A Section 2: Follow the Perturbed Multiple Leaders

Proof of Proposition 1

Proof. Construct a cost sequence inductively for $t = 1, 2, \ldots, T$: Let $S_t \subset \mathcal{A}$ be the deterministic choice of arms the bandit algorithm chooses for round t given the previously chosen cost functions $c_1, \ldots, c_{t-1}$. Now choose $c_t(a) = 1$ if $a \in S_t$, and $c_t(a) = 0$ otherwise. Then the bandit algorithm achieves cost T. The total cost summed over all arms is $\sum_{t=1}^{T} |S_t| \leq BT$, so there must exist at least one $a' \in \mathcal{A}$ such that $\sum_{t=1}^{T} c_t(a') \leq \frac{BT}{N}$. Thus $R_T^* \geq T - \frac{BT}{N} = (1 - \frac{B}{N})T$. □

Proof of Lemma 3

Proof. Let $R_i = \sum_{t=1}^{i} c_t(S_t^*) - \min_{a^* \in \mathcal{A}} \sum_{t=1}^{i} c_t(a^*)$ be the regret at the end of round i. Then the increase in regret in round i is

$$\begin{aligned}
r_i &:= R_i - R_{i-1} \\
&= \sum_{t=1}^{i} \left(c_t(S_t^*) - c_t(a_i^{*,1}) \right) - \sum_{t=1}^{i-1} \left(c_t(S_t^*) - c_t(a_{i-1}^{*,1}) \right) \\
&= c_i(S_i^*) - c_i(a_i^{*,1}) + \left(\sum_{t=1}^{i-1} c_t(a_{i-1}^{*,1}) - \sum_{t=1}^{i-1} c_t(a_i^{*,1}) \right) \\
&\leq c_i(S_i^*) - c_i(a_i^{*,1}) \\
&\leq \mathbb{1}[a_i^{*,1} \notin S_i^*]
\end{aligned}$$

and the result follows by evaluating $\sum_{t=1}^{T} r_t$. □

Lemma 9. *(Proof of independence for Lemma 4) Let $X_1, \ldots, X_K$ be jointly independent continuous random variables. Let $i_1, \ldots, i_k$ and $v_{i_1}, \ldots, v_{i_k}$ be the indices and values of the largest $k < K$ random variables, and let $X := \{X_i | i \notin \{i_1, \ldots, i_k\}$ be the smallest $K - k$ random variables. Then conditional on $(i_j, v_{i_j})_{j \in [k]}$, the values of each $X_i \in X$ are jointly independent. Moreover, the marginal distribution $X_i | (i_j, v_{i_j})_{j \in [k]}$ for $i \notin \{i_1, \ldots, i_k\}$ is $X_i | X_i \leq \min_{j \in [k]} v_{i_j}$.*

Proof. Let $M := \min_{j \in [k]} v_{i_j}$. The conditional joint density function is

$$\begin{aligned}
f(X_1, \ldots, X_K | (i_j, v_{i_j})_{j \in [k]}) &\propto f(X \wedge (i_j, v_{i_j})_{j \in [k]}) \\
&= \prod_{j \in [K] - \{i_1, \ldots, i_k\}} f(X_j) \prod_{j \in \{i_1, \ldots, i_k\}} f(X_j = v_j) \prod_{j \in [K] - \{i_1, \ldots, i_k\}} \mathbb{1}[X_j \leq M] \\
&\propto \prod_{j \in [K] - \{i_1, \ldots, i_k\}} f(X_j) \mathbb{1}[X_j \leq M]
\end{aligned}$$

i.e. the joint density factorizes for each X_j (which implies joint independence), and marginally the density for $X_j \in X$ is $\propto f(X_j)\mathbb{1}[X_j \leq M]$ which gives the required result. □

Proof of Lemma 4

Proof. Fix a round t. Consider the jointly independent random variables $X_a = \tilde{C}_{t-1}(a)$ for $a \in \mathcal{A}$. Condition on the values and identities of the $N - B$ largest of these random variables, i.e. condition on $E = \{(X_{a_{t-1}^{*,j}}, a_{t-1}^{*,j})\}_{j=B+1}^{N}$, and let $M = X_{a_{t-1}^{*,B+1}}$ be the minimum perturbed cost among these non-leading arms. Impose an ordering on $\mathcal{A}$ and let $l_1, \ldots, l_B \in \mathcal{A} - \{a^{*,j}\}_{j=B+1}^{N}$ be the remaining arms (the top B leaders) ordered lexicographically (i.e. not necessarily in order of cumulative perturbed

cost). Then it can be shown that the distribution of the random variables $X_{l_1}, \ldots, X_{l_B}$ conditioned on E is jointly independent, and the marginal distribution of X_{l_j} given E is $X_{l_j}|(X_{l_j} \leq M)$ (see Lemma 9). Now observe that if $X_{l_j} < M - c_t(l_j)$ for any $j \in [B]$, then the event $(\tilde{a}_t^{*,1} \notin \tilde{S}_t^*)$ is impossible. This is because $l_j \in \tilde{S}_t^*$, but for any $a \notin S_t^*$, $\tilde{C}_t(a) \geq M$ but $\tilde{C}_t(l_j) = X_{l_j} + c_t(l_j) < M$ (i.e. l_j cannot be overtaken by any non-top-B-leader in round t). Therefore we have

$$\mathbb{E}\left[\mathbb{1}[\tilde{a}_t^{*,1} \notin \tilde{S}_t^*]|E\right] \leq P\left[\bigwedge_{j=1}^{B} \neg(X_{l_j} < M - c_t(l_j)) \,\middle|\, E\right] \tag{1}$$

$$= \prod_{j=1}^{B} P\left[\neg(X_{l_j} < M - c_t(l_j))|X_{l_j} \leq M\right] \tag{2}$$

$$= \prod_{j=1}^{B} \left(1 - P\left[p(l_j) > C_{t-1}(l_j) + c_t(l_j) - M | p(l_j) \geq C_{t-1}(l_j) - M\right]\right) \tag{3}$$

$$\leq \prod_{j=1}^{B} \left(1 - P\left[p(l_j) > c_t(l_j)\right]\right) \tag{4}$$

$$= \prod_{j=1}^{B} \left(1 - \int_{c_t(l_j)}^{\infty} \varepsilon e^{-\varepsilon x} dx\right) = \prod_{j=1}^{B} (1 - e^{-\varepsilon c_t(l_j)}) \tag{5}$$

$$\leq \prod_{j=1}^{B} (\varepsilon c_t(l_j)) \leq \varepsilon^B c_t(\tilde{S}_t^*) \tag{6}$$

(2) follows by independence, (4) is due to the memorylessness property of the exponential distribution (with equality unless $C_{t-1}(l_j) - M < 0$), and (6) follows because $1 - e^{-x} \leq x$ for $x \geq 0$ and $c_t(S_t^*) \geq \prod_{a \in S_t^*} c_t(a) = \prod_{j=1}^{B} c_t(l_j)$. The final claim follows by taking the expectation over the conditioned event E. □

