommended by a candidate are also played often enough to support their confidence widths. Lemma 1 gives sufficient lower bounds on the master's sampling probabilities for Condition 5. We introduce Condition 5 for OP mode because the default EXP4.P does not generally ensure a horizon-independent coverage ratio χ as its smoothing probability decreases with T. In FE mode, uniform exploration avoids this issue by supplying action samples independently of candidate recommendations.

Guarantees. Let $K_0 := K+1$ denote the number of actions including `null`. The master-comparison cost, including override concentration and the transfer from observed to clean scores, is

$$\mathrm{Reg}_{\mathrm{mst}}(T) := 20h\sqrt{TK_0\log(3M/\delta_{\mathrm{master}})}. \tag{26}$$

The total failure probabilities are

$$\delta_{\mathrm{tot}}^{\mathrm{U,FE}} := \delta_{\mathrm{est}} + \delta_{\mathrm{cov}} + \delta_{\mathrm{master}}, \quad \delta_{\mathrm{tot}}^{\mathrm{U,OP}} := \delta_{\mathrm{tot}}^{\mathrm{U,FE}} + \delta_{\mathrm{cand}}. \tag{27}$$

Below, δ_{tot} denotes $\delta_{\mathrm{tot}}^{\mathrm{U,FE}}$ in FE mode and $\delta_{\mathrm{tot}}^{\mathrm{U,OP}}$ in OP mode.

Theorem 2 (Unknown corruption: regret and resource control)**.** *Run Algorithm 2 with $B^{\mathrm{op}} = B$ and the master specification of Appendix D. Suppose Assumptions 1 and 2 hold, and the pointwise and grid-uniform width interfaces (10) and (24) hold on the corresponding estimator-and-coverage event. Then*

$$\begin{aligned}\mathrm{Reg}(T) \le O\Bigg(&\mathcal{R}_{\mathrm{sh}}^0 + ht_{\mathrm{init}} + h\sum_{t=1}^{T}\epsilon_t + \mathrm{Reg}_{\mathrm{mst}}(T) + \mathcal{D}_T + Z \\ &+ \mathbb{E}[3\Gamma^\star + \psi_{\mathrm{est}}(\Gamma^\star)]\Bigg) + h\delta_{\mathrm{tot}}T.\end{aligned} \tag{28}$$

On every realization, $\mathrm{Viol}_{\mathrm{obs}}(T) = 0$ *and* $\mathrm{Viol}_{\mathrm{cl}}(T) \le C_b(T) \le C_Z/Z \le \Gamma^\star/Z$. *Consequently,* $\mathbb{E}[\mathrm{Viol}_{\mathrm{cl}}(T)] \le \mathbb{E}[\Gamma^\star]/Z$.

Proof overview. In (28), $\mathcal{R}_{\mathrm{sh}}^0 + \psi_{\mathrm{est}}(\Gamma^\star)$ is shared estimation, ht_{init} initialization, $h\sum_t \epsilon_t$ forced exploration, $\mathrm{Reg}_{\mathrm{mst}}(T)$ master adaptation, and $\mathcal{D}_T$ price learning. A common event validates all valid grid radii before the path-dependent $\Gamma^\star$ is selected; two clean-score transfers plus the resource transfer give $3\Gamma^\star$. The remaining price/stopping argument is unchanged; Appendix G gives details.

Corollary 2 (Unknown corruption: concrete shared-data rates)**.** *Under Assumptions 1–2, run Algorithm 2 with $B^{\mathrm{op}} = B$, the shared estimator and precision of Proposition 1, and the EXP4.P specification of Appendix D. With a universal C and the corresponding failure total in (27):*

(i) ***Forced exploration.*** *Under the population-design Assumption 3 (calibration in Appendix E.3), use (43) and a constant forced-exploration probability $\epsilon_t \equiv \epsilon$ chosen by (44). Then*

$$\begin{aligned}\mathrm{Reg}(T) \le C\Bigg[&h^{1/3}A_{\mathrm{FE}}^{2/3}K^{1/3}T^{2/3} + h\sqrt{K(n_0+\ell)T} + ht_{\mathrm{init}} \\ &+ \left(3 + \frac{KH_T}{\kappa_{\mathrm{FE}}\epsilon}\right)\mathbb{E}[\Gamma^\star] + \mathrm{Reg}_{\mathrm{mst}}(T) + \mathcal{D}_T + Z\Bigg] + h\delta_{\mathrm{tot}}T.\end{aligned} \tag{29}$$

(ii) ***On-policy learning.*** *Under Conditions 4 and 5, with $\epsilon_t = 0$,*

$$\begin{aligned}\mathrm{Reg}(T) \le C\Bigg[&\chi A_{\mathrm{op}}\sqrt{KT} + \mathrm{Reg}_{\mathrm{mst}}(T) + h\{t_{\mathrm{init}} + K(n_{\mathrm{cand}} + \chi n_0)\} \\ &+ \left(3 + \frac{K\{\chi H_T + n_{\mathrm{cand}}\}}{\kappa_{\mathrm{op}}}\right)\mathbb{E}[\Gamma^\star] + \mathcal{D}_T + Z\Bigg] + h\delta_{\mathrm{tot}}T.\end{aligned} \tag{30}$$

Relative to Corollary 1, the unknown-corruption bounds add the initialization cost ht_{init} and the master-comparison term $\mathrm{Reg}_{\mathrm{mst}}(T)$; OP further incurs recommendation-coverage costs through χ and n_{cand}. Under the same proportional-budget and stepsize regime, with fixed remaining parameters, dominated initialization and burn-in, and $\delta_{\mathrm{tot}} \le T^{-2}$, the FE and OP bounds simplify to $\widetilde{O}(T^{2/3} + \mathbb{E}[\Gamma^\star]T^{1/3})$ and $\widetilde{O}(\sqrt{T} + \mathbb{E}[\Gamma^\star])$, respectively.

6 Conclusion and Limitations

We develop corruption-robust contextual allocation methods that account for both contaminated predictions and errors in resource accounting, pricing, and stopping. ROPD uses a supplied corruption bound, while Shared-Grid adapts confidence radii on one realized history. Both preserve observed budgets pathwise and control clean resource violation by consumption corruption. The estimator interface separates these allocation guarantees from the choice of sparse fitting method. Forced exploration provides one coverage route; the sharper on-policy rates require informative realized designs, with the additional recommendation-coverage condition for Shared-Grid. Obtaining less conservative calibration and weaker trajectory conditions remains a direction for further work.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, LaTeX restructuring, and code-level consistency checks. All AI-assisted material, including mathematical notation, citations, and experimental descriptions, was reviewed by the authors. The authors take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The appendices provide the materials needed to verify the theoretical and empirical results. They specify the full interaction protocol, corruption model, and complete ROPD and Shared-Grid algorithms, including the common-history construction for unknown corruption. The assumptions and trajectory conditions underlying the stated rates are made explicit, and complete proofs and proof dependencies are provided for the regret, observed-budget feasibility, and clean resource-control guarantees. For sparse estimation, the appendix gives the concrete Lasso specification, its calibration, the deterministic optimization procedure, and the numerical-error certificate used to connect approximate computation to the theoretical confidence bounds. Appendix H describes the experimental environments, corruption mechanisms, baselines, evaluation metrics, implementation choices, and finite-sample comparisons used in Figures 1–4, covering robustness with a supplied corruption bound, adaptation to unknown corruption, reward-versus-consumption corruption effects, and pathwise resource accounting. The corresponding corollaries identify the information, design, and realized-trajectory conditions under which the simplified rates apply.

REFERENCES

Alekh Agarwal, Haipeng Luo, Behnam Neyshabur, and Robert E Schapire. Corralling a band of bandit algorithms. In *Conference on Learning Theory*, pp. 12–38. PMLR, 2017.

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in neural information processing systems*, 29, 2016.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pp. 989–1006, 2014.

Shipra Agrawal, Zizhuo Wang, and Yinyu Ye. A dynamic near-optimal algorithm for online linear programming. *Operations Research*, 62(4):876–890, 2014.

Shipra Agrawal, Nikhil R Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pp. 4–18. PMLR, 2016.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

Sivaraman Balakrishnan, Simon S Du, Jerry Li, and Aarti Singh. Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pp. 169–212. PMLR, 2017.

Hamsa Bastani and Mohsen Bayati. Online decision making with high-dimensional covariates. *Operations Research*, 68(1):276–294, 2020.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 19–26. PMLR, 2011.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. *Advances in neural information processing systems*, 28, 2015.

Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009. doi: 10.1214/08-AOS620.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Sunrit Chakraborty, Saptarshi Roy, and Ambuj Tewari. Thompson sampling for high-dimensional sparse linear contextual bandits. In *International Conference on Machine Learning*, pp. 3979–4008. PMLR, 2023.

Yudong Chen, Constantine Caramanis, and Shie Mannor. Robust sparse regression under adversarial corruption. In *International conference on machine learning*, pp. 774–782. PMLR, 2013.

Ashok Cutkosky, Christoph Dann, Abhimanyu Das, Claudio Gentile, Aldo Pacchiano, and Manish Purohit. Dynamic balancing for model selection in bandits and rl. In *International Conference on Machine Learning*, pp. 2276–2285. PMLR, 2021.

Nikhil R Devanur and Thomas P Hayes. The adwords problem: online keyword matching with budgeted bidders under random permutations. In *Proceedings of the 10th ACM conference on Electronic commerce*, pp. 71–78, 2009.

Ilias Diakonikolas, Daniel Kane, Sushrut Karmalkar, Eric Price, and Alistair Stewart. Outlier-robust high-dimensional sparse estimation via iterative filtering. *Advances in Neural Information Processing Systems*, 32, 2019.

Qin Ding, Cho-Jui Hsieh, and James Sharpnack. Robust stochastic linear contextual bandits under adversarial attacks. In *International Conference on Artificial Intelligence and Statistics*, pp. 7111–7123. PMLR, 2022.

Dylan J Foster, Akshay Krishnamurthy, and Haipeng Luo. Model selection for contextual bandits. *Advances in Neural Information Processing Systems*, 32, 2019.

Avishek Ghosh, Abishek Sankararaman, and Kannan Ramchandran. Model selection for generic contextual bandits. *arXiv preprint arXiv:2107.03455*, 2021.

Anupam Gupta, Tomer Koren, and Kunal Talwar. Better algorithms for stochastic bandits with adversarial corruptions. In *Conference on Learning Theory*, pp. 1562–1578. PMLR, 2019.

Botao Hao, Tor Lattimore, and Mengdi Wang. High-dimensional sparse linear bandits. *Advances in Neural Information Processing Systems*, 33:10753–10763, 2020.

Jiafan He, Dongruo Zhou, Tong Zhang, and Quanquan Gu. Nearly optimal algorithms for linear contextual bandits with adversarial corruptions. *Advances in neural information processing systems*, 35:34614–34625, 2022.

Nicole Immorlica, Karthik Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. *Journal of the ACM*, 69(6):1–47, 2022.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Robust lipschitz bandits to adversarial corruptions. *Advances in Neural Information Processing Systems*, 36:10897–10908, 2023.

Sanath Kumar Krishnamurthy and Susan Athey. Optimal model selection in contextual bandits with many classes via offline oracles. Working paper, Stanford Graduate School of Business, 2021.

Yingkai Li, Edmund Y Lou, and Liren Shan. Stochastic linear optimization with adversarial corruption. *arXiv preprint arXiv:1909.02109*, 2019.

Fang Liu and Ness Shroff. Data poisoning attacks on stochastic bandits. In *International Conference on Machine Learning*, pp. 4042–4050. PMLR, 2019.

Haolin Liu, Artin Tajdini, Andrew Wagenmaker, and Chen-Yu Wei. Corruption-robust linear bandits: Minimax optimality and gap-dependent misspecification. *Advances in Neural Information Processing Systems*, 37:24277–24325, 2024.

Liu Liu, Yanyao Shen, Tianyang Li, and Constantine Caramanis. High dimensional robust sparse regression. In *International Conference on Artificial Intelligence and Statistics*, pp. 411–421. PMLR, 2020.

Shang Liu, Jiashuo Jiang, and Xiaocheng Li. Non-stationary bandits with knapsacks. *Advances in Neural Information Processing Systems*, 35:16522–16532, 2022.

Thodoris Lykouris, Vahab Mirrokni, and Renato Paes Leme. Stochastic bandits robust to adversarial corruptions. In *Proceedings of the 50th annual ACM SIGACT symposium on theory of computing*, pp. 114–122, 2018.

Wanteng Ma, Dong Xia, and Jiashuo Jiang. High-dimensional linear bandits with knapsacks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34008–34037. PMLR, 2024.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012. doi: 10.1214/12-STS400.

Aldo Pacchiano, Christoph Dann, Claudio Gentile, and Peter Bartlett. Regret bound balancing and elimination for model selection in bandits and rl. *arXiv preprint arXiv:2012.13045*, 2020.

Aldo Pacchiano, Christoph Dann, and Claudio Gentile. Best of both worlds model selection. *Advances in Neural Information Processing Systems*, 35:1883–1895, 2022.

Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):601–627, 2020.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *The Journal of Machine Learning Research*, 11:2241–2259, 2010.

Zhimei Ren and Zhengyuan Zhou. Dynamic batch learning in high-dimensional sparse linear contextual bandits. *Management Science*, 70(2):1315–1342, 2024.

Karthik Abinav Sankararaman and Aleksandrs Slivkins. Bandits with knapsacks beyond the worst case. *Advances in Neural Information Processing Systems*, 34:23191–23204, 2021.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Chen-Yu Wei, Christoph Dann, and Julian Zimmert. A model selection approach for corruption robust reinforcement learning. In *International Conference on Algorithmic Learning Theory*, pp. 1043–1096. PMLR, 2022.

Qingyuan Yu, Euijin Baek, Xiang Li, and Qiang Sun. Corruption-robust variance-aware algorithms for generalized linear bandits under heavy-tailed rewards. In *The 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

A Related Work

Contextual bandits with knapsacks and high-dimensional allocation. The CBwK framework joins bandit learning with finite resource constraints (Badanidiyuru et al., 2018). Agrawal & Devanur (2016) study the closest clean linear contextual model: expected rewards and resource consumptions are linear in observed features, and confidence sets are combined with online resource prices to control the knapsacks. Agrawal et al. (2016) develop a more general oracle-based contextual framework, and primal–dual ideas also connect this literature to online packing and dynamic resource allocation (Agrawal & Devanur, 2014; Devanur & Hayes, 2009; Agrawal et al., 2014). Our known-corruption method keeps the basic confidence-plus-price structure, but the feedback used by the estimator, the dual update, and the stopping rule can now differ from the clean process used by the benchmark. Our analysis accounts for the difference between clean and observed consumption in the price updates and the reward lost after stopping. Adversarial and non-stationary BwK study changing reward and consumption environments (Immorlica et al., 2022; Liu et al., 2022), while instance-dependent BwK analyses go beyond worst-case guarantees under clean feedback (Sankararaman & Slivkins, 2021). Our model instead keeps a latent uncorrupted linear contextual model with i.i.d. contexts and studies corrupted observations of that process, so the goal is to recover performance against a clean population-LP benchmark.

Sparse contextual bandits and clean high-dimensional CBwK. Sparse contextual bandit methods use regularization, forced sampling, Thompson sampling, or hard thresholding so regret depends on a small relevant support (Bastani & Bayati, 2020; Hao et al., 2020; Ren & Zhou, 2024; Chakraborty et al., 2023). The closest paper to our sparse allocation setting is Ma et al. (2024), which combines an online hard-thresholding estimator with a primal–dual CBwK method and obtains logarithmic dependence on the ambient feature dimension under clean feedback. Their sharper clean regime and our zero-corruption on-policy result both have $\widetilde{O}(\sqrt{T})$ horizon dependence, with different action, sparsity, curvature, norm, and resource-price factors. This comparison concerns the horizon exponent, not identical dimension dependence. Our setting adds corrupted reward and consumption observations, a direct resource-accounting term, a sequential charge for the later effect of corrupted action-specific samples, and adaptation when the corruption level is unknown. The retained-design RSC condition in our sharper result plays the familiar statistical role of ensuring that action-specific data identify sparse parameters. The additional Shared-Grid recommendation-coverage condition is different: it controls how often candidate recommendations are supported by realized plays.

Corruption-robust bandits. Corruption-robust bandit work first showed how regret can degrade with an adversarial corruption budget in stochastic multi-armed bandits (Lykouris et al., 2018; Gupta et al., 2019). Later work studies linear and contextual models. Ding et al. (2022) allow attacked rewards and contexts without requiring the attack budget. For a known played-action reward-corruption budget C, He et al. (2022) obtain $\widetilde{O}(d\sqrt{T} + dC)$ and also study adaptation without a supplied bound. For fixed-parameter linear bandits, Liu et al. (2024) distinguish C from the worst-action budget C_∞ and establish $\widetilde{O}(d\sqrt{T} + \min\{dC, \sqrt{d}\, C_\infty\})$ with matching lower bounds under the corresponding corruption information. These results concern reward corruption without knapsack constraints and do not by themselves establish optimal dependence on our joint reward–consumption budget. Other work studies linear optimization, Lipschitz bandits, generalized linear models, and explicit poisoning attacks (Li et al., 2019; Kang et al., 2023; Yu et al., 2025; Liu & Shroff, 2019). These papers develop robust confidence sets, weighting, elimination, and attack-adaptive exploration for bandit objectives centered on prediction and action selection. Our setting adds a resource signal that changes both statistical estimates and the state of the allocation system. The Z-scaled budget measures corruption in the action score, and the proof also controls the observed budget account, the dual-price path, and the stopping time.

