ONLINE SUBMODULAR MINIMIZATION WITH SWITCHING COSTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider online submodular minimization with switching costs against an oblivious adversary under full-information and bandit feedback. Our algorithms use shared-threshold rounding to control movement. For a ground set of size n, horizon T, losses in $[-M, M]$, and switching-cost weight λ, the minimax expected regret under full information is $\Theta\big(M \min\{T, \sqrt{(1+\lambda/M)nT}\}\big)$. Under bandit feedback, a chain-traversal exploration scheme achieves $O\big(M \min\big\{T, n(1+\lambda/M)^{1/3}T^{2/3}\big\}\big)$, matching our lower bound up to logarithmic factors when $\lambda \geq M$. To obtain high-probability guarantees, we establish a pathwise upper bound on the realized regret that isolates a cumulative movement the learner can monitor. Using this quantity to trigger a switching-budget restarting wrapper yields high-probability bounds under both feedback models with only logarithmic overhead.

1 INTRODUCTION

We study online submodular minimization (OSM) (Hazan & Kale, 2012), a sequential decision-making problem with submodular losses. The learner repeatedly chooses a subset S_t from a finite ground set $[n] = \{1, \ldots, n\}$ and receives feedback about the loss function f_t. A loss $f : 2^{[n]} \to [-M, M]$ is submodular if, for any $A \subseteq B \subseteq [n]$ and $i \in [n] \setminus B$, $f(A \cup \{i\}) - f(A) \geq f(B \cup \{i\}) - f(B)$. The performance of the learner is often measured by regret, which is the difference between the cumulative loss of the learner's choices and that of the best fixed choice in hindsight. We consider this problem in two feedback settings: full-information and bandit. In the full-information setting, the learner observes the entire loss function after making a decision, while in the bandit setting, the learner only obtains the loss value of the chosen subset.

Applications of online submodular minimization include choosing a set of items in online profit maximization (Hazan & Kale, 2012) and selecting which products to discount in multi-product price optimization (Matsuoka et al., 2021). Switching decisions can be costly in such applications: replacing selected items or updating a price configuration consumes resources beyond the instantaneous loss.

An additional application is the assignment of interacting modules to two execution sites, using the execution-plus-communication cost structure of Stone (1977). Stone also discusses dynamic reassignment when working sets change and the refinement of cost estimates during execution. Here, we consider a sequential model in which each assignment precedes the realization of the period's workload and reassignment incurs a cost. Section B.1 gives the formal model and reduction.

Standard regret does not account for such switching costs. We therefore use regret with switching costs to evaluate the learner. For a time horizon T and a switching-cost weight $\lambda \geq 0$,

$$\mathcal{R}_T(\lambda) := \sum_{t=1}^{T} f_t(S_t) - \min_{S \subseteq [n]} \sum_{t=1}^{T} f_t(S) + \lambda \sum_{t=2}^{T} |S_t \,\Delta\, S_{t-1}|, \tag{1}$$

where the Hamming movement $|S_t \,\Delta\, S_{t-1}|$ is the size of the symmetric difference between consecutive subsets. The last term, $\lambda \sum_{t=2}^{T} |S_t \,\Delta\, S_{t-1}|$, quantifies the total switching cost.

Following Altschuler & Talwar (2018), we assume an oblivious adversary: the loss sequence is fixed in advance, although it may vary arbitrarily across rounds. This setting covers applications where

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.

production costs or demand conditions evolve exogenously without reacting to the learner's past actions. Moreover, if the adversary is adaptive to the learner's past actions, it can force the regret with switching costs to be at least $(T-1)\min\{M,\lambda\}/2$; see Proposition 25.

Our full-information guarantee also yields policy-regret bounds for bounded losses with finite memory (Anava et al., 2015), provided their diagonal losses are submodular and can be evaluated through counterfactual diagonal feedback (Corollary 24).

Controlling switching costs, however, creates a tension with high-probability guarantees. The submodular subgradient method of Hazan & Kale (2012) uses independent threshold rounding and attains the minimax standard-regret rate, including with high probability, but can incur $\Omega(nT)$ total Hamming movement even when the fractional iterates never move (Proposition 26). Instead, sharing a single threshold throughout yields sublinear expected regret with switching costs, but can incur linear regret with probability $\Omega(T^{-1/2})$ on some instance (Proposition 27). Thus, simple variants of existing methods struggle to provide high-probability guarantees for sublinear regret with switching costs.

1.1 Our Contributions

Results. We introduce online submodular minimization with switching costs. Under full-information feedback, shared-threshold SSD together with a fixed action in the trivial regime achieves $O\Big(M\min\{T,\sqrt{(1+\lambda/M)nT}\}\Big)$, matching our lower bound up to constants. Under bandit feedback, cyclic exploration along the maximal chain associated with the fractional iterate gives the expected upper bound $O\big(M\min\{T,n(1+\lambda/M)^{1/3}T^{2/3}\}\big)$ for the regret $\mathcal{R}_T^X(\lambda)$ of the played sets. A lower bound matches this rate up to logarithmic factors for $\lambda \geq M$. We also prove high-probability bounds under both feedback settings. Table 1 summarizes all bounds.

Table 1: Upper and lower bounds for $\mathcal{R}_T(\lambda)$ ($\mathcal{R}_T^X(\lambda)$ under bandit feedback). Here $\bar{\lambda} := \lambda/M$, and w.h.p. means with probability at least $1-\delta$.

Feedback	Result	Regret bound
Full information	Upper, in expectation (Theorem 2)	$O\Big(M\min\{T,\sqrt{(1+\bar{\lambda})nT}\}\Big)$
	Upper, w.h.p. (Theorem 6)	$O\Big(M\min\Big\{T,\sqrt{(1+\bar{\lambda})nT\log(1/\delta)}\Big\}\Big)$
	Lower, in expectation (Theorem 9)	$\Omega\Big(M\min\{T,\sqrt{(1+\bar{\lambda})nT}\}\Big)$
Bandit	Upper, in expectation (Theorem 3)	$O\big(M\min\{T,n(1+\bar{\lambda})^{1/3}T^{2/3}\}\big)$
	Upper, w.h.p. (Theorem 8)	$O\Big(M\min\{T,n(1+\bar{\lambda})^{1/3}T^{2/3}\log^{1/3}(3/\delta)\}\Big)$
	Lower, in expectation[a] (Theorem 10)	$\widetilde{\Omega}\Big(M\min\Big\{T,\max\Big\{n\sqrt{T},\bar{\lambda}^{1/3}nT^{2/3}\Big\}\Big\}\Big)$

Explicit finite-time high-probability bounds are given in Equations (35) and (45).
[a]Only logarithmic factors in T are suppressed. The bound holds for all $n \geq 1$, $T \geq 2$, and $\lambda \geq 0$; see Proposition 22 for explicit constants.

Techniques. Our algorithms run the submodular subgradient descent (SSD) of Hazan & Kale (2012) on the Lovász extension and round its fractional iterates with a single random threshold drawn once. For bandit feedback, a natural baseline is the roundwise SSD formulation of Lattimore (2026, Algorithm 13.1), modified to share the exploitation threshold across rounds. On each exploration round, it independently samples one set uniformly from the maximal chain induced by the fractional iterate. The sampled set may differ from each neighboring played set in up to n elements, so a single exploration round can incur $O(n)$ movement. Our key improvement is to incur only $O(n)$ movement per exploration block of $n+1$ rounds, rather than per exploration round. We keep the fractional

iterate fixed within each block and traverse its induced maximal chain in cyclic order, starting from a uniformly random set in the chain. The total movement is $O(n)$: adjacent sets in the chain differ by one element, there is at most one transition from $[n]$ to $\emptyset$, and each block-boundary transition incurs at most n units of Hamming movement. The random starting point makes every set in the chain equally likely at each round. With exploration probability γ, the modified roundwise SSD has γT expected exploration rounds, whereas our blockwise method has at most $\gamma T/(n+1)$ expected exploration blocks. The corresponding exploration movement bounds are therefore $O(n\gamma T)$ and $O(\gamma T)$, respectively, yielding the improved dimension dependence of cyclic-chain bandit SSD.

Although shared-threshold rounding controls expected movement, it does not by itself yield high-probability regret guarantees (Proposition 27). To overcome this limitation, we establish a pathwise upper bound on realized regret that isolates a cumulative movement quantity observable to the learner (Lemma 4). The restarting wrapper of Altschuler & Talwar (2018) uses this quantity as its epoch monitor. We apply the wrapper to shared-threshold SSD under full-information feedback and to cyclic-chain SSD under bandit feedback, obtaining high-probability guarantees in both settings.

For the bandit lower bound, we extend the multi-scale random-walk construction of Dekel et al. (2014) by embedding independent hidden signs in all coordinates. Comparing environments that differ in one sign bounds the information about that sign by movement in the corresponding coordinate. Conditioning the Gaussian construction on a high-probability admissibility event gives bounded modular losses while preserving a switching-dependent lower bound with linear dependence on n.

2 Related Work

Online submodular minimization. For unconstrained online submodular minimization, Hazan & Kale (2012) gave computationally efficient full-information and bandit algorithms with expected-regret bounds $O(M\sqrt{nT})$ and $O(MnT^{2/3})$, respectively, and also established high-probability regret guarantees. Their Submodular Follow-The-Perturbed-Leader (FTPL) uses a fixed random modular perturbation, while their subgradient methods use the Lovász extension. Jegelka & Bilmes (2011) studied the constrained variant over combinatorial structures. The algorithm of Fokkema et al. (2024), with the refined analysis of Lattimore (2026, Theorem 11.8 and Section 13.2), yields a computationally efficient high-probability regret bound of $\widetilde{O}(Mn^{5/2}\sqrt{T})$ for OSM without switching costs. This guarantee is complementary for small switching weights: running the same algorithm adds at most $\lambda n(T-1)$ in switching costs. Our matching bandit bounds concern $\lambda \geq M$; the dimension dependence for small λ, including $\lambda = 0$, remains open. Ito (2022) obtained stochastic gap-dependent and best-of-both-worlds guarantees under both feedback settings, and Matsuoka et al. (2021) studied tracking regret against time-varying comparators. In contrast to tracking regret, our comparator remains fixed and the learner's Hamming movement is explicitly penalized. We review classical algorithms for offline submodular function minimization in Section D.

Online submodular maximization. Early work studied online monotone submodular maximization in resource-allocation settings (Streeter & Golovin, 2008). Unlike unconstrained submodular minimization, submodular maximization is generally computationally intractable, so efficient online algorithms are typically evaluated using approximation regret. For unconstrained maximization of nonnegative, not necessarily monotone submodular functions, Roughgarden & Wang (2018) gave a polynomial-time no-$1/2$-regret algorithm and showed that the factor $1/2$ is optimal in the value-oracle model. Tracking-regret formulations cover both constrained monotone and unconstrained maximization variants (Matsuoka et al., 2021). Our work instead concerns exact regret for unconstrained minimization, exploiting convexity of the Lovász extension (Lovász, 1983).

Online learning with switching costs. For finite action sets, Cesa-Bianchi et al. (2013) studied switching costs within a hierarchy of adaptive adversaries. In the multi-armed bandit setting, Dekel et al. (2014) established the $\widetilde{\Theta}(T^{2/3})$ minimax rate caused by switching costs. Koren et al. (2017a;b) studied bandit learning with metric-dependent movement costs. Switching costs proportional to the Hamming distance between consecutive actions have been studied for adversarial combinatorial bandits with fixed-cardinality actions, linear losses and bandit or semi-bandit feedback (Dong & Tan, 2024). In contrast, we consider unconstrained minimization of submodular functions over all

subsets, under bandit or full-information feedback. For prediction with expert advice and adversarial bandits, Altschuler & Talwar (2018) characterized regret under switching budgets and switching costs, and gave the first high-probability bounds for both regret and the number of switches in the expert setting. Their restarting principle serves as the starting point for our high-probability regret analysis. Rouyer et al. (2021) obtained the minimax-optimal expected-regret rate in the oblivious adversarial regime and a gap-dependent bound with $T^{1/3}$ dependence in the stochastically constrained adversarial regime. For positive switching costs, the latter is not logarithmic and is not known to be minimax optimal.

3 Problem Setting and Preliminaries

Let $[n] = \{1, \dots, n\}$. In each round $t = 1, \dots, T$, the learner chooses a subset $S_t \subseteq [n]$, incurs loss $f_t(S_t)$, and then receives feedback about the submodular function $f_t : 2^{[n]} \to [-M, M]$. We assume an oblivious adversary: the sequence $(f_t)_{t=1}^T$ is fixed before the game begins. The switching-cost weight $\lambda \geq 0$ is known in advance, allowing parameter tuning. In both feedback settings, performance is measured by $\mathcal{R}_T(\lambda)$ in Equation (1). All omitted proofs are given in the appendix.

Our proposed method uses the Lovász extension, introduced by Lovász (1983). For $x \in [0,1]^n$ and $\tau \in [0,1]$, let $S_\tau(x) = \{i \in [n] : x_i > \tau\}$. The Lovász extension of f is $\hat{f}(x) := \mathbb{E}_{\tau \sim \mathrm{Unif}[0,1]}[f(S_\tau(x))]$. For any set $S \subseteq [n]$, let $\chi_S \in \{0,1\}^n$ denote its characteristic vector with $(\chi_S)_i = \mathbf{1}\{i \in S\}$. The extension $\hat{f}$ satisfies $\hat{f}(\chi_S) = f(S)$ and is convex if and only if f is submodular (Lovász, 1983). Let π be a permutation of $[n]$ such that $x_{\pi(1)} \geq \cdots \geq x_{\pi(n)}$, and define $B_0 = \emptyset$ and $B_k = \{\pi(1), \dots, \pi(k)\}$ for $k \in [n]$. A subgradient $g \in \mathbb{R}^n$ of $\hat{f}$ at x is given by

$$g_{\pi(k)} := f(B_k) - f(B_{k-1}), \qquad k = 1, \dots, n. \tag{2}$$

Lemma 1. *For every submodular function $f : 2^{[n]} \to [-M, M]$, the subgradient defined in Equation* (2) *satisfies $\|g\|_1 \leq 4M$.*

This bound follows from Jegelka & Bilmes (2011, Lemma 1).

4 Algorithms and Expected-Regret Guarantees

We first give shared-threshold submodular subgradient descent under both feedback models.

4.1 Full-information feedback

Our full-information algorithm uses a fixed action throughout the horizon when $n(1 + \lambda/M) \geq 2T$. Otherwise, it modifies the submodular subgradient descent (SSD) of Hazan & Kale (2012) by reusing a single random threshold throughout the horizon. In this regime, initialize $x_1 = \frac{1}{2}\mathbf{1} \in [0,1]^n$, draw $\tau \sim \mathrm{Unif}[0,1]$ once, and play $S_t = S_\tau(x_t) = \{i \in [n] : x_{t,i} > \tau\}$. After observing f_t, compute the subgradient g_t at x_t as in Equation (2), using tie-breaking independent of τ, and update

$$x_{t+1} = \Pi_{[0,1]^n}(x_t - \eta g_t). \tag{3}$$

The pseudocode for the full-information algorithm is in Algorithm 1.

Theorem 2. *For any $\lambda \geq 0$, the full-information algorithm, with an optimally tuned η in the shared-threshold branch, guarantees*

$$\mathbb{E}[\mathcal{R}_T(\lambda)] \leq \min\left\{2MT, \sqrt{2M(M+\lambda)nT}\right\} = O\Big(M \min\left\{T, \sqrt{(1+\lambda/M)nT}\right\}\Big).$$

For a shared threshold and deterministic $x, y \in [0,1]^n$, $\mathbb{E}_\tau[|S_\tau(x) \,\Delta\, S_\tau(y)|] = \|x - y\|_1$. Since the fractional iterates $(x_t)_{t=1}^T$ are independent of τ, $\mathbb{E}[\sum_{t=2}^T |S_t \,\Delta\, S_{t-1}|] \leq \eta \sum_{t<T} \|g_t\|_1 \leq 4M\eta T$ by Lemma 1. Combining this with the standard projected-subgradient regret bound and tuning η proves Theorem 2.

4.2 BANDIT FEEDBACK

Our starting point is the roundwise reformulation of Hazan–Kale bandit SSD given by Lattimore (2026, Algorithm 13.1), modified to use a single threshold τ on all exploitation rounds. At each round t, independently draw $E_t \sim \mathrm{Ber}(\gamma)$, and let X_t denote the played set. If $E_t = 1$, sample one of the $n+1$ sets in the maximal chain induced by x_t uniformly; otherwise set $X_t = S_\tau(x_t)$. Only exploration losses are used to construct the importance-weighted subgradient estimator. Consequently, $(x_t)_{t=1}^T$ is independent of τ. Sharing τ controls the movement of the sequence $(S_\tau(x_t))_{t=1}^T$. Since $|X_t \,\Delta\, S_\tau(x_t)| \le nE_t$, the triangle inequality gives

$$\sum_{t=2}^{T} |X_t \,\Delta\, X_{t-1}| \le \sum_{t=2}^{T} |S_\tau(x_t) \,\Delta\, S_\tau(x_{t-1})| + 2n \sum_{t=1}^{T} E_t.$$

The second term has expectation $2n\gamma T$.

Our improvement is to incur $O(n)$ movement per exploration *block* of $n+1$ rounds, instead of per exploration *round*. Put $m := \lfloor T/(n+1) \rfloor$, $T' := (n+1)m$, and $r := T - T'$. If $m = 0$, use the fixed-action policy. Henceforth assume $m \ge 1$ and partition the first T' rounds into m consecutive blocks of $n+1$ rounds. Set $x_1 = \frac{1}{2}\mathbf{1}$ and draw $\tau \sim \mathrm{Unif}[0,1]$ once. x_b is updated only at the end of each complete block of $n+1$ rounds. At the start of block b, sort the coordinates of x_b using deterministic tie-breaking independent of τ, obtaining a permutation π_b, and define the induced maximal chain by $B_{b,j} := \{\pi_b(1), \dots, \pi_b(j)\}, \qquad j = 0, \dots, n$. Let

$$v_{b,j} := \mathbf{1}\{j \ne 0\} e_{\pi_b(j)} - \mathbf{1}\{j \ne n\} e_{\pi_b(j+1)}, \qquad j = 0, \dots, n.$$

For every round t in block b, let $g_t(x_b)$ denote the subgradient of $\hat{f}_t$ at x_b obtained from π_b via Equation (2). Independently draw $E_b \sim \mathrm{Ber}(\gamma)$ and $J_b \sim \mathrm{Unif}\{0, \dots, n\}$ and set $J_{b,s} := (J_b + s) \bmod (n+1)$ for $s = 0, \dots, n$. Writing $t_b := (b-1)(n+1) + 1$ for the first round of block b, at round $t_b + s$ we play

$$X_{t_b+s} = \begin{cases} B_{b,J_{b,s}}, & E_b = 1, \\ S_\tau(x_b), & E_b = 0, \end{cases} \qquad s = 0, \dots, n. \tag{4}$$

After the block, we form

$$\widehat{G}_b := \frac{(n+1)E_b}{\gamma} \sum_{s=0}^{n} f_{t_b+s}(X_{t_b+s}) v_{b,J_{b,s}}, \qquad x_{b+1} := \Pi_{[0,1]^n}(x_b - \eta \widehat{G}_b). \tag{5}$$

For the remaining r rounds, we keep the last played action $X_{T'}$ and perform no updates.

The random shift has two roles. First, for every fixed position s, $J_{b,s}$ is uniform on $\{0, \dots, n\}$, so $B_{b,J_{b,s}}$ is uniform over the $n+1$ sets in the chain. Consequently,

$$\mathbb{E}_{E_b, J_b}[\widehat{G}_b \mid x_b] = \sum_{s=0}^{n} g_{t_b+s}(x_b),$$

even when the losses vary within the block.

Second, conditional on $E_b = 1$, the played sets visit the chain in cyclic order. Their internal Hamming movement is at most $2n$, the length of the closed walk

$$B_{b,0}, B_{b,1}, \dots, B_{b,n}, B_{b,0}.$$

Let $S_b := S_\tau(x_b)$ denote the set that would be played throughout block b under exploitation. Connecting the first and last sets of the traversal to S_b costs at most n on each side. Thus, an exploration block adds at most $4n$ movement relative to the sequence $(S_b)_{b=1}^m$.

Since $\mathbb{E}[\sum_{b=1}^m E_b] = \gamma m$, the expected exploration contribution is at most $4n\gamma m \le 4\gamma T$. By comparison, roundwise exploration incurs $O(n\gamma T)$ expected movement.

Finally, because the loss sequence is oblivious and Equation (5) uses only exploration observations, $(x_b)_{b=1}^{m+1}$ is independent of τ. The shared-threshold identity therefore controls the expected movement of $(S_b)_{b=1}^m$. Let $\mathcal{R}_T^X(\lambda)$ denote Equation (1) with X_t in place of S_t.

We run the cyclic-chain bandit algorithm when its derived upper bound is smaller than $2MT$; otherwise, we play $X_t = \emptyset$ throughout the horizon.