Proof of Theorem 2

Proof. Consider a modified version of FPML where $p_t(a) = p_1(a) = p(a)$ for all $t > 1$ (i.e. we keep the random perturbation fixed across rounds). Then this version of FPML picks the set $\tilde{S}_t^*$ in round t, and the regret can be bounded as

$$\sum_{t=1}^{T} c_t(\tilde{S}_t^*) - \min_{a^* \in \mathcal{A}} \sum_{t=1}^{T} c_t(a^*) = \left(\sum_{t=1}^{T} c_t(\tilde{S}_t^*) - \min_{\tilde{a}^* \in \mathcal{A}} \sum_{t=1}^{T} c_t(\tilde{a}^*) - p(\tilde{a}^*)\right) + \left(\min_{\tilde{a}^* \in \mathcal{A}} \sum_{t=1}^{T} c_t(\tilde{a}^*) - p(\tilde{a}^*) - \min_{a^* \in \mathcal{A}} \sum_{t=1}^{T} c_t(a^*)\right)$$

The second term is ≤ 0. The first term can be interpreted as the regret of a modified version of MAB with a 0^{th} round with cost function $-p$, where we are only allowed to pull arms from round $t = 1$. The regret increase incurred in the 0^{th} round is at most $\max_{a \in \mathcal{A}} p(a)$. For the remaining rounds, we use Lemma 3 followed by Lemma 4 to get

$$\mathbb{E}\left[\sum_{t=1}^{T} c_t(S_t^*) - \min_{a^* \in \mathcal{A}} \sum_{t=1}^{T} c_t(a^*)\right] \leq \sum_{t=1}^{T} \mathbb{E}\left[\mathbb{1}[\tilde{a}_t^{*,1} \notin \tilde{S}_t^*]\right] + \mathbb{E}\left[\max_{a \in \mathcal{A}} p(\tilde{a})\right]$$

$$\leq \varepsilon^T \mathbb{E}\left[\sum_{t=1}^{T} c_t(S_t^*)\right] + \frac{(1 + \ln(N))}{\varepsilon}$$

Where the inequality on $\mathbb{E}\left[\max_{a\in\mathcal{A}} p(\tilde{a})\right]$ comes from Kalai and Vempala [2005]. The final step is to argue that the unmodified version of FPML which chooses independent noise $p_t(a)$ in each round also achieves this bound. This is immediate because both versions of the algorithm incur the same expected cost in each round, and $\mathbb{E}\left[\sum_{t=1}^{T} c_t(\textbf{FPML})\right] = \sum_{t=1}^{T} \mathbb{E}\left[c_t(\textbf{FPML})\right]$. Having new random perturbations in each round is not necessary against oblivious adversaries, but is necessary to achieve the regret bound against adaptive adversaries. □

Lower bound:

Proposition 10. *(Lower bounds) In the full feedback setting, any randomized algorithm has* $R_T^* \geq \Omega\left(\left(\frac{1}{4}\right)^B (\log_2(N) - \log_2(B))\right)$ *for* $T \geq \Omega(\log_2(N) - \log_2(B))$.

Proof. First suppose $N = 2^k$ for some $k \in \mathbb{N}$. Let $\mathcal{A}_0 = \mathcal{A}$. In round $t = 1, \ldots, T$, we let $\mathcal{A}_t$ be a uniformally randomly chosen subset of $\mathcal{A}_{t-1}$ of size $\max(2^{k-t}, 1)$, and we let $c_t(a) = 0$ if $a \in \mathcal{A}_t$, 1 otherwise. Suppose an algorithm chooses arms $S_t \subset \mathcal{A}$ in round t. Then

$$\begin{aligned}
P[S_t \cap A_t = \emptyset] &\geq \prod_{i=0}^{B-1} \left(1 - \frac{|\mathcal{A}_t|}{|\mathcal{A}_{t-1}| - i}\right) \\
&\geq \left(1 - \frac{|\mathcal{A}_t|}{|\mathcal{A}_{t-1}| - B}\right)^B \\
&= \left(1 - \frac{1}{2 - B/|\mathcal{A}_t}\right)^B \\
&\geq \left(1 - \frac{3}{4}\right)^B \\
&= \left(\frac{1}{4}\right)^B
\end{aligned}$$

provided that $|\mathcal{A}_t| \geq \frac{3}{2}B$ and $t \leq k$, which holds when $t \leq k - \log_2(B) - \log_2\left(\frac{3}{2}\right)$. If we set $T = \lfloor k - \log_2(B) - \log_2\left(\frac{3}{2}\right)\rfloor$, then the expected cost of any fixed algorithm ALG is $\geq \left(\frac{1}{4}\right)^B T = \Omega\left(\left(\frac{1}{4}\right)^B (\log_2(N) - \log_2(B))\right)$. By construction, the cost of the best expert in hindsight is 0. Since the expected regret is 0, there exists fixed cost functions $c_1, \ldots, c_T$ such that the expected regret of ALG on this on this sequence is $\geq \Omega\left(\left(\frac{1}{4}\right)^B (\log_2(N) - \log_2(B))\right)$. If N is not a power of 2, we can just let $\mathcal{A}_0 \subset \mathcal{A}$ be any subset of size $2^{\lfloor \log_2(N)\rfloor}$ and the asymptotic bounds remain the same. □

Proof of Proposition 5

Proof. Fix deterministic cost functions $c_1, \ldots, c_T$. We first consider the simpler case where the unbiased cost estimators $(\hat{c}_1, \ldots, \hat{c}_T)$ are jointly independent of any random perturbations used by the algorithm, and the algorithm re-uses random perturbations between rounds, i.e. $p_t(a) = p(a)$ for all $a \in \mathcal{A}, t \in [T]$. Afterwards we will show how to reduce the general problem to this special case.

Writing $\mathcal{F}_t$ for the σ-algebra generated by all actions and observations (as well as any other randomness) up to and including round t, for each $t \in [T]$ let $\hat{c}_t$ be a $\mathcal{F}_t$-measurable random function $\mathcal{A} \to [0, K]$ such that $\mathbb{E}[\hat{c}_t(a) \mid \mathcal{F}_{t-1}] = c_t(a)$ for each a. Assume an oblivious adversary and that w.l.o.g. instead of perturbations there is a 'round zero' with costs $(-p(a))_{a\in\mathcal{A}}$ where $p(a) \sim \text{Exp}(\varepsilon)$ independently for each a; define $\mathcal{F}_0 := \sigma((p(a))_{a\in\mathcal{A}}$ to be the σ-algebra generated by these and include it in each $(\mathcal{F}_t)_{t\geq 1}$.