Robust sparse estimation and sequential prediction. Robust sparse regression studies parameter recovery when some responses or samples are corrupted (Chen et al., 2013; Bhatia et al., 2015; Balakrishnan et al., 2017; Diakonikolas et al., 2019; Prasad et al., 2020; Liu et al., 2020). Restricted-eigenvalue and sparse covariance conditions explain when high-dimensional parameters can be identified (Bickel et al., 2009; Raskutti et al., 2010; Negahban et al., 2012; Wainwright, 2019; Bühlmann & Van De Geer, 2011). These tools provide the statistical base of our estimator module, while an online allocation proof additionally requires the sum of reward and consumption widths along

an adaptive action path. We prove this sequential bound by matching the estimator sample set to its coverage event and by charging each corrupted observation through the later inverse action counts in which it still appears. Our concrete analysis uses norm-constrained Lasso with candidate-independent regularization. An enlarged-cone argument and an RSC tolerance yield explicit corruption-aware confidence, with a norm bound handling large corruption. The resulting allocation theorems transfer to any sparse estimator that supplies the same predictable pointwise and cumulative-width guarantees.

Unknown corruption and bandit model selection. Corralling and model-selection methods combine candidates or policy classes under partial feedback (Agarwal et al., 2017; Foster et al., 2019; Pacchiano et al., 2020; 2022; Cutkosky et al., 2021; Krishnamurthy & Athey, 2021; Ghosh et al., 2021). Agarwal et al. (2017) show that a master must avoid starving a candidate of feedback. Foster et al. (2019) adapt to an unknown contextual model class, and Pacchiano et al. (2022) combine model selection with guarantees for fixed-mean and adversarial rewards. Model selection has also been used for corruption-robust reinforcement learning (Wei et al., 2022). Unknown corruption in sparse CBwK adds a data-design problem. Two independently updated corruption candidates would choose different actions and would need different action-specific context histories, so the confidence set of an inactive candidate can depend on data that the learner never collected. The shared-estimator grid avoids reliance on these unobserved histories: every candidate uses the same realized point estimate, and the master compares only their current recommendations. The remaining candidate-coverage condition in the exploration-free result states when these recommendations receive enough realized plays for their shared confidence widths to be summed.

Our paper combines primal–dual resource control, corruption-robust learning, and sparse action-specific estimation in a setting where resource-consumption feedback is itself corrupted. This requires a relation between clean and observed resource use, stopping control through resource prices, sequential tracking of corrupted retained samples, and shared-estimator adaptation on one realized design.

B Complete Model and Feasibility Details

B.1 Interaction Timing and Three Feedback Levels

A *potential outcome* is the reward or consumption that would be produced by an action if that action were selected. Potential outcomes let the analysis compare the chosen action with other actions, but the learner observes only the outcome of a_t. Let $\mathcal{F}_{t-1}$ be the full pre-round analysis history, including earlier contexts, actions, clean potential outcomes, corruption, and learner randomization. The learner's observed history is a subhistory: it does not observe latent clean outcomes or corruption. Assumption 1's context and clean-noise conditions hold given this full history. This distinction makes the clean-noise sums adapted martingales; centering only given the learner's observations would not suffice. All learner decisions use only its observed history, the current context, and fresh independent randomness. The safety check is evaluated from $\mathcal{F}_{t-1}$ before the following timing on an active round:

(i) the learner observes x_t;

(ii) conditional on $\mathcal{F}_{t-1}$ and x_t, the environment generates the clean potential outcomes and fixes the corruption for every action;

(iii) the learner draws a_t using fresh randomness that is conditionally independent of these potential outcomes and corruption, then observes feedback only for a_t;

(iv) the learner updates its estimates, resource prices, and observed-budget account.

The corruption is *non-anticipating*: it may depend on $\mathcal{F}_{t-1}$, the current context, and the generated clean potential outcomes, but not on the learner's fresh round-t random seed. The action and retention eligibility are determined before the current feedback is revealed. This timing preserves the martingale structure of retained clean-noise terms and the conditional unbiasedness of the unknown-corruption importance-weighted comparison.

We use three levels of reward and consumption. The clean conditional mean is the expected outcome given an action and context. The clean realization adds ordinary stochastic noise. The observed feedback applies corruption and clipping to the clean realization and is the only level available to the learner. Table 2 summarizes these quantities.

Table 2: Three levels of reward and resource-consumption feedback.

Quantity	Meaning	Observed by the learner?
$r^0(a,x), b^0(a,x)$	clean conditional means	No
$r_t^{\mathrm{cl}}(a), b_t^{\mathrm{cl}}(a)$	clean noisy realizations before corruption	No
$r_t(a_t), b_t(a_t)$	clipped feedback after corruption	Yes, for the chosen action

Table 3: Feedback, budget, and coverage notation.

Symbol	Meaning
r^0, b^0	Clean conditional reward and consumption means used by the benchmark.
$r^{\mathrm{cl}}, b^{\mathrm{cl}}$	Clean noisy realizations before corruption.
r, b	Clipped corrupted feedback observed for the played action.
c^r, c^b	Post-clipping clean-to-observed deviations.
C_Z	Global effective corruption budget with a worst-action maximum on every round.
$C_Z(t)$	Effective corruption accumulated along the realized action path through round t.
$C_b(t)$	Consumption corruption accumulated along the realized action path through round t.
B^{op}, U_t	Operating budget and observed resource use before round t.
R; s_0, s	Parameter-norm bound; true sparsity and supplied calibration upper bound.
$\kappa_{\mathrm{FE}}, \kappa_{\mathrm{op}}$	Curvature bounds for the concrete estimator's forced and realized designs (Appendix E.3).
$A_{\mathrm{FE}}, A_{\mathrm{op}}$; n_0	Clean-width coefficients; per-action retained-sample threshold.
χ, n_{cand}; t_{init}	Recommendation ratio and slack; safe cyclic initialization length.

The quantity $\|v\|_0$ counts the nonzero coordinates of a vector. Thus s_0 is the true sparsity level, while $s \geq 1$ is a supplied upper bound used for confidence and numerical-precision calibration, not a support constraint in the Lasso fit. Conditional mean zero means $\mathbb{E}[\eta_t^r(a) \mid \mathcal{F}_{t-1}, x_t] = 0$ and $\mathbb{E}[\eta_{t,i}^b(a) \mid \mathcal{F}_{t-1}, x_t] = 0$; the conditional sub-Gaussian assumption supplies the time-uniform concentration bounds used below.

The adversary changes clean realizations rather than clean conditional means. The raw perturbations $\xi_t^r(a)$ and $\xi_{t,i}^b(a)$ have no probability model; their total effect is controlled by the corruption budget. Since both clean and observed outcomes lie in $[0,1]$, $|c_t^r(a)| \leq 1$ and $\|c_t^b(a)\|_\infty \leq 1$. For the null action, all clean outcomes, observed outcomes, and corruption terms are zero. Clipping keeps feedback in the bounded range used by the analysis; it does not reveal which observations were corrupted.

B.2 Benchmark and Corruption Quantities

For each context x, the population-LP policy assigns probability $y(a \mid x)$ to non-null action a and the remaining probability to `null`. It is static across time but context-dependent. With clean conditional means, the benchmark is

$$
\begin{aligned}
V^{\mathrm{UB}} = \max_{y} \quad & T\mathbb{E}_x\left[\sum_{a\in\mathcal{A}} r^0(a,x)y(a \mid x)\right] \\
\text{s.t.} \quad & T\mathbb{E}_x\left[\sum_{a\in\mathcal{A}} b_i^0(a,x)y(a \mid x)\right] \leq B_i \quad (i \in [m]), \\
& y(a \mid x) \geq 0 \quad (a \in \mathcal{A}), \qquad \sum_{a\in\mathcal{A}} y(a \mid x) \leq 1 \quad (\text{for every } x).
\end{aligned}
$$

Under the i.i.d. model, this value upper-bounds the clean expected reward attainable under the population resource constraints (Agrawal & Devanur, 2016; Agrawal et al., 2016). We fix one optimizer $y^\star$, measurable in x, before the interaction.

The global effective corruption budget is $C_Z = \sum_{t=1}^T \max_{a\in\mathcal{A}}\{|c_t^r(a)| + Z\|c_t^b(a)\|_\infty\}$. For later analysis, define the realized-path quantities $C_Z(t) := \sum_{\tau=1}^t (|c_\tau^r(a_\tau)| + Z\|c_\tau^b(a_\tau)\|_\infty)$ and $C_b(t) := \sum_{\tau=1}^t \|c_\tau^b(a_\tau)\|_\infty$. The first measures realized reward and consumption corruption in Lagrangian

units, while the second keeps only realized consumption corruption. For the retained set $\mathcal{S}_a^{\mathrm{est}}(t)$, also define $C^r_{a,\mathrm{est}}(t) := \sum_{\tau\in\mathcal{S}_a^{\mathrm{est}}(t)} |c^r_\tau(a)|$ and $C^b_{a,i,\mathrm{est}}(t) := \sum_{\tau\in\mathcal{S}_a^{\mathrm{est}}(t)} |c^b_{\tau,i}(a)|$. The worst-action maximum in C_Z controls the corruption path before the fresh action draw, whereas $C_Z(t)$, $C_b(t)$, and the retained-channel quantities describe the realized path. They are analysis quantities that the learner cannot compute, and $C_Z(t) \le C_Z$.

B.3 Observed and Clean Feasibility

Proposition 2 (Three feasibility levels)**.** *Under (7), for every resource i:*

(i) $\sum_{t=1}^T b_{t,i}(a_t) \le B_i^{\mathrm{op}}$ *(exact observed feasibility);*
(ii) $\sum_{t=1}^T b^{\mathrm{cl}}_{t,i}(a_t) \le B_i^{\mathrm{op}} + C_b(T)$ *(clean sample-path control);*
(iii) if $C_b(T) \le \Gamma_b$, $B_i > \Gamma_b$ *for every i, and* $B^{\mathrm{op}} = B - \Gamma_b\mathbf{1}$, *then* $\sum_{t=1}^T b^{\mathrm{cl}}_{t,i}(a_t) \le B_i$ *(strict clean sample-path feasibility).*

Proof of Proposition 2. Before a non-null action, the rule in (7) and $b_{t,i}(a_t) \le 1$ imply that the updated observed use is at most B_i^{op}. Null actions use no resources, which proves (i). Since $b^{\mathrm{cl}}_{t,i}(a_t) = b_{t,i}(a_t) - c^b_{t,i}(a_t)$,

$$\sum_{t=1}^T b^{\mathrm{cl}}_{t,i}(a_t) \le \sum_{t=1}^T b_{t,i}(a_t) + \sum_{t=1}^T |c^b_{t,i}(a_t)| \le B_i^{\mathrm{op}} + C_b(T),$$

which proves (ii). Substituting $B_i^{\mathrm{op}} = B_i - \Gamma_b$ proves (iii). □

For the default choice $B^{\mathrm{op}} = B$, Proposition 2 gives

$$\mathrm{Viol}_{\mathrm{obs}}(T) = 0, \quad \mathrm{Viol}_{\mathrm{cl}}(T) \le C_b(T) \le C_Z/Z.$$

The last inequality follows because the per-round maximum defining C_Z dominates $Z\|c^b_t(a_t)\|_\infty$. Hence $C_Z \le \Gamma$ also certifies $\mathrm{Viol}_{\mathrm{cl}}(T) \le \Gamma/Z$.

The one-unit safety margin is already included in (7), so it is not subtracted again from B^{op}. Let $V^{\mathrm{UB}}(B')$ denote the value of the population LP with budget vector B'. If the buffered algorithm uses $B^{\mathrm{op}} = B - \Gamma_b \cdot \mathbf{1}$ but is evaluated against the original budget B, its regret also includes the reward lost by reducing the benchmark budget $V^{\mathrm{UB}}(B) - V^{\mathrm{UB}}(B - \Gamma_b \cdot \mathbf{1})$, where $\mathbf{1}$ is the all-ones vector.

Remark 1 (Conditional-mean feasibility). Proposition 2 controls clean realized consumptions $b^{\mathrm{cl}}_{t,i}(a_t)$. A bound on the sum of clean conditional means $\sum_t b_i^0(a_t, x_t)$ must also remove clean stochastic noise. Under Assumption 1, a martingale concentration argument gives an additional buffer of order $\sqrt{T\log(m/\delta)}$; adding this standard concentration buffer gives the corresponding conditional-mean feasibility statement.

C Full Known-Corruption Algorithm and Concrete Rates

Algorithm 3 expands the main-paper Algorithm 1 into its estimation, exploration, dual, and accounting updates, using the concrete estimator of Section 3.1 and Appendix E.2. Another module satisfying the interface can replace its estimation calls without changing the allocation, price, or safety updates. In forced-exploration mode, only samples generated by the independent forced draw enter the estimator; in on-policy-estimation mode, every realized non-null sample enters and $\epsilon_t = 0$. Conditional on the retained history, the supplied bound Γ changes the confidence radii but does not enter the point-estimation objective; it may still affect the realized design through action selection.

Algorithm 3 Robust Optimistic Primal–Dual Learning with Known Corruption (ROPD)

Require: Horizon T; actions $\mathcal{A} = [K]$ and $\mathcal{A}_0 = \mathcal{A} \cup \{\texttt{null}\}$; budgets $0 < B^{\mathrm{op}} \le B$; deterministic pathwise bound $\Gamma \ge C_Z$; scale Z; dual stepsize η_q.
Coverage inputs: mode $\mathsf{mode} \in \{\mathsf{forced}, \mathsf{on\text{-}policy}\}$ and exploration probabilities $\epsilon_t \in [0,1]$, with $\epsilon_t = 0$ in on-policy mode.
Estimator inputs: s, R, σ_r, σ_b, calibrated (κ, υ, n_0), failure levels $\delta_{\mathrm{est}}, \delta_{\mathrm{cov}}$, and precision $s\rho_n^2/(8\kappa)$.
Set $\mathcal{H}^{\mathrm{est}}_{a,0} = \varnothing$ for all $a \in \mathcal{A}$, $U_1 = 0$, $q_{1,j} = 1/(m+1)$ for $j \in [m+1]$, and $\lambda_1 = q_{1,1:m}$.
for $t = 1, \dots, T$ **do**
  if $U_{t,i} > B_i^{\mathrm{op}} - 1$ for some $i \in [m]$ **then**
    Play $a_t = \texttt{null}$; keep $(U_{t+1}, q_{t+1}, \lambda_{t+1}) = (U_t, q_t, \lambda_t)$ and $\mathcal{H}^{\mathrm{est}}_{a,t} = \mathcal{H}^{\mathrm{est}}_{a,t-1}$ for all a; **continue**.
  end if
  Observe x_t.
  for $a \in \mathcal{A}$ **do**
    Set $n = |\mathcal{H}^{\mathrm{est}}_{a,t-1}|$. If $n = 0$ or no justified calibration is supplied, use zero coefficients; otherwise obtain the channel fits with Appendix E.8's certified solver, reusing unchanged fits.
    Use the supplied calibration and certified Lasso precision. Use unit channel radii if $n < n_0$ or calibration is unavailable; otherwise use (45)–(46).
    Compute $\beta^{Z,\Gamma}_{a,t}$ and $U^{\Gamma}_{a,t}$ from (18).
  end for
  Set $U^{\Gamma}_{\texttt{null},t} = 0$ and choose $\widehat{a}_t \in \arg\max_{a \in \mathcal{A}_0} U^{\Gamma}_{a,t}$ using the fixed tie-breaking rule.
  Draw $e_t \sim \mathrm{Bernoulli}(\epsilon_t)$. If $e_t = 1$, draw a_t uniformly from $\mathcal{A}$; otherwise set $a_t = \widehat{a}_t$.
  if $a_t \neq \texttt{null}$ **then**
    Observe $r_t(a_t)$ and $b_t(a_t)$.
  else
    Set $r_t(\texttt{null}) = 0$ and $b_t(\texttt{null}) = 0$.
  end if
  Set $\mathcal{H}^{\mathrm{est}}_{a,t} = \mathcal{H}^{\mathrm{est}}_{a,t-1}$ for every $a \in \mathcal{A}$.
  if $a_t \neq \texttt{null}$ and [$\mathsf{mode} = \mathsf{on\text{-}policy}$ or $e_t = 1$] **then**
    Append $(x_t, r_t(a_t), b_t(a_t))$ to $\mathcal{H}^{\mathrm{est}}_{a_t,t}$.
  end if
  Set $g_t = b_t(a_t) - B^{\mathrm{op}}/T$ and $\widetilde{g}_t = (g_t, 0) \in \mathbb{R}^{m+1}$.
  Set $q_{t+1,j} = \dfrac{q_{t,j}\exp(\eta_q \widetilde{g}_{t,j})}{\sum_{\ell=1}^{m+1} q_{t,\ell}\exp(\eta_q \widetilde{g}_{t,\ell})}, \forall j \in [m+1]$; then set $\lambda_{t+1} = q_{t+1,1:m}$ and $U_{t+1} = U_t + b_t(a_t)$.
end for

Table 4: Inputs to the complete known-corruption method.

Input	Definition and role
B^{op}	Operating budget used by the observed-consumption safety account; the default is B.
Γ	Deterministic supplied bound satisfying $C_Z \le \Gamma$ on every admissible realization.
Z	Positive Lagrangian scale satisfying Assumption 2; $Z = \frac{T}{B_{\min}}$ is a computable sufficient choice.
η_q	Multiplicative-weights stepsize for the resource-price distribution.
ϵ_t	Probability of a uniform non-null forced-exploration override; it is zero in on-policy mode.
n_0	Action-specific retained-sample threshold before nontrivial sparse confidence is used.
(κ, υ)	Supplied RSC calibration justified by (43) or Condition 4.
s, R, σ_r, σ_b	Supplied sparsity, parameter-norm, and noise bounds for Lasso confidence.
$\delta_{\mathrm{est}}, \delta_{\mathrm{cov}}$	Failure levels used in ℓ and design calibration.