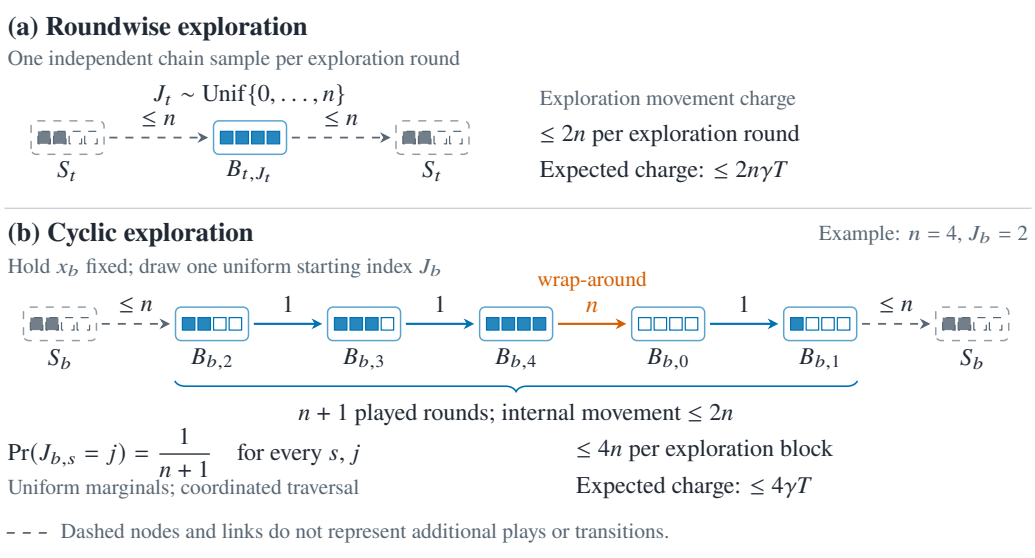


Figure 1: **Exploration movement: independent sampling versus cyclic traversal.** Filled cells denote selected elements in chain order ($n = 4$). (a) Independent sampling costs at most $2n$ movement per exploration round, yielding expected charge at most $2n\gamma T$. (b) A uniformly shifted cyclic traversal visits all $n + 1$ chain sets while keeping every position uniform. Including the two dashed boundary terms gives a bound of $4n$ per exploration block and expected charge at most $4\gamma T$. Dashed nodes and links do not represent additional plays or transitions. Exploitation-sequence movement is accounted for separately.

Theorem 3. *For every $\lambda \ge 0$, cyclic-chain bandit shared-threshold SSD, with suitably chosen $\eta > 0$ and $\gamma \in (0, 1]$, guarantees*

$$\mathbb{E}[\mathcal{R}_T^X(\lambda)] = O\Big(M \min\Big\{T, n(1 + \lambda/M)^{1/3} T^{2/3}\Big\}\Big).$$

A regret bound for arbitrary $\eta > 0$ and $\gamma \in (0, 1]$, together with its optimization and proof, is given in Appendix A.5. Pseudocode is given in Algorithm 2.

5 High-Probability Regret Guarantees

We apply the restarting wrapper of Altschuler & Talwar (2018) to a base algorithm. Following their terminology, we call the resulting method a *batched algorithm* and denote its output at round t by A_t. Here $A_t = S_t$ under full-information feedback and $A_t = X_t$ under bandit feedback. For epoch e starting at t_e, let $A_s^{(e)} := A_{t_e+s-1}$ and $C_s^{(e)} := \sum_{r=2}^{s} |A_r^{(e)} \,\Delta\, A_{r-1}^{(e)}|$.

The wrapper monitors a feedback-specific quantity $V_s^{(e)}$ to determine when to restart. Given a budget $B > 0$, the wrapper restarts the base algorithm after this quantity first exceeds B; let I_e and $L_e := |I_e|$ denote the resulting epoch and its length. At each restart, a fresh threshold $\tau^{(e+1)}$ is drawn and we reset $x_{t_{e+1}} = \frac{1}{2}\mathbf{1}$.

The feedback-specific analysis relates the switching cost to this monitor: for every fixed comparator $S \subseteq [n]$, it provides a decomposition of the form

$$\sum_{t \in I_e} \big(f_t(A_t) - f_t(S)\big) + \lambda C_{L_e}^{(e)} \le \kappa V_{L_e}^{(e)} + R_e(S), \tag{6}$$

where $\kappa \ge 0$ is feedback-specific and $R_e(S)$ is a residual term. Let $\mathcal{F}_{e-1}$ denote the history before epoch e. We assume that $V_s^{(e)}$ is nonnegative, adapted, observable, and nondecreasing in s for integer $s \ge 0$, and that, for a nondecreasing function Φ,

$$0 = V_0^{(e)} \le V_s^{(e)} \le V_{s+1}^{(e)}, \qquad \mathbb{E}[V_s^{(e)} \mid \mathcal{F}_{e-1}] \le \Phi(s). \tag{7}$$

Choosing $B = 4\Phi(H)$ with $H := 4T/N_\delta$ and $N_\delta = \Theta(\log(1/\delta))$ gives $E = O(\log(1/\delta))$ with high probability; see Proposition 15.

To apply the restarting wrapper to our methods, we establish the following inequality to obtain a pathwise upper bound on realized regret that isolates a cumulative movement.

Lemma 4. *Let $\eta > 0$, $a_t \in [-G, G]$, $u_1 = 1/2$, and $u_{t+1} = \Pi_{[0,1]}(u_t - \eta a_t)$. Let $\tau \sim \mathrm{Unif}[0,1]$ be shared across time, and set $z_t = \mathbf{1}\{u_t > \tau\}$. For any horizon L, assume that $\Pr(\exists t \le L : u_t = \tau) = 0$. Then, for every $b \in \{0,1\}$, almost surely,*

$$\sum_{t=1}^{L} a_t(z_t - b) \le \frac{1}{2\eta} + G\left(1 + \sum_{t=1}^{L-1} |z_{t+1} - z_t|\right).$$

Applied coordinatewise, this inequality bounds the realized regret by cumulative Hamming movement. The proof is deferred to Appendix A.6.

5.1 Full-information feedback

We call the wrapper applied to full-information shared-threshold SSD batched SSD. The wrapper monitors $V_s^{(e)} := C_s^{(e)}$.

Proposition 5. *For batched SSD, for every epoch e and every $S \subseteq [n]$, almost surely, $\sum_{t\in I_e}\big(f_t(S_t) - f_t(S)\big) + \lambda C_{L_e}^{(e)} \le (2M+\lambda)C_{L_e}^{(e)} + \frac{n}{2\eta} + 2Mn$.*

Thus Equation (6) holds with $\kappa = 2M+\lambda$ and $R_e(S) = n/(2\eta)+2Mn$. Since the algorithm shares a threshold throughout an epoch, we have $\mathbb{E}[C_s^{(e)} \mid \mathcal{F}_{e-1}] \le 4M\eta s$, so Equation (7) holds with $\Phi(s) = 4M\eta s$ and the wrapper restarts once the epoch movement exceeds $16M\eta H$. The final full-information method compares the resulting explicit high-probability bound with the deterministic bound $2MT$ of the fixed-action policy and uses whichever guarantee is smaller.

Theorem 6. *For every $\delta \in (0, 1/2)$, the full-information high-probability algorithm, using batched SSD with suitably chosen parameters or the fixed action in the trivial regime, satisfies, with probability at least $1-\delta$,*

$$\mathcal{R}_T(\lambda) = O\Big(M \min\Big\{T, \sqrt{(1+\lambda/M)nT\log(1/\delta)}\Big\}\Big).$$

5.2 Bandit feedback

For bandit feedback, the restarting wrapper operates on the $(n+1)$-round blocks of Section 4.2, rather than on rounds. We retain the block construction and notation from that section. At the start of epoch e, we set $x_1^{(e)} = \frac{1}{2}\mathbf{1}$ and draw $\tau^{(e)} \sim \mathrm{Unif}[0,1]$. Let $t_q^{(e)} := t_e + (q-1)(n+1)$ be the first round of the qth block of epoch e. The block is played and its estimator $\widehat{G}_q^{(e)}$ is formed exactly as in Section 4.2, using $x_q^{(e)}$ and $\tau^{(e)}$. Define $S_q^{(e)} := S_{\tau^{(e)}}(x_q^{(e)})$, the set that would be played throughout block q under exploitation. After forming $\widehat{G}_q^{(e)}$, define

$$\widetilde{x}_{q+1}^{(e)} := \Pi_{[0,1]^n}\big(x_q^{(e)} - \eta\widehat{G}_q^{(e)}\big), \qquad \widetilde{S}_{q+1}^{(e)} := S_{\tau^{(e)}}(\widetilde{x}_{q+1}^{(e)}).$$

The wrapper monitors

$$V_q^{(e)} := \sum_{j=1}^{q} \big|\widetilde{S}_{j+1}^{(e)} \,\Delta\, S_j^{(e)}\big|, \qquad V_0^{(e)} := 0. \tag{8}$$

After block q, both $S_q^{(e)}$ and $\widetilde{S}_{q+1}^{(e)}$ are computable from $x_q^{(e)}$, $\widehat{G}_q^{(e)}$, and $\tau^{(e)}$. Hence the updated monitor $V_q^{(e)}$ is observable even when block q is an exploration block. The budget is checked only after the block has ended. If $V_q^{(e)}$ does not exceed the budget, set $x_{q+1}^{(e)} := \widetilde{x}_{q+1}^{(e)}$. Otherwise, discard $\widetilde{x}_{q+1}^{(e)}$ and begin a new epoch at the next block. Thus every exploration block completes its cyclic traversal before a restart.

For an epoch containing L_e blocks, let $N_e := \sum_{q=1}^{L_e} E_q^{(e)}$ be its number of exploration blocks, and let C_e^X be the Hamming movement between consecutive played sets within epoch e. For

$q = 1, \dots, L_e$, define $G_q^{(e)} := \sum_{s=0}^{n} g_{t_q^{(e)}+s}(x_q^{(e)})$ and, for every fixed $S \subseteq [n]$, $Z_e(S) := \sum_{q=1}^{L_e} \left\langle G_q^{(e)} - \widehat{G}_q^{(e)}, \chi_{S_q^{(e)}} - \chi_S \right\rangle$.

Put $K_\gamma := \frac{2(n+1)M}{\gamma}$, and $Q_\lambda := 2M(n+1) + 4\lambda n$. Then $\|\widehat{G}_q^{(e)}\|_\infty \le K_\gamma$, while Q_λ bounds the additional loss and movement contribution of each exploration block.

Proposition 7. *For every epoch e and every fixed $S \subseteq [n]$, almost surely,*

$$\sum_{t \text{ in epoch } e} \big(f_t(X_t) - f_t(S)\big) + \lambda C_e^X \le (K_\gamma + \lambda) V_{L_e}^{(e)} + Q_\lambda N_e + \frac{n}{2\eta} + nK_\gamma + Z_e(S).$$

Thus the generic residual in Equation (6) is $R_e(S) = Q_\lambda N_e + n/(2\eta) + nK_\gamma + Z_e(S)$, with $\kappa = K_\gamma + \lambda$.

Let $\mathcal{F}_{e-1}$ denote the history before the threshold $\tau^{(e)}$ is drawn. Consider the process that starts from the same state and accepts every proposed update without enforcing the restart budget, and let $V_q^{\mathrm{cont},(e)}$ denote its monitor. Then

$$\mathbb{E}\Big[V_q^{\mathrm{cont},(e)} \mid \mathcal{F}_{e-1}\Big] \le 2\eta n(n+1)Mq =: \Phi(q).$$

This process agrees with the actual epoch up to and including the first budget-crossing block, after which it continues without restarting.

Set $H := 4m/N_\delta$ and $\beta := 4\Phi(H)$, with N_δ as in Equation (38). Applying Proposition 15 with m blocks in place of T rounds gives $\mathbb{P}(E > N_\delta) \le \delta/3$. Each monitor increment is at most n, so whenever $E \le N_\delta$, $\sum_{e=1}^{E} V_{L_e}^{(e)} \le N_\delta(\beta + n)$.

We then sum Proposition 7 over epochs and apply concentration inequalities to the total number of exploration blocks $\sum_{e=1}^{E} N_e$ and the cumulative estimation error $\sum_{e=1}^{E} Z_e(S)$. Movement across epoch boundaries adds at most $\lambda n(E-1)$ to the switching cost; the final r rounds add at most $2Mr$ to the loss regret and incur no further movement.

Finally, we compare the restarted procedure with the deterministic policy $X_t = \emptyset$, whose regret is at most $2MT$.

Theorem 8. *For every $\delta \in (0,1)$, the resulting bandit policy, with an appropriately chosen learning rate, exploration probability, and restart budget, satisfies, with probability at least $1-\delta$,*

$$\mathcal{R}_T^X(\lambda) = O\Big(M \min\Big\{T, n(1+\lambda/M)^{1/3} T^{2/3} \log^{1/3}(3/\delta)\Big\}\Big).$$

The proof, including the blockwise restart rule and all finite-time terms, is given in Appendix A.7.

6 Lower Bounds

Theorem 9. *For every randomized full-information OSM algorithm and every $\lambda \ge 0$, there is a deterministic oblivious sequence of modular, hence submodular, losses $f_t : 2^{[n]} \to [-M, M]$ such that $\mathbb{E}[\mathcal{R}_T(\lambda)] = \Omega\big(M \min\{T, \sqrt{(1+\lambda/M)nT}\}\big)$.*

The proof applies a switching-cost lower bound for two-expert problems separately to each coordinate. In the two-expert construction, losses remain constant over blocks of length proportional to λ/M when λ is large relative to M. Combining this construction with the lower bound of Hazan & Kale (2012, Theorem 14) for full-information OSM without switching costs ($\lambda = 0$) proves the theorem. The full proof is deferred to Appendix A.8.

Under bandit feedback, the following bound captures both the ordinary loss and the switching costs.

Theorem 10 (Bandit-feedback lower bound)**.** *Let $n \ge 1$, $T \ge 2$, $M > 0$, and $\lambda \ge 0$. For every randomized bandit-feedback OSM algorithm, there is a deterministic oblivious sequence of modular losses $f_t : 2^{[n]} \to [-M, M]$ such that*

$$\mathbb{E}[\mathcal{R}_T^X(\lambda)] = \widetilde{\Omega}\Big(M \min\Big\{T, \max\Big\{n\sqrt{T},\ n(\lambda/M)^{1/3} T^{2/3}\Big\}\Big\}\Big). \tag{9}$$

Here $\widetilde{\Omega}$ hides only logarithmic factors in T.

Proof sketch. Draw a random vector $\theta = (\theta_1, \dots, \theta_n) \in \{-1, 1\}^n$ whose coordinates are i.i.d. Rademacher random variables, and set

$$f_t(S) = W_t + \varepsilon \sum_{i \in S} \theta_i, \qquad S^*_\theta = \{i : \theta_i = -1\},$$

where W_t is the multi-scale Gaussian random walk of Dekel et al. (2014):

$$W_0 = 0, \qquad W_t = W_{\rho(t)} + \xi_t, \qquad \rho(t) = t - 2^{\nu_2(t)},$$

where $2^{\nu_2(t)}$ is the largest power of two dividing t and the ξ_t are i.i.d. $N(0, \sigma^2)$ random variables. Write $x_t = \chi_{X_t}$ and $x^*_\theta = \chi_{S^*_\theta}$. Since $\theta_i(x_{t,i} - x^*_{\theta,i}) = |x_{t,i} - x^*_{\theta,i}|$ for every i, the common noise term W_t cancels and

$$f_t(X_t) - f_t(S^*_\theta) = \varepsilon \sum_{i=1}^{n} \theta_i (x_{t,i} - x^*_{\theta,i}) = \varepsilon \|x_t - x^*_\theta\|_1.$$

Define $D := \sum_{t=1}^{T} \|x_t - x^*_\theta\|_1$, and $C := \sum_{t=2}^{T} \|x_t - x_{t-1}\|_1$. Summing the preceding equality gives the pathwise identity $\mathcal{R}^X_T(\lambda) = \varepsilon D + \lambda C$.

The key step relates information about each hidden variable θ_i to movement in coordinate i. Before conditioning the walk to remain bounded, let P_θ be the observation-history law for a fixed deterministic learner, and let $\theta^{(i)}$ be obtained by replacing θ_i with $-\theta_i$. If $Y_t = f_t(X_t)$, with the conventions $x_0 = 0$ and $Y_0 = 0$, then

$$Y_t \mid Y_{1:t-1} \sim N\big(Y_{\rho(t)} + \varepsilon \langle \theta, x_t - x_{\rho(t)} \rangle, \sigma^2\big)\,.$$

The Gaussian chain rule therefore gives $\mathrm{KL}(P_\theta \| P_{\theta^{(i)}}) = \frac{2\varepsilon^2}{\sigma^2} \mathbb{E}_{P_\theta} \sum_{t=1}^{T} (x_{t,i} - x_{\rho(t),i})^2$
$\le \frac{2\varepsilon^2 L}{\sigma^2}(1 + \mathbb{E}_{P_\theta} C_i)$. Here $C_i := \sum_{t=2}^{T} |x_{t,i} - x_{t-1,i}|$ and $L := 1 + \lfloor \log_2 T \rfloor$. The inequality follows because each coordinate change is counted at most L times by the intervals $(\rho(t), t]$. Let $G := \{\max_{t \le T} |W_t| \le M/2\}$. Pinsker's inequality (Cover & Thomas, 2006, Lemma 11.6.1), together with a transfer from the Gaussian distribution to the distribution conditioned on G, shows that

$$\mathbb{E}[C \mid G] \le c \frac{n\sigma^2}{\varepsilon^2 L} \quad \Longrightarrow \quad \mathbb{E}[D \mid G] = \Omega(nT),$$

for a universal constant $c > 0$. Otherwise, the switching-cost term is already $\Omega(\lambda n \sigma^2 / (\varepsilon^2 L))$. With the parameter choices in Proposition 22, this yields $\widetilde{\Omega}\big(M \min\{T, n(\lambda/M)^{1/3} T^{2/3}\}\big)$.

Choosing $n\varepsilon \le M/2$ ensures that the loss functions conditioned on G take values in $[-M, M]$; the conditioning costs only logarithmic factors in T. Combining the resulting bound with the bandit lower bound without switching costs from Dani et al. (2007, Theorem 4.3) proves the theorem. The proof details are given in Proposition 22.

Together with Theorem 3, the theorem characterizes the minimax expected regret up to logarithmic factors when $\lambda \ge M$. For smaller λ, the bandit OSM dimension gap remains.

7 Conclusion

We established expected and high-probability guarantees for online submodular minimization with switching costs under both full-information and bandit feedback. In the full-information setting, our upper and lower bounds characterize the minimax expected regret. Under bandit feedback, we introduced cyclic-chain bandit SSD, which amortizes exploration movement by traversing a maximal chain over blocks of $n + 1$ rounds; together with our lower bound, it characterizes the minimax expected regret up to logarithmic factors when $\lambda \ge M$. Finally, for both feedback models, we derived a pathwise upper bound on realized regret that isolates an observable cumulative movement quantity. Using this quantity as the restart monitor yields high-probability guarantees with only logarithmic overhead.

An open question is the minimax dependence on dimension for small switching weights under bandit feedback, including the ordinary $\lambda = 0$ problem. Other promising directions include extensions to distributive lattices, unknown or time-varying switching weights, and a confidence-uniform algorithm whose parameters do not depend on the failure probability δ.

AI USE STATEMENT

In this work, we used generative AI tools to help develop parts of the conceptual framework, formulate some mathematical claims, provide ingredients for proving mathematical claims, assist in writing proofs, provide feedback on the experimental design, implement the proposed methods, assist with translation, and interpret experimental results. We did not use generative AI tools to generate synthetic datasets, develop the theoretical model, or propose or refine hypotheses. Tasks involving dataset cleaning or reformatting and qualitative or thematic data analysis were not applicable to this work. Additionally, we used generative AI tools to search for and identify relevant literature, draft portions of the manuscript, and edit the manuscript for clarity and readability. We reviewed all AI-assisted work, independently checked the AI-assisted mathematical claims and proofs, verified the identified literature, and tested the AI-assisted implementations. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The proofs and algorithmic details are provided in the appendices. Section C specifies the loss-generation procedures, hyperparameters, sources of randomness, and evaluation protocol used in our experiments. An code package containing the experimental implementations, configurations, recorded results, and reproduction instructions is provided in the supplementary material.

REFERENCES

Jason Altschuler and Kunal Talwar. Online learning over a finite action set with limited switching. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1569–1573. PMLR, 2018.

Oren Anava, Elad Hazan, and Shie Mannor. Online learning for adversaries with memory: Price of past mistakes. In *Advances in Neural Information Processing Systems*, volume 28, pp. 784–792. Curran Associates, Inc., 2015.

Nicolò Cesa-Bianchi, Ofer Dekel, and Ohad Shamir. Online learning with switching costs and other adaptive adversaries. In *Advances in Neural Information Processing Systems*, volume 26, pp. 1160–1168. Curran Associates, Inc., 2013.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, second edition, 2006.

Varsha Dani, Sham M. Kakade, and Thomas P. Hayes. The price of bandit information for online optimization. In *Advances in Neural Information Processing Systems*, volume 20, pp. 345–352. Curran Associates, Inc., 2007.

Ofer Dekel, Jian Ding, Tomer Koren, and Yuval Peres. Bandits with switching costs: $T^{2/3}$ regret. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pp. 459–467. ACM, 2014.

Yanyan Dong and Vincent Y. F. Tan. Adversarial combinatorial bandits with switching costs. *IEEE Transactions on Information Theory*, 70(7):5213–5227, 2024.