Writing $\hat{C}_i(\cdot) \coloneqq \sum_{t=1}^{i} \hat{c}_t(\cdot)$ for cumulative estimated reward and $\hat{C}^*_i(\cdot) \coloneqq \sum_{t=0}^{i} \hat{c}_t(\cdot) = \hat{C}_i(\cdot) - p(\cdot)$ for the same but including the 'round zero' random initializations, define

$$R'_i \coloneqq \sum_{t=1}^{i} c_t(S_t) - \min_{a \in \mathcal{A}} \hat{C}^*_i(a)$$

for each $i \in [T]$. Let S_i be the set of arms chosen by the algorithm at round i and a^*_i be the best of these by perturbed estimated cost. We follow the argument from Lemma 3:

$$R'_i - R'_{i-1} = c_i(S_i) - \hat{c}_i(a^*_i) + \left(\sum_{t=1}^{i-1} \hat{c}_t(a^*_{i-1}) - \sum_{t=1}^{i-1} \hat{c}_t(a^*_i) \right) \leq c_i(S_i) - \hat{c}_i(a^*_i)$$

and so

$$\mathbb{E}[R'_i - R'_{i-1} \mid \mathcal{F}_{i-1}] \leq c_i(S_i) - \mathbb{E}[\hat{c}_i(a^*_i) \mid \mathcal{F}_{i-1}] = c_i(S_i) - c_i(a^*_i) \leq \mathbb{1}[a^*_i \notin S_i].$$

Hence (using the tower law)

$$\begin{aligned} \mathbb{E}[R'_T \mid \mathcal{F}_0] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[R'_t - R'_{t-1} \mid \mathcal{F}_{t-1}] + R'_0 \right] &\leq \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}[a^*_t \notin S_t] + R_0 \mid \mathcal{F}_0 \right] \\ &= \mathbb{E}[|\mathcal{I}| \mid \mathcal{F}_0] + \max_{a \in \mathcal{A}} p(a), \end{aligned}$$

where $\mathcal{I} \coloneqq \{t \in [T] : a^*_t \notin S_t\}$. Noting that by Jensen's inequality

$$\begin{aligned} \mathbb{E}[R'_T \mid \mathcal{F}_0] \geq \sum_{t=1}^{T} c_t(S_t) - \min_{a \in \mathcal{A}} \mathbb{E}[\hat{C}_T(a) \mid \mathcal{F}_0] &= \sum_{t=1}^{T} c_t(S_t) - \min_{a \in \mathcal{A}} (C_T(a) - p(a)) \\ &\geq \sum_{t=1}^{T} c_t(S_t) - C_T(a^*) + p(a^*) \end{aligned}$$

(where a^* is the best-in-hindsight arm) hence gives that the algorithm regret satisfies

$$\mathbb{E}[R^*_T \mid \mathcal{F}_0] \leq \mathbb{E}[|\mathcal{I}| \mid \mathcal{F}_0] + \max_{a \in \mathcal{A}} p(a) - p(a^*).$$

Since $\mathbb{E}[p(a^*)] = 1/\varepsilon$ as for any fixed action, and $\max_{a \in \mathcal{A}} p(a)$ is the maximum of $|\mathcal{A}|$ i.i.d. $\mathrm{Exp}(\varepsilon)$ random variables, so has expectation at most $(1 + \ln |\mathcal{A}|)/\varepsilon$ as argued in Kalai and Vempala [2005], taking expectations gives

$$\mathbb{E}[R^*_T] \leq \frac{\ln |\mathcal{A}|}{\varepsilon} + \mathbb{E}[|\mathcal{I}|].$$

It remains to upper-bound $\mathbb{E}[|\mathcal{I}|]$.

Fix $t \in [T]$ and let $V \coloneqq \min_{a \in \mathcal{A} - S_t} \hat{C}^*_{t-1}(a)$. So for any a, $a \in S_t$ iff $\hat{C}^*_{t-1}(a) < V$. Define $E_a \coloneqq \{\hat{C}^*_{t-1}(a) < V - K\}$; if this holds then a must have been ahead of every action $a' \notin S_t$ by at least K and therefore *cannot* be overtaken by any such action, since the estimated costs are all upper-bounded by K. So

$$\{a \text{ overtaken by some } a' \notin S_t\} \subset E^c_a.$$

Note that

$$\begin{aligned} \{a^*_t \notin S_t\} &= \{\exists a' \in \mathcal{A} - S_t : \forall a \in S_t,\ a' \text{ overtakes } a \text{ at round } t\} \\ &= \bigcup_{a' \in \mathcal{A} - S_t} \bigcap_{a \in S_t} \{a' \text{ overtakes } a \text{ at round } t\} \\ &\subset \bigcap_{a \in S_t} \bigcup_{a' \in \mathcal{A} - S_t} \{a' \text{ overtakes } a \text{ at round } t\} = \bigcap_{a \in S_t} \{a \text{ overtaken by some } a' \notin S_t\}. \end{aligned}$$

Let $\mathcal{G}_t := \sigma(S_t, (\hat{C}^*_{t-1}(a))_{a \notin S_t})$ be the σ-algebra generated by the random set S_t and the current perturbed estimated cumulative costs of the actions not in it. So we have

$$\begin{aligned}
\mathbb{P}\left(a^*_t \notin S_t \mid \mathcal{G}_t\right) &\leq \mathbb{P}\left(\bigcap_{a \in S_t} \{a \text{ overtaken by some } a' \notin S_t\} \mid \mathcal{G}_t\right) \\
&\leq \mathbb{P}\left(\bigcap_{a \in C} E^c_a \mid \mathcal{G}_t\right) \\
&= \mathbb{P}\left(\bigcap_{a \in S_t} \{\hat{C}^*_{t-1}(a) < V - K\} \mid \mathcal{G}_t\right).
\end{aligned}$$

But, since $V = \min_{a \in \mathcal{A} - S_t} \hat{C}^*_{t-1}(a)$, applying Lemma 9 gives us that

$$\mathbb{P}\left(\bigcap_{a \in S_t} \{\hat{C}^*_{t-1}(a) < V - K\} \mid \mathcal{G}_t\right) = \prod_{a \in S_t} \mathbb{P}\left(\hat{C}^*_{t-1}(a) < V - K \mid \hat{C}^*_{t-1}(a) \leq V\right).$$

By the memoryless property of the exponential distribution, each term here just becomes

$$1 - \mathbb{P}\left(p(a) \geq \hat{C}_{t-1}(a) - V + K \mid p(a) \geq \hat{C}_{t-1}(a) - V\right) \leq 1 - \mathbb{P}(p(a) \geq K) = 1 - \mathrm{e}^{-K\varepsilon}.$$

Where we have used the assumption that the perturbation $p(a)$ is independent of $\hat{C}_{t-1}$. Thus $\mathbb{P}(a^*_t \notin S_t \mid \mathcal{G}_t) \leq (1 - \mathrm{e}^{-K\varepsilon})^B$. Since this expression is deterministic and so trivially independent from the σ-algebra $\mathcal{G}_t$, this immediately implies that $\mathbb{P}(a^*_t \notin S_t) \leq (1 - \mathrm{e}^{-K\varepsilon})^B$.

The result then follows, since $\mathbb{E}[|\mathcal{I}|] = \sum_{t=1}^T \mathbb{P}(a^*_t \notin S_t) \leq T(1 - \mathrm{e}^{-K\varepsilon})^B$.