The pre-action check and $b_{t,i}(a_t) \le 1$ give exact observed feasibility, $\sum_{t=1}^{T} b_{t,i}(a_t) \le B_i^{\mathrm{op}}$. If a valid bound $C_b(T) \le \Gamma_b$ is known and $B_i > \Gamma_b$ for every resource, the choice $B^{\mathrm{op}} = B - \Gamma_b \mathbf{1}$ gives strict clean sample-path feasibility by Proposition 2; comparison with the original budget adds $V^{\mathrm{UB}}(B) - V^{\mathrm{UB}}(B - \Gamma_b \mathbf{1})$. The exact dual stepsize calculation is

$$\eta_q \asymp \frac{\sqrt{\log(m+1)}}{(1 + B_{\max}/T)\sqrt{T}} \quad \text{gives} \quad \mathcal{D}_T = O\left(Z(1 + B_{\max}/T)\sqrt{T\log(m+1)}\right). \tag{31}$$

C.1 Concrete Corollaries and Exploration Choice

Corollary 3 (Known FE with a calibrated estimator)**.** *Under Assumptions 1, 2, and 3, run Algorithm 3 with $B^{\mathrm{op}} = B$, a deterministic pathwise bound $C_Z \le \Gamma$, constant $0 < \epsilon \le 1/2$, and (43). Use the*

regularization, trivial early radii, and certified numerical precision of Appendix E.2. For a universal C and $\delta_{\mathrm{tot}} = \delta_{\mathrm{est}} + \delta_{\mathrm{cov}}$,

$$\mathrm{Reg}(T) \le C\left[\frac{hK(n_0+\ell)}{\epsilon} + A_{\mathrm{FE}}\sqrt{\frac{KT}{\epsilon}} + h\epsilon T + \mathcal{D}_T + \Gamma + \frac{K\Gamma H_T}{\kappa_{\mathrm{FE}}\epsilon} + Z\right] + h\delta_{\mathrm{tot}}T. \tag{32}$$

The bound holds for every $T \ge 1$, even if coverage is not reached. The choice (44) gives the main-text bound (20), by the calculation below.

Corollary 4 (Known OP with a calibrated estimator)**.** *Under Assumptions 1, 2 and Condition 4, run Algorithm 3 with $B^{\mathrm{op}} = B$, $\epsilon_t = 0$, and a deterministic pathwise bound $C_Z \le \Gamma$. Use the supplied OP calibration and the same certified Lasso rule on all non-null samples. Then (21) holds with $\delta_{\mathrm{tot}} = \delta_{\mathrm{est}} + \delta_{\mathrm{cov}}$.*

Corollary 5 (Clean known-corruption case)**.** *Under Corollary 4's conditions with $C_Z = \Gamma = 0$, the Γ-dependent contribution in (21) vanishes. The burn-in, numerical/statistical width, dual term, stopping boundary, and $h\delta_{\mathrm{tot}}T$ remain.*

These claims follow by substituting Lemma 7 or Proposition 4 into the modular known-corruption theorem; Appendix F proves the allocation step. In particular, no sparsity-constrained optimizer or empirical sparse-eigenvalue oracle is used.

For clarity about the FE choice, put $a = A_{\mathrm{FE}}\sqrt{KT}$ and $b = hK(n_0+\ell)$. The three clean terms are $b/\epsilon + a/\sqrt{\epsilon} + hT\epsilon$. If the cap in (44) is inactive, its two lower bounds imply

$$\frac{b}{\epsilon} \le \sqrt{bhT}, \qquad \frac{a}{\sqrt{\epsilon}} \le a^{2/3}(hT)^{1/3}, \qquad hT\epsilon \le \sqrt{bhT} + a^{2/3}(hT)^{1/3}.$$

This proves the $T^{2/3}$ term together with the square-root burn-in contribution in (20). For the capped case, write

$$r_1 = \left(\frac{A_{\mathrm{FE}}^2 K}{h^2 T}\right)^{1/3}, \qquad r_2 = \sqrt{\frac{K(n_0+\ell)}{T}}.$$

If $\max\{r_1, r_2\} \ge 1/2$, then the sum of the two clean terms in the main bound is $hT(r_1+r_2) \ge hT/2 \ge T/2$. The unconditional range bound $\mathrm{Reg}(T) \le T$ therefore proves the same main-text inequality after choosing $C \ge 2$. Thus it is valid for all horizons, not only the uncapped regime; no informative rate is asserted when these terms are of order T. The argument also proves (29), since its other terms are nonnegative. The rule balances clean costs only and is implementable with either known or unknown corruption; the corruption term is evaluated at the chosen ϵ.

D Full Shared-Grid Method and Concrete Rates

D.1 Grid, Valid Point, and Shared Notation

Use the main-paper grid $\mathcal{G}$ and its size $|\mathcal{G}|$ from (22). After the corruption path is realized, the analysis selects $\Gamma^\star = \min\{\Gamma \in \mathcal{G} : \Gamma \ge C_Z\}$, so

$$\Gamma^\star = 0 \text{ if } C_Z = 0, \qquad C_Z \le \Gamma^\star \le \max\{1, 2C_Z\} \text{ otherwise.} \tag{33}$$

The algorithm runs every radius in $\mathcal{G}$, while $\Gamma^\star$ serves as the valid comparator in the analysis.

Table 5: Shared-Grid notation defined before the complete procedure.

Symbol	Definition and role
t_{init}	Length of the fixed cyclic non-null initialization; it is divisible by K.
$\lvert\mathcal{G}\rvert$	Number of corruption candidates in $\mathcal{G}$.
K_0	Number of actions including the null action, $K_0 = \lvert\mathcal{A}_0\rvert = K+1$.
$p_{\min}$	Master smoothing probability that gives every action positive sampling probability.
α_m	EXP4.P confidence-bonus coefficient.
$N_a^{\mathrm{rec},\Gamma}(t)$	Post-initialization active-round count of candidate Γ recommending action a before t.
n_{cand}	Additive prefix slack in coverage of candidate recommendations.
χ	Multiplicative ratio between candidate recommendations and realized action counts.

For later reference, $N_a^{\mathrm{rec},\Gamma}(t) := \sum_{\tau=t_{\mathrm{init}}+1}^{t-1} I_\tau \mathbf{1}\{\widehat{a}_\tau^\Gamma = a\}$, and Condition 5 states $N_a^{\mathrm{rec},\Gamma}(t) \le n_{\mathrm{cand}} + \chi N_a(t)$ uniformly for $t \le T+1$ with probability at least $1-\delta_{\mathrm{cand}}$.

D.2 Shared Estimator Sample Sets

Let $e_t = 1$ mark a forced-exploration override, and let the first t_{init} rounds follow a cyclic non-null schedule fixed before the current context and feedback. Before round t, the shared estimator for action a uses

$$\mathcal{I}_{a,t}^{\mathrm{est}} = \begin{cases} \{\tau < t : I_\tau = 1,\ a_\tau = a,\ [\tau \le t_{\mathrm{init}} \text{ or } e_\tau = 1]\}, & \mathsf{mode} = \mathsf{forced}, \\ \{\tau < t : I_\tau = 1,\ a_\tau = a\}, & \mathsf{mode} = \mathsf{on\text{-}policy}. \end{cases} \tag{34}$$

Thus forced mode retains the fixed initialization and later forced samples, while on-policy mode retains every realized non-null sample. The fixed eligibility rule is known before the current clean noise is observed.

D.3 Master Quantities and Payoff Normalization

We use the high-probability EXP4.P master (Beygelzimer et al., 2011). Let $\mathcal{E}_{\mathrm{M}} := \mathcal{G} \cup \{\mathrm{unif}\}$ and $M = |\mathcal{G}| + 1$. Each grid point has one-hot advice $\xi_{t,\Gamma}(a) = \mathbf{1}\{\widehat{a}_t^\Gamma = a\}$; the extra expert has $\xi_{t,\mathrm{unif}}(a) = 1/K_0$. Its inclusion matches the uniform-expert hypothesis in Beygelzimer et al. (2011, Theorem 2). This expert is distinct from the action smoothing below and requires no additional fit. With $K_0 = K+1$, define

$$p_{\min} := \min\left\{\frac{1}{2K_0}, \sqrt{\frac{\log M}{K_0 T}}\right\}, \qquad \alpha_m := \sqrt{\frac{\log(3M/\delta_{\mathrm{master}})}{K_0 T}}. \tag{35}$$

The master uses the potential observed score and normalized payoff

$$F_t^{\mathrm{obs}}(a;\lambda_t) := r_t(a) - Z\lambda_t^\top b_t(a), \qquad Y_t(a) := \frac{F_t^{\mathrm{obs}}(a;\lambda_t) + Z}{1+Z}, \qquad a \in \mathcal{A}_0. \tag{36}$$

Because $F_t^{\mathrm{obs}}(a;\lambda_t) \in [-Z, 1]$, $Y_t(a) \in [0,1]$. The high-probability EXP4.P comparison gives $\mathrm{Reg}_{\mathrm{mst}}(T) = 20h\sqrt{TK_0 \log(3M/\delta_{\mathrm{master}})}$; when the standard nontrivial-horizon condition fails, the score-range bound is already no larger than this order.

D.4 Complete Shared-Grid Procedure

Algorithm 4 Complete Shared-Grid Procedure for Unknown Corruption

Require: Horizon T; initialization t_{init}; actions $\mathcal{A}_0$; B^{op}; scale $Z = T/B_{\min}$; coverage mode $\in \{\text{forced}, \text{on-policy}\}$; rates $0 \le \epsilon_t \le 1/2$ with $\epsilon_t = 0$ on-policy; stepsize η_q; estimator inputs $(s, R, \sigma_r, \sigma_b, \kappa, \upsilon, n_0)$; failure levels $\delta_{\text{est}}, \delta_{\text{cov}}, \delta_{\text{master}}$ and the certified Lasso precision.

Form $\mathcal{G}$ and $\mathcal{E}_{\text{M}}$ as above; set $w_{1,j} = 1$ for $j \in \mathcal{E}_{\text{M}}$, $q_{1,j} = 1/(m+1)$, $\lambda_1 = q_{1,1:m}$, $U_1 = 0$, and initialize one shared estimator history per action.
for $t = 1, \dots, T$ **do**
  if $U_{t,i} > B_i^{\text{op}} - 1$ for some $i \in [m]$ **then**
    Play $a_t = \texttt{null}$; set $I_t = e_t = 0$; freeze all states; **continue**.
  end if
  Set $I_t = 1$ and observe x_t.
  if $t \le t_{\text{init}}$ **then**
    Play the next non-null action in the fixed cyclic schedule; set $e_t = 0$ and keep the master weights fixed.
  else
    For each action, set $n = |\mathcal{I}_{a,t}^{\text{est}}|$. Use zero coefficients if $n = 0$ or no justified calibration is supplied; otherwise fit or reuse the certified channel estimates. All candidates share these fits.
    for $\Gamma \in \mathcal{G}$ **do**
      Compute $\beta_{a,t}^{Z,\Gamma}$ and $\widehat{a}_t^{\Gamma} \in \arg\max_{a \in \mathcal{A}_0} \{\widehat{F}_{a,t}(\lambda_t) + \beta_{a,t}^{Z,\Gamma}\}$.
    end for
    Set $W_t = \sum_{j \in \mathcal{E}_{\text{M}}} w_{t,j}$ and $p_t(a) = (1 - K_0 p_{\min}) \sum_{j \in \mathcal{E}_{\text{M}}} (w_{t,j}/W_t) \xi_{t,j}(a) + p_{\min}$ for $a \in \mathcal{A}_0$.
    In forced mode draw $e_t \sim \text{Bernoulli}(\epsilon_t)$; otherwise set $e_t = 0$. If $e_t = 1$, draw a_t uniformly from $\mathcal{A}$; otherwise draw $a_t \sim p_t$.
  end if
  Observe $(r_t(a_t), b_t(a_t))$ for $a_t \ne \texttt{null}$; for $a_t = \texttt{null}$, set both to zero.
  if $a_t \ne \texttt{null}$ and [$t \le t_{\text{init}}$ or mode = on-policy or $e_t = 1$] **then**
    Append $(x_t, r_t(a_t), b_t(a_t))$ once to the shared history for a_t.
  end if
  if $t > t_{\text{init}}$ and $e_t = 0$ **then**
    Set $Y_t(a_t) = [r_t(a_t) - Z\lambda_t^{\top} b_t(a_t) + Z]/(1+Z)$ and $\widehat{Y}_t(a) = \mathbf{1}\{a = a_t\} Y_t(a_t)/p_t(a_t)$.
    for $j \in \mathcal{E}_{\text{M}}$ **do**
      Set $\widehat{y}_{t,j} = \sum_{a \in \mathcal{A}_0} \xi_{t,j}(a) \widehat{Y}_t(a)$ and $\widehat{v}_{t,j} = \sum_{a \in \mathcal{A}_0} \xi_{t,j}(a)/p_t(a)$.
      Update $w_{t+1,j} = w_{t,j} \exp\{(p_{\min}/2)(\widehat{y}_{t,j} + \alpha_m \widehat{v}_{t,j})\}$.
    end for
  else
    Keep every master weight unchanged.
  end if
  Update the observed budget account and resource prices: $U_{t+1} = U_t + b_t(a_t)$ and (q_{t+1}, λ_{t+1}) by (16).
end for

Algorithm 4 uses one realized history for all corruption candidates. Initialization and forced overrides leave the master weights unchanged, every master-controlled action is drawn from the explicitly smoothed distribution p_t used in its importance-weighted payoff, and the common dual update always uses the played action's observed consumption.

D.5 One Shared Estimator and Candidate Coverage

All candidates use the same point estimates, sample sets, regularization, numerical precision, supplied calibration, and realized design; only their radii and current recommendations differ. At most one eligible observation is appended on a played round. The common time-uniform noise event therefore covers the grid, and the monotonicity of the radius in Γ makes it valid for the path-dependent $\Gamma^{\star}$ selected after the run.

Condition 5 links each candidate's recommendation count to the realized shared history used by the on-policy width calculation. Forced exploration supplies candidate-independent coverage. The following lemma gives a sufficient sampling condition. These probability guarantees concern the actual shared trajectory; a post-run diagnostic alone does not establish them.

Lemma 1 (Recommendation-aligned sampling implies candidate coverage)**.** *Suppose that on every post-initialization active round in on-policy mode, a candidate-recommended action receives*

conditional sampling probability at least $q_{\mathrm{cov}} > 0$: whenever $\widehat{a}_t^{\Gamma} = a$,

$$\mathbb{P}(a_t = a \mid \mathcal{Q}_t) \geq q_{\mathrm{cov}},$$

where $\mathcal{Q}_t$ contains the history, current context, shared estimates, prices, and all candidate recommendations before the fresh action draw. Then, with probability at least $1 - \delta_{\mathrm{cand}}$, simultaneously for every $\Gamma \in \mathcal{G}$, $a \in \mathcal{A}$, and $t \leq T + 1$,

$$N_a^{\mathrm{rec},\Gamma}(t) \leq \frac{2}{q_{\mathrm{cov}}} N_a(t) + \frac{8}{q_{\mathrm{cov}}} \log \frac{2|\mathcal{G}|KT}{\delta_{\mathrm{cand}}}. \tag{37}$$

Consequently, Condition 5 holds with

$$\chi = \frac{2}{q_{\mathrm{cov}}}, \qquad n_{\mathrm{cand}} = \frac{8}{q_{\mathrm{cov}}} \log \frac{2|\mathcal{G}|KT}{\delta_{\mathrm{cand}}}.$$

Proof. Fix (Γ, a, t) and, for post-initialization rounds $\tau < t$, let

$$Q_\tau := I_\tau \mathbf{1}\{\widehat{a}_\tau^{\Gamma} = a\}, \qquad X_\tau := Q_\tau \mathbf{1}\{a_\tau = a\}.$$

Conditional on $\mathcal{Q}_\tau$, $\mathbb{E}[X_\tau \mid \mathcal{Q}_\tau] \geq q_{\mathrm{cov}} Q_\tau$. For every deterministic $u > 0$, a multiplicative martingale lower-tail bound gives

$$\Pr\left(\sum_{\tau<t} X_\tau < \frac{q_{\mathrm{cov}}}{2} \sum_{\tau<t} Q_\tau, \quad \sum_{\tau<t} Q_\tau \geq u \right) \leq \exp\left(-\frac{q_{\mathrm{cov}} u}{8} \right).$$

Set

$$u = \frac{8}{q_{\mathrm{cov}}} \log \frac{2|\mathcal{G}|KT}{\delta_{\mathrm{cand}}}.$$

A union bound over candidates, actions, and prefixes implies that, simultaneously for all (Γ, a, t), either $N_a^{\mathrm{rec},\Gamma}(t) < u$, or

$$\sum_{\tau<t} X_\tau \geq \frac{q_{\mathrm{cov}}}{2} N_a^{\mathrm{rec},\Gamma}(t).$$

Since $\sum_{\tau<t} X_\tau \leq N_a(t)$, in the latter case $N_a^{\mathrm{rec},\Gamma}(t) \leq 2N_a(t)/q_{\mathrm{cov}}$. Combining the two cases proves (37). □

The smoothed EXP4.P distribution satisfies $p_t(a) \geq p_{\min}$, so Lemma 1 gives explicit finite-horizon coverage parameters with $q_{\mathrm{cov}} = p_{\min}$. The same result shows directly how stronger recommendation alignment improves n_{cand} and χ; when q_{cov} is bounded below uniformly over the horizon, these parameters match the fixed-coverage regime used in the $\widetilde{O}(\sqrt{T})$ summary.

Appendix E.7 gives a bounded model and sampling regime satisfying the new RSC and count conditions. This establishes nonemptiness of the conditions; their validity for an adaptive OP algorithm concerns that algorithm's own trajectory. In particular, using only the default $p_{\min}$ yields horizon-dependent recommendation-coverage parameters.