Hidde Fokkema, Dirk van der Hoeven, Tor Lattimore, and Jack J. Mayo. Online newton method for bandit convex optimisation extended abstract. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1713–1714. PMLR, 2024.

David A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

Martin Grötschel, László Lovász, and Alexander Schrijver. The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1(2):169–197, 1981.

Elad Hazan and Satyen Kale. Online submodular minimization. *Journal of Machine Learning Research*, 13(93):2903–2922, 2012.

Shinji Ito. Revisiting online submodular minimization: Gap-dependent regret bounds, best of both worlds and adversarial robustness. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9678–9694. PMLR, 2022.

Satoru Iwata, Lisa Fleischer, and Satoru Fujishige. A combinatorial strongly polynomial algorithm for minimizing submodular functions. *Journal of the ACM*, 48(4):761–777, 2001.

Stefanie Jegelka and Jeff A. Bilmes. Online submodular minimization for combinatorial structures. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 345–352. International Machine Learning Society, 2011.

Tomer Koren, Roi Livni, and Yishay Mansour. Bandits with movement costs and adaptive pricing. In *Proceedings of the 30th Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 1242–1268. PMLR, 2017a.

Tomer Koren, Roi Livni, and Yishay Mansour. Multi-armed bandits with metric movement costs. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4119–4128. Curran Associates, Inc., 2017b.

Tor Lattimore. *Bandit Convex Optimisation*. Cambridge University Press, 2026.

Yin Tat Lee, Aaron Sidford, and Sam Chiu-Wai Wong. A faster cutting plane method and its implications for combinatorial and convex optimization. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, pp. 1049–1065. IEEE, 2015.

László Lovász. Submodular functions and convexity. In Achim Bachem, Martin Grötschel, and Bernhard Korte (eds.), *Mathematical Programming: The State of the Art*, pp. 235–257. Springer, Berlin, Heidelberg, 1983.

Tatsuya Matsuoka, Shinji Ito, and Naoto Ohsaka. Tracking regret bounds for online submodular optimization. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3421–3429. PMLR, 2021.

Francesco Orabona and Dávid Pál. Optimal non-asymptotic lower bound on the minimax regret of learning with expert advice. arXiv preprint arXiv:1511.02176, 2015.

James B. Orlin. A faster strongly polynomial time algorithm for submodular function minimization. *Mathematical Programming*, 118(2):237–251, 2009.

Tim Roughgarden and Joshua R. Wang. An optimal learning algorithm for online unconstrained submodular maximization. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1307–1325. PMLR, 2018.

Chloé Rouyer, Yevgeny Seldin, and Nicolò Cesa-Bianchi. An algorithm for stochastic and adversarial bandits with switching costs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9127–9135. PMLR, 2021.

Alexander Schrijver. A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *Journal of Combinatorial Theory, Series B*, 80(2):346–355, 2000.

Harold S. Stone. Multiprocessor scheduling with the aid of network flow algorithms. *IEEE Transactions on Software Engineering*, SE-3(1):85–93, 1977.

Matthew Streeter and Daniel Golovin. An online algorithm for maximizing submodular functions. In *Advances in Neural Information Processing Systems*, volume 21, pp. 1577–1584. Curran Associates, Inc., 2008.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, second edition, 2026.

A Deferred Proofs

We first give the auxiliary subgradient bound and the algorithmic pseudocode, then prove the expected guarantees, establish the pathwise regret decompositions, and analyze the restarting wrapper. We conclude with the lower-bound constructions.

A.1 Proof of the subgradient norm bound

Proof of Lemma 1. Recall that $B_0 = \emptyset$, $B_k = \{\pi(1), \dots, \pi(k)\}$, and $g_{\pi(k)} = f(B_k) - f(B_{k-1})$ as in Equation (2). Let $k_1 < \cdots < k_s$ be the indices for which $g_{\pi(k_\ell)} \geq 0$, and define

$$Q_0 := \emptyset, \qquad Q_\ell := \{\pi(k_1), \dots, \pi(k_\ell)\}, \quad \ell = 1, \dots, s.$$

Since $Q_{\ell-1} \subseteq B_{k_\ell - 1}$, submodularity gives

$$g_{\pi(k_\ell)} = f(B_{k_\ell}) - f(B_{k_\ell - 1}) \leq f(Q_\ell) - f(Q_{\ell-1}).$$

Hence the positive marginals telescope as

$$\sum_{k : g_{\pi(k)} \geq 0} g_{\pi(k)} \leq f(Q_s) - f(\emptyset) \leq M - f(\emptyset).$$

Moreover, $\sum_{k=1}^{n} g_{\pi(k)} = f([n]) - f(\emptyset)$. Therefore,

$$\begin{aligned} \|g\|_1 &= 2 \sum_{k : g_{\pi(k)} \geq 0} g_{\pi(k)} - \sum_{k=1}^{n} g_{\pi(k)} \\ &\leq 2M - f(\emptyset) - f([n]) \leq 4M, \end{aligned}$$

where the last inequality uses $f(\emptyset), f([n]) \geq -M$. □

A.2 Pseudocode for the full-information algorithm

Algorithm 1: Full-information algorithm

Input: loss bound $M > 0$; switching-cost weight $\lambda \geq 0$; learning rate $\eta > 0$
if $n(1 + \lambda/M) \geq 2T$ **then**
  Play the fixed action $S_t = \emptyset$ for $t = 1, \ldots, T$;
else
  Set $x_1 = \frac{1}{2}\mathbf{1}$ and draw $\tau \sim \mathrm{Unif}[0, 1]$ once;
  for $t = 1, \ldots, T$ **do**
    Play $S_t = \{i \in [n] : x_{t,i} > \tau\}$;
    Incur $f_t(S_t)$ and observe f_t;
    Construct a permutation π_t sorting x_t in nonincreasing order, with tie-breaking independent of τ;
    Set $B_{t,0} = \emptyset$;
    for $k = 1, \ldots, n$ **do**
      $B_{t,k} = B_{t,k-1} \cup \{\pi_t(k)\}$;
      $g_{t,\pi_t(k)} = f_t(B_{t,k}) - f_t(B_{t,k-1})$;
    Update $x_{t+1} = \Pi_{[0,1]^n}(x_t - \eta g_t)$;

A.3 Pseudocode for the bandit algorithm

Algorithm 2: Cyclic-chain bandit shared-threshold submodular subgradient descent

Input: horizon T; dimension n; loss bound $M > 0$; switching-cost weight $\lambda \geq 0$; learning rate $\eta > 0$; exploration probability $\gamma \in (0, 1]$
Initialize: $m = \lfloor T/(n+1) \rfloor$, $T' = (n+1)m$, $r = T - T'$, and

$$U_{\mathrm{cyc}}(\eta, \gamma) = \frac{n}{8\eta} + 2\eta n M \left(\frac{(n+1)M}{\gamma} + \lambda \right) T' + (2M + 4\lambda)\gamma T' + 2Mr$$

if $m = 0$ *or* $U_{\mathrm{cyc}}(\eta, \gamma) \geq 2MT$ **then**
  Play the fixed action $X_t = \emptyset$ for $t = 1, \ldots, T$;
else
  Set $x_1 = \frac{1}{2}\mathbf{1}$ and draw $\tau \sim \mathrm{Unif}[0, 1]$ once;
  for $b = 1, \ldots, m$ **do**
    Set $t_b = (b-1)(n+1) + 1$ and construct the chain $(B_{b,j})_{j=0}^{n}$ induced by x_b;
    Set $v_{b,j} = \mathbf{1}\{j \neq 0\} e_{\pi_b(j)} - \mathbf{1}\{j \neq n\} e_{\pi_b(j+1)}$ for $j = 0, \ldots, n$;
    Draw $E_b \sim \mathrm{Ber}(\gamma)$ and $J_b \sim \mathrm{Unif}\{0, \ldots, n\}$ independently;
    for $s = 0, \ldots, n$ **do**
      Set $J_{b,s} = (J_b + s) \bmod (n+1)$;
      Play $X_{t_b+s} = B_{b,J_{b,s}}$ if $E_b = 1$ and $X_{t_b+s} = S_\tau(x_b)$ otherwise;
      Observe $f_{t_b+s}(X_{t_b+s})$;
    Construct $\widehat{G}_b$ and update x_{b+1} according to Equation (5);
  if $r > 0$ **then**
    Keep $X_t = X_{T'}$ for $t = T' + 1, \ldots, T$ and perform no update;

A.4 Proof of the expected-regret bound with switching costs under full-information feedback

We first analyze the shared-threshold branch of Algorithm 1. Because the loss sequence is oblivious and the chain tie-breaking does not depend on τ, the continuous trajectory $(x_t, g_t)_{t=1}^{T}$ is independent of the rounding threshold. Moreover, $S_t = B_{t,k}$ for some $k \in \{0, \ldots, n\}$, where $(B_{t,k})_{k=0}^{n}$ is the

chain used to define g_t. The same vector g_t is a subgradient of $\hat{f}_t$ at χ_{S_t}. Hence, for every $S \subseteq [n]$,

$$f_t(S_t) - f_t(S) = \hat{f}_t(\chi_{S_t}) - \hat{f}_t(\chi_S) \leq \langle g_t, \chi_{S_t} - \chi_S \rangle. \tag{10}$$

In addition to $\|g_t\|_1 \leq 4M$ from Lemma 1, we use the coordinatewise bound $|g_{t,i}| \leq 2M$, which follows directly from the definition of g_t.

For every fixed comparator $S \subseteq [n]$, convexity and the projected-gradient inequality, together with $x_1 = \frac{1}{2}\mathbf{1}$, give

$$\begin{aligned}\sum_{t=1}^{T}\big(\hat{f}_t(x_t) - \hat{f}_t(\chi_S)\big) &\leq \frac{\|x_1 - \chi_S\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_2^2 \\ &\leq \frac{n}{8\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t\|_\infty \|g_t\|_1 \\ &\leq \frac{n}{8\eta} + 4M^2\eta T, \end{aligned} \tag{11}$$

where the last inequality uses the coordinatewise bound and Lemma 1. Since x_t is independent of τ, the definition of the Lovász extension gives $\mathbb{E}_\tau[f_t(S_\tau(x_t))] = \hat{f}_t(x_t)$ for every realization of x_t.

For deterministic $x, y \in [0,1]^n$, the shared-threshold rounding satisfies

$$\mathbb{E}_\tau[|S_\tau(x) \,\Delta\, S_\tau(y)|] = \|x - y\|_1, \tag{12}$$

since, for each $i \in [n]$,

$$\mathbb{P}_\tau(i \in S_\tau(x) \,\Delta\, S_\tau(y)) = \mathbb{P}_\tau(\tau \text{ lies between } x_i \text{ and } y_i) = |x_i - y_i|,$$

and summing over i proves Equation (12). Applying this identity with $x = x_t$ and $y = x_{t-1}$, and using Equation (3) yields

$$\begin{aligned}\mathbb{E}[\sum_{t=2}^{T}|S_t \,\Delta\, S_{t-1}|] &= \sum_{t=2}^{T}\mathbb{E}[\|x_t - x_{t-1}\|_1] \\ &\leq \eta\sum_{t=1}^{T-1}\|g_t\|_1 \leq 4M\eta(T-1). \end{aligned} \tag{13}$$

Combining the expected-loss bound with Equation (13), for every $\eta > 0$,

$$\mathbb{E}[\mathcal{R}_T(\lambda)] \leq \frac{n}{8\eta} + 4M(M+\lambda)\eta T. \tag{14}$$

The right-hand side of Equation (14) is minimized at

$$\eta^* = \sqrt{\frac{n}{32M(M+\lambda)T}}.$$

Thus shared-threshold SSD with the optimal learning rate satisfies the explicit bound

$$\mathbb{E}[\mathcal{R}_T(\lambda)] \leq \sqrt{2M(M+\lambda)nT} = M\sqrt{2(1+\lambda/M)nT}. \tag{15}$$

When $n(1+\lambda/M) < 2T$, this bound is smaller than $2MT$, and Algorithm 1 uses shared-threshold SSD with $\eta = \eta^*$. When $n(1+\lambda/M) \geq 2T$, the algorithm instead plays $\emptyset$ on every round. Its movement is then zero, and boundedness of the losses gives

$$\mathcal{R}_T(\lambda) = \sum_{t=1}^{T} f_t(\emptyset) - \min_{S\subseteq[n]}\sum_{t=1}^{T} f_t(S) \leq 2MT.$$

Combining the two regimes yields Theorem 2.

A.5 Proof of the expected-regret bound with switching costs under bandit feedback

Put $m := \lfloor T/(n+1) \rfloor$, $T' := (n+1)m$, and $r := T - T'$. We first analyze the m complete blocks. For arbitrary $\eta > 0$ and $\gamma \in (0, 1]$, we prove

$$\begin{aligned}\mathbb{E}[\mathcal{R}_T^X(\lambda)] &\le U_{\mathrm{cyc}}(\eta, \gamma) \\ &:= \frac{n}{8\eta} + 2\eta n M \left(\frac{(n+1)M}{\gamma} + \lambda \right) T' \\ &\quad + (2M + 4\lambda)\gamma T' + 2Mr. \end{aligned} \tag{16}$$

When $m = 0$ or $U_{\mathrm{cyc}}(\eta, \gamma) \ge 2MT$, Algorithm 2 plays a fixed action and satisfies $\mathcal{R}_T^X(\lambda) \le 2MT$ deterministically. We therefore analyze the cyclic learning branch below, where $m \ge 1$ and $U_{\mathrm{cyc}}(\eta, \gamma) < 2MT$.

Let $\mathcal{F}_{b-1}$ be the history just before drawing (E_b, J_b) in block b. It includes the shared threshold and all observations from earlier blocks, so x_b, π_b, and the chain $(B_{b,j})_{j=0}^n$ are $\mathcal{F}_{b-1}$-measurable. Recall that

$$v_{b,j} := \mathbf{1}\{j \ne 0\} e_{\pi_b(j)} - \mathbf{1}\{j \ne n\} e_{\pi_b(j+1)}, \qquad j = 0, \dots, n,$$

and, with $J_{b,s} := (J_b + s) \bmod (n+1)$,

$$\widehat{G}_b := \frac{(n+1)E_b}{\gamma} \sum_{s=0}^{n} f_{t_b+s}(B_{b,J_{b,s}}) v_{b,J_{b,s}}.$$

For $t = t_b + s$, let $g_t(x_b)$ be the chain subgradient of $\hat{f}_t$ induced by π_b, and set

$$G_b := \sum_{s=0}^{n} g_{t_b+s}(x_b).$$

Lemma 11 (Block estimator)**.** *For every complete block b,*

$$\mathbb{E}[\widehat{G}_b \mid \mathcal{F}_{b-1}] = G_b, \tag{17}$$

$$\mathbb{E}[\|\widehat{G}_b\|_2^2 \mid \mathcal{F}_{b-1}] \le \frac{4n(n+1)^2 M^2}{\gamma}, \tag{18}$$

$$\mathbb{E}[\|\widehat{G}_b\|_1 \mid \mathcal{F}_{b-1}] \le 2n(n+1)M. \tag{19}$$

Proof. Conditional on $\mathcal{F}_{b-1}$, for each fixed s the index $J_{b,s}$ is uniform on $\{0, \dots, n\}$. Moreover,

$$\sum_{j=0}^{n} f_t(B_{b,j}) v_{b,j} = \sum_{k=1}^{n} \big(f_t(B_{b,k}) - f_t(B_{b,k-1}) \big) e_{\pi_b(k)} = g_t(x_b). \tag{20}$$

Since E_b and J_b are independent and $\mathbb{E}[E_b] = \gamma$,

$$\begin{aligned}\mathbb{E}[\widehat{G}_b \mid \mathcal{F}_{b-1}] &= \frac{n+1}{\gamma} \gamma \sum_{s=0}^{n} \frac{1}{n+1} \sum_{j=0}^{n} f_{t_b+s}(B_{b,j}) v_{b,j} \\ &= \sum_{s=0}^{n} g_{t_b+s}(x_b) = G_b. \end{aligned}$$

Thus the estimator remains unbiased when the losses vary within a block. Notice that this is a blockwise statement: after conditioning on J_b, one cannot assert per-round unbiasedness.

On an exploration block the cyclic traversal visits every $j \in \{0, \dots, n\}$ exactly once. For each coordinate $i = \pi_b(k)$, only $v_{b,k}$ and $v_{b,k-1}$ contribute, once with sign $+1$ and once with sign -1. Since the corresponding loss values lie in $[-M, M]$,

$$\|\widehat{G}_b\|_\infty \le \frac{2(n+1)M}{\gamma}, \qquad \|\widehat{G}_b\|_2^2 \le \frac{4n(n+1)^2 M^2}{\gamma^2}, \qquad \|\widehat{G}_b\|_1 \le \frac{2n(n+1)M}{\gamma}.$$

The estimator is zero on an exploitation block. Multiplying the last two bounds by the exploration probability γ proves Equations (18) and (19). □

Proposition 12 (Expected loss regret over complete blocks)**.** *For every fixed $S \subseteq [n]$,*

$$\mathbb{E}\left[\sum_{t=1}^{T'}\big(f_t(X_t) - f_t(S)\big)\right] \le \frac{n}{8\eta} + \frac{2\eta n(n+1)M^2T'}{\gamma} + 2M\gamma T'. \tag{21}$$

Proof. Define the convex block loss

$$F_b(x) := \sum_{s=0}^{n} \hat{f}_{t_b+s}(x).$$

By construction, $G_b \in \partial F_b(x_b)$. Hence, with $u := \chi_S$, convexity and Equation (17) give

$$\begin{aligned}\mathbb{E}\left[\sum_{b=1}^{m}(F_b(x_b) - F_b(u))\right] &\le \mathbb{E}\left[\sum_{b=1}^{m}\langle G_b, x_b - u\rangle\right] \\ &= \mathbb{E}\left[\sum_{b=1}^{m}\langle \widehat{G}_b, x_b - u\rangle\right].\end{aligned}$$

The projected-gradient inequality, summed over blocks, yields pathwise

$$\sum_{b=1}^{m}\langle \widehat{G}_b, x_b - u\rangle \le \frac{\|x_1 - u\|_2^2}{2\eta} + \frac{\eta}{2}\sum_{b=1}^{m}\|\widehat{G}_b\|_2^2.$$

Since $x_1 = \frac{1}{2}\mathbf{1}$, $\|x_1 - u\|_2^2 = n/4$. Taking expectations and applying Equation (18), together with $(n+1)m = T'$, gives

$$\mathbb{E}\left[\sum_{b=1}^{m}(F_b(x_b) - F_b(u))\right] \le \frac{n}{8\eta} + \frac{2\eta n(n+1)M^2T'}{\gamma}. \tag{22}$$

The continuous block trajectory $(x_b)_{b=1}^{m+1}$ is independent of τ. Indeed, exploitation observations are not used in the update, while exploration actions and observations do not depend on τ. Therefore, for every block and every realization of x_b,

$$\mathbb{E}_\tau\left[\sum_{s=0}^{n} f_{t_b+s}(S_\tau(x_b)) \;\middle|\; x_b\right] = F_b(x_b).$$

Also, $F_b(u) = \sum_{s=0}^{n} f_{t_b+s}(S)$. If $E_b = 0$, the played action equals $S_\tau(x_b)$ throughout the block. If $E_b = 1$, their loss difference is at most $2M$ at each of the $n+1$ rounds. Consequently,

$$\sum_{s=0}^{n}\big(f_{t_b+s}(X_{t_b+s}) - f_{t_b+s}(S_\tau(x_b))\big) \le 2M(n+1)E_b.$$

Taking expectations, summing over blocks, and using $\mathbb{E}[\sum_{b=1}^{m} E_b] = \gamma m$ adds $2M(n+1)\gamma m = 2M\gamma T'$ to Equation (22), proving the claim. □

Proposition 13 (Expected movement under cyclic exploration)**.** *Let $C_T^X := \sum_{t=2}^{T} |X_t \,\Delta\, X_{t-1}|$. Then*

$$\mathbb{E}[C_T^X] \le 2\eta n M T' + 4\gamma T'. \tag{23}$$

Proof. Write $S_b := S_\tau(x_b)$ for the virtual exploitation action of block b. The closed chain walk

$$B_{b,0}, B_{b,1}, \ldots, B_{b,n}, B_{b,0}$$

has total Hamming length $2n$: the forward part adds each element once, and the closing edge removes all n elements. Every cyclic shift used by the algorithm therefore has internal movement at most $2n$. At the entrance to and exit from an exploration block, routing through its virtual action

S_b costs at most n on each side. Comparing the actual action sequence with the virtual sequence, which is constant inside each block, gives the pathwise bound

$$C_{T'}^X \le \sum_{b=2}^{m} |S_b \,\Delta\, S_{b-1}| + 4n \sum_{b=1}^{m} E_b. \tag{24}$$

The same estimate covers consecutive exploration blocks; charging an entrance and an exit to each such block only overcounts their common boundary.