We now show how to reduce the general problem to a simpler case where the unbiased cost estimates $(\hat{c}_1, \ldots, \hat{c}_T)$ are jointly independent of the perturbations used by the algorithm, and the algorithm re-uses random perturbations between rounds, i.e. $p_t(a) = p(a)$ for all $a \in \mathcal{A}, t \in [T]$. Consider the general problem. Let p_t be the noise perturbations of the algorithm in round t, so $p_t(a) \sim \frac{1}{\varepsilon}\text{Exp}$. Let $\hat{c}_{:t} = (\hat{c}_1, \ldots, \hat{c}_{t-1})$ and $S(\hat{c}_{:t}, p_t) \subset \mathcal{A}$ be the B lowest cost-perturbed arms given $\hat{c}_{:t}$, p_t (i.e. the arms chosen by the algorithm in round t if cost vectors $\hat{c}_{:t}$ are observed and noise perturbation p_t is chosen). We are guaranteed that $\mathbb{E}[\hat{c}_t | p_1, \hat{c}_1, p_2, \hat{c}_2, \ldots, p_t] = c_t$. The expected regret is

$$\mathbb{E}_{p_1, \hat{c}_1, p_2, \hat{c}_2, \ldots, p_T}\left[\sum_{t=1}^T c_t\left(S(\hat{c}_{:t}, p_t)\right)\right] - \min_{a \in \mathcal{A}} \sum_{t=1}^T c_t(a)$$

Focusing on just the first term, and letting $\{p'_t\}_{t=0}^T$ be independent random noise perturbations where $p'_t(a) \sim \frac{1}{\varepsilon}\text{Exp}$, we have

$$\mathbb{E}_{p_1,\hat{c}_1,p_2,\hat{c}_2,\ldots,p_T}\left[\sum_{t=1}^{T} c_t\left(S(\hat{c}_{:t},p_t)\right)\right]$$

$$\begin{aligned}
&= \sum_{t=1}^{T} \mathbb{E}_{p_1,\hat{c}_1,p_2,\hat{c}_2,\ldots,p_T}\left[c_t\left(S(\hat{c}_{:t},p_t)\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{p_1,\hat{c}_1,p_2,\hat{c}_2,\ldots,p_t}\left[c_t\left(S(\hat{c}_{:t},p_t)\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{p_1,\hat{c}_1,p_2,\hat{c}_2,\ldots,p_t}\mathbb{E}_{p'_t}\left[c_t\left(S(\hat{c}_{:t},p'_t)\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{p'_t}\left[\mathbb{E}_{p_1,\hat{c}_1,p_2,\hat{c}_2,\ldots,p_t} c_t\left(S(\hat{c}_{:t},p'_t)\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{p'_t}\left[\mathbb{E}_{p_1,\hat{c}_1,p_2,\hat{c}_2,\ldots,p_T} c_t\left(S(\hat{c}_{:t},p'_t)\right)\right] \\
&= \sum_{t=1}^{T} \mathbb{E}_{p'_0}\left[\mathbb{E}_{p_1,\hat{c}_1,p_2,\hat{c}_2,\ldots,p_T} c_t\left(S(\hat{c}_{:t},p'_0)\right)\right] \\
&= \mathbb{E}_{p'_0}\left[\mathbb{E}_{p_1,\hat{c}_1,p_2,\hat{c}_2,\ldots,p_T} \sum_{t=1}^{T} c_t\left(S(\hat{c}_{:t},p'_0)\right)\right]
\end{aligned}$$

Therefore the final expected regret is equal to

$$\mathbb{E}_{p'_0}\left[\mathbb{E}_{p_1,\hat{c}_1,p_2,\hat{c}_2,\ldots,p_T} \sum_{t=1}^{T} c_t\left(S(\hat{c}_{:t},p'_0)\right)\right] - \min_{a\in\mathcal{A}} \sum_{t=1}^{T} c_t(a) \tag{7}$$

Note the expression $\sum_{t=1}^{T} c_t\left(S(\hat{c}_{:t},p'_0)\right)$ is precisely the cost incurred by the algorithm when observing cost estimates $\hat{c}_{:T}$ and using random perturbations p_0 in each round, where $\mathbb{E}[\hat{c}_t|p_0,p_1,\hat{c}_1,\ldots,p_{t-1}] = c_t$. We therefore conclude that the expected regret is equal to the expected regret of the algorithm in the special case where (a) the algorithm fixes an initial perturbation p'_0 and uses this randomness for all subsequent rounds and (b) where p'_0 is jointly independent of $\hat{c}_{:T}$. □

B Section 3: Generalized regret bounds for Online Submodular Function Maximization

Proof of Theorem 6

Before proving the theorem, we give a modification to the original result from Streeter and Golovin [2008]. The problem setting they considered was slightly more general:

Definition 4. *Let an action now be an activity-duration pair $a = (\nu,\tau) \in \mathcal{V}\times(0,\infty) = \mathcal{A}$ for some fixed finited set of* activities $\mathcal{V}$.[*] *The* length $\ell(S)$ *of a schedule $S \in \mathcal{S}$ is now the sum of the durations of all the actions in S. Write $S_{\langle i\rangle}$ for the prefix of length i of schedule S.*

The algorithm OG they introduced, which takes a budget B and experts algorithm $\mathcal{E}$, is given in Algorithm 3 using our notation for ease of reference.

We first prove a lemma generalizing Theorem 6 in Streeter and Golovin [2008]:

[*] We will enforce integer durations so that there are only finitely many possible actions to choose from given a duration constraint.

Algorithm 2 $\mathrm{OG}_{\mathrm{hybrid}}(B,\widetilde{B})$

Require: $B \geq \widetilde{B} \geq 1$. Assume for simplicity that $\widetilde{B} \mid B$; define $L \coloneqq B/\widetilde{B}$.
Let $\mathcal{B}_1, \ldots, \mathcal{B}_L$ be instances of FPML, each with budget $\widetilde{B}$.
for rounds $t = 1, \ldots, T$ **do**
Let $S_{t,0} = \langle\rangle$ be the empty schedule.
for $i = 1, \ldots, L$ **do**
Use $\mathcal{B}_i$ to choose $\widetilde{B}$ actions $a^t_{(i-1)\widetilde{B}+1}, \ldots, a^t_{i\widetilde{B}}$.
Set $S_{t,i} \coloneqq S_{t,i-1} \oplus \langle a^t_{(i-1)\widetilde{B}+1}, \ldots a^t_{i\widetilde{B}}\rangle$.
end for
Set $S_t \coloneqq S_{t,L}$; receive the job f_t.
for $i = 1, \ldots, L$ **do**
For each action $a \in \mathcal{A}$ feed back the cost $c_t^{(i)}(a) \coloneqq 1 - (f_t(\langle a^*_{1,t}, \ldots, a^*_{i-1,t}, a\rangle) - f_t(\langle a^*_{1,t}, \ldots, a^*_{i-1,t}\rangle))$ to FPML instance $\mathcal{B}_i$.
Define $a^*_{i,t} \coloneqq \arg\min_{j\in[\widetilde{B}]} c_t^{(i)}(a_{(i-1)\widetilde{B}+j})$.
end for
end for

Algorithm 3 $\mathrm{OG}(B, \mathcal{E})$

Require: $B \geq 1$.
Let $\mathcal{E}_1, \ldots, \mathcal{E}_B$ be instances of experts algorithm $\mathcal{E}$ (e.g. Hedge).
for rounds $t = 1, \ldots, T$ **do**
Let $S_{t,0} = \langle\rangle$ be the empty schedule.
for $i = 1, \ldots, B$ **do**
Use $\mathcal{E}_i$ to choose an action $a^t_i = (\nu, \tau) \in \mathcal{A}$.
With probability $1/\tau$ set $S_{t,i} \coloneqq S_{t,i-1} \oplus \langle a^t_i\rangle$, otherwise set $S_{t,i} \coloneqq S_{t,i-1}$.
end for
Set $S_t \coloneqq S_{t,B}$; receive the job f_t.
For each $i \in [B]$ and each action $a = (\nu, \tau) \in \mathcal{A}$ feed back the cost $c_t^{(i)}(a) \coloneqq (f_t(S_{t,i} \oplus \langle a\rangle) - f_t(S_{t,i}))/\tau$ to experts algorithm $\mathcal{E}_i$.
end for

Lemma 11. *Let f be any job and let $\bar{G} = \langle \bar{g}_1, \bar{g}_2, \ldots\rangle$ be an infinite 'greedy' schedule satisfying*

$$\frac{f(\bar{G}_j \oplus \bar{g}_j) - f(\bar{G}_j)}{\bar{\tau}_j} \geq \max_{(\nu,\tau)\in\mathcal{V}\times(0,\infty)} \left(\frac{f(\bar{G}_j \oplus \langle(\nu,\tau)\rangle) - f(\bar{G}_j)}{\tau}\right) - \varepsilon_j, \qquad j \geq 1$$

for additive errors $\varepsilon_1, \varepsilon_2, \ldots \geq 0$, where $\bar{g}_j = (\bar{v}_j, \bar{\tau}_j)$ and $\bar{G}_j = \langle \bar{g}_1, \ldots, \bar{g}_{j-1}\rangle$ for each $j \geq 1$.