D.6 Complete Concrete Shared-Data Rates

Corollary 6 (Unknown FE with shared estimation)**.** *Under Assumptions 1, 2, and 3, run Algorithm 4 with $B^{\mathrm{op}} = B$, fixed safe cyclic initialization length t_{init}, constant $0 < \epsilon \leq 1/2$, and (43). Use the complete grid, uniform-advice master expert, and Appendix E.2's common regularization and certified precision on the initialization and later forced samples. For a universal C and $\delta_{\mathrm{tot}} = \delta_{\mathrm{tot}}^{\mathrm{U,FE}}$,*

$$\begin{aligned} \mathrm{Reg}(T) \leq C \Bigg[& h t_{\mathrm{init}} + \frac{hK(n_0 + \ell)}{\epsilon} + A_{\mathrm{FE}} \sqrt{\frac{KT}{\epsilon}} + h\epsilon T \\ & + \mathrm{Reg}_{\mathrm{mst}}(T) + \mathcal{D}_T + Z + \left(3 + \frac{KH_T}{\kappa_{\mathrm{FE}} \epsilon} \right) \mathbb{E}[\Gamma^\star] \Bigg] + h\delta_{\mathrm{tot}} T. \end{aligned} \tag{38}$$

Choosing (44) gives (29), including its capped regime, by the calculation in Appendix C.

Corollary 7 (Unknown OP with shared estimation)**.** *Under Assumptions 1, 2 and Conditions 4, 5, run Algorithm 4 with $B^{\mathrm{op}} = B$, $\epsilon_t = 0$, its stated initialization, complete grid, and uniform-advice expert. Fit every realized non-null sample using the supplied OP calibration and the same certified Lasso rule. Then (30) holds with failure total $\delta_{\mathrm{tot}}^{\mathrm{U,OP}}$.*

The initialization costs at most ht_{init} in score regret. Statistical widths depend on the actual retained sample sets; recommendation coverage controls hypothetical candidate recommendations, not only played actions. Direct corruption contributes $3\mathbb{E}[\Gamma^\star]$, separately from its repeated effect on confidence widths. The master and the dual-price terms remain explicit. Appendix G proves these statements on one joint event before substituting the path-dependent valid grid point.

E Lasso Confidence, Design Calibration, and Cumulative Widths

E.1 Retained Histories and the Confidence Interface

The regression sample set, not the recommendation count, determines each fit. With e_τ the independent forced-exploration indicator, the three retained histories are

$$\begin{aligned}\mathcal{S}_a^{\mathrm{FE}}(t) &= \{\tau < t : I_\tau e_\tau = 1,\ a_\tau = a\},\\ \mathcal{S}_a^{\mathrm{OP}}(t) &= \{\tau < t : a_\tau = a\},\\ \mathcal{S}_a^{\mathrm{sh,FE}}(t) &= \{\tau \le t_{\mathrm{init}} : \tau < t,\ I_\tau = 1,\ a_\tau = a\} \cup \{t_{\mathrm{init}} < \tau < t : I_\tau e_\tau = 1,\ a_\tau = a\}.\end{aligned} \tag{39}$$

These sets are defined for $a \in \mathcal{A}$; let $N_a^{\mathrm{FE}}(t) := |\mathcal{S}_a^{\mathrm{FE}}(t)|$ and $N_a^{\mathrm{est}}(t) := |\mathcal{S}_a^{\mathrm{est}}(t)|$ for the selected mode. In OP it equals the realized non-null action count $N_a(t)$; in FE it generally does not. The candidate-recommendation count $N_a^{\mathrm{rec},\Gamma}(t)$ is a different object from either retained count.

E.2 Regularization and Prediction Calibration

Section 3.1 defines the norm-constrained objective and states Proposition 1. Here we specify its inputs and calibration. Supply $R > 0$, an integer $\max\{1, s_0\} \le s \le d$, and noise bounds $\sigma_r, \sigma_b \ge 0$. Bounded clean outcomes permit $\sigma_r = \sigma_b = 1/2$ by Hoeffding's lemma. The fit does not require a support constraint; s is used for confidence and numerical-precision calibration. For one action and channel, n, X, y, and $\theta^\star$ have the local meanings given in Section 3.1; the corresponding estimate $\widehat{\theta}$ is $\widehat{\theta}_{a,t-1}$. At zero samples set it to zero. Exact fits select the minimum-Euclidean-norm minimizer, while approximate fits use the deterministic certified solver in Appendix E.8.

A design calibration consists of supplied $\kappa, \upsilon > 0$ and an integer n_0. For failure levels $\delta_{\mathrm{est}}, \delta_{\mathrm{cov}} \in (0,1)$ and $n \ge 1$, set

$$\ell = \log \frac{16eK(m+1)dT}{\min\{\delta_{\mathrm{est}}, \delta_{\mathrm{cov}}\}}, \quad \omega_n = \frac{\upsilon\ell}{n}, \quad \rho_n = 4(\sigma + R\sqrt{\kappa\upsilon})\sqrt{\frac{2\ell}{n}}, \quad n_0 \ge \left\lceil \frac{256 s\upsilon\ell}{\kappa} \right\rceil. \tag{40}$$

Use $\sigma = \sigma_r$ or σ_b for the respective channel, writing ρ_n^r or ρ_n^b accordingly. The required restricted strong convexity (RSC) event is

$$\frac{1}{n}\|Xu\|_2^2 \ge \kappa\|u\|_2^2 - \omega_n\|u\|_1^2 \quad \text{for every } u \in \mathbb{R}^d \text{ and every retained prefix } n \ge n_0. \tag{41}$$

The tolerance is essential when $n < d$. The positive regularization floor in (40) controls corruption even if clean noise is zero. It depends on the design calibration, not on a candidate corruption bound.

Under this calibration, Proposition 1 gives simultaneous prediction on every unit-ball context. Its bound holds for all $C_n \ge 0$, with no corruption-dependent sample threshold, and permits an objective gap of $s\rho_n^2/(8\kappa)$. Exact optimization replaces the coefficient 7 by 6.

Appendix E.5 proves the sequential noise bound, enlarged-cone argument, and large-corruption case. Appendix E.4 gives the resulting reward and resource radii. All candidates share the retained data, regularization, numerical-precision rule, and fitted coefficients.

Computation and information. Appendix E.8 gives a constrained proximal solver with a computable optimality-gap certificate for exactly (8). Calibration is a separate requirement: the learner uses

model-supplied bounds justified in Appendix E.3, not an oracle for an unknown empirical RE constant. The on-policy guarantee remains conditional on supplied trajectory bounds. Convex fitting alone does not certify informative confidence for an arbitrary design. At most $(m+1)T$ changed fits are needed; the grid adds candidate scoring, not candidate-specific fits.

E.3 Population and On-Policy Design Conditions

Assumption 3 and Condition 4 instantiate the main-text interface for the estimator in Appendix E.2; they are not imposed on another estimator that already meets that interface. Recall $h = 1 + Z$. For a calibrated pair (κ, υ), define the clean-width coefficient

$$A(\kappa, \upsilon) := \frac{28\sqrt{2s\ell}}{\kappa}\left[\sigma_r + Z\sigma_b + hR\sqrt{\kappa\upsilon}\right], \qquad H_T := 1 + \log(1+T). \tag{42}$$

This coefficient includes noise, sparsity, norm, design, and numerical-accuracy costs. Horizon summaries use the price stepsize specified after (17).

Assumption 3 is stated in Section 3.2.

The condition imposes a lower bound in all population directions, rather than only in $2s$-sparse directions. Under context-independent forced selection, Lemma 5 proves (41) simultaneously for retained prefixes with the conservative calibration

$$D_x = 4096L_x^4/\kappa_x^2, \quad \kappa_{\mathrm{FE}} = \kappa_x/2, \quad \upsilon_{\mathrm{FE}} = 16\kappa_{\mathrm{FE}}D_x, \quad n_0 = \lceil 4096 s D_x \ell \rceil. \tag{43}$$

Write $A_{\mathrm{FE}} = A(\kappa_{\mathrm{FE}}, \upsilon_{\mathrm{FE}})$. These formulas do not require $n \geq d$, but their constants can be conservative. Since $\|x\|_2 \leq 1$, necessarily $\kappa_x \leq 1/d$; dimension dependence through curvature must not be suppressed. Appendix E.7 gives a bounded Rademacher-feature instance with D_x independent of d and explains its compatible nonnegative feedback model.

On-policy coverage. Retain all non-null observations and write $N_a(t) = \sum_{\tau<t} \mathbf{1}\{a_\tau = a\}$. Let $\mathcal{E}_{\mathrm{RSC}}(\kappa_{\mathrm{op}}, \upsilon_{\mathrm{op}}, n_0)$ denote (41) for every action-specific realized design with $N_a(t) \geq n_0$, using $\omega_n = \upsilon_{\mathrm{op}}\ell/n$. Condition 4 in Section 3.2 states the required probability bound for this event.

Set $A_{\mathrm{op}} = A(\kappa_{\mathrm{op}}, \upsilon_{\mathrm{op}})$. Population calibration does not imply this condition for adaptive actions: selection, budget stopping, and corruption can change each action's realized design. A post-run diagnostic is not a learner-supplied lower bound, and a selected-support eigenvalue does not certify (41). Context-independent uniform actions with sufficient budget give a nonempty illustrative trajectory class, not a guarantee that our on-policy algorithm behaves that way (Appendix E.7). Candidate-recommendation coverage remains the separate Condition 5 for Shared-Grid OP.

Exploration calibration. For the concrete forced-exploration implementation, choose a constant probability

$$\epsilon = \min\left\{\frac{1}{2}, \max\left\{\left(\frac{A_{\mathrm{FE}}^2 K}{h^2 T}\right)^{1/3}, \sqrt{\frac{K(n_0+\ell)}{T}}\right\}\right\}. \tag{44}$$

This balances clean statistical, burn-in, and exploration costs using supplied calibration only; it is independent of the corruption bound and is also used by Shared-Grid. The full bounds for arbitrary constant $0 < \epsilon \leq 1/2$ appear in Appendices C and D.

E.4 Channel Radii and Score Confidence

All point estimates use (8) and (40). At zero samples the estimate is zero and both channel radii are one. For a supplied, justified calibration and $n = N_a^{\mathrm{est}}(t) \geq n_0$, the concrete radii are

$$\beta_{a,t}^{r,\Gamma} = \min\{1, 7\rho_n^r\sqrt{s}/\kappa + 4\Gamma/(\kappa n)\}, \tag{45}$$

$$\beta_{a,i,t}^{b,\Gamma} = \min\{1, 7\rho_n^b\sqrt{s}/\kappa + 4\Gamma/(Z\kappa n)\}. \tag{46}$$

Otherwise they are one. Null-action estimates and radii are zero. These prescriptions are defined even when a calibration event fails; confidence is asserted on that event, not by a data-dependent claim that an uncomputed RSC constant is positive.

Let $\beta^{r,0}$ and $\beta^{b,0}$ be the same formulas with $\Gamma = 0$. The clean combined width and its corruption envelope obey

$$\beta_{a,t}^{Z,\Gamma} \le f(n) + \frac{8\Gamma}{\kappa(n \vee 1)} \mathbf{1}\{n \ge n_0\}, \qquad f(n) := \begin{cases} h, & n < n_0, \\ \min\{h, A(\kappa, \upsilon)/\sqrt{n}\}, & n \ge n_0. \end{cases} \tag{47}$$

Here f is nonincreasing, $h = 1+Z$, and A is defined in (42). The bound follows from $\min\{1, u+v\} \le \min\{1, u\}+v$ and holds on any realized history, independently of confidence validity. Before burn-in, the entire radius is charged to the clean part. The nonnegative envelope $8\Gamma/(\kappa n)$ need not itself be clipped.

Lemma 2 (From channel prediction to score confidence)**.** *On the noise-and-design event of Proposition 1, (10) holds simultaneously for all valid Γ, and for every $\lambda \in \Lambda$,*

$$|\widehat{F}_{a,t}(\lambda) - F^0(a, x_t; \lambda)| \le \beta_{a,t}^{Z,\Gamma}. \tag{48}$$

Proof. The clean means are in $[0, 1]$ because the clean outcomes are in that interval and their noise is conditionally mean-zero. Projection onto $[0, 1]$ cannot increase distance to a clean mean. On a retained history,

$$\sum_{\tau \in \mathcal{S}_a^{\mathrm{est}}(t)} |c_\tau^r(a)| \le C_Z \le \Gamma, \qquad \sum_{\tau \in \mathcal{S}_a^{\mathrm{est}}(t)} |c_{\tau,i}^b(a)| \le C_Z/Z \le \Gamma/Z.$$

These inequalities concern observed-minus-clean deviations after clipping, not the raw perturbations. Apply Proposition 1 to each channel, including the trivial early radii, and use $\|\lambda\|_1 \le 1$. The same fits and noise event work for every valid Γ, including a path-dependent grid choice. No union bound over corruption values is needed. □

E.5 Sequential Noise and the Corrupted Lasso Basic Inequality

Lemma 3 (Coordinate noise control for retained prefixes)**.** *For the pre-feedback retention rules in (39), with probability at least $1 - \delta_{\mathrm{est}}$, every action, response channel, and obtained prefix of size $1 \le n \le T$ satisfies*

$$\|X^\top \eta/n\|_\infty \le \sigma\sqrt{2\ell/n} \le \rho_n/2. \tag{49}$$

The claim remains valid for adaptively selected and budget-stopped OP histories.

Proof. Fix an action, channel, and coordinate j. Let Q_t indicate that this action's observation is retained. Conditional on the past, current context, and the learner's fresh action/retention seed, $Q_t x_{t,j}$ is determined and the clean noise still has its stated conditional sub-Gaussian bound: the fresh seed is conditionally independent of the potential outcomes and corruption. The retention rule cannot inspect the current feedback. Thus, for every real u,

$$\mathbb{E}\left[\exp\{uQ_t x_{t,j}\eta_t - u^2\sigma^2 Q_t/2\} \mid \text{past, context, seed}\right] \le 1.$$

Iterated conditioning makes the product an exponential supermartingale in calendar time. Stop it at the nth retained observation or at T, whichever comes first. On the event that the nth observation occurs and its score exceeds b, its value is at least $\exp(ub - u^2\sigma^2 n/2)$. Optional stopping and Markov's inequality, with $u = b/(n\sigma^2)$, give a tail at most $\exp[-b^2/(2n\sigma^2)]$ on this event. Apply both signs and take $b = \sigma\sqrt{2n\ell}$. Union over $dK(m+1)T$ choices gives failure at most $2dK(m+1)Te^{-\ell} \le \delta_{\mathrm{est}}$. When $\sigma = 0$, the conditional sub-Gaussian property implies $\eta_t = 0$ almost surely and the assertion is immediate. The supermartingale argument retains adaptive selection and stopping throughout. □

Lemma 4 (Shared-regularization Lasso under response corruption)**.** *Consider $y = X\theta^\star + \eta + \zeta$, with row norms at most one, $\|\theta^\star\|_2 \le R$, and $|J| \le s$ for $J = \mathrm{supp}(\theta^\star)$. Let $\widehat{\theta}$ be feasible for (8) and have objective suboptimality at most $\varepsilon_{\mathrm{opt}} \ge 0$. Suppose*

$$Q(u) := \|Xu\|_2^2/n \ge \kappa\|u\|_2^2 - \omega\|u\|_1^2 \quad \text{for every } u,$$
$$\|X^\top \eta/n\|_\infty \le \rho/2, \qquad 0 < \omega \le \kappa/(256s), \qquad \rho \ge 4R\sqrt{\kappa\omega}.$$

Writing $C_n = \|\zeta\|_1$, one has

$$\|\widehat{\theta} - \theta^\star\|_2 \le \min\left\{2R, \frac{6\rho\sqrt{s}}{\kappa} + \frac{4C_n}{\kappa n} + 2\sqrt{\frac{\varepsilon_{\rm opt} + 4\omega\varepsilon_{\rm opt}^2/\rho^2}{\kappa}}\right\}. \tag{50}$$

In particular, $\varepsilon_{\rm opt} \le s\rho^2/(8\kappa)$ gives (9). The error vector may be dense; the proof controls its ℓ_1 norm through the penalized basic inequality.