Since the block trajectory is independent of τ, the shared-threshold identity and the coordinatewise nonexpansiveness of projection imply

$$\begin{aligned}\mathbb{E}\left[\sum_{b=2}^{m} |S_b \,\Delta\, S_{b-1}|\right] &= \mathbb{E}\left[\sum_{b=2}^{m} \|x_b - x_{b-1}\|_1\right] \\ &\le \eta \sum_{b=1}^{m-1} \mathbb{E}[\|\widehat{G}_b\|_1] \\ &\le 2\eta n(n+1)M(m-1) \le 2\eta n M T'.\end{aligned}$$

Moreover,

$$4n\mathbb{E}\left[\sum_{b=1}^{m} E_b\right] = 4n\gamma m \le 4\gamma T',$$

because $n \le n+1$. The action is held fixed during the final r rounds, so they add no movement. Taking expectations in Equation (24) proves the proposition. □

For the final r rounds, the fixed action incurs at most $2Mr$ additional loss relative to any fixed comparator and no additional movement. Adding λ times Equation (23) to Equation (21) therefore proves Equation (16). Together with the fixed-action branch, this shows that for every $\eta > 0$ and $\gamma \in (0, 1]$, Algorithm 2 satisfies

$$\mathbb{E}[\mathcal{R}_T^X(\lambda)] \le \min\{2MT, U_{\mathrm{cyc}}(\eta, \gamma)\}. \tag{25}$$

It remains to tune the parameters. Using $T' \le T$ and $r < n+1$, and optimizing Equation (16) over η for fixed γ gives

$$\mathbb{E}[\mathcal{R}_T^X(\lambda)] \le n\sqrt{M\left(\frac{(n+1)M}{\gamma} + \lambda\right)T} + (2M + 4\lambda)\gamma T + 2M(n+1). \tag{26}$$

By subadditivity of the square root, the first term is at most

$$nM\sqrt{\frac{(n+1)T}{\gamma}} + n\sqrt{M\lambda T}.$$

Put $a := nM\sqrt{(n+1)T}$ and $b := (2M + 4\lambda)T$. If $\gamma_0 := (a/(2b))^{2/3} \le 1$, choosing $\gamma = \gamma_0$ in Equation (26) and using $n + 1 \le 2n$ yields

$$\mathbb{E}[\mathcal{R}_T^X(\lambda)] = O\Big(nM^{2/3}(M+\lambda)^{1/3}T^{2/3} + n\sqrt{M\lambda T} + Mn\Big). \tag{27}$$

If $\gamma_0 > 1$, then

$$T < \frac{n^2(n+1)M^2}{16(M+2\lambda)^2} \le \frac{n^3}{8},$$

which is a regime in which the trivial $2MT$ bound from playing one fixed set throughout is sufficient.

Finally, let

$$A := nM^{2/3}(M+\lambda)^{1/3}T^{2/3}.$$

If $A \ge MT$, the comparison in Algorithm 2 guarantees expected regret at most $2MT = O(A)$. If $A < MT$, then $T > n^3(1 + \lambda/M)$, so in particular $\lambda < MT$, and

$$n\sqrt{M\lambda T} \le nM^{2/3}\lambda^{1/3}T^{2/3} \le A, \qquad Mn \le A.$$

Thus Equation (27), together with the fixed-action fallback, proves Theorem 3.

A.6 PROOFS OF THE PATHWISE REGRET BOUNDS

The restarting-wrapper analysis below relies on pathwise regret bounds within each epoch. We establish them here for both feedback models.

Proof of Lemma 4. First consider $b = 0$. Partition $\{t \in [L] : z_t = 1\}$ into its maximal contiguous intervals $I_j = [s_j, r_j]$, $j = 1, \dots, K$. Thus, $z_t = 1$ for every $t \in I_j$; moreover, whenever the adjacent indices lie in $[L]$, maximality gives $z_{s_j-1} = 0$ and $z_{r_j+1} = 0$. For $t = s_j, \dots, r_j - 1$, both u_t and u_{t+1} exceed τ, so projection onto the lower endpoint cannot occur. If projection onto the upper endpoint occurs, then $u_t - \eta a_t \geq u_{t+1} = 1$, and hence $a_t \leq (u_t - u_{t+1})/\eta$, and equality holds without projection. Thus, all terms except the last one telescope; bounding that last term by G gives

$$\sum_{t=s_j}^{r_j} a_t = \sum_{t=s_j}^{r_j-1} a_t + a_{r_j} \leq \frac{u_{s_j} - u_{r_j}}{\eta} + G < \frac{u_{s_j} - \tau}{\eta} + G.$$

If $s_j = 1$, this is at most $1/(2\eta) + G$. Otherwise, maximality gives $u_{s_j-1} \leq \tau < u_{s_j}$. Since projection is nonexpansive,

$$u_{s_j} - \tau \leq u_{s_j} - u_{s_j-1} \leq \eta |a_{s_j-1}| \leq \eta G.$$

Substituting this inequality into the preceding display shows that, whenever $s_j > 1$,

$$\sum_{t=s_j}^{r_j} a_t \leq \frac{\eta G}{\eta} + G = 2G.$$

Put $c := \sum_{t=1}^{L-1} |z_{t+1} - z_t|$, which counts the threshold crossings of u_t across τ, equivalently, the number of times z_t changes value. If $K = 0$, the claim is immediate. If $s_1 = 1$, then $2K - 1 \leq 1 + c$; hence

$$\sum_{t=1}^{L} a_t z_t \leq \frac{1}{2\eta} + G + 2G(K-1) \leq \frac{1}{2\eta} + G(1+c).$$

If $s_1 > 1$, then $2K \leq 1 + c$, and therefore

$$\sum_{t=1}^{L} a_t z_t \leq 2GK \leq G(1+c) \leq \frac{1}{2\eta} + G(1+c).$$

For $b = 1$, let $v_t := 1 - u_t$, $h_t := -a_t$, and $w_t := 1 - z_t$. Then $v_1 = 1/2$ and $v_{t+1} = \Pi_{[0,1]}(v_t - \eta h_t)$. Outside the probability-zero event that $u_t = \tau$ for some $t \leq L$, w_t is the threshold rounding of v_t at $1 - \tau$. Applying the $b = 0$ case gives the same bound, since

$$h_t w_t = a_t(z_t - 1), \qquad |w_{t+1} - w_t| = |z_{t+1} - z_t|.$$

Thus, for either $b \in \{0, 1\}$, summing over $t = 1, \dots, L$ yields

$$\sum_{t=1}^{L} a_t(z_t - b) \leq \frac{1}{2\eta} + G\left(1 + \sum_{t=1}^{L-1} |z_{t+1} - z_t|\right),$$

as claimed. □

Proof of Proposition 5. Fix an epoch e and restrict the following argument to $t \in I_e$. By Equation (10), it suffices to control $\sum_{t \in I_e} \langle g_t, \chi_{S_t} - \chi_S \rangle$. Apply Lemma 4 coordinatewise with $a_t = g_{t,i}$, $G = 2M$, $z_t = \mathbf{1}\{i \in S_t\}$, and $b = \mathbf{1}\{i \in S\}$, after reindexing the rounds in I_e as $t = 1, \dots, L_e$. Summing over $i \in [n]$ gives

$$\sum_{t \in I_e} \big(f_t(S_t) - f_t(S)\big) \leq 2MC_{L_e}^{(e)} + \frac{n}{2\eta} + 2Mn.$$

Adding $\lambda C_{L_e}^{(e)}$ proves the claim. □

Proof of Proposition 7. Fix an epoch e and write $S_b := S_{\tau^{(e)}}(x_b)$ and $d_b := \chi_{S_b} - \chi_S$ for every block b in epoch e. The threshold set S_b is one of the chain sets induced by x_b. Hence every chain subgradient $g_{t_b+s}(x_b)$ is also a subgradient of $\hat{f}_{t_b+s}$ at χ_{S_b}, and therefore

$$\sum_{s=0}^{n} \big(f_{t_b+s}(S_b) - f_{t_b+s}(S)\big) \le \langle G_b, d_b \rangle.$$

On an exploitation block, $X_{t_b+s} = S_b$ throughout the block. On an exploration block, the loss at each of its $n+1$ rounds can exceed the loss of S_b by at most $2M$. Thus

$$\sum_{t \text{ in epoch } e} \big(f_t(X_t) - f_t(S)\big) \le \sum_{b \text{ in epoch } e} \langle \widehat{G}_b, d_b \rangle + 2M(n+1)N_e + Z_e(S). \tag{28}$$

By the definition of the cyclic estimator, $\|\widehat{G}_b\|_\infty \le K_\gamma$ almost surely. Conditional on the history before epoch e, the iterate sequence obtained without restarting is independent of its fresh continuous threshold; hence the equality events in Lemma 4 have probability zero, and its pathwise bound applies to the stopped prefix. Applying the lemma coordinatewise to the block iterates, with $a_q = \widehat{G}_{b_q,i}$ and $G = K_\gamma$, gives

$$\begin{aligned}\sum_{b \text{ in epoch } e} \langle \widehat{G}_b, d_b \rangle &\le \frac{n}{2\eta} + nK_\gamma + K_\gamma \sum_{q=1}^{L_e-1} |S_{q+1}^{(e)} \,\Delta\, S_q^{(e)}| \\ &\le \frac{n}{2\eta} + nK_\gamma + K_\gamma V_{L_e}^{(e)}. \end{aligned} \tag{29}$$

Here the last monitor increment, from the final block to its proposed next virtual set, only enlarges the upper bound.

It remains to compare the actual and virtual movements. An exploration block has internal movement of at most $2n$ because it follows a cyclic ordering of the chain. Entering this traversal from the preceding virtual action and leaving it for the following virtual action cost at most n each. Charging these quantities separately for every exploration block is valid even when two such blocks are adjacent, and yields

$$C_e^X \le V_{L_e}^{(e)} + 4nN_e. \tag{30}$$

Multiplying Equation (30) by λ and adding it to Equations (28) and (29) proves the proposition because $Q_\lambda = 2M(n+1) + 4\lambda n$. □

A.7 Restarting analysis

For the epoch-count argument, couple each actual epoch to a process that uses the same randomness but continues without restarting. For integer $s \ge 0$, let $V_s^{(e)}$ denote its monitor after s local rounds. This process agrees with the actual epoch up to and including the budget-crossing round. If it extends beyond the horizon, set all subsequent losses to zero.

Given a budget $B > 0$, let t_e be the first round of epoch e and define its untruncated length by

$$\widetilde{L}_e := \min\{s \in \mathbb{Z}_{\ge 1} : V_s^{(e)} > B\}, \qquad \min \emptyset := \infty.$$

Thus, the budget-crossing round belongs to epoch e. Its truncated length and round set are

$$L_e := \min\{\widetilde{L}_e, T - t_e + 1\}, \qquad I_e := \{t_e, \dots, t_e + L_e - 1\}.$$

Set $t_1 := 1$ and $t_{e+1} := t_e + L_e$, and define the number of epochs by

$$E := \max\{e \ge 1 : t_e \le T\}.$$

In particular, movement from the last action of epoch e to the first action of epoch $e+1$ is not included in $C_{L_e}^{(e)}$ and is charged separately.

Lemma 14 (Conditional long-epoch bound)**.** *Let $H > 0$. Suppose Equation* (7) *holds for every integer $s \ge 0$ and $\Phi(H) > 0$. If the wrapper restarts when $V_s^{(e)}$ first exceeds $4\Phi(H)$, then*

$$\mathbb{P}\left(\widetilde{L}_e \ge H \mid \mathcal{F}_{e-1}\right) \ge \frac{3}{4}. \tag{31}$$

Proof. Here the budget is $B = 4\Phi(H)$. If $H \le 1$, then $\widetilde{L}_e \ge 1 \ge H$ almost surely, so the claim is immediate. Suppose henceforth that $H > 1$. By the minimality of $\widetilde{L}_e$, on the event $\{\widetilde{L}_e < H\}$ we have

$$V^{(e)}_{\widetilde{L}_e} > 4\Phi(H).$$

Since $\widetilde{L}_e$ is integer, $\widetilde{L}_e < H$ implies $\widetilde{L}_e \le \lfloor H \rfloor$. The monotonicity of $V^{(e)}_s$ therefore gives

$$\{\widetilde{L}_e < H\} \subseteq \{V^{(e)}_{\lfloor H \rfloor} > 4\Phi(H)\}.$$

The conditional Markov inequality, applied to the nonnegative random variable $V^{(e)}_{\lfloor H \rfloor}$, and Equation (7) give

$$\mathbb{P}(\widetilde{L}_e < H \mid \mathcal{F}_{e-1}) \le \mathbb{P}(V^{(e)}_{\lfloor H \rfloor} > 4\Phi(H) \mid \mathcal{F}_{e-1}) \le \frac{\mathbb{E}[V^{(e)}_{\lfloor H \rfloor} \mid \mathcal{F}_{e-1}]}{4\Phi(H)} \le \frac{\Phi(\lfloor H \rfloor)}{4\Phi(H)} \le \frac{1}{4}.$$

Taking complements proves Equation (31). □

Proposition 15 (Tail bound on the number of epochs)**.** *For any $\rho \in (0,1)$, define*

$$N_\rho := \max\left\{4, \left\lceil \frac{2\log(1/\rho)}{\log 3} \right\rceil \right\}, \qquad H_\rho := \frac{4T}{N_\rho}. \tag{32}$$

Suppose that every epoch satisfies

$$\mathbb{P}\left(\widetilde{L}_e \ge H_\rho \,\middle|\, \mathcal{F}_{e-1}\right) \ge \frac{3}{4}. \tag{33}$$

Then

$$\mathbb{P}\big(E > N_\rho\big) \le 3^{-N_\rho/2} \le \rho. \tag{34}$$

Proof. For epochs started by time T, put $X_e := \mathbf{1}\{\widetilde{L}_e \ge H_\rho\}$; for all remaining epochs, set $X_e := 1$. On the event $E > N_\rho$, the maximality in the definition of E gives $t_{N_\rho+1} \le T$. Hence the first N_ρ epochs are untruncated, and

$$\sum_{e=1}^{N_\rho} \widetilde{L}_e = t_{N_\rho+1} - 1 < T.$$

Since $\widetilde{L}_e \ge H_\rho X_e$ and $T = H_\rho N_\rho/4$, this implies

$$\sum_{e=1}^{N_\rho} X_e < \frac{N_\rho}{4}.$$

The variables X_e need not be independent, but Equation (33) gives $\mathbb{E}[X_e \mid \mathcal{F}_{e-1}] \ge 3/4$. For every $\theta > 0$,

$$\mathbb{E}[e^{-\theta X_e} \mid \mathcal{F}_{e-1}] \le \frac{1}{4} + \frac{3}{4}e^{-\theta}.$$

Iterating this inequality with the tower property and applying the exponential Markov inequality yields the Bernoulli lower-tail bound

$$\mathbb{P}\left(\sum_{e=1}^{N_\rho} X_e < \frac{N_\rho}{4}\right) \le \exp(-N_\rho D_{\mathrm{KL}}(1/4 \,\|\, 3/4)) = 3^{-N_\rho/2}.$$

Thus, the definition of N_ρ gives

$$3^{-N_\rho/2} \le \exp\left(-\log\frac{1}{\rho}\right) = \rho.$$

□

Full information. We apply Proposition 15 with $\rho = \delta$, and write $N_\delta := N_\rho$ and $H := H_\rho$ in this paragraph. For a fresh epoch, let $C_s^{\mathrm{cont},(e)}$ be the cumulative movement over s local rounds if the algorithm continues without restarting, and set $V_s^{(e)} := C_s^{\mathrm{cont},(e)}$. The conditional expected-movement bound

$$\mathbb{E}[C_s^{\mathrm{cont},(e)} \mid \mathcal{F}_{e-1}] \le 4M\eta s, \qquad s \in \mathbb{Z}_{\ge 0},$$

verifies Equation (7) with $\Phi(s) = 4M\eta s$. Thus, setting the budget $B := 16M\eta H = 4\Phi(H)$ and restarting when $C_s^{(e)}$ first exceeds B allows us to apply Lemma 14 and Proposition 15. The stopping round overshoots B by at most n, and each epoch boundary contributes at most another n. Summing Proposition 5 over epochs gives, on the event $E \le N_\delta$,

$$\begin{aligned}\mathcal{R}_T(\lambda) &\le 64M(2M+\lambda)\eta T \\ &\quad + N_\delta\left(\frac{n}{2\eta} + 2Mn + 2(2M+\lambda)n\right).\end{aligned} \tag{35}$$

The η-dependent terms on the right-hand side are minimized at

$$\eta_\delta^* := \sqrt{\frac{N_\delta n}{128M(2M+\lambda)T}}.$$

Substitution gives, with probability at least $1-\delta$,

$$\mathcal{R}_T(\lambda) \le U_\delta := \sqrt{128M(2M+\lambda)nTN_\delta} + 2(3M+\lambda)nN_\delta. \tag{36}$$

The high-probability algorithm plays the fixed action $S_t = \emptyset$ on every round when $U_\delta \ge 2MT$ and otherwise runs batched SSD with $\eta = \eta_\delta^*$. Since the fixed-action policy satisfies $\mathcal{R}_T(\lambda) \le 2MT$ deterministically, the combined algorithm obeys $\mathcal{R}_T(\lambda) \le \min\{2MT, U_\delta\}$ with probability at least $1-\delta$. Put $q_\delta := (1+\lambda/M)nN_\delta$. The optimized bound is $O\big(M(\sqrt{Tq_\delta} + q_\delta)\big)$. If $q_\delta \le T$, then $q_\delta \le \sqrt{Tq_\delta}$; otherwise, the fixed-action bound is $O(MT)$. Hence, uniformly over the problem parameters,

$$\min\{2MT, U_\delta\} = O\Big(M\min\{T, \sqrt{Tq_\delta}\}\Big).$$

Finally, $N_\delta = \Theta(\log(1/\delta))$ for $\delta \in (0, 1/2)$, which proves Theorem 6.

Bandit feedback. Put $m := \lfloor T/(n+1)\rfloor$, $T' := (n+1)m$, and $r := T - T'$. If $m = 0$, the fixed-action policy gives $\mathcal{R}_T^X(\lambda) \le 2MT$, so assume below that $m \ge 1$. The wrapper now uses blocks as its time unit. At the beginning of each epoch it sets $x_1^{(e)} = \frac{1}{2}\mathbf{1}$ and draws a fresh threshold $\tau^{(e)}$. It then runs complete cyclic blocks and performs the update only after receiving all observations in a block.

More precisely, after the qth block of epoch e, the wrapper computes

$$\widetilde{x}_{q+1}^{(e)} := \Pi_{[0,1]^n}(x_q^{(e)} - \eta\widehat{G}_q^{(e)}), \qquad \widetilde{S}_{q+1}^{(e)} := S_{\tau^{(e)}}(\widetilde{x}_{q+1}^{(e)}),$$

and updates the observable monitor

$$V_q^{(e)} := \sum_{j=1}^{q} |\widetilde{S}_{j+1}^{(e)} \,\Delta\, S_j^{(e)}|.$$

If this quantity exceeds the budget, $\widetilde{x}_{q+1}^{(e)}$ is discarded and a new epoch begins at the next block; otherwise it is accepted. Notice that $\widehat{G}_q^{(e)} = 0$ on an exploitation block, so such a block contributes zero to the monitor. The virtual set $S_q^{(e)}$ is maintained on exploration blocks but need not be played or queried.

Let $\mathcal{H}_e$ be the history immediately before the fresh threshold of epoch e is drawn, and consider the process that continues without restarting, accepting every proposed update $x_{q+1}^{(e)} = \widetilde{x}_{q+1}^{(e)}$. Denote its monitor by $V_q^{\mathrm{cont},(e)}$. The loss sequence is oblivious, and the estimator uses only exploration

observations, so the iterate sequence in this process is independent of $\tau^{(e)}$ conditional on $\mathcal{H}_e$. The shared-threshold identity, nonexpansiveness of projection, and Equation (19) therefore give

$$\mathbb{E}[V_q^{\mathrm{cont},(e)} \mid \mathcal{H}_e] = \mathbb{E}\left[\sum_{j=1}^{q} \|\widetilde{x}_{j+1}^{(e)} - x_j^{(e)}\|_1 \,\middle|\, \mathcal{H}_e\right]$$
$$\le \eta \sum_{j=1}^{q} \mathbb{E}[\|\widehat{G}_j^{(e)}\|_1 \mid \mathcal{H}_e] \le 2\eta n(n+1)Mq =: \Phi(q). \tag{37}$$

The actual epoch is a prefix of this process. In particular, we do not condition on the event that the epoch has survived, since that event can depend on the shared threshold.