Then for any $L, B_0 \in \mathbb{N}$ and for $B_1 \coloneqq \sum_{j=1}^{L} \bar{\tau}_j$,

$$f(\bar{G}_{\langle B_1\rangle}) > \left(1 - \mathrm{e}^{-B_1/B_0}\right) f(S^*_{B_0}) - \sum_{j=1}^{L} \varepsilon_j \bar{\tau}_j$$

where $S^*_{B_0} \coloneqq \arg\max_{S\in\mathcal{S}:\ell(S)=B_0} f(S)$ is the best schedule of length B_0 for f.

Proof. For each $j \in \mathbb{N}$ write $\Delta_j \coloneqq f(S^*_{B_0}) - f(\bar{G}_j)$. By Fact 2 from Streeter and Golovin [2008], for any $j \in \mathbb{N}$, $b > 0$ and $S \in \mathcal{S}$ with $\ell(S) \leq b$,

$$f(S) \leq f(\bar{G}_j) + b \cdot (s_j + \varepsilon_j),$$

where

$$s_j \coloneqq \max_{(v,\tau)\in\mathcal{V}\times(0,\infty)} \frac{f(\bar{G}_j \oplus \langle(n,\tau)\rangle) - f(\bar{G}_j)}{\tau} = \frac{f(\bar{G}_j \oplus \bar{g}_j) - f(\bar{G}_j)}{\bar{\tau}_j} = \frac{f(\bar{G}_{j+1}) - f(\bar{G}_j)}{\bar{\tau}_j},$$

so in particular for any j

$$f(S^*_{B_0}) = \max_{S\in\mathcal{S}:\ell(S)=B_0} f(S) \le f(\hat{G}_j) + B_0\cdot(s_j+\varepsilon_j) \tag{8}$$

$$= f(\hat{G}_j) + B_0\left(\frac{f(\bar{G}_{j+1}) - f(\bar{G}_j)}{\bar{\tau}_j} + \varepsilon_j\right) \tag{9}$$

$$= f(\hat{G}_j) + B_0\left(\frac{\Delta_j - \Delta_{j+1}}{\bar{\tau}_j} + \varepsilon_j\right), \tag{10}$$

giving $\Delta_j \le B_0\left(\frac{\Delta_j-\Delta_{j+1}}{\bar{\tau}_j}+\varepsilon_j\right)$.

Rearranging gives $\Delta_{j+1} \le \Delta_j\left(1-\frac{\bar{\tau}_j}{B_0}\right) + \bar{\tau}_j\varepsilon_j$ for each j, and unrolling this inequality and using that $1-\frac{\bar{\tau}_j}{B_0} < 1\ \forall j$ as in Streeter and Golovin [2008] gives us

$$\Delta_{L+1} \le \Delta_1\left(\prod_{j=1}^{L} 1-\frac{\bar{\tau}_j}{B_0}\right) + \sum_{j=1}^{L}\bar{\tau}_j\varepsilon_j.$$

By definition $B_1 = \sum_{j=1}^{L}\bar{\tau}_j\varepsilon_j$, and maximizing the product above subject to this constraint results in $\bar{\tau}_j = \frac{B_1}{L}$ for all j. Thus

$$\prod_{j=1}^{L} 1-\frac{\bar{\tau}_j}{B_0} \le \prod_{j=1}^{L} 1-\frac{B_1/L}{B_0} = \left(1+\frac{(-B_1/B_0)}{L}\right)^L < \mathrm{e}^{-B_1/B_0}$$

and so

$$f(S^*_{B_0}) - f(\bar{G}_{L+1}) = \Delta_{L+1} < \Delta_1\mathrm{e}^{-T_1/T_0} + \sum_{j=1}^{L}\bar{\tau}_j\varepsilon_j \le f(S^*_{B_0})\mathrm{e}^{-B_1/B_0} + \sum_{j=1}^{L}\bar{\tau}_j\varepsilon_j,$$

giving $f(\bar{G}_{\langle T_1\rangle}) = f(\bar{G}_{L+1}) > (1-\mathrm{e}^{-B_1/B_0})f(S^*_{B_0}) - \sum_{j=1}^{L}\bar{\tau}_j\varepsilon_j$ as required. □

Next we prove a generalized regret bound for the original OG algorithm:

Lemma 12. *For $B \ge B'\log T$ the algorithm OG, run using Hedge as the subroutine experts algorithm, produces a sequence of schedules $S_1,\dots,S_B$ with regret*

$$\mathbb{E}\left[\sum_{t=1}^{T} f_t(S^*_{B'}) - \sum_{t=1}^{T} f_t(S_t)\right] = \mathcal{O}\left(\mathbb{E}\left[\sum_{j=1}^{B} R_{T,1}(\mathcal{E}_j)\right]\right)$$

relative to $S^*_{B'} \coloneqq \arg\max_{S\in\mathcal{S}:\ell(S)=B'}\sum_{t=1}^{T} f_t(S)$, the best-in-hindsight fixed schedule of length B', where $R_{T,1}(\mathcal{E}_j)$ is the 1-regret incurred by the j^{th} experts algorithm.

In particular, when run with Hedge as the subroutine experts algorithm, this is $\mathcal{O}\left(\sqrt{BT\log N}\right)$.

Proof. Consider the quantity $\rho_{B,B'} \coloneqq \left(1-\mathrm{e}^{-B/B'}\right)\sum_{t=1}^{T} f_t(S^*_{B'}) - \sum_{t=1}^{T} f_t(S_t)$. As argued in Streeter and Golovin [2008], we may view the sequence of actions $a_i^1,\dots,a_i^T$ selected by each experts algorithm $\mathcal{E}_i$ as a single 'meta-action' $\tilde{a}_i \in \mathcal{A}^T$; so the schedules $S_1,\dots,S_T$ output by **OG** can be viewed as a single 'meta-schedule' $\tilde{S} = \langle\tilde{a}_1,\dots,\tilde{a}_B\rangle$ over $\mathcal{A}^T$ which is a version of the greedy schedule $\bar{G}_{B+1}$ for the job $f = \frac{1}{T}\sum_{t=1}^{T} f_t$, and it may be assumed that each meta-action $\tilde{a}_t$ takes unit time per job. Thus we may write

$$\rho_{B,B'} = T\left[\left(1-\mathrm{e}^{-B/B'}\right)f(S^*_{B'}) - f(\tilde{S})\right]$$