Proof. Put $\Delta = \widehat{\theta} - \theta^\star$, $r = \|\Delta\|_2 \le 2R$, and $\gamma = C_n/n$. Feasibility of $\theta^\star$ and the objective-gap inequality yield the basic inequality

$$\frac{1}{2}Q(\Delta) \le \frac{\eta^\top X\Delta}{n} + \frac{\zeta^\top X\Delta}{n} + \rho(\|\theta^\star\|_1 - \|\theta^\star + \Delta\|_1) + \varepsilon_{\rm opt} \le \frac{3\rho}{2}\|\Delta_J\|_1 - \frac{\rho}{2}\|\Delta_{J^c}\|_1 + \gamma r + \varepsilon_{\rm opt}. \tag{51}$$

The corruption bound uses $|\langle x_j, \Delta\rangle| \le r$, and the penalty difference uses sparsity of $\theta^\star$. Optimality compares the penalized objectives, so the penalty difference remains in the basic inequality. Since $Q \ge 0$,

$$\|\Delta_{J^c}\|_1 \le 3\|\Delta_J\|_1 + \frac{2\gamma}{\rho}r + \frac{2\varepsilon_{\rm opt}}{\rho}, \qquad \|\Delta\|_1 \le (4\sqrt{s} + 2\gamma/\rho)r + 2\varepsilon_{\rm opt}/\rho. \tag{52}$$

Squaring with $(a+b)^2 \le 2a^2 + 2b^2$ gives

$$\|\Delta\|_1^2 \le (64s + 16\gamma^2/\rho^2)r^2 + 8\varepsilon_{\rm opt}^2/\rho^2.$$

First suppose $\gamma \le \rho\sqrt{\kappa/(64\omega)}$. The two tolerance coefficients are each at most $\kappa/4$, so

$$Q(\Delta) \ge \kappa r^2/2 - 8\omega\varepsilon_{\rm opt}^2/\rho^2.$$

Combining with (51) and dropping its negative term gives

$$\kappa r^2/4 \le (3\rho\sqrt{s}/2 + \gamma)r + \varepsilon_{\rm opt} + 4\omega\varepsilon_{\rm opt}^2/\rho^2.$$

For $ar^2 \le br + c$ with $a > 0, b, c \ge 0$, $r \le b/a + \sqrt{c/a}$, proving the second bound in (50) in this case. If $\gamma > \rho\sqrt{\kappa/(64\omega)}$, then the regularization floor implies $\gamma > R\kappa/2$. Consequently $r \le 2R < 4\gamma/\kappa$, which proves the same bound without absorbing the enlarged cone. This case distinction uses the unknown true corruption only in analysis; neither the algorithm nor burn-in tests it. Finally, when $\varepsilon_{\rm opt} \le s\rho^2/(8\kappa)$, one has $4\omega\varepsilon_{\rm opt}/\rho^2 \le 1$. The square-root term is at most $2\sqrt{2\varepsilon_{\rm opt}/\kappa} \le \rho\sqrt{s}/\kappa$. Cauchy–Schwarz gives simultaneous prediction on every new unit-ball context. With exact optimization that term is zero. □

Proof of Proposition 1. Use Lemma 3. At every calibrated prefix, (40) implies $\omega_n \le \kappa/(256s)$ and $\rho_n \ge 4R\sqrt{\kappa\omega_n}$, even when $\sigma = 0$. Apply Lemma 4 on the actual retained design. The lemma holds for arbitrary retained corruption vectors and therefore simultaneously for all admissible corruption sums on that history. □

E.6 A Primitive RSC Calibration for Forced Samples

Lemma 5 (Bounded sub-Gaussian design calibration)**.** *Under Assumptions 1 and 3, the FE retained histories, including a safe context-independent cyclic initialization, satisfy (41) with (43), simultaneously for all obtained prefixes $n \ge n_0$, with failure probability at most $\delta_{\rm cov}/2$.*

Proof. Write $\Sigma = \mathbb{E}[xx^\top]$. For a unit vector v, put $W = \langle v, x\rangle^2$. The exponential-moment assumption gives $\mathbb{E}W^p \le 2p!L_x^{2p}$ and $\mathbb{E}|W - \mathbb{E}W|^p \le 2^{p+1}p!L_x^{2p}$ for $p \ge 2$. Expanding the exponential power series yields Bernstein's bound

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{j=1}^n W_j - \mathbb{E}W\right| > z,\ n \text{ obtained}\right) \le 2\exp\left(-\frac{nz^2}{32L_x^4 + 4L_x^2 z}\right). \tag{53}$$

For independent samples this is the ordinary Chernoff calculation. For the actual retained samples the identical calculation uses the exponential supermartingale with predictable selection and stopping at the nth retained observation, as in Lemma 3. FE selection and cyclic initialization are independent of the current context; safety is determined before that context. Hence the conditional moment bound is unchanged by selection. No conditioning on the final count or on survival is needed.

Jensen's inequality implies $\kappa_x \le L_x^2 \log 2 \le L_x^2$. With $z = \kappa_x/8$, the exponent in (53) is at least n/D_x for $D_x = 4096 L_x^4/\kappa_x^2$. For each support of size q, a $1/4$-net of its unit sphere has size at most 9^q. Union over supports and the net implies

$$|v^\top(\widehat{\Sigma}_n - \Sigma)v| \le (\kappa_x/4)\|v\|_2^2 \quad \text{for all } q\text{-sparse } v$$

except with probability at most $2\exp[-n/D_x + q\log(9ed/q)]$. To see the factor two from the net explicitly, for the symmetric restriction M to any support, its operator norm is at most the maximum net quadratic form plus $\frac{1}{2}\|M\|_{\mathrm{op}}$.

For each $n \ge 8D_x\ell$, let $k = \lfloor n/(8D_x\ell)\rfloor$ and $q = \min\{2k, d\}$. Since $\log(9ed/q) \le \ell$ and $q \le n/(4D_x\ell)$, the preceding failure bound is at most $2e^{-3n/(4D_x)} \le 2e^{-6\ell}$. If $q = d$, the net result gives $\widehat{\Sigma}_n \succeq 3\kappa_x I_d/4$ and the desired RSC follows. Otherwise partition any u into disjoint blocks of k coordinates, in decreasing absolute order. On the union of any two blocks the symmetric operator norm of $\widehat{\Sigma}_n - \Sigma$ is at most $\kappa_x/4$. Also

$$\sum_j \|u_{J_j}\|_2 \le \|u\|_2 + \|u\|_1/\sqrt{k}.$$

Indeed, the norm of each block after the first is at most the preceding block's ℓ_1 norm divided by $\sqrt{k}$. Bilinearity and the two-block restriction therefore give

$$|u^\top(\widehat{\Sigma}_n - \Sigma)u| \le \frac{\kappa_x}{4}\left(\sum_j \|u_{J_j}\|_2\right)^2 \le \frac{\kappa_x}{2}\|u\|_2^2 + \frac{\kappa_x}{2k}\|u\|_1^2.$$

It follows that $u^\top\widehat{\Sigma}_n u \ge \kappa_{\mathrm{FE}}\|u\|_2^2 - \kappa_{\mathrm{FE}}\|u\|_1^2/k$. The bound $k \ge n/(16D_x\ell)$ proves the asserted tolerance. Union over at most KT obtained prefixes costs at most $2KTe^{-6\ell} \le \delta_{\mathrm{cov}}/2$. The threshold $n_0 = \lceil 4096 s D_x \ell\rceil$ is above $8D_x\ell$ and satisfies the absorption requirement in (40). □

The full population lower bound and all-direction sub-Gaussian envelope are used in the block-extension step above; positivity only on $2s$-sparse directions would not justify that step. RSC with a tolerance is a standard framework for high-dimensional regularized estimation (Negahban et al., 2012); the selected-sample concentration and corruption calculation above supply the specific steps used here.

E.7 A Nonempty Bounded-Context Instance

Let $x = (1, \epsilon_2, \ldots, \epsilon_d)/\sqrt{d}$, where the ϵ_j are independent Rademacher signs. Then $\|x\|_2 = 1$ and $\mathbb{E}xx^\top = I_d/d$. One may take $\kappa_x = 1/d$ and $L_x = 4/\sqrt{d}$. To verify the exponential envelope, write $\sqrt{d}\langle u, x\rangle = u_1 + V$, where $V = \sum_{j\ge 2} u_j\epsilon_j$ is sub-Gaussian with scale at most $\|u\|_2$. For $\alpha > 0$ and an independent standard normal g, the identity $\mathbb{E}_{V,g}e^{\sqrt{2\alpha}gV} = \mathbb{E}_V e^{\alpha V^2}$ implies $\mathbb{E}e^{V^2/(8\|u\|_2^2)} \le (1 - 1/4)^{-1/2}$. Since $(u_1 + V)^2 \le 2u_1^2 + 2V^2$, the required expectation is at most $e^{1/8}(1 - 1/4)^{-1/2} < 2$. Thus D_x is independent of d and the sufficient sample threshold is of order $s\ell$, albeit with conservative numerical constants. It can be smaller than d when $d/(s\ell)$ is sufficiently large; no claim of a small practical burn-in follows from these constants.

For compatible sparse means, take $s_0 = s$ and choose $\theta^\star = \sqrt{d}(ce_1 + \sum_{j=2}^s b_j e_j)$ with $0 < c - \sum|b_j| \le c + \sum|b_j| < 1$ and $d(c^2 + \sum b_j^2) \le R^2$. Such choices exist for every $R > 0$ by scaling c, b_j down together. Add independent bounded mean-zero noise of amplitude less than the distance of the means to the endpoints of $[0, 1]$, separately for each channel. This gives sparse parameters with bounded, potentially context-dependent reward and consumption means. In this normalized class curvature is $1/d$; the coefficient A must retain that dependence.

For illustration, context-independent uniform non-null sampling and budgets $B_i \geq T+1$ prevent safety stopping and give the same design event on all action histories. A time-uniform Chernoff bound gives $N_a(t) \geq (t-1)/(2K) - c\log(KT/\delta)$. Moreover, $N_a^{\mathrm{rec},\Gamma}(t) \leq t-1 \leq 2KN_a(t) + cK\log(KT/\delta)$ for any recommendations. Thus this sampling regime provides a nonempty illustrative class with informative action-specific designs and recommendation coverage. It is not a guarantee that the proposed OP algorithms generate the same trajectories under corruption and budget stopping.

E.8 A Reproducible Solver and Certified Numerical Error

For $n \geq 1$, the following solver minimizes the constrained objective (8). Inputs are the retained matrix X, clipped response vector $y \in [0,1]^n$, R, and the channel-specific ρ_n. There is no unpenalized intercept and no automatic centering, feature standardization, or response rescaling. A constant feature, when used, is part of x and subject to both the penalty and norm bound. Any deliberate preprocessing must first be reflected in R, noise bounds, and design calibration.

Let $g(\theta) = \|y - X\theta\|_2^2/(2n)$ and $F(\theta) = g(\theta) + \rho_n\|\theta\|_1$. Its gradient is 1-Lipschitz because $\|X^\top X/n\|_{\mathrm{op}} \leq 1$. Start at $\theta^{(0)} = 0$ and use step size one. For iteration j set

$$v = \theta^{(j)} - \nabla g(\theta^{(j)}), \qquad z = \mathrm{sign}(v)(|v| - \rho_n)_+, \qquad \theta^{(j+1)} = z\min\{1, R/\|z\|_2\}, \tag{54}$$

with the last expression interpreted as zero when $z = 0$. The soft threshold and scaling together are the proximal map of $\rho_n\|\cdot\|_1 + \iota_{\{\|\cdot\|_2 \leq R\}}$, where the last term is the convex indicator (zero on the ball, $+\infty$ outside). To check this, set $a = (\|z\|_2/R - 1)_+$, and choose $w_j = \mathrm{sign}(z_j)$ if $z_j \neq 0$, and $w_j = v_j/\rho_n$ otherwise. Then $w \in \partial\|\theta^{(j+1)}\|_1$, $v - \theta^{(j+1)} = \rho_n w + a\theta^{(j+1)}$, and $a(\|\theta^{(j+1)}\|_2 - R) = 0$, which are sufficient convex optimality conditions for the proximal problem.

At the new iterate form $v_F = \nabla g(\theta^{(j+1)}) + \rho_n w$ and the computable certificate

$$G_j = \langle v_F, \theta^{(j+1)}\rangle + R\|v_F\|_2 \geq F(\theta^{(j+1)}) - \min_{\|\theta\|_2 \leq R} F(\theta) \geq 0. \tag{55}$$

The first inequality follows by minimizing the supporting affine minorant of F over the ball. Stop at the first certified $G_j \leq \varepsilon_{\mathrm{opt},n} := s\rho_n^2/(8\kappa)$. This is a deterministic selection convention for approximate fits; it need not return the minimum-norm exact minimizer. The latter convention is well defined because the minimizer set is nonempty, compact, and convex and the squared Euclidean norm is strictly convex. Measurability of the exact convention follows, for example, by taking limits of the unique minimizers of $F(\theta) + \epsilon\|\theta\|_2^2$ on the ball as $\epsilon \downarrow 0$. The finite algorithm uses measurable arithmetic and a first-passage stopping rule.

Proposition 3 (Solver termination and precision accounting)**.** *In exact arithmetic the stopping rule above terminates. A sufficient upper bound on the iterations is* $1 + \lceil 16R^2/\varepsilon_{\mathrm{opt},n}^2\rceil$. *Each iteration costs* $O(nd)$ *arithmetic operations. The returned fit satisfies Proposition 1, including its numerical-error contribution.*

Proof. Proximal descent and the 1-Lipschitz gradient give $F(\theta^{(j)}) - F(\theta^{(j+1)}) \geq \frac{1}{2}\|\theta^{(j+1)} - \theta^{(j)}\|_2^2$. Since $F(0) \leq 1/2$ and $F \geq 0$, the sum of these squared steps is at most one. The proximal optimality equation implies

$$\|v_F + a\theta^{(j+1)}\|_2 \leq \|\theta^{(j)} - \theta^{(j+1)}\|_2 + \|\nabla g(\theta^{(j+1)}) - \nabla g(\theta^{(j)})\|_2 \leq 2\|\theta^{(j+1)} - \theta^{(j)}\|_2.$$

Using complementarity and $\|\theta^{(j+1)}\|_2 \leq R$ in (55) gives $G_j \leq 4R\|\theta^{(j+1)} - \theta^{(j)}\|_2$. If all of the first J certificates exceeded $\varepsilon_{\mathrm{opt},n}$, the squared-step sum would exceed $J\varepsilon_{\mathrm{opt},n}^2/(16R^2)$. This is impossible for the stated J. The certificate implies the objective-gap condition in Lemma 4. □

The iteration bound is conservative; it establishes certified termination rather than practical speed. For floating-point implementation, verify feasibility and evaluate an upper bound on (55) using directed rounding or explicitly bounded arithmetic residuals. An approximately valid subgradient with a certified Euclidean discrepancy e adds at most $2Re$ to the certificate. Accept an iterate only when a certified upper bound on its Euclidean norm is at most R. A fixed iteration limit or small successive-step difference without these error bounds is not the certificate used in our guarantee. If the tolerance is not met, use trivial radii until a certified fit is available; the concrete rates assume the stated certificate is obtained.

For a different certified objective gap $\varepsilon_{\mathrm{opt},a,t}$, the general additional prediction term is

$$E_{a,t} = 2\sqrt{\left(\varepsilon_{\mathrm{opt},a,t} + 4\omega_n \varepsilon^2_{\mathrm{opt},a,t}/\rho_n^2\right)/\kappa}.$$

Apply this separately by channel and add twice the recommendation-weighted sum of $E^r + Z \max_i E_i^b$ to the cumulative width interface. Under the prescribed tolerance these sums are already included in the coefficient 7 and hence in A; there is no omitted fixed numerical error accumulated over T rounds.

Full online cost and calibration. For a fixed channel, all corruption candidates use the same X, y, ρ_n and solver stopping criterion. Only an action receiving a newly retained observation requires a new fit; cached fits for other actions retain their certified objective gaps for the unchanged objectives, since ρ_n depends on n, not calendar time. There are at most $(m+1)T$ changed fits. A conservative total arithmetic cost is

$$O\left((m+1)d\sum_{a=1}^{K}\sum_{n=1}^{N_a^{\mathrm{est}}(T+1)} n\left[1 + \frac{R^2}{\varepsilon^2_{\mathrm{opt},n}}\right] + T\{K(m+1)d + K|\mathcal{G}| + m\}\right),$$

using the largest channel-specific iteration bound in the first term. Computing all fitted scores costs $O(K(m+1)d)$ per round; scoring the grid costs $O(K|\mathcal{G}|)$, and the master and price updates add $O(|\mathcal{G}| + K + m)$. Storing retained raw samples costs $O(T(d+m))$. No delayed refitting or unanalysed stale estimator is used. In FE, κ_x, L_x and the model bounds are supplied and (43) is computed directly. In OP, the supplied pair must be justified for the actual trajectory by Condition 4; the method does not compute the smallest RE constant. Without justified bounds, trivial radii remain valid but these informative-width rates do not follow.

E.9 Counting and Summing Statistical Widths

Lemma 6 (Sequential scalar charging). *For $z_t \in [0,1]$, $Z_t = \sum_{u<t} z_u$, and $S = \sum_t z_t$, one has*

$$\sum_t \frac{z_t}{Z_t \vee 1} \le 2[1 + \log(1+S)], \qquad \sum_t \frac{z_t}{\sqrt{Z_t \vee 1}} \le 4\sqrt{S}. \tag{56}$$

If nonnegative integer counts A_n have cumulative sums $\sum_{j=0}^{n} A_j \le b + \chi(n+1)$, then every nonincreasing nonnegative sequence f satisfies

$$\sum_{n=0}^{N} A_n f(n) \le b f(0) + \chi \sum_{n=0}^{N} f(n). \tag{57}$$

Proof. For the first sum, the terms with $Z_t < 1$ have total mass less than two and denominator one. For the remaining terms $Z_t \ge 1$ and $z_t/Z_t \le 1$, so $z_t/Z_t \le 2\log(1 + z_t/Z_t)$. These logarithms telescope from a starting cumulative mass at least one, giving a sum at most $2\log(1+S)$; the asserted bound follows. For the square-root sum,

$$\sqrt{1+v+z} - \sqrt{1+v} = \frac{z}{\sqrt{1+v+z} + \sqrt{1+v}},$$

whose denominator is at most $4\sqrt{v \vee 1}$. Telescoping gives at most $4(\sqrt{1+S} - 1) \le 4\sqrt{S}$. Finally apply discrete summation by parts to the cumulative sums of A_n; all differences $f(n) - f(n+1)$ are nonnegative. Substitution of $b + \chi(n+1)$ gives (57). □

Lemma 7 (Forced counts and cumulative widths). *Let $0 < \epsilon \le 1/2$ be constant. On an event of failure at most $\delta_{\mathrm{cov}}/2$, every active round with $\epsilon(t-1) \ge 8K\ell$ satisfies*

$$N_a^{\mathrm{FE}}(t) \ge \epsilon(t-1)/(2K) \quad (a \in \mathcal{A}). \tag{58}$$

Together with Lemma 5, this gives the estimator/coverage event of failure at most $\delta_{\mathrm{est}} + \delta_{\mathrm{cov}}$. For an arbitrary sequence of recommendations, the FE cumulative interface holds with a universal constant C and

$$\mathcal{R}^0_{\mathrm{est}} \le C\left[\frac{hK(n_0+\ell)}{\epsilon} + A_{\mathrm{FE}}\sqrt{\frac{KT}{\epsilon}}\right], \qquad \psi_{\mathrm{est}}(\Gamma) \le \frac{CK\Gamma H_T}{\kappa_{\mathrm{FE}}\epsilon}. \tag{59}$$

For shared FE after initialization the same bounds hold over subsequent rounds, with the separate initialization cost charged in the allocation theorem.

