

This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Electronic Companion to “Semi-Bandit Learning for Monotone Stochastic Optimization”

EC.1. Missing Proofs

Proof of Lemma 3. For any $j \in [n]$ and $t \in [T]$, let B_j^t denote the event that the stated condition *does not* hold for j and t . Using Theorem 3 with $\delta = 2/(nT)^3$, we obtain for any $j \in [n]$, $t \in [T]$, and $m \geq 1$,

$$\Pr(B_j^t \wedge N_j^t = m) \leq \delta.$$

So, for any j and t , we have by union bound,

$$\Pr(B_j^t) \leq \sum_{m=1}^T \Pr(B_j^t \wedge N_j^t = m) \leq \sum_{m=1}^T \delta = \delta T.$$

Again, by union bound,

$$\Pr\left(\bigvee_{j=1}^n \bigvee_{t=1}^T B_j^t\right) \leq \sum_{j,t} \Pr(B_j^t) \leq nT^2\delta \leq \frac{1}{nT}.$$

The last inequality uses $n \geq 2$. This completes the proof. □

Proof of Theorem 8. By the assumption, we have:

$$\begin{aligned} & \left| \int_a^b g dF_P - \int_a^b g dF_Q \right| \\ &= |g(b)(F_P(b) - F_Q(b)) \\ &\quad - g(a)(F_P(a) - F_Q(a)) - \int_a^b (F_P - F_Q) dg| \\ &\leq |g(b)(F_P(b) - F_Q(b))| \\ &\quad + |g(a)(F_P(a) - F_Q(a))| + \left| \int_a^b F_P - F_Q dg \right| \\ &\leq 2g_{\max} \cdot d_{\text{KS}}(P, Q) + \left| \int_a^b F_P - F_Q dg \right|. \end{aligned}$$

The equality follows from the integration by parts formula (Theorem 6).

Using the fact that g is a BV function, we apply the Jordan decomposition (Theorem 5) to write $g = g^+ - g^-$, where g^+ and g^- are non-decreasing functions, and $V_a^b(g) = g^+(b) + g^-(b) - g^+(a) - g^-(a)$. We now have

$$\begin{aligned}
& \left| \int_a^b F_P(x) - F_Q(x) dg(x) \right| \\
&= \left| \int_a^b F_P(x) - F_Q(x) dg^+(x) - \int_a^b F_P(x) - F_Q(x) dg^-(x) \right| \\
&\leq \int_a^b |F_P(x) - F_Q(x)| dg^+(x) + \int_a^b |F_P(x) - F_Q(x)| dg^-(x) \\
&\leq \sup_{x \in \mathbb{R}} |F_P(x) - F_Q(x)| \cdot (g^+(b) - g^+(a)) \\
&\quad + \sup_{x \in \mathbb{R}} |F_P(x) - F_Q(x)| \cdot (g^-(b) - g^-(a)) \\
&= V_a^b(g) \cdot d_{\text{KS}}(P, Q),
\end{aligned}$$

The first inequality is due to the triangle inequality. This establishes the theorem. $\square$

Proof of Theorem 3. We will now show that the distribution $\mathcal{E}$ returned by Algorithm 2 stochastically dominates $\mathcal{D}$ and is also a good approximation in terms of TV-Dist.

$\mathcal{E}$ is a valid distribution. We will show that $\bar{p}_i \geq 0$ for all $i \in [k]$ and $\sum_{i=1}^k \bar{p}_i = 1$. If $\bar{p}_k = 1$ then it is easy to see that $\mathcal{E}$ is valid: all other probabilities $\bar{p}_i = 0$. If $\bar{p}_k < 1$ then $\hat{p}_k < 1 - \epsilon$, which implies $\sum_{i=1}^{k-1} \hat{p}_i = 1 - \hat{p}_k > \epsilon$. So, index $y \in [k-1]$, and

$$\begin{aligned}
\sum_{i=1}^k \bar{p}_i &= \bar{p}_y + \sum_{i=y+1}^{k-1} \bar{p}_i + \bar{p}_k \\
&= \sum_{i=1}^y \hat{p}_i - \epsilon + \sum_{i=y+1}^{k-1} \hat{p}_i + \hat{p}_k + \epsilon = 1.
\end{aligned}$$

Stochastic dominance and TV bound. Instead of using the classical Bernstein inequality, we need to use the “empirical Bernstein” because we only get access to the sample variance. This approach has also been used in Azar et al. (2017) for UCB-VI with Bernstein bonus.

THEOREM EC.1. [Maurer and Pontil (2009)] Let $X_1, X_2, \dots, X_m$ be i.i.d. random variables such that $\mathbf{E}[X_i] = \mu$, and $|X_i| \leq 1$. Then, we have with probability at least $1 - \delta$:

$$\left| \frac{1}{m} \sum_{j=1}^m X_j - \mu \right| \leq \sqrt{\frac{2V_m \log 2/\delta}{m}} + \frac{7 \log 2/\delta}{3m}.$$

where $V_m = \frac{1}{m-1} \sum_{j=1}^m (X_j - \frac{1}{m} \sum_{i=1}^m X_i)^2$ is the sample variance.

For each $i \in [k]$, the empirical probability $\hat{p}_i$ in Algorithm 2 is the average of m independent Bernoulli random variables with mean p_i ; the sample variance of this random variable is $\frac{m}{m-1} \hat{p}_i(1 - \hat{p}_i)$. Theorem EC.1 shows that with probability $1 - \frac{\delta}{k}$ we have:

$$|\hat{p}_i - p_i| < \epsilon_i := \sqrt{\frac{2\hat{p}_i(1 - \hat{p}_i) \log \frac{2k}{\delta}}{m}} + \frac{7 \log \frac{2k}{\delta}}{3m}.$$

Henceforth, we assume the “good” event that $|\hat{p}_i - p_i| < \epsilon_i$ for all $i \in [k]$. This event occurs with probability at least $1 - \delta$, by a union bound argument. Under this event, we will show that $\mathcal{D} \preceq_{\text{SD}} \mathcal{E}$ and $\text{TV-Dist}(\mathcal{E}, \mathcal{D}) \leq \frac{3}{2}\epsilon$, where $\mathcal{E}$ and ϵ are defined in Algorithm 2. This would complete the proof of Theorem 3.

Bounding TV distance between $\mathcal{D}$ and the empirical distribution. Let $\hat{\mathcal{E}}$ denote the distribution with probabilities $\{\hat{p}_i\}_{i=1}^k$, which is just the empirical distribution. We will show that (under the good event),

$$\text{TV-Dist}(\hat{\mathcal{E}}, \mathcal{D}) \leq \frac{1}{2}\epsilon. \quad (\text{EC.1})$$

Indeed, by the definition of total variation distance, we have:

$$\begin{aligned} \text{TV-Dist}(\hat{\mathcal{E}}, \mathcal{D}) &= \frac{1}{2} \sum_{i=1}^k |\hat{p}_i - p_i| \leq \frac{1}{2} \sum_{i=1}^k \epsilon_i \\ &\leq \frac{7k \log 2k/\delta}{6m} + \frac{1}{2} \sum_{i=1}^k \sqrt{\frac{2 \log(2k/\delta)}{m} \hat{p}_i} \\ &\leq \frac{7k \log 2k/\delta}{6m} + \frac{k}{2} \sqrt{\frac{2 \log(2k/\delta)}{m} \frac{1}{k} \sum_{i=1}^k \hat{p}_i} \\ &= \frac{7k \log 2k/\delta}{6m} + \frac{k}{2} \sqrt{\frac{2 \log(2k/\delta)}{km}} \leq \frac{1}{2}\epsilon. \end{aligned}$$

The second inequality uses the definition of ϵ_i and $\hat{p}_i \leq 1$. The third inequality is by Jensen’s inequality for the concave function $\sqrt{x}$. The last equality uses the fact that $\sum_{i=1}^k \hat{p}_i = 1$ and $\hat{p}_i \in [0, 1]$ for all $i \in [k]$. This proves (EC.1).