(after extending the domain of f appropriately). Applying Lemma 11 with $L = B$, $B_0 = B'$, $B_1 = \sum_{j=1}^{B}\bar{\tau}_j = B$ (by the unit-time assumption) then immediately gives

$$\rho_{B,B'} < T\sum_{j=1}^{B}\bar{\tau}_j\varepsilon_j = T\sum_{j=1}^{B}\varepsilon_j.$$

Taking expectations,

$$\mathbb{E}[\rho_{B,B'}] \leq T \sum_{j=1}^{B} \mathbb{E}[\varepsilon_j] = T \sum_{j=1}^{B} \mathbb{E}\left[\frac{R_{T,1}(\mathcal{E}_j)}{T}\right]$$

where $R_{T,1}(\mathcal{E}_j)$ is the 1-regret incurred by the j^{th} experts algorithm; here we used that $\mathbb{E}[\varepsilon_j] = \mathbb{E}[R_{T,1}(\mathcal{E}_j)/n]$ as argued in Streeter and Golovin [2008]. So $\mathbb{E}[\rho_{B,B'}] \leq \mathbb{E}\left[\sum_{j=1}^{B} R_{T,1}(\mathcal{E}_j)\right]$.

The result then follows quickly: since $B \geq B' \log T$, so $\mathrm{e}^{-B/B'} \leq \mathrm{e}^{-\ln T} = T^{-1}$. Thus

$$\rho_{B,B'} \geq \left(1 - T^{-1}\right) \sum_{t=1}^{T} f_t(S^*_{B'}) - \sum_{t=1}^{T} f_t(S_t) = \mathcal{R}_{B'} - T^{-1} \sum_{t=1}^{T} f_T(S^*_{B'})$$

where $\mathcal{R}_{B'} \coloneqq \sum_{t=1}^{T} f_t(S^*_{B'}) - \sum_{t=1}^{T} f_t(S_t)$ is the regret of interest. Consequently,

$$\mathcal{R}_{B'} \leq \rho_{B,B'} + T^{-1} \sum_{t=1}^{T} f_t(S^*_{B'}) \leq \rho_{B,B'} + T^{-1} \cdot T = \rho_{B,B'} + 1.$$

and the result follows.

The bound $\mathbb{E}\left[\sum_{j=1}^{B} R_{T,1}(\mathcal{E}_j)\right] = \mathcal{O}(\sqrt{BT \log N})$ when using Hedge was shown in Streeter and Golovin [2008]. □

Finally we prove the theorem on $\text{OG}_{\text{hybrid}}$:

Proof of Theorem 6. Note first that under Assumption 1, any job f, any schedule $S \in \mathcal{S}$ and any sub-schedule S' of S (i.e. the actions of S' appear in order in S) satisfy

$$f(S) \geq f(S');$$

this is immediate using monotonicity and induction.

Suppose for each $i \in [L]$ there is a fictional experts algorithm (classical full feedback multi-armed bandit algorithm) $\mathcal{E}_i$ which picks $a^*_{i,t}$ at each round t, and consider a hypothetical instance of the standard algorithm **OG** run with time allowance L and these fictional experts algorithms $\mathcal{E}_1, \ldots, \mathcal{E}_L$ as subroutines.

Since $L \geq B' \log T$ (by our assumption that $B \geq B' \widetilde{B} \log T$), by Lemma 12 the B'-regret of our **OG** instance is upper-bounded in expectation by $\sum_{i=1}^{L} \mathcal{R}_1(\mathcal{E}_i)$, where $\mathcal{R}_1(\mathcal{E}_i)$ is the total 1-regret experienced by $\mathcal{E}_i$.

But the payoff received by this **OG** instance at each round t is $f(\langle a^*_{1,t}, \ldots, a^*_{L,t} \rangle)$, which by Appendix B is upper-bounded by $f(S_t)$, the payoff of $\textbf{OG}_{\textbf{hybrid}}$, since the actions $a^*_{1,t}, \ldots, a^*_{L,t}$ appear in order in S_t. So the B'-regret $\mathcal{R}_{B'}$ of $\textbf{OG}_{\textbf{hybrid}}$ must be at most that of our fictional **OG** instance, giving the upper bound

$$\mathbb{E}[\mathcal{R}_{B'}] \leq \sum_{i=1}^{L} \mathbb{E}[\mathcal{R}_1(\mathcal{E}_i)].$$

It remains to argue how large each of the regret of each of these 'fictional' experts algorithms $\mathcal{E}_i$ is. Writing $a^{**}_i = \arg\min_{a \in \mathcal{A}} \sum_{t=1}^{T} c^{(i)}_t(a)$ for the best-in-hindsight fixed action under the costs passed to these subroutines, the regret incurred by $\mathcal{E}_i$ is therefore

$$\mathcal{R}_1(\mathcal{E}_i) = \sum_{t=1}^{T} c^{(i)}_t(a^*_{i,t}) - \sum_{t=1}^{T} c^{(i)}_t(a^{**}_i) \tag{11}$$

$$= \sum_{t=1}^{T} \max_{j \in [\widetilde{B}]} c^{(i)}_t(a^t_{(i-1)\widetilde{B}+j}) - \sum_{t=1}^{T} c^{(i)}_t(a^{**}_i) = \mathcal{R}_1(\mathcal{B}_i). \tag{12}$$

where $\mathcal{R}_1(\mathcal{B}_i)$ is the 1-regret incurred by multitasking bandit algorithm $\mathcal{B}_i$. So by Appendix B

$$\mathbb{E}[\mathcal{R}_{B'}] \leq \sum_{i=1}^{L} \mathbb{E}[\mathcal{R}_1(\mathcal{B}_i)] = L\mathbb{E}[\mathcal{R}_1(\mathcal{B})] = \frac{B\mathbb{E}[\mathcal{R}_1(\mathcal{B})]}{L},$$

where $\mathbb{E}[\mathcal{R}_1(\mathcal{B})]$ is the expected 1-regret of any of the instances $\mathcal{B}_1, \ldots, \mathcal{B}_L$ of $\mathcal{B}$. □

Table 2: Sample means and standard deviations of normalized validation scores of FPML, $\text{OG}_{\text{hybrid}}$ and OG over black-box optimizers.

(a) $B = 1$

	Mean	StD
Best in hindsight	0.574	0
FPML	0.426	0.0202
Exp3	0.351	0.0194

(b) $B = 2$

	Mean	StD
Best in hindsight	0.710	0
FPML	0.652	0.0194
$\textbf{OG}_{\textbf{hybrid}}$ $((B_1, B_2) = (1, 2))$	0.577	0.0187
OG	0.519	0.0179

(c) $B = 3$

	Mean	StD
Best in hindsight	0.779	0
FPML	0.751	0.0151
$\textbf{OG}_{\textbf{hybrid}}$ $((B_1, B_2) = (1, 3))$	0.657	0.0191
OG	0.617	0.0166

(d) $B = 4$

	Mean	StD
Best in hindsight	0.836	0
FPML	0.813	0.0108
$\textbf{OG}_{\textbf{hybrid}}$ $((B_1, B_2) = (2, 2))$	0.756	0.0149
$\textbf{OG}_{\textbf{hybrid}}$ $((B_1, B_2) = (1, 4))$	0.716	0.0178
OG	0.689	0.0151

(e) $B = 5$

	Mean	StD
Best in hindsight	0.874	0
FPML	0.855	0.0094
$\textbf{OG}_{\textbf{hybrid}}$ $((B_1, B_2) = (1, 5))$	0.756	0.0150
OG	0.734	0.0140