Set

$$\ell := \log\frac{3}{\delta}, \qquad N_\delta := \max\left\{4, \left\lceil \frac{2\ell}{\log 3} \right\rceil\right\}, \qquad H := \frac{4m}{N_\delta}, \tag{38}$$

and use the budget

$$\beta := 4\Phi(H) = 8\eta n(n+1)MH = \frac{32\eta nMT'}{N_\delta}. \tag{39}$$

Let $\widetilde{L}_e := \inf\{q \ge 1 : V_q^{\mathrm{cont},(e)} > \beta\}$ be the untruncated block length of epoch e. If $H \le 1$, then $\widetilde{L}_e \ge H$ automatically. If $H > 1$, the event $\{\widetilde{L}_e < H\}$ implies $V_{\lfloor H \rfloor}^{\mathrm{cont},(e)} > \beta$; hence the conditional Markov inequality and Equation (37) yield

$$\mathbb{P}(\widetilde{L}_e < H \mid \mathcal{H}_e) \le \frac{\Phi(\lfloor H \rfloor)}{4\Phi(H)} \le \frac{1}{4}.$$

Applying Proposition 15 with m blocks in place of T rounds shows that the number E of epochs satisfies

$$\mathbb{P}(E > N_\delta) \le 3^{-N_\delta/2} \le \frac{\delta}{3}. \tag{40}$$

Each monitor increment is at most n. Thus the budget-crossing block gives $V_{L_e}^{(e)} \le \beta + n$, and on $\{E \le N_\delta\}$,

$$\sum_{e=1}^{E} V_{L_e}^{(e)} \le N_\delta(\beta + n) = 32\eta nMT' + nN_\delta. \tag{41}$$

We next control the exploration blocks. Their total number over the m complete global blocks is

$$N_{\mathrm{exp}} := \sum_{b=1}^{m} E_b \sim \mathrm{Bin}(m, \gamma),$$

regardless of the restart times. Bernstein's inequality gives, with probability at least $1 - \delta/3$,

$$N_{\mathrm{exp}} \le \gamma m + \sqrt{2\gamma(1-\gamma)m\ell} + \frac{2}{3}\ell. \tag{42}$$

Finally, fix a comparator S and write

$$\zeta_b(S) := \left\langle G_b - \widehat{G}_b, \chi_{S_b} - \chi_S \right\rangle.$$

Let $\mathcal{F}_{b-1}$ contain the history through block $b-1$, the restart decision and any reset before block b, and the threshold for the current epoch, but not (E_b, J_b). Hence S_b and its chain are measurable, and Equation (17) gives

$$\mathbb{E}[\zeta_b(S) \mid \mathcal{F}_{b-1}] = 0.$$

Since $\|\chi_{S_b} - \chi_S\|_1 \le n$ and $\|\widehat{G}_b\|_\infty \le K_\gamma E_b$, where $K_\gamma = 2(n+1)M/\gamma$, we have

$$|\langle \widehat{G}_b, \chi_{S_b} - \chi_S \rangle| \le nK_\gamma E_b,$$
$$|\zeta_b(S)| \le nK_\gamma(1+\gamma) \le 2nK_\gamma,$$
$$\mathbb{E}[\zeta_b(S)^2 \mid \mathcal{F}_{b-1}] \le \gamma n^2 K_\gamma^2. \tag{43}$$

The total conditional variance over the m blocks is therefore at most

$$v := m\gamma n^2 K_\gamma^2 = \frac{4n^2(n+1)M^2T'}{\gamma}.$$

Freedman's inequality (Freedman, 1975, Theorem 1.6) implies, with probability at least $1-\delta/3$,

$$\sum_{b=1}^{m} \zeta_b(S) \le 2\sqrt{2}\, nM\sqrt{\frac{(n+1)T'\ell}{\gamma}} + \frac{4}{3} nK_\gamma \ell. \tag{44}$$

Only the fixed hindsight minimizer $S^\star$ is needed below, so no union bound over the 2^n comparators is required.

Sum Proposition 7 over the epochs and charge at most $\lambda n(E-1)$ for movement across epoch boundaries. The final r rounds hold the last action fixed and contribute at most $2Mr$ in regret. Consequently, on the intersection of the three events above, which has probability at least $1-\delta$,

$$\begin{aligned}\mathcal{R}_T^X(\lambda) \le{}& 32\eta nM(K_\gamma+\lambda)T' + N_\delta\left(\frac{n}{2\eta} + 2nK_\gamma + 2\lambda n\right)\\ &+ Q_\lambda\left(\gamma m + \sqrt{2\gamma(1-\gamma)m\ell} + \frac{2}{3}\ell\right)\\ &+ 2\sqrt{2}\, nM\sqrt{\frac{(n+1)T'\ell}{\gamma}} + \frac{4}{3}nK_\gamma\ell + 2Mr,\end{aligned} \tag{45}$$

where $Q_\lambda = 2M(n+1) + 4\lambda n$. This is the claimed finite-time blockwise bound.

It remains to tune the parameters. Put

$$A_\lambda := M + \lambda, \qquad a := \frac{A_\lambda}{M} = 1 + \frac{\lambda}{M}.$$

In the nontrivial regime $T \ge 2n^3 a N_\delta$, choose

$$\gamma_0 := \left(\frac{n^2(n+1)M^2N_\delta}{A_\lambda^2 T}\right)^{1/3}, \qquad \eta_\delta(\gamma) := \frac{1}{8}\sqrt{\frac{N_\delta}{MT(K_\gamma+\lambda)}}, \tag{46}$$

and set $\gamma = \gamma_0$ and $\eta = \eta_\delta(\gamma_0)$. The stated regime ensures $\gamma_0 \le 1$. Since $T' \le T$, the two leading η-dependent terms satisfy

$$32\eta nM(K_\gamma+\lambda)T' + \frac{N_\delta n}{2\eta} \le 8n\sqrt{N_\delta MT(K_\gamma+\lambda)}.$$

Using $Q_\lambda \le 4(n+1)A_\lambda$ and $\sqrt{K_\gamma+\lambda} \le \sqrt{2(n+1)M/\gamma} + \sqrt{\lambda}$, the principal terms in Equation (45) are

$$O\Big(F + n\sqrt{\lambda MTN_\delta}\Big), \qquad F := nM^{2/3}A_\lambda^{1/3}T^{2/3}N_\delta^{1/3}. \tag{47}$$

The square-root term in Equation (44) is of the same order. Up to universal constants, the remaining terms are

$$\frac{n(n+1)MN_\delta}{\gamma} + \lambda nN_\delta + A_\lambda\sqrt{(n+1)\gamma TN_\delta} + A_\lambda(n+1)N_\delta + Mn.$$

They are all $O(F)$ in the stated regime. Indeed, writing $s := T/(n^3 a N_\delta) \ge 2$ and $k := ((n+1)/n)^{1/3} \in [1, 2^{1/3}]$ gives $\gamma_0 = k/(as^{1/3})$ and $F = Mn^3 a N_\delta s^{2/3}$; for example,

$$\frac{n(n+1)MN_\delta/\gamma_0}{F} = \frac{n+1}{kn^2 s^{1/3}}, \qquad \frac{A_\lambda\sqrt{(n+1)\gamma_0 TN_\delta}}{F} = \sqrt{\frac{(n+1)k}{n^3}}\, s^{-1/3},$$

and

$$\frac{n\sqrt{\lambda MTN_\delta}}{F} = \sqrt{\frac{a-1}{a}}\,\frac{1}{\sqrt{n}\, s^{1/6}}.$$

The other ratios are smaller by inspection. Thus the cyclic policy has regret $O(F)$ in the nontrivial regime.

If $T < 2n^3 a N_\delta$, use one fixed set throughout the horizon. This gives $\mathcal{R}_T^X(\lambda) \le 2MT$ deterministically, while $F/(MT) > 2^{-1/3}$. Combining the two regimes and using $N_\delta = \Theta(\log(3/\delta))$ proves Theorem 8.

A.8 LOWER BOUND ANALYSIS

Lemma 16 (Rademacher maximum lower bound)**.** *Let $B \geq 1$ and $N \geq 2$. For each $b \in [B]$ and $i \in [N]$, let $\epsilon_{b,i}$ be an independent Rademacher random variable. Then there exists a universal constant $c > 0$ such that*

$$\mathbb{E}\left[\max_{i\in[N]} \sum_{b=1}^{B} \epsilon_{b,i}\right] \geq c \min\{B, \sqrt{B \log N}\}.$$

Proof. Write $S_i = \sum_{b=1}^{B} \epsilon_{b,i}$. By Orabona & Pál (2015, Theorem 7), there exist universal constants $N_0 \geq 2$ and $c_0 > 0$ such that, whenever $B \geq 7$ and $N_0 \leq N \leq \exp(B/3)$,

$$\mathbb{E}\left[\max_{i\in[N]} S_i\right] \geq c_0 \sqrt{B \log N}.$$

Here we absorbed the additive $-2\sqrt{B}$ term in that theorem by choosing N_0 sufficiently large.

For all $B \geq 1$ and $N \geq 2$,

$$\mathbb{E}\left[\max_{i\in[N]} S_i\right] \geq \mathbb{E}[\max\{S_1, S_2\}] = \frac{1}{2}\mathbb{E}[S_1 + S_2] + \frac{1}{2}\mathbb{E}|S_1 - S_2| = \Omega(\sqrt{B}),$$

where the last step uses $\mathbb{E}[S_1 + S_2] = 0$ and Khintchine's inequality (Vershynin, 2026, Exercise 2.36). This proves the claim when $N < N_0$, after decreasing the universal constant, and also when $B < 7$.

It remains to consider $B \geq 7$, $N \geq N_0$, and $N > \exp(B/3)$. Set $N' = \lfloor \exp(B/3) \rfloor$. If $N' \geq N_0$, applying the preceding bound to the first N' coordinates and using monotonicity in N yields

$$\mathbb{E}\left[\max_{i\in[N]} S_i\right] = \Omega\Big(\sqrt{B \log N'}\Big) = \Omega(B).$$

If $N' < N_0$, then B is bounded by a universal constant, and the $\Omega(\sqrt{B})$ bound above implies an $\Omega(B)$ bound after one final decrease of the universal constant. □

Lemma 17 (Experts lower bound with switching costs)**.** *For every $H \geq 1$, $N \geq 2$, $M > 0$, and $c > 0$, and every experts algorithm producing $p_1, \ldots, p_H \in \Delta_N := \Big\{p \in [0,1]^N \mid \sum_{i=1}^{N} p_i = 1\Big\}$, there exists an oblivious loss sequence $\ell_1, \ldots, \ell_H \in [-M, M]^N$ such that*

$$\mathbb{E}_{\mathcal{A}}\left[\sum_{t=1}^{H} \langle \ell_t, p_t\rangle - \min_{i\in[N]} \sum_{t=1}^{H} \ell_{t,i} + c\sum_{t=2}^{H} \frac{1}{2}\|p_t - p_{t-1}\|_1\right] = \Omega\Big(\min\{MH, \sqrt{cMH \log N}\}\Big).$$

Proof. Fix a possibly randomized experts algorithm $\mathcal{A}$. For any loss sequence $\ell_{1:H}$, let $R_H(\ell_{1:H})$ denote the quantity inside the expectation in the statement.

It suffices to construct a finitely supported distribution $\mathcal{D}$ over oblivious loss sequences such that

$$\mathbb{E}_{\ell_{1:H}\sim\mathcal{D},\mathcal{A}}[R_H(\ell_{1:H})] = \Omega\Big(\min\{MH, \sqrt{cMH \log N}\}\Big),$$

because

$$\max_{\ell_{1:H}\in\mathrm{supp}(\mathcal{D})} \mathbb{E}_{\mathcal{A}}[R_H(\ell_{1:H})] \geq \mathbb{E}_{\ell_{1:H}\sim\mathcal{D}}\mathbb{E}_{\mathcal{A}}[R_H(\ell_{1:H})].$$

Set $L := \min\left\{H, \left\lfloor \frac{c}{2M} \right\rfloor + 1\right\}, B := \left\lfloor \frac{H}{L} \right\rfloor$. Divide the first BL rounds into B blocks and set all remaining losses to zero. Independently for each block b, draw $\sigma_b \in \{-1, 1\}^N$ with independent Rademacher coordinates, and use the constant loss

$$\ell_{b,s} = M\sigma_b, \qquad s = 1, \ldots, L,$$

throughout that block.

Write $p_{b,s}$ for the learner's decision at local round s of block b, i.e., $p_{b,s} = p_{(b-1)L+s}$, and define the block-wise total variation

$$V_b := \frac{1}{2}\sum_{s=2}^{L}\|p_{b,s} - p_{b,s-1}\|_1.$$

Since $\|\sigma_b\|_\infty = 1$, Hölder's inequality and the triangle inequality give

$$\begin{aligned}\sum_{s=1}^{L}\langle \ell_{b,s}, p_{b,s}\rangle &= ML\langle \sigma_b, p_{b,1}\rangle + M\sum_{s=2}^{L}\langle \sigma_b, p_{b,s} - p_{b,1}\rangle \\ &\geq ML\langle \sigma_b, p_{b,1}\rangle - M\sum_{s=2}^{L}\|p_{b,s} - p_{b,1}\|_1 \\ &\geq ML\langle \sigma_b, p_{b,1}\rangle - 2M(L-1)V_b.\end{aligned}$$

Thus,

$$\sum_{s=1}^{L}\langle \ell_{b,s}, p_{b,s}\rangle + cV_b \geq ML\langle \sigma_b, p_{b,1}\rangle + \big(c - 2M(L-1)\big)V_b.$$

Since $p_{b,1}$ is chosen before σ_b is revealed, $p_{b,1}$ is independent of σ_b. Combined with $\mathbb{E}[\sigma_b] = 0$, this gives

$$\mathbb{E}[\langle \sigma_b, p_{b,1}\rangle] = 0.$$

Moreover, the choice of L ensures $c - 2M(L-1) \geq 0$. Thus the expected sum of the learner's loss and the within-block movement cost cV_b is nonnegative. Movement costs between blocks and after the last complete block are also nonnegative and may be dropped.

The best fixed expert has expected loss

$$-ML\,\mathbb{E}\left[\max_{i\in[N]}\sum_{b=1}^{B}(-\sigma_{b,i})\right].$$

Lemma 16 therefore gives

$$\mathbb{E}[R_H] = \Omega\Big(ML\min\{B, \sqrt{B\log N}\}\Big).$$

Since $1 \leq L \leq H$, we have $B = \lfloor H/L\rfloor \geq H/(2L)$, which implies $ML\sqrt{B} \geq M\sqrt{LH/2}$ and $MLB \geq MH/2$. If $L = H$, the bound immediately yields $\Omega(MH)$. Otherwise, $L = \lfloor c/(2M)\rfloor + 1 > c/(2M)$, and substituting this lower bound on L gives $M\sqrt{LH/2} = \Omega(\sqrt{cMH})$. Combining both cases yields

$$\mathbb{E}[R_H] = \Omega\Big(\min\{MH, \sqrt{cMH\log N}\}\Big).$$

□

Theorem 18 (Lower bound with a fixed cost per action change)**.** *Assume $T \geq n$. For every randomized full-information OSM algorithm, there is an oblivious sequence of modular losses $f_t : 2^{[n]} \to [-M, M]$ such that*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(S_t) - \min_{S\subseteq[n]}\sum_{t=1}^{T} f_t(S) + \lambda\sum_{t=2}^{T}\mathbf{1}\{S_t \neq S_{t-1}\}\right] = \Omega\Big(\min\{MT, \sqrt{\lambda MnT}\}\Big).$$

Proof. For $\lambda = 0$, the claimed lower bound is trivial. Henceforth assume $\lambda > 0$. Fix a randomized full-information OSM algorithm $\mathcal{A}$, and set

$$m := \left\lfloor \frac{T}{n}\right\rfloor, \qquad \mathcal{T}_k := \{(k-1)m+1, \ldots, km\}, \qquad k = 1, \ldots, n.$$

We construct the losses on these intervals recursively. Suppose that the losses before $\mathcal{T}_k$ have already been fixed. On $\mathcal{T}_k$, consider the two-expert algorithm induced by $\mathcal{A}$ through the marginal decisions

$$p_{k,s} := \big(\mathbb{P}_{\mathcal{A}}(k \notin S_{(k-1)m+s}), \mathbb{P}_{\mathcal{A}}(k \in S_{(k-1)m+s})\big) \in \Delta_2, \qquad s = 1, \ldots, m.$$

Lemma 17, applied with $N = 2$, horizon m, loss scale M, and switching weight λ, provides a deterministic sequence $\ell_{k,1}, \dots, \ell_{k,m} \in [-M, M]^2$ such that

$$\sum_{s=1}^{m} \langle \ell_{k,s}, p_{k,s} \rangle - \min_{z \in \{0,1\}} \sum_{s=1}^{m} \ell_{k,s}(z) + \frac{\lambda}{2} \sum_{s=2}^{m} \|p_{k,s} - p_{k,s-1}\|_1 = \Omega\Big(\min\{Mm, \sqrt{\lambda M m}\}\Big). \tag{48}$$

Fix this sequence and continue to interval $k+1$. After all n intervals have been treated, define, for $t = (k-1)m + s \in \mathcal{T}_k$,

$$f_t(S) := \ell_{k,s}\big(\mathbf{1}\{k \in S\}\big), \qquad S \subseteq [n],$$

and set $f_t \equiv 0$ for the remaining rounds $t = nm+1, \dots, T$. This recursive construction is carried out after $\mathcal{A}$ is fixed and does not depend on the realization of its internal randomness. Hence the resulting loss sequence is fixed before play and is oblivious. Each f_t is modular and takes values in $[-M, M]$.

By construction,

$$\mathbb{E}_{\mathcal{A}}\left[\sum_{t=1}^{T} f_t(S_t)\right] = \sum_{k=1}^{n} \sum_{s=1}^{m} \langle \ell_{k,s}, p_{k,s} \rangle,$$

while the best fixed subset decomposes coordinatewise as

$$\min_{S \subseteq [n]} \sum_{t=1}^{T} f_t(S) = \sum_{k=1}^{n} \min_{z \in \{0,1\}} \sum_{s=1}^{m} \ell_{k,s}(z).$$

For an internal transition $t = (k-1)m+s$, where $k \in [n]$ and $s = 2, \dots, m$, put $Z_{k,s} := \mathbf{1}\{k \in S_t\}$. Then

$$\begin{aligned} \mathbb{E}[\mathbf{1}\{S_t \neq S_{t-1}\}] &\geq \mathbb{P}(Z_{k,s} \neq Z_{k,s-1}) \\ &\geq |\mathbb{P}(Z_{k,s} = 1) - \mathbb{P}(Z_{k,s-1} = 1)| \\ &= \frac{1}{2} \|p_{k,s} - p_{k,s-1}\|_1. \end{aligned}$$

Switching costs at interval boundaries and during the remaining zero-loss rounds are nonnegative and may be dropped. Combining the preceding displays with Equation (48) and summing over the n intervals gives

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(S_t) - \min_{S \subseteq [n]} \sum_{t=1}^{T} f_t(S) + \lambda \sum_{t=2}^{T} \mathbf{1}\{S_t \neq S_{t-1}\}\right] = \Omega\Big(\min\{Mnm, \sqrt{\lambda M n (nm)}\}\Big).$$

Since $T \geq n$ implies $T/2 \leq nm \leq T$, the last display is $\Omega(\min\{MT, \sqrt{\lambda M n T}\})$. □

Corollary 19 (Lower bound with Hamming switching costs)**.** *Under the assumptions of Theorem 18, there is an oblivious sequence of modular losses satisfying*

$$\mathbb{E}[\mathcal{R}_T(\lambda)] = \Omega\Big(\min\{MT, \sqrt{\lambda M n T}\}\Big).$$

Proof. Use $|S_t \,\Delta\, S_{t-1}| \geq \mathbf{1}\{S_t \neq S_{t-1}\}$ and apply Theorem 18. □

Lemma 20 (Standard full-information OSM lower bound)**.** *For every randomized full-information OSM algorithm, there is an oblivious sequence of modular losses $f_t : 2^{[n]} \to [-M, M]$ such that*

$$\mathbb{E}[\mathcal{R}_T(0)] = \Omega\Big(M \min\{T, \sqrt{nT}\}\Big).$$

Proof. The construction in the proof of Hazan & Kale (2012, Theorem 14) assigns coordinates cyclically to yield an expected regret lower bound of $\Omega(M \sum_{i=1}^{n} \sqrt{q_i})$, where q_i is the number of rounds to which coordinate i is assigned. If $T \geq n$, then each $q_i \geq \lfloor T/n \rfloor$, giving $\sum_{i=1}^{n} \sqrt{q_i} = \Omega(\sqrt{nT})$. If $T < n$, exactly T coordinates have $q_i = 1$ (and the rest zero), giving $\sum_{i=1}^{n} \sqrt{q_i} = T$. Combining both cases yields $\Omega(M \min\{T, \sqrt{nT}\})$. □

For $T \geq n$, choose whichever loss sequence from Corollary 19 and Lemma 20 gives the larger lower bound. The lower bound from Lemma 20 also applies to $\mathcal{R}_T(\lambda)$ because the switching cost is nonnegative. Since

$$\begin{aligned}&\max\Big\{M\min\{T,\sqrt{nT}\},\min\{MT,\sqrt{\lambda MnT}\}\Big\}\\&\quad=\Theta\Big(M\min\Big\{T,\sqrt{(1+\lambda/M)nT}\Big\}\Big),\end{aligned}$$

we obtain

$$\mathbb{E}[\mathcal{R}_T(\lambda)]=\Omega\Big(M\min\Big\{T,\sqrt{(1+\lambda/M)nT}\Big\}\Big).$$

When $T<n$, Lemma 20 alone gives $\Omega(MT)$, which proves the same bound. This proves Theorem 9.