Bounding TV distance between $\hat{\mathcal{E}}$ and $\mathcal{E}$. We consider two cases:

- If $\bar{p}_k = 1$ then we have $\hat{p}_k \geq 1 - \epsilon$ and $\bar{p}_i = 0$ for $i \in [k-1]$, which implies:

$$\begin{aligned} \text{TV-Dist}(\hat{\mathcal{E}}, \mathcal{E}) &= \frac{1}{2} \cdot \sum_{i=1}^k |\bar{p}_i - \hat{p}_i| \\ &= \frac{1}{2} \left(|1 - \hat{p}_k| + \sum_{i=1}^{k-1} \hat{p}_i \right) \leq \epsilon. \end{aligned}$$

- If $\bar{p}_k < 1$ then $\bar{p}_k - \hat{p}_k = \epsilon$ and using the definition of $\bar{p}$ from (14),

$$\begin{aligned} \text{TV-Dist}(\hat{\mathcal{E}}, \mathcal{E}) &= \frac{1}{2} \cdot \sum_{i=1}^k |\bar{p}_i - \hat{p}_i| \\ &= \frac{1}{2} \left(\epsilon + \left| \sum_{i=1}^y \hat{p}_i - \epsilon - \hat{p}_y \right| + \sum_{i=1}^{y-1} \hat{p}_i \right) = \epsilon, \end{aligned}$$

where the last equality used the choice of index y to conclude $|\sum_{i=1}^y \hat{p}_i - \epsilon - \hat{p}_y| = \epsilon - \sum_{i=1}^{y-1} \hat{p}_i$.

Hence, in either case, we have $\text{TV-Dist}(\hat{\mathcal{E}}, \mathcal{E}) \leq \epsilon$, which combined with (EC.1) implies $\text{TV-Dist}(\mathcal{E}, \mathcal{D}) \leq \text{TV-Dist}(\hat{\mathcal{E}}, \mathcal{E}) + \text{TV-Dist}(\hat{\mathcal{E}}, \mathcal{D}) < \frac{3}{2}\epsilon$.

Showing $\mathcal{D} \preceq_{\text{SD}} \mathcal{E}$. Again, we consider two cases:

- If $\bar{p}_k = 1$ then $\mathcal{E}$ puts all its mass on the highest value a_k : so it is clear that $\mathcal{D} \preceq_{\text{SD}} \mathcal{E}$.
- If $\bar{p}_k < 1$, consider any $v \in \mathbb{R}$. If $v \leq a_y$,

$$\Pr(Z \geq v) = 1 \geq \Pr(X \geq v),$$

$X \sim \mathcal{D}$ and $Z \sim \mathcal{E}$. If $v > a_y$, then let $i \geq y + 1$ be the smallest index such that $a_i \geq v$. We have

$$\begin{aligned} \Pr(Z \geq v) &= \sum_{j=i}^k \bar{p}_j = \sum_{j=i}^{k-1} \hat{p}_j + \hat{p}_k + \epsilon \\ &\geq \sum_{j=i}^k (p_j - \epsilon_j) + \epsilon \geq \Pr(X \geq v), \end{aligned}$$

where the first inequality follows due to the good event. The second inequality uses the same analysis in (EC.1). Hence, $\mathcal{D} \preceq_{\text{SD}} \mathcal{E}$. □

EC.2. Sampling-based Algorithms

A stronger monotonicity condition is that of *strong monotonicity* (Guo et al. 2021), defined next.

DEFINITION EC.1 (STRONG MONOTONICITY). A stochastic problem is *strongly up-monotone* if for any instance $\mathcal{I}$ and probability distributions $\mathbf{D}$ and $\mathbf{E}$ with $\mathbf{D} \preceq_{\text{SD}} \mathbf{E}$, we have $\mathbb{E}_{\mathbf{x} \sim \mathbf{E}}[f(\sigma_{\mathbf{D}}, \mathbf{x})] \leq \text{OPT}_{\mathcal{I}}(\mathbf{D})$, where $\sigma_{\mathbf{D}}$ is the optimal policy for instance $\mathcal{I}$ under distribution $\mathbf{D}$. Strong down-monotonicity is defined similarly.

We note that strong monotonicity implies monotonicity. In prior work, Guo et al. (2021) gave optimal sample complexity bounds for stochastic optimization problems that exhibit the strong monotonicity condition. They also proved this property for problems including Prophet inequality and Pandora’s box. We make use of this property in §5, when we apply our result to these problems.

Below, we assume that $f_{max} = 1$ by scaling.

THEOREM EC.2 (Theorem 17 Guo et al. (2021)). *For any strongly monotone stochastic problem, suppose the number of samples is at least:*

$$C \cdot \frac{n}{\epsilon^2} \log\left(\frac{n}{\epsilon}\right) \log\left(\frac{nT}{\epsilon}\right)$$

where $C > 1$ is a sufficiently large constant. Then, there is an algorithm that gets an ϵ -additive approximation to the optimum with probability at least $1 - \frac{1}{T^2}$.

Below, we discuss what regret bounds can be achieved via this sampling-based approach, in order to compare to our results. In the full-feedback model, one can obtain regret bounds of $\sqrt{nT \log T}$, which is nearly optimal. However, for *semi-bandit* feedback that we consider (and for bandit feedback), such a “reduction” from sample-complexity bounds only provides a sub-optimal $\tilde{O}(T^{2/3})$ regret. Therefore, we need new ideas to get the optimal $\tilde{O}(\sqrt{T})$ regret bound in the semi-bandit model, which we do in this paper.

Full-feedback. Recall that, in the full-feedback model, we get one sample from each r.v. in every time-step. We now describe the algorithm. The algorithm is straightforward: for each $t = 1, 2, \dots, T$, we use estimates from the prior $t - 1$ time steps to obtain a policy to use for the t^{th} time-step.

As a consequence of Theorem EC.2, with probability at least $1 - \frac{1}{T^2}$, we have that $\text{OPT}(\hat{\mathbf{D}}_t)$ is an ϵ_t -additive approximation to $\text{OPT}(\mathbf{D})$ where $\epsilon_t \leq \sqrt{\frac{Cn \log(nT)}{t}}$. (To keep calculation simple, we ignore the $\log \frac{1}{\epsilon}$ dependence in Theorem EC.2; so we are actually assuming a slightly stronger sample-complexity bound.) By union bound, this is true for all $t = 1, \dots, T$ with probability at least $1 - \frac{1}{T}$. We consider this to be a *good* event, G . Under this event, the total regret, say R_T , can be bounded as follows.

$$\mathbb{E}[R_T \mid G] \leq \sum_{t=1}^T \sqrt{\frac{Cn \log(nT)}{t}} = O\left(\sqrt{nT \log(nT)}\right).$$

By law of total expectation we have

$$\mathbb{E}[R_T] \leq \mathbb{E}[R_T \mid G] \cdot \Pr(G) + \mathbb{E}[R_T \mid \bar{G}] \cdot \Pr(\bar{G}) \leq O\left(\sqrt{nT \log(nT)}\right) + T \cdot \frac{1}{T} = O\left(\sqrt{nT \log(nT)}\right).$$

Semi-bandit feedback. A significant challenge in the semi-bandit feedback model arises from our lack of control over the r.v.s from which we get samples. One potential strategy to address this issue is to artificially generate samples for an item i by probing item i first in the algorithm’s policy for that period. However, this approach comes with an inherent drawback – probing item i first may result in a poor policy, and so we

suffer a high regret in such periods. The standard *explore-then-exploit* algorithm that first gets $T^{2/3}$ samples for each r.v., and then plays the optimal policy (for the empirical distribution) for the remaining time steps only achieves regret $\tilde{O}(T^{2/3})$. We are not aware of any generic approach that reduces sample-complexity bounds to regret minimization in the semi-bandit have $o(T^{2/3})$ regret.

Remark 1. The sampling approach for semi-bandit feedback also applies to many problems (e.g., prophet inequality, Pandora’s box and series testing) in the *bandit* feedback model. Basically, for each item i we need a policy where its objective corresponds to the value of r.v. X_i . So, we can directly get $\tilde{O}(T^{2/3})$ regret for these problems even with bandit feedback.

Remark 2. We note that Guo et al. (2021) also give (slightly worse) sample complexity bounds for a broader class of problems that need not satisfy any monotonicity property, but have a finite support-size k . Using the approaches described above, we can convert these sample-complexity guarantees to obtain $O\left(\sqrt{nkT \log(nT)}\right)$ regret under full-feedback and $O\left((nkT)^{2/3}(\log T)^{1/3}\right)$ regret under semi-bandit (and bandit) feedback.