(f) $B = 6$

	Mean	StD
Best in hindsight	0.901	0
FPML	0.888	0.0072
$\textbf{OG}_{\textbf{hybrid}}$ $((B_1, B_2) = (3, 2))$	0.836	0.0111
$\textbf{OG}_{\textbf{hybrid}}$ $((B_1, B_2) = (2, 3))$	0.814	0.0143
$\textbf{OG}_{\textbf{hybrid}}$ $((B_1, B_2) = (1, 6))$	0.785	0.0137
OG	0.767	0.0157

Proof of Proposition 7

Sketch proof. This is a special case of the more general result that the expected regret relative to the best-in-hindsight fixed set of size B' is at most

$$\frac{1 - B'^{-1} + \ln(N/B')}{\varepsilon} + T \sum_{j=0}^{B'-1} \binom{B}{j} \mathrm{e}^{-j\varepsilon} (1 - \mathrm{e}^{-\varepsilon})^{B-j} + \texttt{err}_{B'}$$

where $\texttt{err}_{B'}$ is the difference in cumulative cost between the best-in-hindsight set of B' actions and the set of the top B' actions in hindsight on the given problem instance.

The proof of this is a simple adaptation of the 1-regret argument, using "an action not in the top B enters the best B'-set" as the event of interest; use the harmonic series form of the expectation of the max of exponential random variables to get a lower bound, and use a binomial counting argument to bound the probability of the event. □

C Experiments

C.1 Full comparison of $\text{OG}_{\text{hybrid}}$

In Table 2 we give a more detailed comparison of FPML and OG with various instantiations of $\text{OG}_{\text{hybrid}}$ on the hyperparameter-selection task from Section 4. Specifically, we include for each B and each possible pair (B_1, B_2) s.t. $B_1 B_2 = B$ a version of $\text{OG}_{\text{hybrid}}$ with B_2 internal boxes and arm budget B_1 per box. As can be seen, in all cases decreasing the greediness and adding more arms per box is beneficial in this application.

C.2 Synthetic tasks

In this section we evaluate our algorithms on three synthetic tasks. In all cases,

- let S^* be the best-in-hindsight set of B arms;

Table 3: Cost distributions for round types A and B in the first synthetic environment; Beta distributions are parameterized by mean and variance, not shape.

Arm	A-rounds	B-rounds	Resulting mean
Actions 1 to 5	$\text{Beta}(0.4, 0.01)$	Always 1	0.7
Actions 6 to 10	$\text{Beta}(0.6, 0.01)$	Always 1	0.8
Actions 11 to 15	Always 1	$\text{Beta}(0.8, 0.01)$	0.9

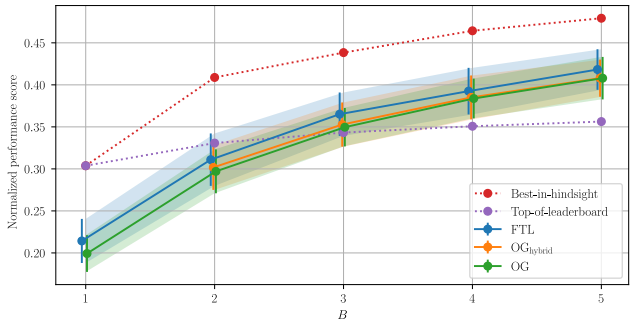

Figure 1: Performances (1-cost) on the first synthetic task. $\text{OG}_{\text{hybrid}}$ is run with FPML boxes each with arm budget 1.

- let S_{greedy} be the greedy choice of B arms in hindsight;
- let S_{top} be the top B arms in hindsight.

Task 1: The first environment is one where $S^* = S_{\text{greedy}}$ and this set does better than S_{top}; greediness is better than picking the top B arms. There are $|\mathcal{A}| = 15$ available arms and two types of round, A and B, which occur with equal probability; costs are distributed within each round according to Table 3. So the best fixed arm set of any size up to 10 will be split evenly across arms $\{1, 2, 3, 4, 5\}$ and arms $\{11, 12, 13, 14, 15\}$—and will be the greedy choice—but for $B \leq 5$ the top B arms will always be in $\{1, 2, 3, 4, 5\}$. We see in Fig. 1 that FPML does not outperform the greedy algorithms on this task.

Task 2: The second environment is one where (approximately) $S^* = S_{\text{greedy}} = S_{\text{top}}$; greediness is good but no better than picking the top B arms. There are $|\mathcal{A}| = 10$ available arms and costs are distributed according to Table 4; because there are no groups of anticorrelated actions, the performance gap between the best set and the top B arms is trivially small. The results in Fig. 2 show that FPML outperforms the greedy algorithms on this task.

Task 3: The third environment is one where $S^* = S_{\text{top}}$ and this set does better better than S_{greedy}; greediness is worse than just picking the top B arms. Suppose there are $|\mathcal{A}| = 4$ available arms and a budget of $B = 3$. Costs are deterministic and listed in Table 5 for some parameter δ which we set to 0.01. The top 3 arms are $S_{\text{top}} = \{1, 3, 4\}$ and this is also the best-in-hindsight set S^*, incurring minimum cost 0 at each round. A quick calculation shows that the greedy choice S_{greedy} is either $\{1, 2, 3\}$ or $\{1, 2, 4\}$, though, and either of these sets incur an average minimum cost of $1/8 - \delta/4$, substantially higher. Our empirical results in Table 6 show this gap in practice.

C.3 Geometric resampling

The *geometric resampling* technique used in the second and third partial feedback versions of FPML in the experiments is adapted from Neu and Bartók [2013]. At each round cost estimates

$$\hat{c}_t(a) := \begin{cases} \frac{c_t(a)}{\hat{q}_{t,a}} & \text{if } a \text{ was pulled,} \\ 0 & \text{otherwise} \end{cases}$$

Table 4: Cost distributions in the second synthetic environment.

Arm	Distribution
1	$\mathrm{Beta}(0.4, 0.01)$
2	$\mathrm{Beta}(0.45, 0.01)$
3	$\mathrm{Beta}(0.5, 0.01)$
4	$\mathrm{Beta}(0.55, 0.01)$
5	$\mathrm{Beta}(0.6, 0.01)$
6	$\mathrm{Beta}(0.65, 0.01)$
7	$\mathrm{Beta}(0.7, 0.01)$
8	$\mathrm{Beta}(0.75, 0.01)$
9	$\mathrm{Beta}(0.8, 0.01)$
10	$\mathrm{Beta}(0.85, 0.01)$

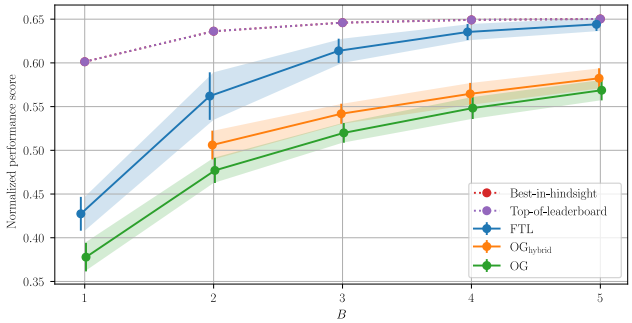

Figure 2: Performances (1-cost) on the second synthetic task. $\mathrm{OG}_{\mathrm{hybrid}}$ is run with FPML boxes each with arm budget 1.