A.8.1 Bandit-feedback lower bound

Lemma 21 (Ordinary bandit OSM lower bound)**.** *For every randomized bandit-feedback OSM algorithm, there is an oblivious sequence of modular losses $f_t : 2^{[n]} \to [-M, M]$ such that*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(X_t) - \min_{S\subseteq[n]}\sum_{t=1}^{T} f_t(S)\right] = \Omega\Big(M\min\{T, n\sqrt{T}\}\Big).$$

Proof. Fix a randomized bandit-feedback OSM algorithm $\mathcal{A}$. The proof of Dani et al. (2007, Theorem 4.3) establishes, for every $0<\varepsilon<1$, an expected-regret lower bound

$$\Omega\big(\min\{\varepsilon T, n^2/\varepsilon\}\big)$$

on the decision set $\{0,1\}^n$. Its signed-coordinate loss vectors define modular losses after the identification $x=\chi_S$. Choose $\varepsilon=\min\{1/2, n/\sqrt{T}\}$. If $n/\sqrt{T}\leq 1/2$, the displayed bound is $\Omega(n\sqrt{T})$. Otherwise, it is $\Omega(\min\{T,n^2\})$, which is $\Omega(\min\{T,n\sqrt{T}\})$ because $n>\sqrt{T}/2$. Thus, in either case, the randomized loss construction in that proof satisfies, writing $R_T(\mathcal{A},L)$ for the corresponding unscaled regret,

$$\mathbb{E}_L\mathbb{E}_{\mathcal{A}}[R_T(\mathcal{A},L)]=\Omega\Big(\min\{T,n\sqrt{T}\}\Big).$$

By averaging over L, some deterministic realization of the loss sequence satisfies the same lower bound in expectation over the internal randomness of $\mathcal{A}$. Finally, scaling that realization by M gives modular losses in $[-M,M]$ and proves the claim. □

Proposition 22 (Explicit switching-dependent bandit lower bound)**.** *Let $n\geq 1$, $T\geq 2$, $M>0$, and $\lambda>0$. Set*

$$L=1+\lfloor\log_2 T\rfloor,\qquad h=\ln(128T^2),\qquad p=\frac{1}{64T},\qquad \sigma^2=\frac{M^2}{8Lh},$$

and

$$\varepsilon=\min\left\{\frac{M}{2n},\frac{\sigma}{8\sqrt{L}},\left(\frac{\lambda\sigma^2}{64LT}\right)^{1/3}\right\}. \tag{49}$$

For every randomized bandit-feedback OSM algorithm, there is a deterministic oblivious sequence of modular losses $f_t : 2^{[n]} \to [-M, M]$ such that

$$\begin{aligned}\mathbb{E}[\mathcal{R}_T^X(\lambda)]&\geq\frac{nT\varepsilon}{4}\\&=\min\left\{\frac{MT}{8},\frac{nMT}{32\sqrt{8}\,L\sqrt{h}},\frac{nM^{2/3}\lambda^{1/3}T^{2/3}}{32L^{2/3}h^{1/3}}\right\}.\end{aligned} \tag{50}$$

Proof. **An admissible distribution of losses.** We use the multi-scale Gaussian random walk of Dekel et al. (2014), with a hidden sign in every coordinate. Let θ be uniform on $\{-1,1\}^n$, independently of $\xi_1,\dots,\xi_T \overset{\text{iid}}{\sim} N(0,\sigma^2)$, and define

$$\rho(t)=t-2^{\nu_2(t)},\qquad W_0=0,\qquad W_t=W_{\rho(t)}+\xi_t,$$

where $\nu_2(t)$ is the largest integer k such that 2^k divides t. Our adversary samples $(\theta, \xi_{1:T})$ *conditioned on*

$$G = \left\{ \max_{t \le T} |W_t| \le M/2 \right\},$$

and chooses the entire loss sequence before play as

$$f_t(S) = W_t + \varepsilon \sum_{i \in S} \theta_i. \tag{51}$$

Every realization in G gives modular losses in $[-M, M]$, since $n\varepsilon \le M/2$. No clipping is applied. Each vertex of the parent map ρ has depth at most L, so $\mathrm{Var}(W_t) \le L\sigma^2$. A Gaussian tail bound and a union bound give

$$q := \mathbb{P}(G^c) \le 2T \exp\left(-\frac{M^2}{8L\sigma^2} \right) = p < 1. \tag{52}$$

Thus the conditional distribution is well defined and supported entirely on admissible losses. The event G depends only on the noise, so conditioning preserves the uniform distribution of θ and its independence from the learner's internal randomness. Below we use the *unconditioned* Gaussian law only as a reference distribution for the information analysis.

Regret and Hamming movement. The set $S_\theta^* = \{i : \theta_i = -1\}$ minimizes every loss in Equation (51). Write $x_t = \chi_{X_t}$ and $x_\theta^* = \chi_{S_\theta^*}$, and set

$$D = \sum_{t=1}^{T} \sum_{i=1}^{n} \mathbf{1}\{x_{t,i} \ne x_{\theta,i}^*\}, \qquad C_i = \sum_{t=2}^{T} |x_{t,i} - x_{t-1,i}|, \qquad C = \sum_{i=1}^{n} C_i.$$

The common noise cancels against the comparator, giving the pathwise identity

$$\mathcal{R}_T^X(\lambda) = \varepsilon D + \lambda C. \tag{53}$$

This identity holds both before and after conditioning, and $0 \le C, D \le nT$.

Information under the unconditioned Gaussian law. First fix a deterministic learner, extending its behavior arbitrarily to observations outside $[-M, M]$. For fixed θ, let P_θ be the law of the full observation history under the unconditioned construction. Let $\theta^{(i)}$ denote θ with coordinate i flipped. With $Y_t = f_t(X_t)$ and the analytic conventions $x_0 = 0$, $Y_0 = 0$, conditioning on the full past gives

$$Y_t \mid Y_{1:t-1} \sim N\big(Y_{\rho(t)} + \varepsilon \langle \theta, x_t - x_{\rho(t)} \rangle, \sigma^2\big).$$

Indeed, x_t is determined by the past and ξ_t is independent of it. At the same history, the two hypotheses induce the same actions, and their conditional means differ by $2\varepsilon\theta_i(x_{t,i} - x_{\rho(t),i})$. The chain rule for relative entropy therefore yields

$$\begin{aligned} \mathrm{KL}(P_\theta \| P_{\theta^{(i)}}) &= \frac{2\varepsilon^2}{\sigma^2} \mathbb{E}_{P_\theta} \sum_{t=1}^{T} (x_{t,i} - x_{\rho(t),i})^2 \\ &\le \frac{2\varepsilon^2}{\sigma^2} \mathbb{E}_{P_\theta} \sum_{t=1}^{T} \sum_{s=\rho(t)+1}^{t} |x_{s,i} - x_{s-1,i}| \\ &\le \frac{2\varepsilon^2 L}{\sigma^2} \left(1 + \mathbb{E}_{P_\theta} C_i\right). \end{aligned} \tag{54}$$

The first inequality uses $(a-b)^2 = |a-b|$ for $a, b \in \{0,1\}$ and the triangle inequality along the path from $\rho(t)$ to t. For the second inequality, observe that for each edge length 2^k, the intervals $(\rho(t), t]$ are disjoint: their left endpoints are multiples of 2^{k+1}. There are at most L possible lengths, so every transition appears in at most L such intervals. The virtual transition at time 1 contributes $|x_{1,i}| \le 1$; it accounts for the additive 1 in Equation (54), and is not charged as an actual switch.

Let bars denote expectations also averaged over uniform θ, and define

$$v_{\theta,i} := \mathrm{TV}(P_\theta, P_{\theta^{(i)}}), \qquad \bar{v} = \frac{1}{n 2^n} \sum_{\theta \in \{-1,1\}^n} \sum_{i=1}^{n} v_{\theta,i}.$$

The optimal bit $x^*_{\theta,i}$ changes when θ_i is flipped. For fixed t, i, θ, let $A_{t,i,\theta}$ be the event $\{x_{t,i} \neq x^*_{\theta,i}\}$ in the common history space. Since the learner is deterministic and the two optimal bits are opposite, the error event under $\theta^{(i)}$ is $A^c_{t,i,\theta}$. Hence

$$\begin{aligned} &P_\theta(x_{t,i} \neq x^*_{\theta,i}) + P_{\theta^{(i)}}(x_{t,i} \neq x^*_{\theta^{(i)},i}) \\ &\quad = 1 + P_\theta(A_{t,i,\theta}) - P_{\theta^{(i)}}(A_{t,i,\theta}) \geq 1 - \mathrm{TV}(P_\theta, P_{\theta^{(i)}}). \end{aligned}$$

For each i, the map $\theta \mapsto \theta^{(i)}$ is a bijection of $\{-1, 1\}^n$. Thus summing either error probability over θ, i, t gives $2^n \overline{D}_P$. Averaging the two equal sums yields

$$\begin{aligned} \overline{D}_P &= \frac{1}{2^{n+1}} \sum_\theta \sum_{i=1}^n \sum_{t=1}^T \left[P_\theta(x_{t,i} \neq x^*_{\theta,i}) + P_{\theta^{(i)}}(x_{t,i} \neq x^*_{\theta^{(i)},i}) \right] \\ &\geq \frac{T}{2^{n+1}} \sum_\theta \sum_{i=1}^n [1 - \mathrm{TV}(P_\theta, P_{\theta^{(i)}})] = \frac{nT}{2}(1 - \bar{v}). \end{aligned} \tag{55}$$

For each θ, i, Pinsker's inequality and Equation (54) give

$$v_{\theta,i} \leq \sqrt{\frac{1}{2} \mathrm{KL}(P_\theta \| P_{\theta^{(i)}})} \leq \frac{\varepsilon\sqrt{L}}{\sigma} \sqrt{1 + \mathbb{E}_{P_\theta} C_i}.$$

Averaging over θ, i and using the concavity of the square root yields

$$\begin{aligned} \bar{v} &\leq \frac{\varepsilon\sqrt{L}}{\sigma} \frac{1}{n2^n} \sum_\theta \sum_{i=1}^n \sqrt{1 + \mathbb{E}_{P_\theta} C_i} \\ &\leq \frac{\varepsilon\sqrt{L}}{\sigma} \sqrt{1 + \frac{1}{n2^n} \sum_\theta \sum_{i=1}^n \mathbb{E}_{P_\theta} C_i} = \frac{\varepsilon\sqrt{L}}{\sigma} \sqrt{1 + \frac{\overline{C}_P}{n}}. \end{aligned} \tag{56}$$

The last equality uses $C = \sum_i C_i$. In particular, information summed over all hidden coordinates is controlled by total Hamming movement, without partitioning the horizon across coordinates.

Transfer to the admissible distribution. Let P^G_θ denote the history law conditioned on G, and let $\overline{C}_G, \overline{D}_G$ denote the corresponding expectations, also averaged over uniformly distributed θ. Since G is independent of θ, the mixture identity

$$P_\theta = (1 - q) P^G_\theta + q P^{G^c}_\theta$$

and $0 \leq C, D \leq nT$ imply

$$\overline{C}_P \leq \overline{C}_G + nTp, \qquad \overline{D}_G \geq \overline{D}_P - nTp. \tag{57}$$

For example, $\overline{D}_P - \overline{D}_G = q(\overline{D}_{G^c} - \overline{D}_G) \leq qnT \leq pnT$. If $q = 0$, the same bounds hold with identical conditional and unconditional laws. We do not apply the Gaussian KL calculation to the conditioned law, whose increments need not be independent Gaussian variables.

We split according to movement under the *conditional* law. If $\overline{C}_G \leq n\sigma^2/(16\varepsilon^2 L)$, then Equations (56) and (57) and $\varepsilon \leq \sigma/(8\sqrt{L})$ imply

$$\bar{v}^2 \leq \frac{\varepsilon^2 L}{\sigma^2}(1 + Tp) + \frac{1}{16} \leq \frac{1 + Tp}{64} + \frac{1}{16} < \frac{1}{9}.$$

Consequently,

$$\overline{D}_G \geq nT(1/3 - p) \geq nT/4, \qquad \mathbb{E}[\mathcal{R}^X_T(\lambda) \mid G] \geq nT\varepsilon/4.$$

Otherwise, using $\varepsilon^3 \leq \lambda\sigma^2/(64LT)$ gives

$$\mathbb{E}[\mathcal{R}^X_T(\lambda) \mid G] \geq \lambda \overline{C}_G > \frac{\lambda n \sigma^2}{16\varepsilon^2 L} \geq 4nT\varepsilon.$$

Both cases give the claimed conditional lower bound. Transferring C and D separately avoids subtracting a regret error proportional to λnTp.

Randomized learners and a deterministic loss sequence. The conditional environment distribution is independent of the learner. Fixing each internal random seed and then averaging therefore gives the same bound for every randomized learner. For any such learner, averaging over the conditional environment yields at least one deterministic admissible loss sequence whose expected regret over the learner's randomness is at least $nT\varepsilon/4$. Substituting Equation (49) proves Equation (50). □

Proof of Theorem 10. The ordinary lower bound in Lemma 21 also holds for regret with switching costs because those costs are nonnegative. For $\lambda > 0$, choose the harder of that construction and the construction in Proposition 22; the two sequences need not coincide. Since $L\sqrt{h} = O((\log(2T))^{3/2})$ and $n \geq 1$, the explicit switching lower bound implies

$$\widetilde{\Omega}\Big(M \min\{T, n(\lambda/M)^{1/3}T^{2/3}\}\Big).$$

Combining the bounds and using $\max\{\min\{T, a\}, \min\{T, b\}\} = \min\{T, \max\{a, b\}\}$ proves the theorem. For $\lambda = 0$, the ordinary lower bound alone suffices and has no logarithmic loss. □

B Extensions and Limitations

B.1 Online assignment to two execution sites

We formalize the two-site assignment example in the introduction using the execution-plus-communication cost structure of the two-processor module-assignment model of Stone (1977). Stone also discusses dynamic reassignment and the collection of execution statistics to refine cost estimates. Here, we specify a sequential model in which the assignment for a period is made before that period's workload, and hence its costs, are known, and evaluate regret including reassignment costs. The two destinations represent execution sites rather than necessarily individual processors.

There are n modules (or services) and two execution sites, denoted by P_1 and P_2. A set $S \subseteq [n]$ represents the assignment in which the modules in S are placed on P_1 and those in $[n] \setminus S$ are placed on P_2. For every period t, let $a_{t,i}^{(p)} \geq 0$ be the total cost of executing module i at site P_p, for $p \in \{1, 2\}$, under the workload realized in that period. Let $c_{t,ij} \geq 0$ be the aggregate communication cost between modules i and j if they are assigned to different sites. We assume every assignment is feasible; coupled capacity and deadline constraints are outside this illustrative model. The operational cost of assignment S is

$$C_t(S) := \sum_{i \in S} a_{t,i}^{(1)} + \sum_{i \notin S} a_{t,i}^{(2)} + \sum_{1 \leq i < j \leq n} c_{t,ij} \mathbf{1}\{|S \cap \{i, j\}| = 1\}. \tag{58}$$

The first two terms are the site-dependent execution costs. The last is the cost of references that cross the site boundary. Directed intermodule references can be included by letting $c_{t,ij}$ aggregate the costs in both directions. We assume the costs have been normalized so that $C_t(S) \in [0, M]$ for every t and S.

Proposition 23 (Reduction to online submodular minimization)**.** *For every period t, the operational cost C_t in Equation* (58) *is submodular. If moving one module between sites costs λ, then the cumulative cost of assignments $S_1, \ldots, S_T$ relative to the best fixed assignment is exactly*

$$\sum_{t=1}^{T} C_t(S_t) - \min_{S \subseteq [n]} \sum_{t=1}^{T} C_t(S) + \lambda \sum_{t=2}^{T} |S_t \,\Delta\, S_{t-1}|, \tag{59}$$

which is Equation (1) *with $f_t = C_t$.*

Proof. The execution-cost part of C_t is modular. For each pair $i < j$, the map

$$S \longmapsto \mathbf{1}\{|S \cap \{i, j\}| = 1\}$$

is the cut function of a single nonnegative-weight edge and is submodular. Hence C_t, being the sum of a modular function and nonnegative weighted cut functions, is submodular. A module changes sites precisely when its membership in S_t differs from its membership in S_{t-1}, so the number of migrated modules at the start of period t is $|S_t \,\Delta\, S_{t-1}|$. The best fixed assignment never migrates, which gives Equation (59). □

For a single known period, Stone augments the module-interconnection graph with two terminal nodes representing P_1 and P_2. Terminal edges encode the processor-dependent execution costs, while an edge between two module nodes encodes their communication cost. Every terminal cut then corresponds to an assignment, and its weight is exactly the assignment cost. Thus, if the coefficients in Equation (58) were known before S_t was chosen and there were only one period, the optimum could be found by a minimum-cut computation. This static solvability does not make the online problem vacuous.

In the sequential protocol, at the beginning of period t the learner knows only the feedback from periods $1, \dots, t-1$ and selects S_t before the coefficients $\{a_{t,i}^{(p)}, c_{t,ij}\}$ are realized under the current workload. Under full-information feedback, all these coefficients (equivalently, value-oracle access to C_t) become available after the decision. Under bandit feedback, only the realized scalar $C_t(S_t)$ is observed. We assume that the entire workload sequence is fixed in advance, consistently with the oblivious-adversary model in the main text.

The online timing captures two difficulties absent from a known static instance. First, the current minimum cut cannot be computed before placement because its weights are not yet known. Second, independently optimizing every realized period after the fact could migrate many modules, so prediction and adaptation must be balanced against reconfiguration. The fixed comparator in Equation (59) asks whether the learner can approach the best stable placement in hindsight while retaining the option to adapt at a controlled cost; it does not compare against a clairvoyant time-varying placement.

The sequential model is motivated by Stone's discussion of dynamic operation. Stone (1977, Section IV) proposes dynamic reassignment when program working sets change, notes that the required cost statistics may need to be collected during execution, and suggests refining future assignments as data arrive. The paper later explicitly lists the cost of moving a module between processors as an open modeling question (Stone, 1977, Section VI). We use this discussion as motivation for the stated feedback protocol and the regret criterion with reassignment costs, rather than attributing these guarantees to Stone. If the complete current-period cost were revealed before the placement decision, the model would have lookahead and would not be the online submodular minimization protocol studied here.

We omit an initial migration charge in Equation (59), matching the convention in the main text. If an initial placement S_0 is prescribed, adding $\lambda|S_1 \,\Delta\, S_0|$ changes all guarantees by at most λn.

B.2 An application to losses with finite memory

Following the counterfactual-feedback model of Anava et al. (2015), we allow the incurred loss to depend on a finite window of past decisions. The reduction below only requires value-oracle access to the diagonal of the loss function, rather than access to every counterfactual history.

Corollary 24 (Policy regret for losses with finite memory)**.** *Let $M > 0$ and let $m \in \{0, \dots, T-1\}$ be a known memory length. Suppose that the functions $F_t : (2^{[n]})^{m+1} \to [-M, M]$, $t = 1, \dots, T$, are fixed before play. At round t, the learner chooses $S_t \subseteq [n]$ and incurs $F_t(S_{t-m}, \dots, S_t)$, with an arbitrary fixed initial history $S_{1-m}, \dots, S_0$ when $m > 0$. Assume that the diagonal loss*

$$g_t(S) := F_t(S, \dots, S)$$

is submodular and that, after each round t, the learner can evaluate $g_t(S)$ for any $S \subseteq [n]$. Define the policy regret against the best constant history by

$$\mathrm{PReg}_T := \sum_{t=1}^{T} F_t(S_{t-m}, \dots, S_t) - \min_{S \subseteq [n]} \sum_{t=1}^{T} F_t(S, \dots, S). \tag{60}$$

For every realized decision sequence, writing $\mathcal{R}_T^g(\lambda)$ for Equation (1) evaluated on $(g_t)_{t=1}^T$, we have

$$\mathrm{PReg}_T \le \mathcal{R}_T^g(2Mm) + 2Mm. \tag{61}$$

Consequently, running full-information shared-threshold SSD on g_t with switching-cost weight $\lambda = 2Mm$ and the learning rate from Theorem 2 gives

$$\mathbb{E}[\mathrm{PReg}_T] \le O\Big(M\sqrt{(1+m)nT}\Big) + 2Mm. \tag{62}$$

Proof. Put $D_r := |S_r \,\Delta\, S_{r-1}|$ for $r \geq 2$. For $t > m$, if there is no change in the window $S_{t-m}, \ldots, S_t$, then $F_t(S_{t-m}, \ldots, S_t) = g_t(S_t)$. Otherwise their difference is at most $2M$, so

$$F_t(S_{t-m}, \ldots, S_t) - g_t(S_t) \leq 2M \sum_{r=t-m+1}^{t} \mathbf{1}\{S_r \neq S_{r-1}\} \leq 2M \sum_{r=t-m+1}^{t} D_r.$$

Each D_r appears in at most m such windows. The first m rounds contribute at most $2Mm$, independently of the initial history. Thus, with empty sums interpreted as zero when $m = 0$,

$$\sum_{t=1}^{T} F_t(S_{t-m}, \ldots, S_t) \leq \sum_{t=1}^{T} g_t(S_t) + 2Mm \sum_{r=2}^{T} D_r + 2Mm.$$

Subtracting $\min_S \sum_t g_t(S)$ proves Equation (61). The sequence $(g_t)_{t=1}^T$ is fixed in advance, submodular, and bounded in $[-M, M]$, and the assumed feedback permits all the oracle queries of full-information SSD. Applying Theorem 2 with $\lambda = 2Mm$ proves Equation (62). □

Feedback and comparator. The reduction requires no Lipschitz assumption on F_t. It does require counterfactual diagonal evaluations: neither the scalar observation $F_t(S_{t-m}, \ldots, S_t)$ nor access only to the current-action slice $S \mapsto F_t(S_{t-m}, \ldots, S_{t-1}, S)$ generally reveals $g_t(S)$. Thus this corollary is an application of our full-information guarantee, not a direct application of our bandit guarantee. The comparator in Equation (60) uses the same set throughout its history, including before round one; the term $2Mm$ accounts for the learner's possibly different initial history. Because the multivariate functions F_t are fixed in advance and the comparator is evaluated on its own history, this result does not contradict the adaptive-adversary limitation below.