EC.3. Lower Bounds

We now state some regret lower bound results for monotone problems under the *semi-bandit* and *binary* feedback models. We highlight particular applications for which lower bounds are already known, and observe that since these problems can be captured by our framework, the corresponding lower bounds also apply to our setting.

THEOREM EC.3. *Any semi-bandit learning algorithm for stochastic monotone problems has regret at least $\Omega\left(\sqrt{nT}\right)$.*

Proof. We observe that the classical multi-armed bandit (MAB) problem with Bernoulli rewards is a special case of our setting. Indeed, consider the stochastic problem with n independent Bernoulli r.v.s $\{X_i\}_{i=1}^n$ where a policy is allowed to probe only one random variable and receives reward X_i for probing i . This problem is clearly down-monotone, and it is equivalent to MAB. As shown in Lattimore and Szepesvári (2020), any algorithm for MAB has regret at least $\Omega(\sqrt{nT})$. $\square$

Hence, our upper bounds match this lower bound up to a factor of $\sqrt{n}$ and logarithmic terms. The next result, also observed in Singla and Wang (2024), shows that stochastic problems under *binary* feedback with continuous distributions are fundamentally harder to learn than with discrete distributions.

THEOREM EC.4. *Any learning algorithm for stochastic monotone problems with binary feedback and continuous distributions has regret at least $\Omega\left(T^{\frac{2}{3}}\right)$.*

Proof. Consider the fixed-order SPM under a uniform matroid of rank one, which is a down-monotone stochastic problem. Leme et al. (2023) proved an $\Omega(1/\epsilon^3)$ query complexity lower bound for this problem under continuous distributions with binary-threshold queries. This directly implies an $\Omega(T^{2/3})$ regret lower

bound for online learning for SPM with binary feedback. Suppose (for a contradiction) that there is an online algorithm with regret $o(T^{\frac{2}{3}})$. Then after T rounds, we can achieve $o(T^{-\frac{1}{3}})$ per-round regret, which means that we can obtain a policy with ϵ additive error in $T = o(1/\epsilon^3)$ rounds, which contradicts the sample complexity result. $\square$

EC.4. Censored Feedback

In this section, we provide complete details regarding our generalization to settings with more restrictive feedback. To keep the exposition self-contained, we begin by revisiting the motivation, setup, and main results. Recall that our goal is to extend the learning framework to accommodate feedback models where the algorithm does not have direct access to the full realization x_i of a probed item i . In contrast to the *semi-bandit* setting discussed so far, we now consider two (more restrictive) feedback models: censored and binary. The first model, called *censored feedback*, assumes that the algorithm only observes the realization of each random variable up to a chosen “threshold”. The second model, termed *binary feedback*, is even more limited: the algorithm only receives binary information indicating whether/not the random variable’s realization is above its chosen threshold.

We also focus on a restricted class of “threshold based” stochastic problems. We assume that all random variables are discrete with a finite support set $\{a_1, \dots, a_k\}$. A solution/policy for a threshold-based stochastic problem is given by a decision tree σ where each internal node is labeled by a pair (c, i) where $i \in [n]$ is an item and $c \in [k]$ corresponds to a threshold value. When the policy σ reaches node (c, i) , it probes item i with threshold a_c and observes the following partial realization of X_i (depending on the feedback model).

- In *censored feedback*, the policy observes the truncated realization of $\min\{X_i, a_c\}$ and the branches out of node (c, i) correspond to this observation.
- In *binary feedback*, the policy observes the indicator $\mathbf{I}_{X_i \geq a_c}$ and the branches out of node (c, i) correspond to this observation.

Given a policy σ and realization $\mathbf{x}$, we use $S(\sigma, \mathbf{x})$ to denote the sequence of item-threshold pairs that are probed. We also make the following assumption on the objective function:

ASSUMPTION 2. *Given any threshold policy σ , the function value at each leaf node of σ is a constant (which may differ across leaf nodes).*

We note that this assumption is more restrictive compared to the previous results (Theorems 1 and 10). This is because at any leaf node of policy σ we may not know the precise realizations of probed items. Nevertheless, there are some interesting problems that satisfy this assumption.

In the online setting, at time $t \in [T]$, if the algorithm implements policy σ^t and the realizations are $\mathbf{x}^t \sim \mathbf{D}$ then it observes the tuple

$$\{\min(x_i^t, a_c) : (c, i) \in S(\sigma^t, \mathbf{x}^t)\}$$

under censored feedback, and the tuple

$$\left\{ \mathbf{I}_{x_i^t \geq a_c} : (c, i) \in S(\sigma^t, \mathbf{x}^t) \right\}$$

under binary feedback. The goal is to minimize α -regret as before.

Censored feedback application: Single Resource Revenue Management A classical instance of *censored feedback* arises in the single resource revenue management setting (Gallego and Topaloglu 2019, 4–6). Consider an airline with C units of seat capacity available for sale across multiple fare classes. The customer demand for each fare class is an independent random variable. We allocate portions of the capacity to different fare classes. When we choose to offer m seats to a specific fare class, we observe only the number of tickets sold—namely, the minimum of the demand and m . The airline’s revenue (which is the objective function) also depends only on the number sold. This corresponds to censored feedback.

Binary feedback application: Sequential Posted Pricing (SPM). Consider a seller who wants to sell a single item to a group of n buyers who have random valuations. The buyers arrive sequentially, and the seller presents each buyer i with a take-it-or-leave-it price p_i . The first buyer i whose value exceeds the posted price p_i buys the item and the seller receives p_i dollars as revenue. Although the seller observes whether/not a particular buyer buys the item, the exact valuations remain unknown. This corresponds to binary feedback.

Naturally, one might attempt to incorporate censored and binary feedback into the *semi-bandit* framework by introducing additional independent copies of each random variable (one for each threshold). However, this approach can be problematic for several reasons. First, it may significantly increase the complexity of the original problem, making it difficult to find a good approximation result even in the offline setting with known distributions. Second, it can fail to preserve the monotone property of the original problem, which is crucial for our analysis.

Instead, we show how to extend our techniques from prior sections to the new *censored/binary* feedback settings. Our main results here are the following.

THEOREM 11 (Censored Feedback). *Consider a stochastic problem $\mathcal{P}$ that is either up-monotone or down-monotone, and satisfies Assumption 2. Suppose that $\mathcal{P}$ has an α -approximation algorithm via threshold policies. Then, there is a polynomial time censored feedback learning algorithm for $\mathcal{P}$ (with unknown distributions) that has α -regret $O(nk f_{\max} \sqrt{kT \log(knT)})$. Here, n is the number of items, k is the maximum support size, $f_{\max}$ is the maximal value of the objective function and T is the number of time periods.*

THEOREM 12 (Binary Feedback). *Consider a stochastic problem $\mathcal{P}$ that is either up-monotone or down-monotone, and satisfies Assumption 2. Suppose that $\mathcal{P}$ has an α -approximation algorithm via threshold policies. Then, there is a polynomial time binary feedback learning algorithm for $\mathcal{P}$ (with unknown distributions) that has α -regret $O(n f_{\max} \sqrt{kT \log(knT)})$. Here, n is the number of items, k is the maximum support size, $f_{\max}$ is the maximal value of the objective function and T is the number of time periods.*

The high-level approach remains the same as in the semi-bandit setting. In the following two subsections, we establish the key sampling theorems and stability lemmas tailored to the new feedback models. We primarily focus on minimization up-monotone problems under the censored feedback setting; the same ideas apply to the binary-feedback setting as well. As with semi-bandit feedback, results for all the other combinations (minimization/maximization objective and up/down monotonicity) follow in an identical manner. The primary technical contribution is the sampling theorem, which addresses the challenge posed by limited observations.

EC.4.1. New Sampling Algorithms

In this subsection, we focus on up-monotone stochastic problems. The analysis for down-monotone stochastic problems is similar. We will first present the sampling theorem for the censored feedback setting. Then, we provide a similar result for the binary feedback case. The key challenge in these settings is that, due to the limited nature of censored/binary feedback, we cannot directly control the total variation distance between the empirical and the true underlying distribution. Instead, we design an algorithm that controls the total variation distance at different threshold levels.