Proof. Generate the independent forced coins and uniform draws for all calendar times, including unused draws after stopping. On the event that round t is active, all earlier rounds were active, so the actual pre-t forced count equals this latent count. A multiplicative Chernoff bound at mean $\epsilon(t-1)/K$ gives failure at most $\exp[-\epsilon(t-1)/(8K)]$; union over a, t and the condition $\epsilon(t-1) \geq 8K\ell$ prove the first assertion. No distributional claim is made conditional on being active.

All actions have both the count bound and $n \geq n_0$ by the deterministic calendar threshold

$$t_{\mathrm{cov}} = \min\left\{T+1, 1+\left\lceil \frac{2K(n_0+4\ell)}{\epsilon} \right\rceil\right\}. \tag{60}$$

Before this time charge at most $2h$ per active round. Afterwards (47) and (58) bound the recommended action's statistical width by $A_{\mathrm{FE}}\sqrt{2K/[\epsilon(t-1)]}$ and its corruption envelope by $16K\Gamma/[\kappa_{\mathrm{FE}}\epsilon(t-1)]$. The interface contains this recommendation width with coefficient two. Summing $t^{-1/2}$ and t^{-1} gives the conservative bound (59). If the threshold exceeds T, the displayed burn-in term already bounds the whole horizon. For shared FE use elapsed post-initialization time in the count argument. Initialization samples can only increase that count, and Lemma 5 applies to the combined retained prefix; no comparison of two different fitted estimators is used. □

Lemma 8 (On-policy played-action widths)**.** *For the concrete on-policy module with the stated calibration, every* $\Gamma \geq 0$ *satisfies*

$$\sum_{t=1}^{T} I_t \beta_{a_t,t}^{Z,\Gamma} \leq C\left\{hKn_0 + A_{\mathrm{op}}\sqrt{KT} + \frac{K\Gamma H_T}{\kappa_{\mathrm{op}}}\right\}.$$

Proof. Null actions have zero radius. For a fixed non-null action, its successive plays have pre-round retained counts $0, 1, \ldots, N_a(T+1)-1$. Apply the envelope (47). The first n_0 counts cost at most hn_0. Subsequently, $\sum_{n=1}^{N} n^{-1/2} \leq 2\sqrt{N}$ and $\sum_{n=1}^{N} n^{-1} \leq H_T$. Summing over actions and using $\sum_a \sqrt{N_a(T+1)} \leq \sqrt{KT}$ proves the claim. The RSC and noise events justify the confidence radii; no comparison with an LP-policy sample count is needed. □

Proposition 4 (Known and shared on-policy cumulative interfaces)**.** *For known-corruption OP, the recommendation equals the played action, and the cumulative interface holds with*

$$\mathcal{R}_{\mathrm{est}}^{0} \leq C\{hKn_0 + A_{\mathrm{op}}\sqrt{KT}\}, \qquad \psi_{\mathrm{est}}(\Gamma) \leq \frac{CK\Gamma H_T}{\kappa_{\mathrm{op}}}. \tag{61}$$

For shared OP, on the coverage event of Condition 5, uniformly over candidates the post-initialization cumulative interface holds with

$$\begin{aligned} \mathcal{R}_{\mathrm{sh}}^{0} &\leq C\{hK(n_{\mathrm{cand}} + \chi n_0) + \chi A_{\mathrm{op}}\sqrt{KT}\}, \\ \psi_{\mathrm{est}}(\Gamma) &\leq \frac{CK\Gamma}{\kappa_{\mathrm{op}}}\{\chi H_T + n_{\mathrm{cand}}\}. \end{aligned} \tag{62}$$

The noise and retained-design events additionally ensure that these radii provide the pointwise confidence interface.

Proof. The known case follows from twice Lemma 8, with constant factors absorbed into C.

For the shared case fix a, Γ, and let A_n count active post-initialization recommendations of a made when its pre-round realized count is n. At the time immediately after the last such recommendation with count at most n, Condition 5 gives

$$\sum_{j=0}^{n} A_j \leq n_{\mathrm{cand}} + \chi(n+1),$$

because that recommendation round can increase the realized count by at most one. Apply (57) to the nonincreasing clean-width envelope f. Its initial value is at most h and

$$\sum_{n=0}^{N} f(n) \leq h(n_0+1) + 2A_{\mathrm{op}}\sqrt{N}.$$

Sum through the final, possibly unplayed count level for each action, then use $\sum_a \sqrt{N_a(T+1)} \le \sqrt{KT}$ and $n_0 \ge 1$. This yields the stated clean-width bound after absorbing universal constants.

For the corruption part, dominate $\mathbf{1}\{n \ge n_0\}/n$ by the nonincreasing sequence $1/(n \vee 1)$, whose partial sum is at most $2H_T$. The same summation inequality gives a charge of order $n_{\mathrm{cand}} + \chi H_T$ per action. Multiplying by the corruption coefficient in (47), summing over actions, and absorbing universal constants proves the second line. The count event holds simultaneously for all candidates, so the conclusion is uniform on the shared history. □

E.10 Failure Probabilities and Concrete Substitution

For known FE combine the coordinate-noise event, the forced-design event, and the forced-count event; their failure is at most $\delta_{\mathrm{est}} + \delta_{\mathrm{cov}}$. For known OP combine noise with Condition 4. For shared OP add the candidate event. None of these events is obtained by conditioning the clean martingale increments on successful estimation. Instead, bound the score gap pathwise on the good event and by h on its complement. This produces the explicit $h\delta_{\mathrm{tot}}T$ expected-regret contribution. Mean-zero clean-feedback terms are removed using the original filtration before this event split, as in Appendix F.

Equations (59), (61), and (62) provide every input to the modular allocation theorems. They include the chosen solver's numerical error through A, retain the actual curvature and tolerance dependence, and have no corruption-dependent burn-in. If κ is small or the calibrated threshold exceeds available data, the bound can be larger than T; the unconditional range bound $\mathrm{Reg}(T) \le T$ still applies. Neither convexity nor an unsuccessful calibration implies sublinear regret.

F Proof of the Known-Corruption Guarantee

This appendix expands the main-text proof overview into five blocks: score confidence and comparison with the population LP; cumulative widths and overrides; dual control and insertion of the benchmark; the relation between clean and observed resource use and stopping cancellation; and final assembly with concrete substitutions. We use $B^{\mathrm{op}} = B$, predictable estimates satisfying the main-text interface, clipped observed feedback, and the clean conditional-mean score $F^0(a, x_t; \lambda_t)$. The estimation objective enters only in the concrete substitutions at the end.

Let I_t indicate that the pre-action safety condition holds at the start of round t. Active rounds form a prefix because the algorithm plays `null` and freezes its state after the first failure. Let $\tau := \sum_{t=1}^{T} I_t$. For the null action, set $\beta^{Z,\Gamma}_{\mathtt{null},t} = 0$, $\widehat{F}_{\mathtt{null},t}(\lambda) = 0$, and $F^0(\mathtt{null}, x; \lambda) = 0$.

Table 6: Notation used in the known-corruption proof.

Symbol	Meaning
I_t	Indicator that round t is active under the pre-action safety rule.
τ	Number of active rounds, $\tau = \sum_t I_t$.
R_D	Dual regret against the best fixed resource-price vector in Λ.
$\mathcal{E}_{\mathrm{conf}}$	Common event supplying the pointwise confidence interface.
$\mathcal{R}^0_{\mathrm{est}}$	Clean cumulative recommendation-width bound supplied by the estimator interface.
$\psi_{\mathrm{est}}(\Gamma)$	Additional cumulative width caused by retained response corruption.

F.1 Score Confidence and Population-LP Comparison

Lemma 9 (Score confidence). *On the event $\mathcal{E}_{\mathrm{conf}}$, suppose the reward and consumption estimates satisfy the componentwise interface (10). Then every active round t, action $a \in \mathcal{A}_0$, and price vector $\lambda_t \in \Lambda$ satisfy*

$$\left|\widehat{F}_{a,t}(\lambda_t) - F^0(a, x_t; \lambda_t)\right| \le \beta^{Z,\Gamma}_{a,t}. \tag{63}$$

Proof. For $a = \texttt{null}$, both scores are zero. For $a \in \mathcal{A}$, apply the componentwise confidence interface directly. The reward error is at most $\beta^{r,\Gamma}_{a,t}$, and the priced consumption error is at most

$$Z \sum_{i=1}^{m} \lambda_{t,i} \beta^{b,\Gamma}_{a,i,t} \le Z \|\lambda_t\|_1 \max_i \beta^{b,\Gamma}_{a,i,t} \le Z \max_i \beta^{b,\Gamma}_{a,i,t}.$$

Adding the two terms gives $\beta^{Z,\Gamma}_{a,t}$, as stated. □

F.1.1 Optimistic comparison with the population-LP policy

Lemma 10 (Optimistic comparison)**.** *On $\mathcal{E}_{\mathrm{conf}}$, every active round satisfies*

$$F^0(\widehat{a}_t, x_t; \lambda_t) \ge \sum_{a \in \mathcal{A}} y^\star(a \mid x_t) F^0(a, x_t; \lambda_t) - 2\beta^{Z,\Gamma}_{\widehat{a}_t,t}. \tag{64}$$

If e_t is the forced-exploration indicator in Algorithm 3, the played action satisfies

$$F^0(a_t, x_t; \lambda_t) \ge \sum_{a \in \mathcal{A}} y^\star(a \mid x_t) F^0(a, x_t; \lambda_t) - 2\beta^{Z,\Gamma}_{\widehat{a}_t,t} - (1+Z)e_t. \tag{65}$$

Proof. Score confidence and maximization imply, for every $a \in \mathcal{A}_0$,

$$F^0(a, x_t; \lambda_t) \le U^\Gamma_{a,t} \le U^\Gamma_{\widehat{a}_t,t} \le F^0(\widehat{a}_t, x_t; \lambda_t) + 2\beta^{Z,\Gamma}_{\widehat{a}_t,t}.$$

The last expression is nonnegative because it is at least $U^\Gamma_{\widehat{a}_t,t} \ge U^\Gamma_{\texttt{null},t} = 0$. Extend $y^\star(\cdot \mid x_t)$ to a distribution on $\mathcal{A}_0$ by assigning its remaining mass to $\texttt{null}$. Multiplying the action-wise inequality by this distribution and summing proves (64), since the null clean score is zero.

A forced action can reduce the clean score by at most $1 + Z$, because all clean scores lie in $[-Z, 1]$; when $e_t = 0$, the played action equals the recommendation. This proves (65) pathwise. □

F.2 Cumulative Widths and Overrides

Define the cumulative clean-score gap on active rounds

$$\mathcal{E}_{\mathrm{opt}} := \sum_{t=1}^{T} I_t \left[\sum_{a \in \mathcal{A}} y^\star(a \mid x_t) F^0(a, x_t; \lambda_t) - F^0(a_t, x_t; \lambda_t) \right]. \tag{66}$$

Lemma 11 (Cumulative clean-score gap)**.** *Under the cumulative estimator interface, with joint estimator and coverage failure probability at most δ_{tot},*

$$\mathbb{E}[\mathcal{E}_{\mathrm{opt}}] \le \mathcal{R}^0_{\mathrm{est}} + \psi_{\mathrm{est}}(\Gamma) + (1+Z) \sum_{t=1}^{T} \epsilon_t + (1+Z)\delta_{\mathrm{tot}} T. \tag{67}$$

Here $\mathcal{R}^0_{\mathrm{est}}$ is the clean cumulative-width term, while $\psi_{\mathrm{est}}(\Gamma)$ is the extra width caused by corrupted retained samples.

Proof. Sum (65) over active rounds and apply the cumulative estimator interface (12) directly. It bounds the width sum by $\mathcal{R}^0_{\mathrm{est}} + \psi_{\mathrm{est}}(\Gamma)$; no estimator-specific support property is needed here. For the concrete module, (47) charges pre-certification radii to clean burn-in. Lemma 7 in FE and Proposition 4 in OP provide the concrete interface bounds. Finally, I_t is fixed before the fresh exploration coin is drawn, so

$$\mathbb{E}\left[\sum_{t=1}^{T} I_t e_t \right] \le \sum_{t=1}^{T} \epsilon_t.$$

On the complement of the joint estimator–coverage event, each clean Lagrangian comparison is at most $1 + Z$. Its contribution is at most $(1+Z)\delta_{\mathrm{tot}} T$. The clean-to-observed resource discrepancy enters later through Lemma 14, separate from the estimator term. □

F.3 Dual Control and Insertion of the LP Benchmark

On an active round, let

$$g_t := b_t(a_t) - \frac{B}{T}, \qquad \widetilde{g}_t := (g_t, 0) \in \mathbb{R}^{m+1}.$$

Define

$$R_D := \max_{\lambda \in \Lambda} \sum_{t=1}^{T} I_t \lambda^\top g_t - \sum_{t=1}^{T} I_t \lambda_t^\top g_t. \tag{68}$$

Lemma 12 (Resource-price update bound)**.** *The multiplicative-weights update in Algorithm 3 satisfies*

$$R_D \le \frac{\log(m+1)}{\eta_q} + \eta_q \left(1 + \frac{B_{\max}}{T}\right)^2 T. \tag{69}$$

So $ZR_D \le \mathcal{D}_T$ for the definition in (17).

Proof. The first m coordinates of q_t equal λ_t, and the last coordinate is the unused mass $1 - \|\lambda_t\|_1$. Every $\lambda \in \Lambda$ corresponds to $q = (\lambda, 1 - \|\lambda\|_1)$ in the $(m+1)$-simplex. Also,

$$\|\widetilde{g}_t\|_\infty \le 1 + \frac{B_{\max}}{T}.$$

Hoeffding's lemma for the finite distribution q_t bounds its log-moment-generating function by its mean plus $\eta_q^2 \|\widetilde{g}_t\|_\infty^2/2$. Summing the log-potential increments yields, for every such q and any $\eta_q > 0$,

$$\sum_{t=1}^{T} I_t (q - q_t)^\top \widetilde{g}_t \le \frac{\mathrm{KL}(q\|q_1)}{\eta_q} + \eta_q \sum_{t=1}^{T} I_t \|\widetilde{g}_t\|_\infty^2.$$

The algorithm freezes q_t after stopping, so this is exactly the update on the active prefix. Since q_1 is uniform, $\mathrm{KL}(q\|q_1) \le \log(m+1)$. Taking the maximum over λ proves (69). □

F.3.1 Insertion of the population-LP benchmark

Lemma 13 (Population-LP conversion)**.** *The cumulative clean-score gap on active rounds satisfies*

$$\begin{aligned} \mathrm{Reg}(T) \le \mathbb{E}[\mathcal{E}_{\mathrm{opt}}] + V^{\mathrm{UB}}\, \mathbb{E}\Big[1 - \frac{\tau}{T}\Big] \\ + Z\mathbb{E}\left[\sum_{t=1}^{T} I_t \lambda_t^\top \left(\frac{B}{T} - b^0(a_t, x_t)\right)\right]. \end{aligned} \tag{70}$$

Proof. Because active rounds form a prefix, I_t is measurable before x_t is drawn. Optimality of $y^\star$ gives

$$V^{\mathrm{UB}} = \mathbb{E}\left[\sum_{t=1}^{T} I_t \sum_{a \in \mathcal{A}} y^\star(a \mid x_t) r^0(a, x_t)\right] + V^{\mathrm{UB}}\, \mathbb{E}\Big[1 - \frac{\tau}{T}\Big]. \tag{71}$$

Expanding (66), using (71), and then applying population-LP feasibility yield (70). The key measurability point is that $I_t \lambda_t$ is predictable before x_t. For every resource coordinate,

$$\mathbb{E}_x\left[\sum_a y^\star(a \mid x) b_i^0(a, x)\right] \le \frac{B_i}{T}$$

can be multiplied by $I_t \lambda_{t,i}$ and averaged without changing its direction. This step is where population-LP feasibility enters the proof. □

F.4 Relating Resource Use and Controlling Stopping

The population LP uses clean conditional-mean consumption, while the algorithm updates prices and stops using observed consumption. This step connects the two resource processes.

Lemma 14 (Relating clean and observed resource use)**.** *The cumulative clean-score gap on active rounds satisfies*

$$\begin{aligned}\mathrm{Reg}(T) \le \mathbb{E}[\mathcal{E}_{\mathrm{opt}}] + V^{\mathrm{UB}}\,\mathbb{E}\Big[1-\frac{\tau}{T}\Big] \\ + Z\mathbb{E}\left[\sum_{t=1}^{T} I_t\lambda_t^\top\left(\frac{B}{T}-b_t(a_t)\right)\right] + \mathbb{E}[C_Z].\end{aligned} \tag{72}$$

Proof. Because $I_t\lambda_t$ is predictable before x_t and a_t is selected before the current feedback using randomness independent of the clean potential outcomes, the conditional mean-zero clean noise has zero expectation. Together with $b_t(a_t) = b_t^{\mathrm{cl}}(a_t) + c_t^b(a_t)$, this implies

$$\begin{aligned}&Z\mathbb{E}\left[\sum_{t=1}^{T} I_t\lambda_t^\top\left(b_t(a_t)-b^0(a_t,x_t)\right)\right] \\ &\quad= Z\mathbb{E}\left[\sum_{t=1}^{T} I_t\lambda_t^\top c_t^b(a_t)\right] \le Z\mathbb{E}\left[\sum_{t=1}^{T} I_t\|c_t^b(a_t)\|_\infty\right] \le \mathbb{E}[C_Z].\end{aligned} \tag{73}$$

Equivalently, pathwise the corruption part is controlled by

$$Z\sum_{t=1}^{T} I_t\left|\lambda_t^\top\left(b_t(a_t)-b_t^{\mathrm{cl}}(a_t)\right)\right| \le C_Z.$$

Substituting the expected inequality into (70) proves (72). This produces one direct C_Z resource-accounting term, separate from the estimator widths. □

F.4.1 Stopping cancellation

Lemma 15 (Stopping cancellation and expected BwK conversion)**.** *Under Assumption 2,*

$$\mathrm{Reg}(T) \le \mathbb{E}[\mathcal{E}_{\mathrm{opt}}] + Z\mathbb{E}[R_D] + \mathbb{E}[C_Z] + Z. \tag{74}$$

Proof. Let

$$G := \sum_{t=1}^{T} I_t\left(b_t(a_t)-\frac{B}{T}\right).$$

The definition of R_D keeps the following maximum:

$$Z\sum_{t=1}^{T} I_t\lambda_t^\top\left(\frac{B}{T}-b_t(a_t)\right) \le ZR_D - Z\max_{\lambda\in\Lambda}\lambda^\top G. \tag{75}$$

The maximum selects the most depleted resource direction and supplies the charge used for the benchmark value after stopping.