B.3 Limitations for regret with switching costs: adaptive adversaries and submodular subgradient descent

We first record why an adversary that is adaptive to the learner's past actions rules out sublinear regret with switching costs. We then give two limitations of Algorithm 2 of Hazan & Kale (2012) when it is used for regret with switching costs. The first concerns its original independent rounding, and the second concerns the natural modification that reuses one random threshold throughout the horizon.

Proposition 25 (Adaptive adversaries force linear regret with switching costs)**.** *Against any algorithm for online submodular minimization (OSM), an adversary adaptive to the learner's past actions can choose modular losses $f_t : 2^{[n]} \to [0, M]$ such that, pathwise,*

$$\mathcal{R}_T(\lambda) \geq \frac{T-1}{2} \min\{M, \lambda\}.$$

In particular, when M and λ are fixed and positive, no algorithm guarantees sublinear regret with switching costs against such an adversary.

Proof. We adapt the two-action construction of Altschuler & Talwar (2018, Appendix A) to modular OSM losses and symmetric-difference switching costs. Fix a coordinate $i \in [n]$, write $x_t = \mathbf{1}\{i \in S_t\}$, and let $a = \min\{M, \lambda\}$. Set $f_1 = 0$. After observing S_{t-1}, the adversary selects, for every $t \geq 2$,

$$f_t(S) = a\, \mathbf{1}\{\mathbf{1}\{i \in S\} = x_{t-1}\}.$$

Each f_t is modular and takes values in $[0, M]$. Let

$$K = \sum_{t=2}^{T} \mathbf{1}\{x_t \neq x_{t-1}\}.$$

The learner incurs loss $\sum_{t=2}^{T} f_t(S_t) = a(T - 1 - K)$, and its Hamming movement satisfies $\sum_{t=2}^{T} |S_t \Delta S_{t-1}| \geq K$. Moreover, $\sum_{t=1}^{T} \big(f_t(\emptyset) + f_t([n])\big) = a(T-1)$, so $\min_{S \subseteq [n]} \sum_{t=1}^{T} f_t(S) \leq a(T-1)/2$. Consequently,

$$\begin{aligned} \mathcal{R}_T(\lambda) &\geq a(T - 1 - K) - \frac{a(T-1)}{2} + \lambda K \\ &= \frac{a(T-1)}{2} + (\lambda - a)K \geq \frac{T-1}{2} \min\{M, \lambda\}. \end{aligned}$$

□

Proposition 26 (Independent thresholds cause linear switching costs)**.** *There is a sequence of submodular losses for which submodular subgradient descent initialized at $x_1 = \frac{1}{2}\mathbf{1}$ and using independent rounding thresholds satisfies*

$$\mathbb{E}\left[\sum_{t=2}^{T} |S_t \,\Delta\, S_{t-1}|\right] = \frac{n(T-1)}{2}.$$

Proof. Let $x_1 = \frac{1}{2}\mathbf{1}$ and take $f_t(S) = 0$ for every $S \subseteq [n]$ and every t. Every Lovász subgradient is zero, so the projected subgradient update leaves $x_t = \frac{1}{2}\mathbf{1}$ for all t. For independent $\tau_t \sim \mathrm{Unif}[0,1]$, the rounded decision is

$$S_t = \{i : x_{t,i} > \tau_t\} = \begin{cases} [n], & \tau_t < \frac{1}{2}, \\ \emptyset, & \tau_t \geq \frac{1}{2}. \end{cases}$$

Let $Z_t = \mathbf{1}\{\tau_t < 1/2\}$. Since Z_t and Z_{t-1} are independent Bernoulli random variables with parameter $1/2$,

$$\mathbb{P}(Z_t \neq Z_{t-1}) = \frac{1}{2}.$$

Moreover, $|S_t \,\Delta\, S_{t-1}| = n\mathbf{1}\{Z_t \neq Z_{t-1}\}$. Therefore,

$$\begin{aligned} \mathbb{E}\left[\sum_{t=2}^{T} |S_t \,\Delta\, S_{t-1}|\right] &= n \sum_{t=2}^{T} \mathbb{P}(Z_t \neq Z_{t-1}) \\ &= \frac{n(T-1)}{2}. \end{aligned}$$

□

Proposition 27 (A shared threshold has polynomial tails for regret and movement)**.** *Let T be even. There is a sequence of one-dimensional modular losses such that submodular subgradient descent initialized at $x_1 = 1/2$, with learning rate $\eta = 1/(4M\sqrt{T})$ and one shared threshold satisfies*

$$\mathbb{P}(\mathcal{R}_T(0) = MT) = \mathbb{P}\left(\sum_{t=2}^{T} |S_t \,\Delta\, S_{t-1}| = T - 1\right) = \frac{1}{2\sqrt{T}}.$$

The two events coincide. Thus linear loss regret and linear Hamming movement occur simultaneously with probability $1/(2\sqrt{T})$. On this event, $\mathcal{R}_T(\lambda) = MT + \lambda(T-1)$ for every $\lambda \geq 0$.

Proof. Take $n = 1$, $x_1 = 1/2$, and define the modular losses by

	$f_t(\emptyset)$	$f_t(\{1\})$
t odd	$-M$	M
t even	M	$-M$

.

Their subgradients are $g_t = 2M$ for odd t and $g_t = -2M$ for even t. Therefore, x_t satisfies

$$x_t = \begin{cases} \frac{1}{2}, & t \text{ odd}, \\ \frac{1}{2} - \frac{1}{2\sqrt{T}}, & t \text{ even}. \end{cases}$$

Draw one $\tau \sim \mathrm{Unif}[0,1]$ and reuse it at every round. The rounded decision is

$$S_t = \{1 \mid x_t > \tau\} = \begin{cases} \{1\}, & x_t > \tau, \\ \emptyset, & x_t \leq \tau. \end{cases}$$

On the event

$$\{\frac{1}{2} - \frac{1}{2\sqrt{T}} \leq \tau < \frac{1}{2}\},$$

which has probability $1/(2\sqrt{T})$, the algorithm alternates between $S_t = \{1\}$ at odd rounds and $S_t = \emptyset$ at even rounds. It therefore incurs loss M at every round and switches at every transition. Since T is even, both fixed decisions have cumulative loss zero. This gives $\mathcal{R}_T(0) = MT$ and $\sum_{t=2}^{T} |S_t \,\Delta\, S_{t-1}| = T - 1$ on this event. Outside this event, the algorithm plays the same set at every round, so its cumulative loss, loss regret, and Hamming movement are all zero. Hence both displayed probabilities equal $1/(2\sqrt{T})$, and the identity for regret with switching costs follows. □

Table 2: Horizon comparison: mean regret with switching costs per round, $\mathcal{R}_T(1)/T$, with pointwise approximate 95% confidence intervals (50 paired trials). Fixed policies are specified in advance. Full-info uses shared-threshold SSD; bandit uses cyclic SSD without fixed-action fallback.

Regime	T	Fixed $\emptyset$	Fixed $[n]$	Full-info	Bandit
Stationary	500	0.528 ± 0.045	0.491 ± 0.042	0.199 ± 0.031	1.557 ± 0.058
	2000	0.528 ± 0.045	0.491 ± 0.042	0.094 ± 0.015	1.217 ± 0.046
	8000	0.528 ± 0.045	0.491 ± 0.042	0.046 ± 0.008	0.948 ± 0.048
Nonstationary	500	0.133 ± 0.023	0.138 ± 0.025	0.104 ± 0.018	1.258 ± 0.052
	2000	0.142 ± 0.024	0.141 ± 0.024	0.026 ± 0.012	0.893 ± 0.034
	8000	0.143 ± 0.028	0.125 ± 0.021	-0.081 ± 0.013	0.626 ± 0.026

C Numerical Experiments

C.1 Experimental overview and results

We examine how threshold coupling and cyclic exploration affect loss regret and switching cost on synthetic submodular losses, including comparisons with policies that never switch. We use graph cuts with unary costs favoring a planted subset in stationary and nonstationary regimes. We set $n = 20$, $M = \lambda = 1$, and $T \in \{500, 2000, 8000\}$, with 50 paired trials per setting and horizon-dependent theoretical parameters. We also report a stationary follow-up at $T = 128000$ with weaker edge weights. Section C.2 specifies loss generation, trial pairing, comparators, and confidence intervals; subsequent subsections describe the implementations and the $T = 500$ exploration-rate comparison.

Comparisons. Under full information we compare Submodular FTPL (Hazan & Kale, 2012), independent-threshold SSD, shared-threshold SSD, and movement-budget restarting versions of FTPL and shared SSD. Under bandit feedback we compare roundwise independent-threshold SSD, roundwise shared-threshold SSD, cyclic shared-threshold SSD, and restarting versions of the latter two. The bandit baseline follows Lattimore's reformulation of the Hazan–Kale algorithm (Lattimore, 2026, Algorithm 13.1), with independent thresholds and an exploration-only estimator; see Algorithm 4. We additionally compare the two prespecified policies that always play $\emptyset$ or $[n]$.

Horizon dependence and fixed policies. Table 2 shows that increasing the horizon improves cyclic bandit SSD's mean regret with switching costs per round, from 1.557 to 0.948 under stationary losses and from 1.258 to 0.626 under nonstationary losses. Nevertheless, cyclic SSD remains worse than both fixed policies at each of these three horizons. All twelve paired mean differences (cyclic minus each fixed policy, across the six settings) have pointwise 95% confidence intervals strictly above zero. Full-information shared-threshold SSD, in contrast, has smaller mean regret with switching costs than both fixed policies at all three horizons, in both regimes. At $T = 8000$, its mean regret with switching costs per round is 0.046 under stationary losses and -0.081 under nonstationary losses. A negative value is possible because an adapting learner can outperform the best fixed comparator.

Loss regret versus switching cost. Figure 2 separates the two terminal contributions; Figures 3 and 4 show their evolution over rounds, with parameters fixed for the $T = 8000$ horizon. Prefix regret need not be monotone, since its fixed hindsight comparator is recomputed at each round. For cyclic SSD under stationary losses, increasing T from 500 to 8000 changes loss regret per round only from 0.437 to 0.442, while switching cost per round falls from 1.121 to 0.506. Under nonstationary losses the corresponding changes are 0.136 to 0.117 and 1.122 to 0.509. Thus, the improvement over these horizons comes primarily from reduced switching cost per round. Any loss-regret advantage over a fixed policy is outweighed by the movement cost in these settings. Cyclic exploration still improves on roundwise shared-threshold SSD: at $T = 8000$, the respective mean values of regret with switching costs per round are 0.948 versus 1.235 (stationary) and 0.626 versus 0.931 (nonstationary). At $T = 500$, the comparison at matched exploration probability in Table 6 also shows substantially less movement under cyclic exploration with a similar number of exploration rounds; the learning rates are optimized separately for the two methods.

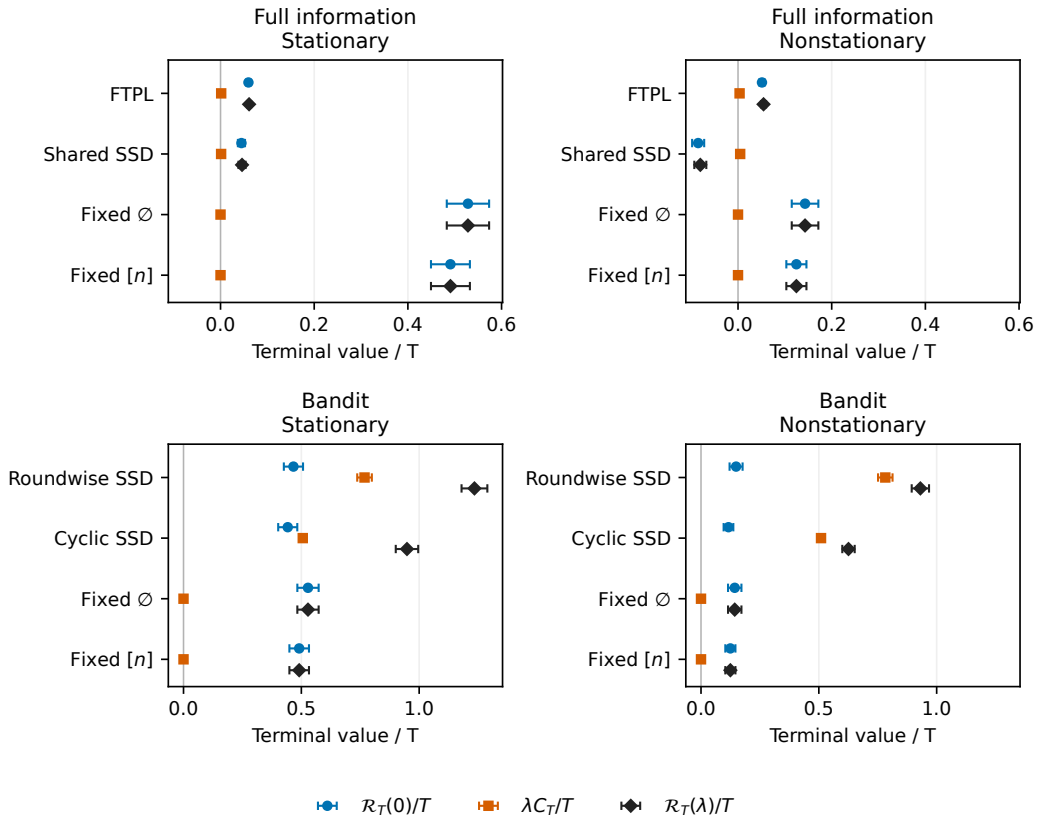


Figure 2: Terminal regret decomposition at $T = 8000$ (50 paired trials). Columns show stationary and nonstationary losses; rows show the feedback models. Symbols separate loss regret $\mathcal{R}_T(0)$, switching cost λC_T, and regret with switching costs $\mathcal{R}_T(\lambda)$, each divided by T; horizontal bars give pointwise approximate 95% confidence intervals for their means. The figure focuses on shared-threshold learners, FTPL, and fixed policies. Restarted counterparts coincide with their bases because no prescribed-budget restart occurred; they are omitted here to avoid duplicate points. Both fixed policies have zero switching cost, so their loss and total regret agree.

Table 3: Stationary follow-up at $T = 128000$ with edge-weight generation coefficient 0.01: 40 additional paired trials, excluding the ten-trial selection pilot. Entries are means divided by T. Since $\lambda = 1$, Hamming movement equals switching cost.

Method	$\mathcal{R}_T(0)/T$	C_T/T	$\mathcal{R}_T(\lambda)/T$
Cyclic bandit SSD	0.44678	0.19328	0.64007
Fixed $\emptyset$	0.82022	0	0.82022
Fixed $[n]$	0.79915	0	0.79915

Longer horizon and weaker cut weights. In a small exploratory sensitivity experiment at $n = 20$, $T = 8000$, and $M = \lambda = 1$, we varied the edge-weight coefficient over $\{0.1, 0.05, 0.01, 0\}$, using ten paired trials per coefficient in each regime. We retained the loss normalization and theoretically tuned learning and exploration rates. Weakening the cut term reduced the disadvantage relative to the fixed policies but did not reverse the comparison, even in the modular case with coefficient zero. A subsequent stationary ten-trial pilot at $T = 128000$ identified a favorable setting with edge weights $w_{t,ij} \sim 0.01\,\mathrm{Gamma}(1, 1)$, retaining positive cut weights. We fixed this setting and evaluated 40 additional seed pairs unused at this setting, without further parameter selection or early stopping. The loss normalization, $n = 20$, and $M = \lambda = 1$ are unchanged; the theoretically selected rates are $\eta \approx 4.2142 \times 10^{-5}$ and $\gamma \approx 0.076654$, without empirical tuning.

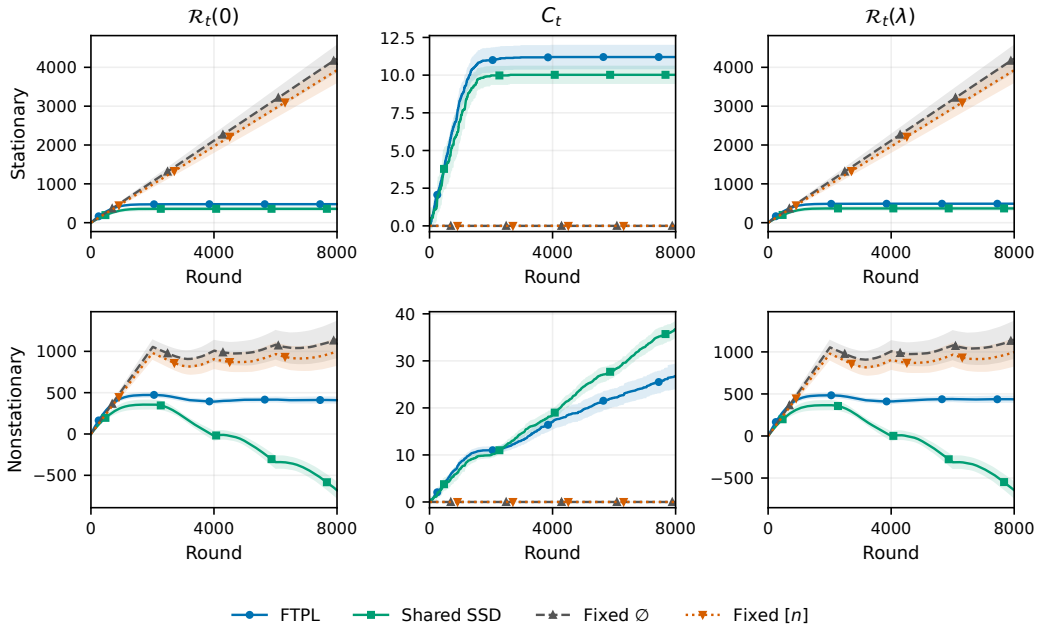


Figure 3: Full-information trajectories for horizon $T = 8000$. The horizontal axis is the round t; columns show cumulative loss regret $\mathcal{R}_t(0)$, cumulative Hamming movement C_t, and regret with switching costs $\mathcal{R}_t(1) = \mathcal{R}_t(0) + C_t$, without division by t or T. Rows show stationary and nonstationary losses. Curves are means over the same 50 trials as in Figure 2; shading denotes pointwise approximate 95% confidence intervals for the means. Regret uses the exact best fixed comparator for each prefix. Fixed-policy movement curves coincide at zero. Restarted counterparts coincide with their bases and are omitted.

In these additional trials, cyclic SSD has smaller mean regret with switching costs than both fixed policies (Table 3). The paired mean differences divided by T (cyclic minus fixed) are -0.18016 against $\emptyset$ and -0.15909 against $[n]$. Bonferroni-adjusted Student-t intervals with simultaneous approximate 95% coverage for these two comparisons are $[-0.27367, -0.08664]$ and $[-0.29652, -0.02166]$, respectively. Movement still accounts for about 30% of its mean regret with switching costs, but the loss-regret improvement now outweighs that cost. This supports an advantage on this particular loss distribution, not on every instance or against the better of the two fixed policies selected after observing each instance. Both the horizon and the edge-weight coefficient differ from the main sweep; the reversal cannot be attributed to increasing T alone. Reducing the coefficient also changes the normalized unary scale because the whole objective is renormalized.

An exploratory stationary run at $T = 32000$ with coefficient 0.01, using the same ten pilot seed pairs and horizon-specific theoretical parameters, gave mean regret with switching costs per round 0.91066 for cyclic SSD, versus 0.87093 and 0.73822 for fixed $\emptyset$ and $[n]$, respectively. The loss-regret improvement was outweighed by movement; however, pointwise 95% Student-t intervals for both paired cyclic-minus-fixed differences included zero, so these ten trials do not resolve the expected ordering.