DEFINITION EC.2 (TRUNCATED DISTRIBUTION). Given a discrete r.v. $X \sim \mathcal{D}$ with support set $\{a_1, \dots, a_k\}$, the truncated distribution $\mathcal{D}_c$ is the distribution of r.v. $X_c := \min(X, a_c)$. It is supported on values $\{a_1, \dots, a_c\}$, with $\Pr[X_c = v] = \Pr[X = v]$ for all $v < a_c$ and $\Pr[X_c = a_c] = \Pr[X \geq a_c]$.

THEOREM EC.5 (Sampling under Censored Feedback). Consider any r.v. $X \sim \mathcal{D}$ with support set $\{a_1, \dots, a_k\}$. There is an efficient algorithm that, given m_c i.i.d. “censored” samples of the form $\min(X, a_c)$ for each $c \in [k]$, computes a distribution $\mathcal{E}$ satisfying the following properties with probability at least $1 - \delta$.

- $\mathcal{E}$ stochastically dominates $\mathcal{D}$.
- For each $c \in [k]$, the total-variation distance $\text{TV-Dist}(\mathcal{E}_c, \mathcal{D}_c) < k \sqrt{\frac{2 \log(2k/\delta)}{m_c}}$.

We first note that when a sample of $X_b := \min(X, a_b)$ is obtained, it also provides a sample of X_c for all $c \leq b$. Consequently, we define $n_c := \sum_{b \geq c} m_b$ as the effective number of samples at level c . For any $c \in [k]$, we also set its confidence width as $\epsilon_c := \sqrt{\frac{\log(2k/\delta)}{2n_c}}$. Note that $\epsilon_1 \leq \dots \leq \epsilon_k$.

For each $c \in [k]$, let $\hat{\mathcal{D}}_c$ denote the empirical distribution of the n_c samples of X_c , and define its probability mass function (pmf) as

$$\hat{p}_c(b) := \Pr_{Y \sim \hat{\mathcal{D}}_c} [Y = a_b], \quad \forall b \leq c.$$

Let $q_c := \Pr_{X \sim \mathcal{D}} [X \geq a_c]$ for $c \in [k]$ denote the tail cdf of the true distribution $\mathcal{D}$. By Hoeffding’s inequality (Hoeffding 1963), for each $c \in [k]$, we have:

$$\Pr[|\hat{p}_c(c) - q_c| > \epsilon_c] < \frac{\delta}{k}.$$

Combined with a union bound, the following “good” event holds w.p. at least $1 - \delta$.

$$|\widehat{p}_c(c) - q_c| \leq \epsilon_c, \quad \forall c \in [k]. \quad (\text{EC.2})$$

We will assume this good event in the rest of the proof.

Our algorithm to construct the desired distribution $\mathcal{E}$ modifies the (truncated) empirical distributions $\widehat{\mathcal{D}}_c$ by increasing $c = 1, 2, \dots, k$. For each $c \in [k]$, we construct a distribution $\mathcal{E}_c$ supported on $\{a_1, \dots, a_c\}$ that stochastically dominates $\mathcal{D}_c$; we construct $\mathcal{E}_{c+1}$ from $\mathcal{E}_c$ by moving some mass from a_c to a_{c+1} . Below, the pmf of $\mathcal{E}_c$ is denoted by $\mathcal{E}_c(b) = \Pr_{Y \sim \mathcal{E}_c}[Y = a_b]$ for $b \leq c$.

Algorithm 4 Constructing Empirical Modified Distribution $\mathcal{E}$

```

1: Set  $\mathcal{E}_1(1) = 1$ 
2: for  $c = 1, 2, \dots, k - 1$  do
3:   Set  $\mathcal{E}_{c+1}(b) \leftarrow \mathcal{E}_c(b)$  for  $b = 1, \dots, c - 1$ 
4:    $\mathcal{E}_{c+1}(c + 1) \leftarrow \min\{\mathcal{E}_c(c), \widehat{p}_{c+1}(c + 1) + \epsilon_{c+1}\}$ 
5:    $\mathcal{E}_{c+1}(c) \leftarrow \mathcal{E}_c(c) - \mathcal{E}_{c+1}(c + 1)$ 
6: end for
7: Output:  $\mathcal{E} = \mathcal{E}_k$ 

```

We complete the proof of the theorem by proving the following claims via induction.

CLAIM EC.1. *For each $c \in [k]$, $\mathcal{E}_c$ is a valid probability distribution.*

The base case ($c = 1$) is trivial by construction: $\mathcal{E}_1$ is a point mass distribution. We will now show that $\mathcal{E}_{c+1}$ is a valid probability distribution assuming that $\mathcal{E}_c$ is. It is easy to see that the pmf $\mathcal{E}_{c+1}(\cdot)$ is non-negative by Steps 3-5 in Algorithm 4. Further,

$$\begin{aligned}
\sum_{b=1}^{c+1} \mathcal{E}_{c+1}(b) &= \sum_{b=1}^{c-1} \mathcal{E}_c(b) + \mathcal{E}_{c+1}(c) + \mathcal{E}_{c+1}(c+1) \\
&= \sum_{b=1}^{c-1} \mathcal{E}_c(b) + \mathcal{E}_c(c) = 1,
\end{aligned}$$

where the first equality is by Step 3, the second equality is by Step 5 and the last equality uses that $\mathcal{E}_c$ is a valid distribution. Therefore, $\mathcal{E}_{c+1}$ is also a valid distribution.

CLAIM EC.2. *Under the good event (EC.2), for each $c \in [k]$, $\mathcal{E}_c$ stochastically dominates $\mathcal{D}_c$.*

The base case ($c = 1$) is trivial as both $\mathcal{D}_1$ and $\mathcal{E}_1$ are point masses with value a_1 . We will show that $\mathcal{D}_{c+1} \preceq_{\text{SD}} \mathcal{E}_{c+1}$ assuming $\mathcal{D}_c \preceq_{\text{SD}} \mathcal{E}_c$. It suffices to show that

$$\sum_{j=b}^{c+1} \mathcal{E}_{c+1}(j) \geq q_b, \quad \forall b = 1, \dots, c+1. \quad (\text{EC.3})$$

For $b \leq c$, we have

$$\begin{aligned} \sum_{j=b}^{c+1} \mathcal{E}_{c+1}(j) &= \sum_{j=b}^{c-1} \mathcal{E}_c(j) + \mathcal{E}_{c+1}(c) + \mathcal{E}_{c+1}(c+1) \\ &= \sum_{j=b}^{c-1} \mathcal{E}_c(j) + \mathcal{E}_c(c) = \sum_{j=b}^c \mathcal{E}_c(j) \geq q_b. \end{aligned}$$

The first two equalities are by Steps 3-5 in Algorithm 4, and the inequality is by $\mathcal{D}_c \preceq_{\text{SD}} \mathcal{E}_c$.

It remains to prove (EC.3) for $b = c+1$. By the good event (EC.2), we have $q_{c+1} \leq \widehat{p}_{c+1}(c+1) + \epsilon_{c+1}$. Moreover, by $\mathcal{D}_c \preceq_{\text{SD}} \mathcal{E}_c$ we have $\mathcal{E}_c(c) \geq q_c \geq q_{c+1}$. So, by Step 5 in Algorithm 4, we have $\mathcal{E}_{c+1}(c+1) = \min \{ \mathcal{E}_c(c), \widehat{p}_{c+1}(c+1) + \epsilon_{c+1} \} \geq q_{c+1}$, which completes the proof of (EC.3), and thus the claim.

CLAIM EC.3. *Under the good event (EC.2), for each $c \in [k]$, we have $\text{TV}(\mathcal{E}_c, \mathcal{D}_c) \leq k \cdot \sqrt{\frac{2 \log(2k/\delta)}{n_c}}$.*

Fix any $c \in [k]$. We first claim that, for all $b \leq c$,

$$\sum_{j=b}^c \mathcal{E}_c(j) = \mathcal{E}_b(b) \leq \widehat{p}_b(b) + \epsilon_b \leq q_b + 2\epsilon_b. \quad (\text{EC.4})$$

The equality $\sum_{j=b}^c \mathcal{E}_c(j) = \mathcal{E}_b(b)$ for $b \leq c$ is by construction of the distributions $\mathcal{E}_c$ s: we only move mass from low to high support values. The inequality is by Step 4 of Algorithm 4, and the last inequality is by the good event (EC.2).