If $\tau = T$, there is no inactive benchmark value and the nonnegative maximum may be dropped. If $\tau < T$, then the pre-action test first fails after τ active rounds. For some resource i,

$$U_{\tau+1,i} = \sum_{t=1}^{T} I_t b_{t,i}(a_t) > B_i - 1,$$

and so

$$G_i = U_{\tau+1,i} - \frac{\tau B_i}{T} > B_i\left(1-\frac{\tau}{T}\right) - 1. \tag{76}$$

Since both 0 and the coordinate vector e_i belong to Λ,

$$\max_{\lambda\in\Lambda}\lambda^\top G \ge \max\{0, G_i\}.$$

Assumption 2 gives $V^{\mathrm{UB}} \le ZB_{\min} \le ZB_i$. If $G_i \ge 0$, then (76) implies

$$V^{\mathrm{UB}}\left(1-\frac{\tau}{T}\right) - Z\max_{\lambda\in\Lambda}\lambda^\top G < Z.$$

If $G_i < 0$, the same display follows because (76) implies $B_i(1-\tau/T) < 1$, while the maximum is nonnegative. In both cases,

$$V^{\mathrm{UB}}\left(1-\frac{\tau}{T}\right) - Z\max_{\lambda\in\Lambda}\lambda^\top G \le Z. \tag{77}$$

Taking expectations in the two pathwise inequalities and combining (72), (75), and (77) proves the stopping-aware conversion. The remaining Z is the stopping boundary. The pre-action rule already gives exact observed budget feasibility. □

F.5 Assembly and Concrete Substitutions

Proof of Theorem 1. Lemma 11 gives

$$\mathbb{E}[\mathcal{E}_{\mathrm{opt}}] \le \mathcal{R}^0_{\mathrm{est}} + \psi_{\mathrm{est}}(\Gamma) + (1+Z)\sum_{t=1}^{T}\epsilon_t + (1+Z)\delta_{\mathrm{tot}}T.$$

Lemma 12 gives $Z\mathbb{E}[R_D] \le \mathcal{D}_T$, and Lemma 15 gives the direct resource transfer and the stopping boundary. Thus, the assembled bound is

$$\boxed{\begin{aligned}\mathrm{Reg}(T) \le O\Bigg(&\mathcal{R}^0_{\mathrm{est}} + (1+Z)\sum_{t=1}^{T}\epsilon_t + \mathcal{D}_T + \mathbb{E}[C_Z] + \psi_{\mathrm{est}}(\Gamma) + Z\Bigg)\\ &+ (1+Z)\delta_{\mathrm{tot}}T.\end{aligned}} \tag{78}$$

Finally, the supplied bound is pathwise, $C_Z \le \Gamma$, and therefore $\mathbb{E}[C_Z] \le \Gamma$. Replacing the explicit $\mathbb{E}[C_Z]$ in (78) by Γ proves (19). The final display keeps the direct transfer Γ separate from the estimator term $\psi_{\mathrm{est}}(\Gamma)$.

Lemma 7 gives (59). Substituting it into the assembled bound proves (32) and Corollary 3; the term $h\epsilon T$ is already present in that bound. The exploration balance in Appendix C, including its range-bound argument for the capped case, then gives (20) and Corollary 1(i).

For OP, the recommendation is the actual played action and the regression count is $N_a(t)$. Substituting (61) from Proposition 4 proves (21) and Corollary 4. These substitutions retain the explicit A, n_0, curvature, and harmonic factors. Setting $\Gamma = 0$ proves Corollary 5. □

If the algorithm instead uses a strictly smaller operating budget, the same five proof blocks apply after replacing B by B^{op}, using a scale valid for the reduced-budget benchmark, and measuring regret first against $V^{\mathrm{UB}}(B^{\mathrm{op}})$. Returning to the original benchmark adds exactly $V^{\mathrm{UB}}(B) - V^{\mathrm{UB}}(B^{\mathrm{op}})$.

G Proof of the Unknown-Corruption Guarantee

The proof compares recommendations computed from one common realized estimator history. It does not compare counterfactual histories that different candidates would have generated. Direct corruption enters twice when transferring the master comparison to clean scores, and once in the resource conversion. The cumulative confidence term is separate.

G.1 One Common Confidence and Width Event

Lemma 16 (Shared estimates and all valid grid points)**.** *For the concrete module of Appendices E.2–E.3, under the FE calibration or the OP trajectory condition, one event of failure at most $\delta_{\mathrm{est}} + \delta_{\mathrm{cov}}$ gives pointwise confidence for every valid $\Gamma \in \mathcal{G}$. On this event the FE grid-uniform width interface holds with (59). For OP, intersect also with Condition 5's event; then (62) supplies that interface, with total failure at most $\delta_{\mathrm{est}} + \delta_{\mathrm{cov}} + \delta_{\mathrm{cand}}$.*

Proof. The estimator sample set (34) is exactly the initialization-augmented FE or all-realized OP set in (39). Lemma 3 covers its obtained prefixes, including stopping. In FE, Lemma 5 applies directly to the combined initialization and later forced samples, and Lemma 7 supplies the count event. In OP, the retained design is precisely the design in Condition 4. For every response channel, the data, ρ_n, and certified objective-gap tolerance are common to all candidates. Lemma 4 is deterministic for any corruption vector on the noise event. Therefore Lemma 2 covers all valid candidates simultaneously, without a candidate-dependent noise theorem or a candidate-dependent fit. The FE width bound holds for arbitrary recommendations. The OP width bound uses Proposition 4 and the additional candidate event. Since $C_Z \le T(1+Z)$, the grid contains a valid point and the same event permits the post hoc substitution of $\Gamma^\star$. □

Thus, on the appropriate common event,

$$2\sum_{t=t_{\mathrm{init}}+1}^{T} I_t \beta^{Z,\Gamma^\star}_{\widehat{a}_t^{\Gamma^\star},t} \le \mathcal{R}^0_{\mathrm{sh}} + \psi_{\mathrm{est}}(\Gamma^\star). \tag{79}$$

Initialization is charged separately, so $\mathcal{R}^0_{\mathrm{sh}}$ is a post-initialization width bound. The parameters of the OP event concern the algorithm's actual shared trajectory, not the trajectory of a candidate run alone.

G.2 Master Comparison with a Uniform-Advice Expert

Let $\mathcal{Q}_t$ include the full analysis history, current context, fitted coefficients, prices, and all recommendations before fresh round-t draws. Let $\mathcal{P}_t$ also include the potential observed payoff vector. Non-anticipation fixes this vector before the independent action seed, so, conditional on $\mathcal{P}_t$ and a non-forced branch, $a_t \sim p_t$.

Lemma 17 (Unbiased expert-payoff estimate)**.** *On each post-initialization, non-forced active round, every expert $j \in \mathcal{E}_{\mathrm{M}}$ obeys*

$$\mathbb{E}[\widehat{y}_{t,j} \mid \mathcal{P}_t, e_t = 0] = \sum_{a \in \mathcal{A}_0} \xi_{t,j}(a) Y_t(a).$$

For a grid candidate this equals $Y_t(\widehat{a}_t^{\Gamma})$; for the uniform expert it is the uniform average.

Proof. All $p_t(a) \ge p_{\min} > 0$. Conditional expectation of $\widehat{Y}_t(a) = \mathbf{1}\{a_t = a\}Y_t(a_t)/p_t(a_t)$ is $Y_t(a)$; summing against the advice gives the result. The denominator is the actual non-forced sampling probability, not the mixture probability including forced overrides. □

Lemma 18 (Master score regret with overrides)**.** *For the parameters (35), let $L_{\mathrm{M}} = \log(3M/\delta_{\mathrm{master}})$. With failure at most $2\delta_{\mathrm{master}}/3$, simultaneously for all $\Gamma \in \mathcal{G}$,*

$$\sum_{t=t_{\mathrm{init}}+1}^{T} I_t[F_t^{\mathrm{obs}}(\widehat{a}_t^{\Gamma};\lambda_t) - F_t^{\mathrm{obs}}(a_t;\lambda_t)] \le 8h\sqrt{TK_0L_{\mathrm{M}}} + h\sum_{t=1}^{T}\epsilon_t. \tag{80}$$

Proof. Suppose first $T \ge 4K_0 \log M$ and $L_{\mathrm{M}} \le K_0 T$. Then the cap on $p_{\min}$ is inactive and (35) matches EXP4.P with confidence $\delta_{\mathrm{master}}/3$. The expert class includes an expert whose advice is uniform on every round, as required by Beygelzimer et al. (2011, Theorem 2 and Algorithm 1). The explicit uniform-advice expert is not replaced by action smoothing. Payoffs are in $[0,1]$, advice and potential payoffs are determined before the fresh master draw, and Lemma 17 verifies importance weighting.

Reindex post-initialization, non-forced active rounds. Their number is at most T; pad this sequence after its final round with zero-payoff rounds, retaining the uniform expert and continuing the same master update. Padding adds zero payoff difference for every expert and does not alter earlier decisions. This is a length-T non-anticipating expert-advice problem, even though the original subsequence and stopping time depend on the preceding history. The cited theorem gives a normalized regret at most $6\sqrt{TK_0L_{\mathrm{M}}}$ with failure at most $\delta_{\mathrm{master}}/3$.

Forced rounds leave the master weights unchanged and cost at most h each. Since I_t is predictable and the forced coin is fresh,

$$\sum_{t>t_{\text{init}}} I_t e_t \le \sum_{t=1}^{T} \epsilon_t + \sqrt{2T\log(3/\delta_{\text{master}})}$$

except with probability $\delta_{\text{master}}/3$, by the exponential bound for bounded martingale differences. Multiplication by h and $K_0 \ge 2$ give (80). If either horizon condition fails, the total observed-score regret is always at most hT. When $T < 4K_0 \log M$ this is at most $2h\sqrt{TK_0L_{\text{M}}}$; when $L_{\text{M}} > K_0 T$ it is at most $h\sqrt{TK_0L_{\text{M}}}$. Thus the same bound holds in the capped or trivial regime without invoking the nontrivial-horizon theorem. □

Lemma 19 (Uniform clean-score transfer)**.** *With failure at most $\delta_{\text{master}}/3$, simultaneously for every fixed $\Gamma \in \mathcal{G}$,*

$$\begin{aligned}&\sum_{t>t_{\text{init}}} I_t[F^0(\widehat{a}_t^{\Gamma}, x_t; \lambda_t) - F^0(a_t, x_t; \lambda_t)] \\ &\quad\le \sum_{t>t_{\text{init}}} I_t[F_t^{\text{obs}}(\widehat{a}_t^{\Gamma}; \lambda_t) - F_t^{\text{obs}}(a_t; \lambda_t)] + 2C_Z + 2h\sqrt{2T\log(3|\mathcal{G}|/\delta_{\text{master}})}.\end{aligned} \tag{81}$$

Proof. For each non-null action,

$$F_t^{\text{obs}}(a; \lambda_t) - F^0(a, x_t; \lambda_t) = \eta_t^r(a) - Z\lambda_t^\top \eta_t^b(a) + c_t^r(a) - Z\lambda_t^\top c_t^b(a).$$

Each of the two action sequences contributes at most C_Z in absolute corruption. For a fixed grid index, the recommendation is pre-feedback measurable and the played action uses an independent fresh seed. Hence the difference of their clean-noise terms is conditionally mean-zero, with absolute value at most $2h$. This is a martingale-difference argument under the original history and current context, not under conditioning on the potential observed payoff vector or a successful confidence event. Azuma's inequality and a union bound over the fixed grid give the displayed remainder. Only after this uniform event is established do we substitute the path-dependent $\Gamma^\star$. □

The preceding two lemmas together imply a clean candidate-to-played comparison bounded by $\text{Reg}_{\text{mst}}(T) + h\sum_t \epsilon_t + 2C_Z$, using the stated choice $\text{Reg}_{\text{mst}}(T) = 20h\sqrt{TK_0\log(3M/\delta_{\text{master}})}$.

G.3 Resource Conversion, Stopping, and Assembly

Proof of Theorem 2. Define

$$\mathcal{E}_{\text{opt}}^{\text{sh}} = \sum_{t=1}^{T} I_t \left[\sum_a y^\star(a \mid x_t) F^0(a, x_t; \lambda_t) - F^0(a_t, x_t; \lambda_t)\right].$$

Split this sum into the initialization prefix, the subsequent population-LP-to-$\Gamma^\star$ recommendation comparison, and the subsequent $\Gamma^\star$-to-played-action comparison. Initialization costs at most ht_{init}. On the joint confidence and coverage event, Lemma 10 and (79) bound the second part by $\mathcal{R}_{\text{sh}}^0 + \psi_{\text{est}}(\Gamma^\star)$. The master and clean-transfer lemmas bound the third by $\text{Reg}_{\text{mst}}(T) + h\sum_t \epsilon_t + 2C_Z$.

For FE the joint event comprises noise, forced design/counts, and the master/transfer events. For OP it instead uses the realized-design event together with candidate-recommendation coverage. Their failure probabilities sum to the corresponding δ_{tot} in (27). The per-round clean-score gap is always at most h, including on the complement. All terms on the good-event upper bound are nonnegative, so taking expectations gives

$$\mathbb{E}[\mathcal{E}_{\text{opt}}^{\text{sh}}] \le \mathcal{R}_{\text{sh}}^0 + \text{Reg}_{\text{mst}}(T) + ht_{\text{init}} + h\sum_t \epsilon_t + \mathbb{E}[\psi_{\text{est}}(\Gamma^\star) + 2C_Z] + h\delta_{\text{tot}}T. \tag{82}$$

The method uses exactly the known algorithm's actual observed-consumption price update and pre-context stopping test. Lemmas 13, 14, and 15 therefore apply to this played sequence without any estimator change:

$$\text{Reg}(T) \le \mathbb{E}[\mathcal{E}_{\text{opt}}^{\text{sh}}] + Z\mathbb{E}[R_D] + \mathbb{E}[C_Z] + Z. \tag{83}$$

Here $Z\mathbb{E}[R_D] \le \mathcal{D}_T$ by Lemma 12. The additional C_Z is the clean-to-observed resource transfer, separate from the two master score transfers. Thus their total is $3\mathbb{E}[C_Z] \le 3\mathbb{E}[\Gamma^\star]$. This proves the expected-regret claim. Proposition 2 gives the pathwise observed and clean resource claims, independently of all confidence and coverage events. □

Proofs of Corollaries 2, 6, and 7. For FE substitute (59) into (82) and then (83). The combined initialization and forced retained history has the RSC calibration proved in Lemma 5; subsequent forced counts provide a lower bound for its sample size. This is a bound on counts and radii, not an assumed ordering of estimators fitted on different samples. Add ht_{init}, $h\epsilon T$, the master, dual, and stopping costs. This yields (38). Applying the exploration calculation in Appendix C gives (29); its chosen ϵ is independent of the unknown pathwise corruption. For OP use (62). Its proof keeps recommendation and retained counts separate and explicitly charges $h\chi Kn_0$ before informative recommendation confidence. Substitution, with $\epsilon_t = 0$, yields (30). Retain the full $h\delta_{\text{tot}}T$ contribution in both cases. If $C_Z = 0$, $\Gamma^\star = 0$ removes only the corruption terms; it does not remove initialization, regularization, or the master comparison. □

H Experiments: Robustness, Adaptation, and Resource Accounting

This section gives finite-sample illustrations of the four mechanisms emphasized by the theory: robustness with a supplied corruption bound, adaptation when the corruption level is unknown, the distinct roles of reward and consumption corruption, and the pathwise resource-accounting guarantee. The experiments use the same clean population-LP benchmark as the analysis and evaluate reward with the clean conditional means, so corrupted feedback affects the learner but not the benchmark. We focus on robustness as the corruption level varies rather than fitting an empirical asymptotic exponent; the latter would mix finite-horizon initialization, master, design, and logarithmic terms that are kept explicit in Corollaries 1 and 2.

H.1 Common environment and evaluation protocol

Three-specialist instance. The main experiments use $K = 3$ non-null actions, $m = 2$ resources, dimension $d = 20$, true and calibration sparsity $s_0 = s = 3$, and parameter-radius bound $R = 1.2$. Budgets are $B = (0.5T, 0.5T)$, hence $Z = 2$. Draw q_1, q_2 independently and uniformly from $\{-1, -0.5, 0.5, 1\}$ and $q_3, z_5, \dots, z_{20}$ independently and uniformly from $\{-1, +1\}$, and set

$$x = (0.50,\ 0.40q_1,\ 0.40q_2,\ 0.20q_3,\ 0.15z_5, \dots, 0.15z_{20}). \tag{84}$$

Then $\|x\|_2^2 \le 0.97$. The clean reward means are

$$\begin{aligned} r_1^0(x) &= 0.55 + 0.08q_1 + 0.02q_3, \\ r_2^0(x) &= 0.55 - 0.08q_1 + 0.02q_3, \\ r_3^0(x) &= 0.57 + 0.08q_2 + 0.02q_3, \end{aligned} \tag{85}$$

and the clean mean consumptions are

$$b_1^0 = (0.60, 0.40), \qquad b_2^0 = (0.40, 0.60), \qquad b_3^0 = (0.50, 0.50). \tag{86}$$

Independent reward and consumption noises are uniform on $[-0.02, 0.02]$. The exact finite-support population LP has value $V^{\text{UB}} = 0.6225T$, action masses $(5/16, 5/16, 6/16)$, and resource use $(0.5, 0.5)$; thus all three actions are clean-LP relevant and both resources bind in expectation. The minimum clean reward gap between the best and second-best action over the supported (q_1, q_2) states is 0.02.