Relation to the guarantees and scope. The learning branches are evaluated without the fixed-action fallback. For the chosen cyclic parameters, the expected-regret upper bound exceeds $2MT$ at $T = 500, 2000, 8000$ but is smaller at $T = 128000$. Thus the theoretical bandit policy selects fixed $\emptyset$ at the three shorter horizons and the learning branch in the follow-up; full-information shared SSD selects its learning branch throughout. The cyclic learner's disadvantage on the shorter horizons is consistent with the theory, which bounds worst-case expected regret rather than guaranteeing superiority over a fixed policy on every instance. Smaller nonstationary regret does not imply an easier tracking problem: redrawing the planted subset also weakens the best fixed comparator. These experiments characterize finite-horizon performance on a bounded cut-loss family; they do not es-

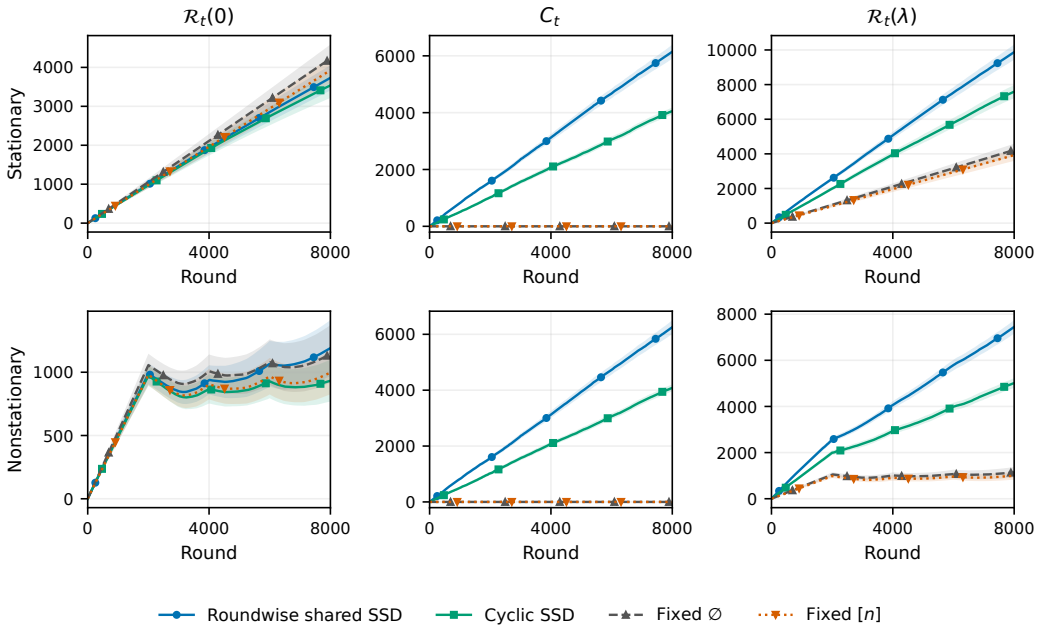


Figure 4: Bandit trajectories for horizon $T = 8000$, with the axes, trial pairing, prefix comparator, and confidence intervals of Figure 3. Roundwise shared-threshold SSD and cyclic SSD are evaluated without fixed-action fallback. The same fixed empty/full policies are shown for reference. Restarted counterparts are omitted because their trajectories coincide with those of the base learners.

tablish worst-case rates or high-probability tail guarantees. Separately, Section C.5 tests the modular construction of Proposition 27, contrasting mean regret with upper quantiles and comparing shared SSD with its batched counterpart.

C.2 Loss generation and evaluation

For each trial we sample an undirected graph on n vertices, independently including each edge with probability 0.25, and a planted binary vector $y \in \{0,1\}^n$ with independent fair coordinates. The graph remains fixed. At each round draw independent $m_{t,i}, p_{t,i} \sim \mathrm{Unif}[0.7, 1.3]$ and $w_{t,ij} \sim 0.1\,\mathrm{Gamma}(1,1)$, where the Gamma distribution has unit shape and scale. The weaker-cut follow-up in Table 3 replaces the coefficient 0.1 by 0.01, with the same normalization below. Set

$$a^{(1)}_{t,i} = p_{t,i}(1 - 0.85y_i), \qquad a^{(0)}_{t,i} = m_{t,i}(0.15 + 0.85y_i),$$

and define the raw objective

$$q_t(S) = \sum_{i \in S} a^{(1)}_{t,i} + \sum_{i \notin S} a^{(0)}_{t,i} + \sum_{(i,j) \in E} w_{t,ij} |\mathbf{1}\{i \in S\} - \mathbf{1}\{j \in S\}|.$$

Writing $U_t = \sum_i \max\{a^{(1)}_{t,i}, a^{(0)}_{t,i}\} + \sum_{(i,j)\in E} w_{t,ij}$, we use $f_t(S) = 2Mq_t(S)/U_t - M$. Thus $|f_t(S)| \le M$ for every subset; $M = 1$ bounds the entire function, not each edge weight. The unary term avoids the automatic empty/full minimizers of a pure nonnegative cut. Since M is fixed, this normalization changes the per-coordinate scale as n grows.

Stationary means that y remains fixed, not that every realized loss is identical. Nonstationary runs independently redraw y every $T/4$ rounds (at rounds 126, 251, and 376 when $T = 500$), holding the number of redraws rather than their frequency fixed across horizons. The complete sequence is generated before learning. All methods and both feedback models receive the same sequence within each trial. Each independent/shared pair uses the same learning rate and continuous trajectory; base/restarted pairs are coupled until the first restart. For the original trajectory plots, we compute the best fixed subset for every prefix exactly using an s–t minimum cut; the horizon-comparison

Algorithm 3: Submodular FTPL (full-information baseline)

Input: Learning rate $\eta_{\mathrm{F}} > 0$, loss bound M
Independently draw $r_i \sim \mathrm{Unif}[-M/\eta_{\mathrm{F}}, M/\eta_{\mathrm{F}}]$ for $i \in [n]$;
Initialize $F_0(S) = 0$;
for $t = 1, \ldots, T$ **do**
  Play $S_t \in \arg\min_{S \subseteq [n]} \{F_{t-1}(S) + \sum_{i \in S} r_i\}$;
  Observe f_t and set $F_t = F_{t-1} + f_t$;

table in Section C.1 evaluates terminal metrics, while its $T = 8000$ trajectories also use exact prefix comparators. The reported quantities are loss regret $\mathcal{R}_T(0)$, total Hamming movement C_T, and $\mathcal{R}_T(\lambda) = \mathcal{R}_T(0) + \lambda C_T$; no initial-action switching charge is included. Bandit learners receive only the selected scalar loss; full loss objects are used solely for evaluation. The prespecified fixed policies always play $\emptyset$ or $[n]$, without using observed losses to choose their set, and incur no switching cost. They are distinct from the best fixed subset in hindsight defining regret.

The main configurations use 50 trials with master seed zero; the exploratory edge-weight sensitivity experiment in Section C.1 uses ten paired trials per setting. We record the separate loss and algorithm seeds. Confidence intervals include variation in both generated losses and algorithmic randomness and are $\bar{z} \pm 1.96 s/\sqrt{50}$ for the main configurations, where s is the sample standard deviation across trials. They are pointwise intervals for a mean, not simultaneous confidence bands or bounds on the realized regret of a single run. The stationary $T = 128000$, coefficient-0.01 follow-up fixes its setting after a ten-trial exploratory pilot (trial IDs 0–9) and adds 40 trials (IDs 10–49) from the original master-seed schedule. These additional seed pairs were previously used in the main horizon sweep, but not at this follow-up setting. Only cyclic bandit SSD and the two fixed policies are evaluated in this follow-up, using an exact terminal min-cut comparator. Its primary comparison uses the additional 40 trials alone; the pooled 50-trial means include the selection pilot and are descriptive. For each fixed policy let d_i be the paired difference in regret with switching costs divided by T. The simultaneous approximate 95% intervals reported for the two comparisons are $\bar{d} \pm t_{39,0.9875} s_d/\sqrt{40}$, using a Bonferroni correction, where $t_{\nu,p}$ is the Student-t quantile.

C.3 Baselines and implementation

Submodular FTPL is Algorithm 1 of Hazan & Kale (2012); the fixed modular perturbation preserves submodularity and permits exact min-cut optimization for our losses. Its restarting version is an experimental application of the movement-budget wrapper, with no high-probability theorem claimed here.

Independent-threshold full-information SSD uses the same initialization, chain subgradient, and projected update as shared-threshold SSD, but redraws $\tau_t \sim \mathrm{Unif}[0, 1]$ at each round. For the bandit comparison, Algorithm 4 follows the reformulation of the Hazan–Kale bandit SSD (Hazan & Kale, 2012) presented in Lattimore (2026, Algorithm 13.1). Lattimore's formulation explicitly separates uniform-chain exploration from threshold-based exploitation and uses only exploration observations in the subgradient estimator, with weight $(n+1)/\gamma$. We use this formulation with $x_1 = \frac{1}{2}\mathbf{1}$, losses in $[-M, M]$, and the experimental parameter choices below. The baseline draws a fresh threshold each round; the shared variant reuses one threshold per epoch. This permits identical exploration draws and continuous trajectories for the independent/shared comparison. For clarity, the original Algorithm 3 of Hazan & Kale (2012) instead importance-weights the full mixed action distribution and randomly selects one coordinate at an interior chain set.

Cyclic SSD uses Algorithm 2. Full-information restarting monitors movement between the sets played within each epoch and resets after its budget is exceeded. Roundwise bandit restarting monitors $\sum_t |S_\tau(x_{t+1}^{\mathrm{proposed}}) \,\Delta\, S_\tau(x_t)|$; cyclic restarting uses the blockwise virtual monitor in Equation (8). When the monitor exceeds its budget, the algorithm discards $x_{t+1}^{\mathrm{proposed}}$ and resets x and the threshold before the next round or complete block. Exploration movement and restart-boundary movement are included in the reported C_T even though they are not charged to the virtual monitor. The final incomplete cyclic block holds the last action fixed.

Algorithm 4: Roundwise bandit SSD following Lattimore's reformulation of Hazan–Kale

Input: $\eta > 0$, $\gamma \in (0, 1]$; independent or shared threshold mode
Set $x_1 = \frac{1}{2}\mathbf{1}$; in shared mode draw $\tau \sim \mathrm{Unif}[0, 1]$;
for $t = 1, \dots, T$ **do**
  Sort the coordinates of x_t in decreasing order with deterministic tie-breaking, obtaining π_t and $B_{t,j} = \{\pi_t(1), \dots, \pi_t(j)\}$ for $j = 0, \dots, n$;
  In independent mode draw $\tau_t \sim \mathrm{Unif}[0, 1]$; in shared mode set $\tau_t = \tau$;
  Independently draw $E_t \sim \mathrm{Ber}(\gamma)$ and $I_t \sim \mathrm{Unif}\{0, \dots, n\}$;
  Play $X_t = B_{t,I_t}$ if $E_t = 1$; otherwise play $S_{\tau_t}(x_t)$;
  Observe only $\ell_t = f_t(X_t)$;
  Set $\widehat{g}_t = \frac{(n+1)E_t\ell_t}{\gamma}[\mathbf{1}\{I_t \neq 0\}e_{\pi_t(I_t)} - \mathbf{1}\{I_t \neq n\}e_{\pi_t(I_t+1)}]$;
  Update $x_{t+1} = \Pi_{[0,1]^n}(x_t - \eta\widehat{g}_t)$;

Table 4: Full information: final metrics under prescribed budgets. Entries are means; $\pm$ gives the approximate 95% confidence interval for mean regret with switching costs.

Method	$\mathcal{R}_T(0)$	C_T	$\mathcal{R}_T(1)$	Restarts
Stationary				
FTPL	118.77	11.28	130.05 ± 8.98	0.00
Batched FTPL	118.77	11.28	130.05 ± 8.98	0.00
Independent SSD	100.63	2695.86	2796.49 ± 80.47	0.00
Shared SSD	89.27	10.02	99.29 ± 15.36	0.00
Batched shared SSD	89.27	10.02	99.29 ± 15.36	0.00
Nonstationary				
FTPL	83.78	16.10	99.88 ± 8.07	0.00
Batched FTPL	83.78	16.10	99.88 ± 8.07	0.00
Independent SSD	36.93	4291.24	4328.17 ± 76.44	0.00
Shared SSD	36.02	15.74	51.76 ± 8.95	0.00
Batched shared SSD	36.02	15.74	51.76 ± 8.95	0.00

C.4 Hyperparameters and additional comparisons

We use $\eta_{\mathrm{F}} = T^{-1/2}$ for FTPL and $\eta = \sqrt{n/[32M(M+\lambda)T]}$ for all full-information SSD variants. For roundwise bandit SSD we minimize

$$\frac{n}{8\eta} + \eta T\left(\frac{(n+1)^2M^2}{\gamma} + 2\lambda nM\right) + 2\gamma T(M + \lambda n).$$

The independent baseline uses the shared method's parameters, thereby isolating threshold coupling. For cyclic SSD we minimize Equation (16) with $T' = (n+1)\lfloor T/(n+1)\rfloor$. For each of 2400 logarithmically spaced γ values in $[10^{-6}, 1]$, η is optimized analytically, and the best pair is selected. We run these learners even when a trivial fixed-action guarantee is smaller: the experiment compares the learning mechanisms, not the minimax fixed-action fallback.

Restarted methods inherit their base learning/exploration parameters, rather than being separately optimized for the high-probability bound. With $\delta = 0.05$, full-information budgets are $8n\eta_{\mathrm{F}}H$ for FTPL and $16M\eta H$ for SSD, where $H = 4T/N_\delta$ and $N_\delta = \max\{4, \lceil 2\log(1/\delta)/\log 3\rceil\}$. For bandit feedback we use $N_\delta = \max\{4, \lceil 2\log(3/\delta)/\log 3\rceil\}$ and virtual-movement budgets $32\eta nMT/N_\delta$ (roundwise) and $32\eta nMT'/N_\delta$ (cyclic). The roundwise wrapper is an additional experimental comparison; the present manuscript's bandit wrapper theorem concerns cyclic SSD.

The effective full-information values are $\eta_{\mathrm{F}} = 0.04472$, $\eta_{\mathrm{SSD}} = 0.02500$, and budgets 2385.14 (FTPL) and 133.33 (SSD). Roundwise bandit uses $(\eta, \gamma) = (0.001100, 0.1077)$; cyclic uses $(0.001717, 0.4896)$. Their virtual budgets are 43.98 and 66.35, respectively.

A bandit ablation fixes the cyclic exploration probability to the roundwise value, reoptimizing its learning rate; this separates the effect of chain traversal from changes in the expected number of exploration rounds.

Table 5: Bandit feedback: final metrics under prescribed budgets. Entries are means; $\pm$ gives the approximate 95% confidence interval for mean regret with switching costs.

Method	$\mathcal{R}_T(0)$	C_T	$\mathcal{R}_T(1)$	Restarts
Stationary				
Independent SSD	253.75	4870.00	5123.75 ± 60.81	0.00
Shared SSD	246.06	958.82	1204.88 ± 58.06	0.00
Cyclic SSD	218.37	560.28	778.65 ± 29.09	0.00
Batched shared SSD	246.06	958.82	1204.88 ± 58.06	0.00
Batched cyclic SSD	218.37	560.28	778.65 ± 29.09	0.00
Nonstationary				
Independent SSD	78.14	4873.78	4951.92 ± 60.87	0.00
Shared SSD	80.08	958.52	1038.60 ± 51.05	0.00
Cyclic SSD	68.22	561.00	629.22 ± 26.19	0.00
Batched shared SSD	80.08	958.52	1038.60 ± 51.05	0.00
Batched cyclic SSD	68.22	561.00	629.22 ± 26.19	0.00

Table 6: Bandit exploration-rate ablation. Cyclic matched uses the roundwise γ and reoptimizes η. Exploration is the mean number of exploration rounds.

Method	γ	Exploration	C_T	$\mathcal{R}_T(1)$
Stationary				
Roundwise shared	0.1077	54.1	958.8	1204.88 ± 58.06
Cyclic default	0.4896	229.3	560.3	778.65 ± 29.09
Cyclic matched	0.1077	53.3	145.4	403.58 ± 31.05
Nonstationary				
Roundwise shared	0.1077	54.1	958.5	1038.60 ± 51.05
Cyclic default	0.4896	229.3	561.0	629.22 ± 26.19
Cyclic matched	0.1077	53.3	145.5	219.32 ± 26.25

At matched γ, cyclic exploration uses similar expected exploration time and much less movement than roundwise exploration. Its lower empirical regret with switching costs than the default cyclic tuning also shows that minimizing a worst-case upper bound need not optimize performance on this loss family.

C.5 Restarting on a polynomial-tail stress test

To compare shared-threshold SSD with and without the movement-budget restart wrapper, we implement the oblivious modular loss sequence of Proposition 27 under full information. We use $n = 1$, $M = \lambda = 1$, even horizons $T \in \{100, 400, 1600, 6400\}$, and 20,000 independent algorithmic trials per horizon (master seed 2027). The choice $n = 1$ isolates threshold rounding without graph or oracle effects. Odd-round losses at $\emptyset, \{1\}$ are $-M, M$; even-round losses are $M, -M$. Both methods use the proposition's $\eta = 1/(4M\sqrt{T})$, rather than the planted-cut tuning above, and share their initial threshold within each paired trial. The batched method uses the prescribed full-information movement budget with $\delta = 10^{-3}$; the two trajectories coincide until its first restart. Unlike the main planted-cut experiments, this construction produces trajectories on which the prescribed budget is exceeded. Regret with switching costs includes movement at restart boundaries.

For shared SSD the exact distribution is particularly transparent:

$$(\mathcal{R}_T(0), C_T) = \begin{cases} (MT, T-1), & \text{with probability } p_T = 1/(2\sqrt{T}), \\ (0, 0), & \text{otherwise.} \end{cases}$$

Consequently $\mathbb{E}[\mathcal{R}_T(0)] = M\sqrt{T}/2$, but its $(1-\delta)$-quantile equals MT whenever $\delta < p_T$. We therefore report 99.9% quantiles, not only 95% intervals for means. This is a rare polynomial tail, not a statement that most runs suffer large regret. At fixed δ, the linear quantile disappears once $p_T \le \delta$; all tested horizons satisfy $p_T > 10^{-3}$.

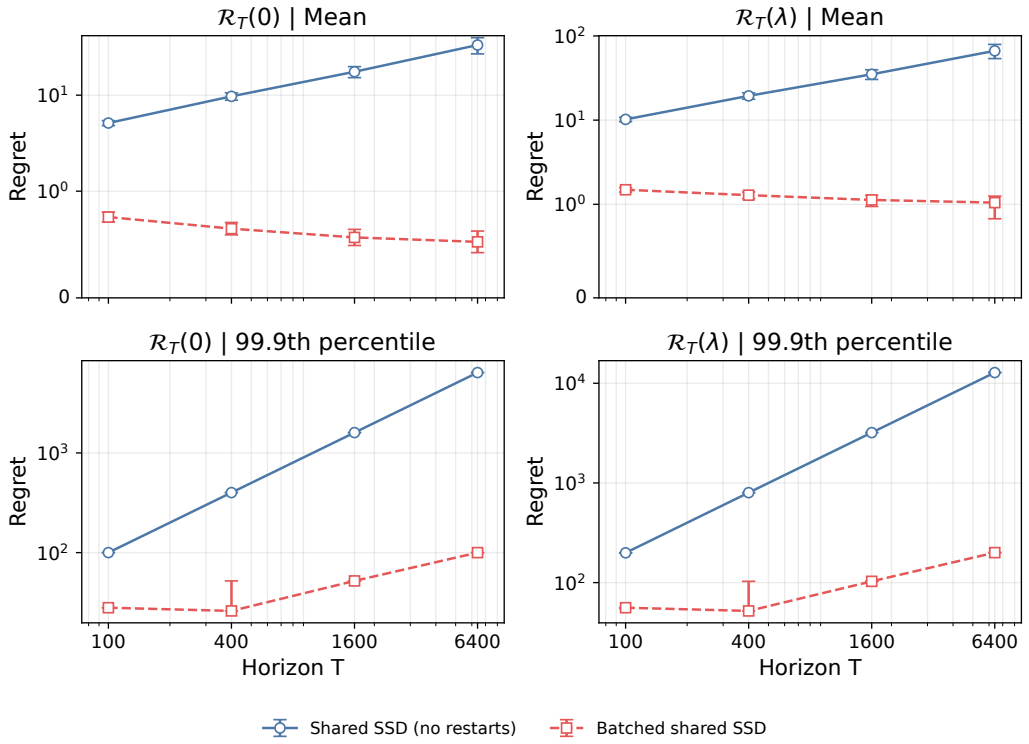


Figure 5: Full-information shared SSD with and without batching on the construction of Proposition 27, using 20,000 paired trials per horizon. Columns show loss regret $\mathcal{R}_T(0)$ and regret with switching costs $\mathcal{R}_T(\lambda)$, with $\lambda = 1$. Top: means with approximate 95% confidence intervals. Bottom: empirical 99.9% quantiles with conservative 95% order-statistic intervals. Both methods use the same learning rate; batching adds the prescribed movement-budget restarts. Horizontal axes are logarithmic and vertical axes are symmetric logarithmic.

At $T = 6400$, the mean regret with switching costs falls from 66.55 without batching to 1.05 with batching; the empirical 99.9% quantile falls from 12799 to 200. The corresponding loss-regret quantiles are 6400 and 100. The batched learner restarts in 104 of 20000 trials (105 restarts in total). Thus a small average restart count does not imply an inactive wrapper: restarts occur on rare trajectories with sustained movement. These empirical quantiles illustrate the effect of restarting on this construction; they do not establish a high-probability guarantee.

The quantile intervals invert the binomial order-statistic coverage and are conservative for discrete distributions. All intervals are pointwise, not simultaneous. The implementation vectorizes independent trials and is tested against the scalar implementations, including movement at restart boundaries. This is not a bandit tail test.

D Additional Related Work

Submodular function minimization. Given value-oracle access to a fixed submodular function, unconstrained submodular function minimization is solvable in polynomial time via the ellipsoid method (Grötschel et al., 1981). Strongly polynomial combinatorial algorithms were subsequently obtained by Schrijver (2000) and Iwata et al. (2001). Later work improved the strongly polynomial running time (Orlin, 2009), and cutting-plane techniques led to further improvements in both weakly and strongly polynomial algorithms (Lee et al., 2015). These offline results optimize one fixed function through repeated oracle queries; in online submodular minimization, by contrast, the loss function changes in each round and performance is measured by regret, with the present work also charging for movement between consecutive decisions.