Now, for any $b \leq c-1$, we have

$$\begin{aligned} &|\mathcal{E}_c(b) - \Pr(\mathbf{X} = a_b)| \\ &= \left| \sum_{j=b}^c \mathcal{E}_c(j) - \sum_{j=b+1}^c \mathcal{E}_c(j) - q_b + q_{b+1} \right| \\ &\leq \left| \sum_{j=b}^c \mathcal{E}_c(j) - q_b \right| + \left| \sum_{j=b+1}^c \mathcal{E}_c(j) - q_{b+1} \right| \\ &\leq 2\epsilon_b + 2\epsilon_{b+1} \leq 4\epsilon_c. \end{aligned} \quad (\text{EC.5})$$

The second inequality uses (EC.4) and the last inequality uses that $\epsilon_i \leq \epsilon_j$ for all $i \leq j$.

We can now bound the TV distance:

$$\begin{aligned}
& \text{TV-Dist}(\mathcal{E}_c, \mathcal{D}_c) \\
&= \frac{1}{2} \sum_{b=1}^{c-1} |\mathcal{E}_c(b) - \Pr(\mathbf{X} = a_b)| + \frac{1}{2} |\mathcal{E}_c(c) - q_c| \\
&\leq 2(c-1)\epsilon_c + \epsilon_c \leq k \sqrt{\frac{2 \log(2k/\delta)}{n_c}}.
\end{aligned}$$

The first inequality uses (EC.5) for $b \leq c-1$ and (EC.4) for c . This implies the TV bound in the theorem because $n_c \geq m_c$. Finally, observe that when the output distribution $\mathcal{E} = \mathcal{E}_k$ is capped by a_c we obtain the distribution $\mathcal{E}_c$.

Combining Claim EC.1, Claim EC.2, and Claim EC.3 completes the proof of Theorem EC.5.

The sampling algorithm and analysis can be extended to binary feedback in a straightforward way. We now describe the changes needed.

DEFINITION EC.3 (BINARY COMPRESSED DISTRIBUTION). Given a discrete r.v. $\mathbf{X} \sim \mathcal{D}$ with support set $\{a_1, \dots, a_k\}$, the binary compressed distribution $\mathcal{D}_c^B$ is the distribution of r.v. $\mathbf{X}_c^B := \mathbf{I}_{\mathbf{X} \geq a_c}$. We have $\Pr(\mathbf{X}_c^B = 0) = \Pr(\mathbf{X} < a_c)$ and $\Pr(\mathbf{X}_c^B = 1) = \Pr(\mathbf{X} \geq a_c)$.

THEOREM EC.6 (Sampling for Binary Feedback). Consider any r.v. $\mathbf{X} \sim \mathcal{D}$ with support set $\{a_1, \dots, a_k\}$. There is an efficient algorithm that, given m_c i.i.d. “binary” samples of the form $\mathbf{I}_{\mathbf{X} \geq a_c}$ for each $c \in [k]$, computes a distribution $\mathcal{E}$ satisfying the following properties with probability at least $1 - \delta$.

- $\mathcal{E}$ stochastically dominates $\mathcal{D}$.
- For $c \in [k]$, the total-variation distance $\text{TV-Dist}(\mathcal{E}_c^B, \mathcal{D}_c^B) < \sqrt{\frac{2 \log(2k/\delta)}{m_c}}$.

For any $c \in [k]$, we denote by $\widehat{\mathcal{D}}_c$ the empirical distribution of the m_c samples of $\mathbf{I}_{\mathbf{X} \geq a_c}$; we use $\hat{p}_c(c) = \Pr_{\mathbf{Y} \sim \widehat{\mathcal{D}}_c}[\mathbf{Y} = 1]$ to keep notation the same as for the censored case. We also define the confidence width $\epsilon_c := \sqrt{\frac{\log(2k/\delta)}{2m_c}}$. As before, let $q_c := \Pr_{\mathbf{X} \sim \mathcal{D}}[\mathbf{X} \geq a_c]$ for $c \in [k]$ denote the tail cdf of $\mathcal{D}$. By Hoeffding’s inequality and union bound, the following “good” event holds w.p. at least $1 - \delta$.

$$|\hat{p}_c(c) - q_c| \leq \epsilon_c, \quad \forall c \in [k]. \quad (\text{EC.6})$$

Given these values $\hat{p}_c(c)$, the sampling algorithm to construct $\mathcal{E}$ remains the same as the censored case (Algorithm 4). Let $\mathcal{E}_c^B$ denote the intermediate distributions: exactly as before, for each $c \in [k]$, $\mathcal{E}_c^B$ is a valid

distribution and $\mathcal{D}_c^B \preceq_{\text{SD}} \mathcal{E}_c^B$ (under the good event). Inequality (EC.4) also continues to hold. We can now bound the TV distance between $\mathcal{E}_c^B$ and $\mathcal{D}_c^B$ for any $c \in [k]$.

$$\begin{aligned} \text{TV-Dist}(\mathcal{E}_c^B, \mathcal{D}_c^B) &= |\mathcal{E}_c^B(1) - \Pr(X_c^B = 1)| \\ &= |\mathcal{E}_c^B(1) - \Pr(X \geq a_c)| = |\mathcal{E}_c^B(1) - q_c| \\ &= \left| \sum_{j=c}^k \mathcal{E}_k(j) - q_c \right| \leq 2\epsilon_c = \sqrt{\frac{2 \log(2k/\delta)}{m_c}}. \end{aligned}$$

where the inequality is by (EC.4). This completes the proof of Theorem EC.6.

EC.4.2. Stability Lemma

We now describe the modifications to the stability lemma under censored feedback.

LEMMA EC.1 (Censored Stability Lemma). *Consider a threshold-based stochastic problem that is up-monotone and satisfies Assumption 2. Suppose that $\mathbf{E} = \{\mathcal{E}_i\}_{i=1}^n$ and $\mathbf{D} = \{\mathcal{D}_i\}_{i=1}^n$ are product distributions such that $\mathbf{D} \preceq_{\text{SD}} \mathbf{E}$, where each distribution $\{\mathcal{D}_i, \mathcal{E}_i\}_{i=1}^n$ is supported on $\{a_1, \dots, a_k\}$. Further, suppose that $\text{TV}(\mathcal{E}_{c,i}, \mathcal{D}_{c,i}) \leq \epsilon_{c,i}$ for each $i \in [n]$ and $c \in [k]$, where $\mathcal{D}_{c,i}$ (resp. $\mathcal{E}_{c,i}$) is the distribution $\mathcal{D}_i$ (resp. $\mathcal{E}_i$) truncated at a_c . If σ is the policy returned by $\text{ALG}(\mathbf{E})$ and σ^* is an optimal policy under $\mathbf{D}$, then*

$$\begin{aligned} f(\sigma) - \alpha \cdot f(\sigma^*) &= \mathbb{E}_{\mathbf{x} \sim \mathbf{D}} [f(\sigma, \mathbf{x}) - \alpha \cdot f(\sigma^*, \mathbf{x})] \\ &\leq f_{\max} \sum_{i=1}^n \sum_{c=1}^k Q_{c,i}(\sigma) \cdot \epsilon_{c,i}, \end{aligned}$$

where $Q_{c,i}(\sigma)$ is the probability (under $\mathbf{D}$) that item i is probed by policy σ with threshold a_c .