Algorithms and practical calibration. All headline policy comparisons use the on-policy mode from Sections 3.3 and 5. We use the norm-constrained Lasso implementation in Section 3.1 with fixed practical coefficients $c_\rho = 0.25$, $c_\beta = 0.75$, $c_C = 1$, unit dual multiplier, start count 20, and certified optimization gap 10^{-8}. Shared-Grid uses the complete dyadic grid, the explicit uniform-advice expert, the EXP4.P specification of Appendix D, and 201 safe cyclic initialization rounds (67 per non-null action). These settings are held fixed across the reported comparisons.

Table 7: Theory-to-experiment map. The experiments are finite-sample illustrations of the mechanisms in the corresponding guarantees; empirical design and master diagnostics do not replace the stated probabilistic coverage conditions.

Experiment	Theory connection	Finite-sample question
Figure 1	Theorem 1, Corollary 1	How absolute regret changes as a valid known corruption level increases.
Figure 2	Theorem 2, Corollary 2	How much additional regret unknown corruption causes after accounting for each method's clean finite-horizon cost.
Figure 3	Joint reward–consumption model and the Reward-Only ablation	How the location of a fixed corruption budget across reward and resource channels changes performance.
Figure 4	Proposition 2 and the pathwise parts of Theorems 1–2	How over- and under-reported consumption affect stopping and clean resource use.

We compare ROPD with a valid supplied bound, a Reward-Only Robust ablation that protects reward confidence but not the consumption channel, and Non-Robust PD with zero corruption allowance. Shared-Grid receives no corruption bound. Attack maps are functions of the potential contexts and are fixed before the learner's fresh action draw; compared methods receive the same contexts, clean potential outcomes, and exogenous attack map within each paired seed.

For a policy π, define the reported percentage Relative Regret by

$$\mathrm{RelReg}(\pi) := 100\,\frac{V^{\mathrm{UB}} - \sum_{t=1}^{T} r^0(a_t, x_t)}{V^{\mathrm{UB}}}, \tag{87}$$

where the null action contributes zero, including after stopping. Each reported cell uses 20 paired replications. Error bars and bands are 95% percentile bootstrap intervals from 2,000 whole-replication resamples; paired differences and corrupted-minus-clean increments are formed within seed before resampling, and rounds are never treated as independent replicates. The final runs used in this section have no excluded seeds, no clipping, and no solver failures.

H.2 Known corruption: performance as C_Z increases

We first study the setting of Theorem 1, where the learner is given a deterministic valid bound $\Gamma \geq C_Z$. Fix $T = 20{,}000$. Corruption targets Action 1 only on contexts where it is uniquely clean-preferred and applies the joint shift

$$(\Delta r_1, \Delta b_{1,1}, \Delta b_{1,2}) = (-0.20, +0.10, +0.10).$$

With $Z = 2$, each attacked opportunity contributes 0.40 units to C_Z. We vary $C_Z \in \{0, 100, 200, 300, 400\}$ by attacking $0, 250, 500, 750, 1000$ spread eligible opportunities.

Figure 1 reports absolute Relative Regret. All three methods coincide at $C_Z = 0$ (mean 0.756%). ROPD has the lowest mean regret at every positive tested corruption level. At $C_Z = 400$, mean Relative Regret is 1.333% for ROPD, 1.383% for Reward-Only Robust, and 2.008% for Non-Robust PD. Thus the separation from the non-robust baseline becomes substantial as corruption grows, while the Reward-Only ablation remains between the full robust method and the non-robust policy throughout the main sweep. This is consistent with the corruption-dependent terms in Corollary 1: the corruption-aware policies trade wider confidence sets for reduced sensitivity to contaminated feedback, and protecting both channels is useful when resource feedback can also be corrupted.



Figure 1: **Known-corruption robustness.** At fixed $T = 20{,}000$, the effective corruption level C_Z increases along the horizontal axis. The vertical axis is absolute Relative Regret against the clean population-LP benchmark. ROPD has the lowest mean regret at every positive tested severity. Bands are the frozen 95% whole-replication bootstrap intervals.

H.3 Unknown corruption: adaptation on one shared history

We next remove the corruption-bound input and evaluate Shared-Grid from Algorithm 2. Because Shared-Grid pays its initialization and master-comparison costs even on clean data, the primary quantity here is the *corruption-induced increase*

$$\Delta \operatorname{RelReg}_M(C_Z) := \operatorname{RelReg}_M(C_Z) - \operatorname{RelReg}_M(0) \tag{88}$$

for method M, formed within paired seeds. This isolates sensitivity to corruption from the method's clean finite-horizon cost. Shared-Grid may still have larger absolute finite-horizon regret because it also pays initialization and master-comparison costs.

Fix $T = 80{,}000$ and vary the unknown corruption budget over $C_Z \in \{0, 1000, 2000, 3000, 4000\}$. The same joint shift $(-0.20, +0.10, +0.10)$ targets clean-preferred Action 1 opportunities; Shared-Grid is never given C_Z. Figure 2 shows that Shared-Grid incurs a smaller corruption-induced increase at every positive severity. At $C_Z = 4000$, the increase is 1.313 percentage points for Shared-Grid versus 2.574 for Non-Robust PD. The paired Shared-minus-Non-Robust differences are -0.262, -0.770, -0.916, and -1.261 percentage points at $C_Z = 1000, 2000, 3000, 4000$, respectively; each pointwise paired 95% interval lies below zero.

These results illustrate the role of the adaptive grid in Theorem 2: even without a supplied corruption level, the shared-data master is markedly less sensitive to increasing corruption on this instance. The result should be read together with the explicit master term in Corollary 2. In absolute terms Shared-Grid still pays a visible clean finite-horizon master cost; the experiment therefore supports adaptation to corruption rather than an absolute dominance claim over the non-robust policy.

H.4 Reward, consumption, and joint corruption

The preceding experiments vary the amount of corruption. We now fix its effective budget and vary *where* it enters the feedback. Set $T = 40{,}000$ and attack the same 1,200 clean-preferred Action 1 opportunities in every corrupted condition. The shifts are

$$\begin{aligned} &\text{Reward:} && (-0.20, 0, 0), \\ &\text{Consumption:} && (0, +0.10, +0.10), \\ &\text{Joint:} && (-0.10, +0.05, +0.05). \end{aligned}$$

Each attacked opportunity contributes 0.20 to C_Z, so all three conditions have exactly $C_Z = 240$, with the same target, attacked contexts, count, and timing within a paired seed.

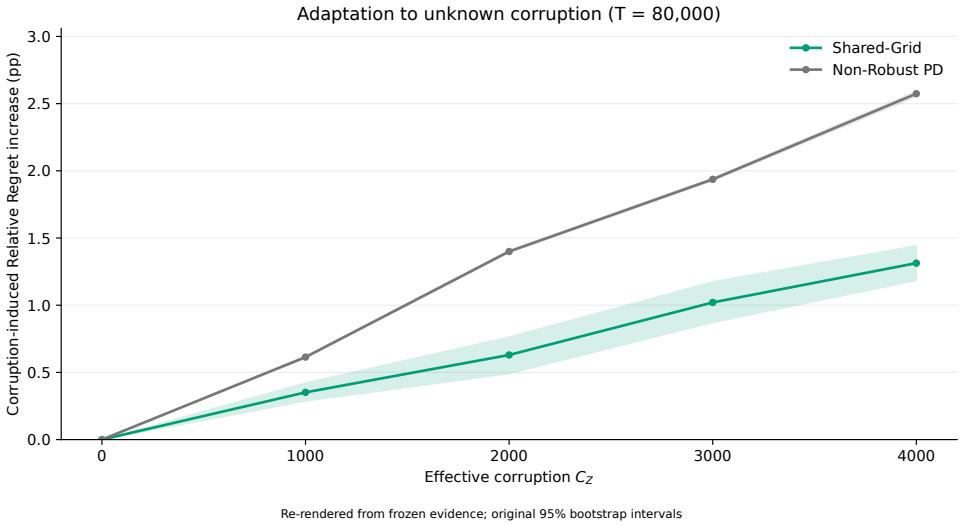


Figure 2: **Unknown-corruption adaptation.** At fixed $T = 80{,}000$, the horizontal axis is the unknown effective corruption level C_Z, and the vertical axis is the paired corruption-induced increase in Relative Regret. Shared-Grid receives no corruption bound and has a smaller increase than Non-Robust PD at every positive tested severity. Bands are 95% paired whole-replication bootstrap intervals.

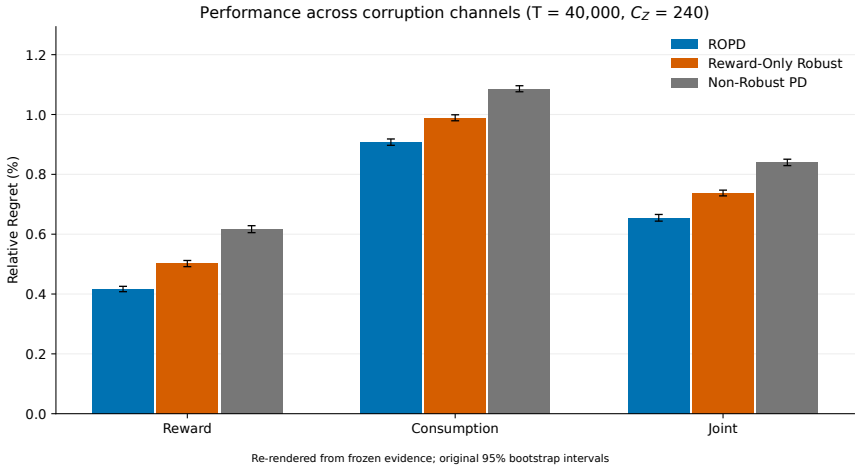


Figure 3: **Corruption channels.** Reward, consumption, and joint attacks use the same 1,200 targeted opportunities and the same effective corruption budget $C_Z = 240$ at $T = 40{,}000$. Bars show absolute Relative Regret. ROPD has the lowest mean regret in all three corrupted environments, followed by Reward-Only Robust and Non-Robust PD. Error bars are the frozen 95% whole-replication bootstrap intervals.

Figure 3 reports absolute Relative Regret under the corrupted environments. ROPD has the lowest mean regret in all three channels: 0.416% under reward corruption, 0.907% under consumption corruption, and 0.655% under joint corruption. Reward-Only Robust attains 0.501%, 0.989%, and 0.738%, respectively, while Non-Robust PD attains 0.617%, 1.086%, and 0.840%. The ordering is consistent across the three equal-C_Z corrupted environments and illustrates why consumption corruption is a separate modeling issue: it changes not only a regression target but also the resource-price and budget-accounting path. ROPD has the lowest absolute finite-horizon Relative Regret in each tested channel; the matched-clean incremental comparison is reported separately below and is more nuanced.

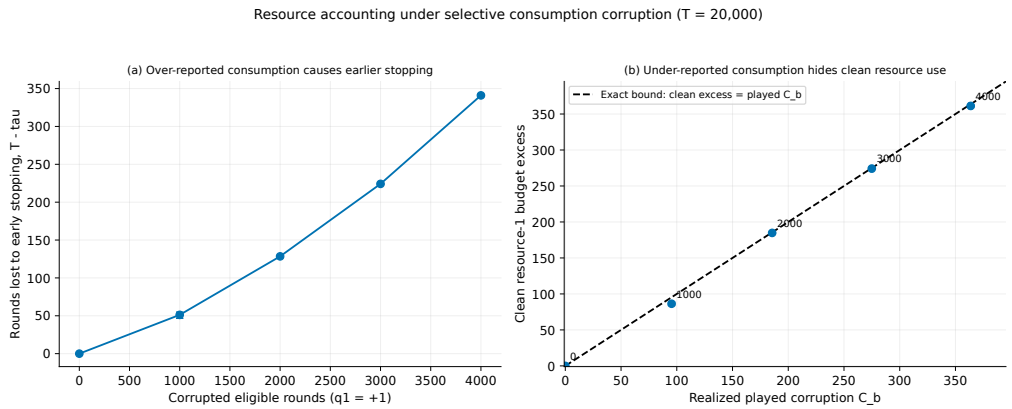


Figure 4: **Resource accounting under consumption corruption.** Over-reported consumption advances stopping (left), while under-reported consumption can hide clean resource use (right). The diagonal in the right panel is the pathwise upper bound $\mathrm{Viol}_{\mathrm{cl}}(T) \le C_b(T)$. Observed-budget violation is zero on every reported trajectory.

H.5 Pathwise resource accounting under corrupted consumption

Finally, we isolate the resource-accounting statement of Proposition 2. This controlled stress test uses the same problem scale ($K = 3$, $m = 2$, $d = 20$, $s = 3$) and perturbs one frequently used action–resource pair by ± 0.10. The goal is not a policy ranking: it is to make the two pathwise effects of consumption corruption visible.

When consumption is over-reported, the observed ledger is depleted too quickly and the safety rule can stop allocation early. Across increasing severities, mean lost rounds rise from zero to 341 at the largest tested perturbation. When consumption is under-reported, the observed ledger remains feasible while clean resource use can exceed the nominal budget. At the largest severity, mean clean resource-1 excess is 361.243 and mean played consumption corruption is $C_b = 363.665$. More importantly, every trajectory satisfies

$$\mathrm{Viol}_{\mathrm{obs}}(T) = 0, \qquad \mathrm{Viol}_{\mathrm{cl}}(T) \le C_b(T) \le C_Z/Z, \tag{89}$$

with no additional additive unit. Thus Figure 4 directly illustrates the distinction between observed feasibility and clean feasibility that motivates the resource component of the theory.

H.6 Additional checks and interpretation

The main experiments above measure finite-horizon performance under known corruption, unknown corruption, different feedback channels, and corrupted resource accounting. We report three additional checks to clarify the interpretation of these results.

Matched clean controls. For the known-corruption algorithms, supplying a positive bound Γ changes the confidence widths even when the observed feedback is clean. We therefore also compare each corrupted run with a clean run using the same supplied bound. This separates the effect of corrupted feedback from the finite-sample effect of the corruption-aware confidence allowance.

For the sweep in Figure 1, both robust methods have smaller matched corruption increments than Non-Robust PD at the two largest tested levels, $C_Z = 300$ and 400. At $C_Z = 400$, the increments are 1.153 percentage points for ROPD, 1.113 for Reward-Only Robust, and 1.252 for Non-Robust PD. The difference between ROPD and Reward-Only is small in this matched comparison, so Figure 1 is interpreted primarily as an absolute finite-horizon performance comparison.

The same control helps interpret Figure 3. With clean feedback and the same supplied $\Gamma = 240$, Relative Regret is 0.353% for ROPD and 0.434% for Reward-Only Robust, compared with 0.546% in the ordinary $\Gamma = 0$ clean setting. Table 8 reports the resulting channel-specific increments.

Table 8: Increase in Relative Regret relative to matched clean controls (percentage points). ROPD and Reward-Only Robust use clean controls with the same supplied $\Gamma = 240$; Non-Robust PD has no corruption-bound input.

Channel	ROPD	Reward-Only Robust	Non-Robust PD
Reward	0.0631	0.0673	0.0709
Consumption	0.5543	0.5551	0.5403
Joint	0.3019	0.3034	0.2939

The matched-control comparison gives a more refined view of the channel effects. In the consumption and joint conditions, the three incremental regrets are numerically close; Non-Robust PD has slightly smaller increments than ROPD by 0.0140 and 0.0080 percentage points, respectively, in the paired comparisons. This does not conflict with the absolute comparison in Figure 3. The robust methods also have lower matched clean regret when supplied with $\Gamma = 240$, and under the corrupted environments themselves ROPD attains the lowest Relative Regret in all three channels. Thus, Figure 3 summarizes finite-horizon performance under corruption, while the matched controls isolate the additional effect of corrupted feedback relative to each method's corresponding clean baseline.

Horizon and design checks. A separate frozen horizon experiment uses $T \in \{40{,}000, 80{,}000, 160{,}000\}$ with corruption of order $T^{2/3}$. The corruption-induced increase for Shared-Grid decreases from 1.612 to 0.920 percentage points over these horizons, while the corresponding increase for Non-Robust PD decreases from 3.530 to 1.851. Clean and corrupted Relative Regret also decrease with T. These results provide a finite-horizon consistency check for the sublinear behavior in Corollaries 1 and 2, rather than an empirical estimate of their asymptotic exponents.

The experiment also has nondegenerate action-specific designs: all three specialist actions have positive clean LP mass, and their population second moments on the true reward supports are strictly positive. For Shared-Grid, candidate recommendations become meaningfully different under corruption; at $C_Z = 4000$, grid candidates disagree on about 61% of active post-initialization rounds. These diagnostics confirm that the finite-sample experiment exercises the learning and adaptation mechanisms studied in the theory.

Summary. The experiments illustrate four distinct parts of the analysis. With a supplied corruption bound, ROPD has lower absolute Relative Regret than the compared baselines throughout the tested corruption sweep. Without such a bound, Shared-Grid incurs substantially less corruption-induced degradation than Non-Robust PD across the tested unknown-corruption levels, while retaining its finite-horizon master cost. Across reward, consumption, and joint corruption, ROPD has the lowest absolute Relative Regret in the tested corrupted environments; the matched-clean incremental comparisons are more nuanced and are interpreted separately above. Finally, the accounting experiment directly illustrates exact observed feasibility and the corruption-controlled clean resource bound.