The analysis is almost the same as for the semi-bandit setting (Lemma 4). The main modification is in labeling the nodes of the decision tree. Each node v in the decision tree σ now corresponds to a pair (c, i) , indicating that item i is probed with threshold a_c . For any item i , in any policy execution we will encounter at most one node labeled by item i and any threshold. So, we can view policy σ as having an item X_v with independent distribution $\mathcal{D}_v \stackrel{d}{=} \mathcal{D}_{c,i}$ at each node $v \in \sigma$ labeled by (c, i) ; here “ $\stackrel{d}{=}$ ” means equal in distribution. Similarly, we use $\mathcal{E}_v \stackrel{d}{=} \mathcal{E}_{c,i}$ for v labeled by (c, i) . Exactly as in the proof of Lemma 4, we index nodes in σ according to the ancestor-descendant partial order and use hybrid product distributions $\mathbf{H}^v$. Again, we will show:

$$|f(\sigma | \mathbf{H}^v) - f(\sigma | \mathbf{H}^{v-1})| \leq f_{\max} \cdot Q_v(\sigma) \cdot \epsilon_v, \quad (\text{EC.7})$$

where $Q_v(\sigma)$ is the probability that policy σ reaches node v under distribution $\mathbf{D}$ and $\epsilon_v = \text{TV}(\mathcal{D}_v, \mathcal{E}_v) = \text{TV}(\mathcal{E}_{c,i}, \mathcal{D}_{c,i}) \leq \epsilon_{c,i}$. We now complete the proof of the lemma using (EC.7). We have

$$\begin{aligned}
& |f(\sigma|\mathbf{D}) - f(\sigma|\mathbf{E})| \\
& \leq \sum_{v=1}^N |f(\sigma|\mathbf{H}^v) - f(\sigma|\mathbf{H}^{v-1})| \\
& \leq f_{\max} \cdot \sum_{v=1}^N Q_v(\sigma) \cdot \epsilon_v \\
& \leq \sum_{i=1}^n \sum_{c=1}^k \epsilon_{c,i} \sum_{v: \text{ labeled by } (c,i)} Q_v(\sigma) \\
& = \sum_{i=1}^n \sum_{c=1}^k \epsilon_{c,i} \cdot Q_{c,i}(\sigma),
\end{aligned}$$

where the final equality uses the fact that $Q_{c,i}(\sigma)$ equals the total probability of reaching some node v labeled by (c, i) . This suffices to prove the lemma, exactly as in Lemma 4.

The proof of (EC.7) is almost identical to that of (9) in Lemma 4. Letting $\mathcal{R}_v$ be the event that σ reaches node v , it suffices to show:

$$\begin{aligned}
& |\mathbb{E}_{\mathbf{x} \sim \mathbf{H}^v} [f(\sigma, \mathbf{x}) \mid \mathcal{R}_v] - \mathbb{E}_{\mathbf{x} \sim \mathbf{H}^{v-1}} [f(\sigma, \mathbf{x}) \mid \mathcal{R}_v]| \\
& \leq f_{\max} \cdot \text{TV}(\mathcal{E}_v, \mathcal{D}_v),
\end{aligned}$$

The proof is identical to that of (13), where we use the assumption that the function value at any leaf-node of σ is deterministic (even though we don't know the exact realizations of the probed r.v.s).

By an identical proof, using distributions $\mathcal{D}_v \stackrel{d}{=} \mathcal{D}_{c,i}^B$ and $\mathcal{E}_v \stackrel{d}{=} \mathcal{E}_{c,i}^B$ at each node $v = (c, i)$, we obtain the following under binary feedback.

LEMMA EC.2 (Binary Stability Lemma). *Consider a threshold-based stochastic problem that is up-monotone and satisfies Assumption 2. Suppose that $\mathbf{E} = \{\mathcal{E}_i\}_{i=1}^n$ and $\mathbf{D} = \{\mathcal{D}_i\}_{i=1}^n$ are product distributions such that $\mathbf{D} \preceq_{\text{SD}} \mathbf{E}$, where each distribution $\{\mathcal{D}_i, \mathcal{E}_i\}_{i=1}^n$ is supported on $\{a_1, \dots, a_k\}$. Further, suppose that $\text{TV}(\mathcal{E}_{c,i}^B, \mathcal{D}_{c,i}^B) \leq \epsilon_{c,i}$ for each $i \in [n]$ and $c \in [k]$, where $\mathcal{D}_{c,i}^B$ (resp. $\mathcal{E}_{c,i}^B$) is the distribution $\mathcal{D}_i$ (resp. $\mathcal{E}_i$) compressed at a_c . If σ is the policy returned by $\text{ALG}(\mathbf{E})$ and σ^* is an optimal policy under $\mathbf{D}$, then*

$$\begin{aligned}
f(\sigma) - \alpha \cdot f(\sigma^*) &= \mathbb{E}_{\mathbf{x} \sim \mathbf{D}} [f(\sigma, \mathbf{x}) - \alpha \cdot f(\sigma^*, \mathbf{x})] \\
&\leq f_{\max} \sum_{i=1}^n \sum_{c=1}^k Q_{c,i}(\sigma) \cdot \epsilon_{c,i},
\end{aligned}$$

where $Q_{c,i}(\sigma)$ is the probability (under $\mathbf{D}$) that item i is probed by policy σ with threshold a_c .

EC.4.3. Overall Regret

We now prove the main theorem for the censored feedback setting (Theorem 11). The overall structure of the proof mirrors that of the semi-bandit setting: we combine a sampling theorem with the stability lemma to bound the regret.

Recall that, $\mathcal{D}_i$ is the true distribution of item i and $\mathcal{E}_i^t$ is the dominating distribution generated by Algorithm 4 in time step t . For any $i \in [n]$, $c \in [k]$, let $N_{c,i}^t$ be the number times the algorithm sampled $\min\{X_i, a_c\}$ before time t .

LEMMA EC.3. *With probability at least $1 - \frac{1}{nT}$, we have $\mathcal{D}_i \preceq_{\text{SD}} \mathcal{E}_i^t$ and the total variation distance $\text{TV}(\mathcal{E}_{c,i}^t, \mathcal{D}_{c,i}) \leq k \sqrt{\frac{6 \log(knT)}{N_{c,i}^t}}$ for $c \in [k]$, $i \in [n]$, and $t \in [T]$.*

For any $c \in [k]$, $i \in [n]$, and $t \in [T]$, let $B_{c,i}^t$ denote the event that the stated condition fails to hold for the triple (c, i, t) . By Theorem EC.5 with $\delta = \frac{2}{k^2 n^3 T^3}$, we have:

$$\Pr(B_{c,i}^t) \leq \sum_{m=1}^T \Pr(B_{c,i}^t \wedge N_{c,i}^t = m) \leq \delta T.$$

Applying another union bound over all values of c , i , and t , we obtain:

$$\begin{aligned} \Pr\left(\bigvee_{c=1}^k \bigvee_{i=1}^n \bigvee_{t=1}^T B_{c,i}^t\right) &\leq \sum_{c=1}^k \sum_{i=1}^n \sum_{t=1}^T \Pr(B_{c,i}^t) \\ &\leq knT^2\delta \leq \frac{1}{nT}, \end{aligned}$$

where the last inequality uses the assumption that $n \geq 2$.

We now define a good event G that corresponds to the condition in Lemma EC.3 holding true for all $c \in [k]$, $i \in [n]$ and $t \in [T]$. Next, we complete the proof assuming that G holds.

Bounding the regret as a sum over time t . To bound the overall regret, it suffices to bound the expected regret $R^t = f(\sigma^t) - \alpha \cdot f(\sigma^*)$ at each time $t \in [T]$. For each time $t \in [T]$, using the good event, we can apply Lemma EC.1 with distributions $\mathbf{E}^t$, $\mathbf{D}$, and parameters $\epsilon_{c,i}^t = k \cdot \sqrt{\frac{6 \log(knT)}{N_{c,i}^t}}$ to obtain

$$R^t \leq f_{\max} \sum_{i=1}^n \sum_{c=1}^k Q_{c,i}(\sigma^t) \cdot \epsilon_{c,i}^t. \quad (\text{EC.8})$$

So, the overall regret is

$$\begin{aligned} \alpha\text{-}R(T) &\leq f_{\max} \sum_{t=1}^T \mathbb{E}_{\mathbf{h}^{t-1}} \left[\sum_{i=1}^n \sum_{c=1}^k Q_{c,i}(\sigma^t) \cdot \epsilon_{c,i}^t \right] \\ &= k f_{\max} \sqrt{6 \log(knT)} \sum_{t=1}^T \mathbb{E}_{\mathbf{h}^{t-1}} \left[\sum_{c,i} \frac{Q_{c,i}(\sigma^t)}{\sqrt{N_{c,i}^t}} \right]. \end{aligned}$$

Recall that $\mathbf{h}^{t-1} = (\mathbf{x}^1, \dots, \mathbf{x}^{t-1})$ is the history until time t . It now suffices to show

$$\sum_{t=1}^T \mathbb{E}_{\mathbf{h}^{t-1}} \left[\sum_{i=1}^n \sum_{c=1}^k \frac{Q_{c,i}(\sigma^t)}{\sqrt{N_{c,i}^t}} \right] \leq 2n\sqrt{kT}. \quad (\text{EC.9})$$

Proving (6) as a sum over decision paths. Similar to the semi-bandit case, we define, for all $i \in [n]$ and $c \in [k]$:

$$Z_i^c(\mathbf{h}^T) := \sum_{t=1}^T \frac{\mathbf{I}[\mathcal{A}_c]}{\sqrt{N_{c,i}^t(\mathbf{h}^{t-1})}}.$$

Here, $\mathcal{A}_c$ denotes the event that i probed by policy $\sigma^t(\mathbf{h}^{t-1})$ with threshold a_c and $\mathbf{I}$ is the indicator function.

By linearity of expectation, we have

$$\begin{aligned} \mathbb{E}_{\mathbf{h}^T} [Z_i^c(\mathbf{h}^T)] &= \sum_{t=1}^T \mathbb{E}_{\mathbf{h}^T} \left[\frac{\mathbf{I}[\mathcal{A}_c]}{\sqrt{N_{c,i}^t(\mathbf{h}^{t-1})}} \right] \\ &= \sum_{t=1}^T \mathbb{E}_{\mathbf{h}^{t-1}, \mathbf{x}^t} \left[\frac{\mathbf{I}[\mathcal{A}_c]}{\sqrt{N_{c,i}^t(\mathbf{h}^{t-1})}} \right] \\ &= \sum_{t=1}^T \mathbb{E}_{\mathbf{h}^{t-1}} \left[\frac{1}{\sqrt{N_{c,i}^t}} \cdot \Pr_{\mathbf{x}^t}[\mathcal{A}_c] \right] \\ &= \sum_{t=1}^T \mathbb{E}_{\mathbf{h}^{t-1}} \left[\frac{Q_{c,i}(\sigma^t)}{\sqrt{N_{c,i}^t}} \right], \end{aligned} \quad (\text{EC.10})$$

where the second equality uses the fact that event $\{i \text{ probed by } \sigma^t \text{ with } c\}$ only depends on $\mathbf{h}^t = (\mathbf{h}^{t-1}, \mathbf{x}^t)$, the third equality uses the fact that $N_{c,i}^t$ only depends on $\mathbf{h}^{t-1}$ and that $\mathbf{x}^t$ is independent of $\mathbf{h}^{t-1}$, and the last equality is by the definition of $Q_{c,i}(\sigma^t)$ and the fact that $\mathbf{x}^t \sim \mathbf{D}$.

Using (EC.10) and adding over $i \in [n]$ and $c \in [k]$, we get $\sum_{t=1}^T \sum_{c=1}^k \sum_{i=1}^n \mathbb{E}_{\mathbf{h}^{t-1}} \left[\frac{Q_{c,i}(\sigma^t)}{\sqrt{N_{c,i}^t}} \right] = \sum_{i=1}^n \sum_{c=1}^k \mathbb{E}_{\mathbf{h}^T} [Z_i^c(\mathbf{h}^T)]$.

Therefore, proving (EC.9) is equivalent to:

$$\sum_{i=1}^n \sum_{c=1}^k \mathbb{E}_{\mathbf{h}^T} [Z_i^c(\mathbf{h}^T)] \leq 2n\sqrt{kT}. \quad (\text{EC.11})$$

Now, for any $i \in [n]$ and $c \in [k]$, we obtain:

$$\begin{aligned} Z_i^c(\mathbf{h}^T) &= \sum_{t=1}^T \frac{\mathbf{I}[i \text{ probed by } \sigma^t(\mathbf{h}^{t-1}) \text{ with } a_c]}{\sqrt{N_{c,i}^t(\mathbf{h}^{t-1})}} \\ &\leq \sum_{t=1}^{N_{c,i}^T} \frac{1}{\sqrt{t}} \leq 2\sqrt{N_{c,i}^T}. \end{aligned}$$

The first inequality uses the fact that if item i gets probed with threshold a_c at time t then $N_{c,i}^t = N_{c,i}^{t-1} + 1$. Using Jensen's inequality and the fact that the total number of probes for one item i is at most T , we have:

$$\sum_{c=1}^k \sqrt{N_{c,i}^T} \leq k \sqrt{\frac{\sum_{c=1}^k N_{c,i}^T}{k}} \leq \sqrt{kT}.$$

This completes the proof of (EC.11) and hence (EC.9).

We note that our good event G holds with probability at least $1 - \frac{1}{nT}$. As in the semi-bandit setting, we can convert the high-probability regret bound into an expected regret bound. This concludes the proof of Theorem 11.

The proof of Theorem 12 is identical: we just need to use the appropriate sampling result (Theorem EC.6) and stability lemma (Lemma EC.2).

EC.5. Missing Applications

EC.5.1. Stochastic Matching and Probing

In the *stochastic matching* problem (Bansal et al. 2012, Chen et al. 2009), there is an undirected graph $G = (V, E)$ with edge-weights $\{w_e\}_{e \in E}$, edge-probabilities $\{p_e\}_{e \in E}$, and vertex bounds $\{t_v\}_{v \in V}$. Each edge e is *active* independently with probability p_e . However, the status (active/inactive) of an edge can only be determined by *probing* it. There is also a constraint on the set of probed edges: for any vertex v , the number of probed edges incident to v must be at most t_v . A solution/policy needs to probe a subset of edges and select a matching M consisting of active edges. The objective is to maximize the expected weight of M . Finally, there is a “query commit” requirement that any probed edge which is active *must* be included in the selected matching M . Observe that in any policy execution, we only see the status of some subset of edges, which corresponds to semi-bandit feedback. This problem has been extensively studied, see e.g., Adamczyk (2011), Bansal et al. (2012), Brubach et al. (2021), Chen et al. (2009). The current best approximation ratio is 0.5 for the unweighted case (Adamczyk 2011) and 0.382 for the weighted case (Brubach et al. 2021). We show in Lemma EC.4 below that the stochastic matching problem is down-monotone. So, Theorem 1

with $k = 2$ (as all r.v.s are binary) implies an online learning algorithm for stochastic matching having 0.382-regret of $O(nW\sqrt{T\log T})$ where $W = \sum_{e \in E} w_e$ is the total weight.

In fact, our result applies to the much more general *stochastic probing problem*, as defined in Gupta and Nagarajan (2013). Here, we have a set E of stochastic items with weights $\{w_e\}_{e \in E}$ and probabilities $\{p_e\}_{e \in E}$. Each item is active independently with probability p_e , and this status can only be determined by probing e . We now have two downward-closed constraints: an *inner* constraint $\mathcal{F}_{in}$ and an *outer* constraint $\mathcal{F}_{out}$. We want to probe a set Q of items subject to the outer constraint (i.e. $Q \in \mathcal{F}_{out}$) and select a subset $S \subseteq Q$ of active (probed) items satisfying the inner constraint (i.e. $S \in \mathcal{F}_{in}$). The objective is to maximize the expected weight of the selected items S . We again have the query-commit requirement that any active probed item must be selected into the solution S . When both inner/outer constraints are k_{in} and k_{out} systems, there is a $\frac{1}{k_{in}+k_{out}}$ approximation for the unweighted case and an $\Omega(\frac{1}{(k_{in}+k_{out})^2})$ approximation for the weighted case (Gupta and Nagarajan 2013). When the k -systems are intersections of matroids, Adamczyk et al. (2016) gave an improved adaptive algorithm with approximation ratio $\frac{1}{k_{in}+k_{out}}$, even for the weighted case.

LEMMA EC.4. *The stochastic probing problem is down-monotone.*

Consider an arbitrary instance $\mathcal{I}$ and product probability distributions $\mathbf{D} = \{p_e\}_{e \in E}$ and $\mathbf{E} = \{\bar{p}_e\}_{e \in E}$ where $\mathbf{D} \preceq_{SD} \mathbf{E}$. As the r.v.s are binary in this problem, the stochastic dominance just means $p_e \leq \bar{p}_e$ for all $e \in E$.

Let σ denote the optimal policy for $\mathcal{I}$ under distribution $\mathbf{D}$, and let $\text{OPT}_{\mathcal{I}}(\mathbf{D})$ denote its expected weight. Recall that a policy is given by a decision tree, where each node is labeled by an item to probe next, and the branches out of a node correspond to the random 0 – 1 realization of the probed item. We construct a policy $\hat{\sigma}$ that works under distribution $\mathbf{E}$, by closely following σ . When policy σ probes item e at any node, policy $\hat{\sigma}$ does the following:

- probe e with probability $p_e/\bar{p}_e$, and follow the active/inactive branch based on its outcome.
- skip directly to the inactive branch with probability $1 - p_e/\bar{p}_e$.

Note that conditioned on reaching such a node, the probability that $\hat{\sigma}$ goes to the active branch is $\frac{p_e}{\bar{p}_e} \cdot \bar{p}_e = p_e$; so the probability that $\hat{\sigma}$ goes to the inactive branch is $1 - p_e$. Therefore, the probability of reaching any node v in policy $\hat{\sigma}$ (under distribution $\mathbf{E}$) equals the probability of reaching node v in policy σ (under distribution $\mathbf{D}$). Thus the expected values of these two policies are the same, $\mathbb{E}_{\mathbf{x} \sim \mathbf{E}}[f(\hat{\sigma}, \mathbf{x})] = \mathbb{E}_{\mathbf{x} \sim \mathbf{D}}[f(\sigma, \mathbf{x})]$. We note that $\hat{\sigma}$ also satisfies the query-commit requirement because σ does (whenever e is probed and found active, it will be included in the solution). Finally, observe that $\hat{\sigma}$ is a randomized policy; however, as noted in Section 1.5, there is always an optimal deterministic policy for these stochastic problems. So, we obtain $\text{OPT}_{\mathcal{I}}(\mathbf{E}) \geq \text{OPT}_{\mathcal{I}}(\mathbf{D})$ as needed.

EC.5.2. Stochastic Covering Knapsack

Here, we have n items with deterministic costs $\{c_i\}_{i \in [n]}$ and (independent) random rewards $\{R_i\}_{i \in [n]}$. There is a known probability distribution $\mathcal{D}_i$ for R_i ; we assume discrete distributions here. Given a target Q , we need to select items until the total observed reward is at least Q (or we run out of items). The objective is to minimize the expected cost. Note that we only observe items until the target is achieved, which corresponds to semi-bandit feedback. There is an adaptive 3-approximation algorithm for this problem (Deshpande et al. 2016). By a proof identical to that of Lemma 8, we get:

LEMMA EC.5. *The stochastic covering knapsack problem is up-monotone.*

So, using Theorem 1, we obtain a polynomial time online learning algorithm for stochastic min-knapsack having 3-regret $O(nC\sqrt{kT\log T})$ where $C = \sum_{i=1}^n c_i$ is the total cost and k is the maximum support size.

Stochastic Q-TSP. This is the stochastic variant of the classical Q-TSP problem (Garg 2005). An instance of the problem comprises a metric (V, d) with a root vertex $r \in V$. There are random rewards $\{R_v\}_{v \in V}$ associated with the vertices. All reward distributions are given as input, but the realized reward R_v is only known when vertex v is visited. We assume that all reward distributions are discrete. Given a target Q , the goal is to find an (adaptive) tour originating from r that collects a total reward at least Q (or runs out of vertices), at the minimum expected length. A constant-factor approximation algorithm for stochastic Q-TSP was obtained in Jiang et al. (2020). Applying Theorem 1, we obtain a polynomial time online learning algorithm for stochastic Q-TSP having $O(1)$ -regret $O(nD\sqrt{kT\log T})$ where $D = \sum_{v \in V} d_{rv}$ corresponds to an upper bound on the (full) TSP cost.

EC.5.3. Stochastic Submodular Optimization

Consider a monotone submodular function $f : 2^E \rightarrow \mathbb{Z}_+$ defined on a groundset E . Constrained submodular maximization involves selecting a subset $S \subseteq E$ of items in a downward-closed family $\mathcal{F} \subseteq 2^E$, that maximizes the function value $f(S)$. This is a fundamental problem in combinatorial optimization and good approximation algorithms are known for a variety of constraints $\mathcal{F}$ such as cardinality, matroids and knapsacks (Călinescu et al. 2011, Chekuri et al. 2014, Kulik et al. 2013, Nemhauser et al. 1978). We consider a stochastic variant, as in Asadpour and Nazerzadeh (2016), Gupta et al. (2017), and Bradac et al. (2019), where each item $i \in E$ is *active* independently with some probability p_i (and inactive otherwise).¹ Moreover, the function value is accrued only on active items. The active/inactive status of any item is only known after selecting it. The goal now is to select a subset $S \in \mathcal{F}$ sequentially (and adaptively) so as to maximize the expected function value from the selected active items, i.e., $\mathbb{E}_{S,A} [f(S \cap A)]$ where $A \subseteq E$ is the random set of active items. Note that we can only select (and observe) a constrained subset of items, which corresponds

¹ These papers also handle a more general setting where items have an arbitrary outcome space (not just binary).

to semi-bandit feedback. When $\mathcal{F}$ is a single matroid constraint, Asadpour and Nazerzadeh (2016) obtained a $(1 - \frac{1}{e} - \epsilon)$ -approximation algorithm for the stochastic submodular maximization problem. Later, Gupta et al. (2017) and Bradac et al. (2019) showed that a ρ -approximation algorithm for *deterministic* submodular maximization under any constraint $\mathcal{F}$ can be used to obtain a $\frac{\rho}{2}$ -approximation algorithm for the stochastic problem under constraint $\mathcal{F}$. It is easy to see that this problem is down-monotone. So, Theorem 1 with $k = 2$ (binary r.v.s) implies a $(1 - \frac{1}{e} - \epsilon)$ -regret of $O(nF\sqrt{T\log T})$ for stochastic submodular maximization over a matroid constraint, where F is the maximum function value. Combined with the result of Bradac et al. (2019), we also obtain online learning algorithms for more general constraints $\mathcal{F}$. Notably, our result improves upon the prior result of Gabillon et al. (2013) that obtains $(1 - \frac{1}{e})$ -regret of $\tilde{O}(nFT^{2/3})$ in the semi-bandit feedback model for the stochastic submodular maximization problem when $\mathcal{F}$ corresponds to a cardinality constraint.

Stochastic Submodular Cover. In the classic submodular cover problem (Wolsey 1982), there is a monotone submodular function $f : 2^E \rightarrow \mathbb{Z}_+$ defined on a groundset E , where each item $i \in E$ has some cost c_i . We assume that the function is integer-valued with maximal value Q . The goal is to select a subset $S \subseteq E$ that “covers” the function, i.e., $f(S) = Q$. The objective is to minimize the total cost $\sum_{i \in S} c_i$ of the chosen items. Note that the submodular cover problem generalizes set cover. It is well-known that this problem can be approximated within a factor $1 + \ln(Q)$ (Wolsey 1982), and no better approximation is possible (Dinur and Steurer 2014). We consider a stochastic version, where each item $i \in E$ is *active* independently with probability p_i , and the function value is accrued only on active items. The goal now is to select items sequentially (and adaptively) so as to cover function f (or run out of items to select), at the minimum expected cost. This problem has been studied extensively (Golovin and Krause 2017, Hellerstein et al. 2021, Im et al. 2016), and Hellerstein et al. (2021) obtained a $1 + \ln(Q)$ approximation algorithm for it. Using the fact that the function f is monotone, it is easy to see that stochastic submodular cover is up-monotone. So, using Theorem 1 with $k = 2$, we obtain a polynomial time online learning algorithm for stochastic submodular cover having $(1 + \ln Q)$ -regret $O(nC\sqrt{T\log T})$ where $C = \sum_{i=1}^n c_i$ is the total cost.