Table 5: Costs in the third synthetic environment, for some parameter $\delta \in (0, 1/2)$.

Arm	Reward at rounds $i \equiv k \mod 4$ for...				Average cost
	$k=1$	$k=2$	$k=3$	$k=4$	
1	$1-\delta$	$1-\delta$	0	0	$1/2-\delta/2$
2	$1/2-\delta$	$1/2-\delta$	1	1	$3/4-\delta/2$
3	0	1	0	1	$1/2$
4	1	0	1	0	$1/2$

Table 6: Means and standard deviations over 50 trials of performances (1-cost) for various combinations of FPML and online greedy algorithms in the third synthetic environment, with $\delta = 0.01$.

Algorithm	Mean	StD
Best-in-hindsight	1.000	0
Top-of-leaderboard	1.000	0
FPML	0.964	0.0145
$\mathbf{OG}_{\mathbf{hybrid}}$ $((B_1, B_2) = 1, 3))$	0.823	0.0200
OG	0.799	0.0202

are made, where $\hat{q}_{t,a}$ is an estimate of the probability $q_{t,a} \coloneqq \mathbb{P}(\text{arm } a \text{ pulled at round } t)$. These estimates are made by sampling $\frac{1}{\hat{q}_{t,a}} \sim \text{Geom}(q_{t,a})$, which is done by repeating the algorithm's execution at this round and counting how many trials are needed until a is pulled again. In practice, the number of repetitions must be capped and this introduces some bias to the estimates, but this is not problematic in practice. In fact, there is a bias variance trade-off, because $K = \max_{a \in \mathcal{A}} |\hat{c}_t(a)|$ is bounded by the number of samples we take. Therefore more samples lead to lower bias but higher variance. Using bounds similar to those of Proposition 5 as a guide (the bounds of Proposition 5 were subsequently refined after the experiments were concluded), we picked the number of samples to be $\left(N\left(\frac{TN}{\ln(N)}\right)^{\tilde{B}}\right)^{1/(2\tilde{B}+1)}$, so $K = \left(N\left(\frac{TN}{\ln(N)}\right)^{\tilde{B}}\right)^{1/(2\tilde{B}+1)}$ and $\varepsilon = \left(\frac{\ln(N)}{T}\left(\frac{\ln(N)}{TN}\right)^{\tilde{B}}\right)^{1/(2\tilde{B}+1)}$, where $\tilde{B}$ is the budget of each **FPML-partial** box.

These estimators make complete use of the information received at each round, unlike the simple one-arm uniform sampling, B arms exploiting version of FPML with partial feedback mentioned in Section 2.1. Moreover, the construction of cost estimates means no explicit exploration is necessary; an arm that hasn't been pulled for several rounds will be overtaken in estimated cumulative cost by ones that have, and so will eventually be pulled again, thus inducing a self-stabilizing property that would not occur if we used the same technique to estimate rewards $r_t(a) \coloneqq 1 - c_t(a)$ instead.

C.4 Methods

Reward definitions: For the black-box optimization experiments in Section 4, the *reward* (1−cost) for each black-box optimizer on each machine learning task (i.e. round) was defined as follows. This approach was inspired heavily by the Bayesmark package used in the 2020 NeurIPS BBO Challenge and which we based our implementation on [Uber, 2020].

Fix a round t and an optimizer a. Let $\texttt{opt}_t$ be an estimate of the global minimum classification/regression loss achievable (at validation, not test) on the task corresponding to round t. Define $\overline{\texttt{rand}}_t$ to be the mean performance of a random hyperparameter search on this task (i.e. the smallest loss achieved using any hyperparameter in the random search, averaged over trials).[*] Finally, let $\overline{\texttt{loss}}_t(a)$ be the actual averaged minimum loss of the optimizer a on this problem.

The reward is then defined as

$$r_i(a) \coloneqq \frac{\overline{\texttt{loss}}_t(a) - \texttt{opt}_t}{\overline{\texttt{rand}}_t(a) - \texttt{opt}_t}.$$

Conceptually, the reward is 0 when optimizer a performs as badly as a random search, and 1 when it performs as well as is possible on this task.

As per usual, the reward for a bandit algorithm selecting multiple optimizers at each round is then calculated as the maximum of the rewards of each optimizer (equivalent to the minimum of costs).

Bayesian optimizers used: The nine black-box optimization algorithms we ran the experiments in Section 4 over were as follows:

1. Hyperopt [Bergstra et al., 2015]
2. The `AUCBanditMetaTechniqueA` technique from OpenTuner [Ansel et al., 2014]
3. The `PSO_GA_Bandit` technique from OpenTuner [Ansel et al., 2014]
4. The `PSO_GA_DE` technique from OpenTuner [Ansel et al., 2014]
5. PySOT [Eriksson et al., 2019]
6. Scikit-Optimize [Head et al., 2018] using base estimator `GBRT` and acquisition objective `gp_hedge`
7. Scikit-Optimize [Head et al., 2018] using base estimator `GP` and acquisition objective `gp_hedge`
8. Scikit-Optimize [Head et al., 2018] using base estimator `GP` and acquisition objective `LCB`

[*] In reality this is estimated using a more statistically efficient technique than actually performing the random search, as in the Bayesmark package.

9. Random search

The default settings of each package were used.

D Section 5: Linear Programming

Proof of Proposition 8 (Note: this proof was given for $B = 1$ in Arora et al. [2012] with slightly tighter bounds, and essentially remains unchanged for $B \geq 1$).

Proof. We run the FPML oracle with budget B, $N = n$ arms, and $\varepsilon = ((\ln(N) + 1)/T)^{1/(B+1)}$. In round $t \in [T]$ we do the following: Let d_t be the joint distribution over N arms returned the FPML oracle in this round. We pass d_t to the (ρ, B)-bounded oracle, and receive either a vector $x_t \in P$ or that no x_t exists which satisfies the oracle problem. Let us first suppose that we always receive an x_t for each round. Then define the cost function $c_t(i) := A_i x_t - b_i \in [-\rho, \rho]$ and pass this to FPML. After T rounds, and by scaling and translating the cost functions to lie in $[0, 1]$, Theorem 2 implies that $\forall j \in [N]$

$$\frac{\sum_{t=1}^{T} \mathbb{E}_{(i_1,\ldots,i_B)\sim d_t} \left[\min_{i\in\{i_1,\ldots,i_B\}} A_i x - b_i\right]}{T} \leq \frac{4\rho T^{\frac{1}{B+1}} (1 + \ln(N))^{\frac{B}{B+1}}}{T} + \frac{\sum_{t=1}^{T} A_j x_t - b_j}{T}$$

By assumption of the (ρ, B)-bounded oracle, the left hand side is ≥ 0. When $T \geq \left(\frac{1}{\varepsilon}\right)^{\frac{B+1}{B}} (4\rho)^{\frac{B+1}{B}} (1 + \ln(N))$, it follows that $x := \frac{\sum_{t=1}^{T} x_t}{T}$ satisfies $\forall j \in [N], A_j x \geq b_j - \varepsilon$. Since P is convex, $x \in P$ and we are done. Now suppose that in some round t we were told the oracle problem was not solvable. We claim that we can conclude that the problem is not feasible and we are done. This is because if $\exists x \in P$ s.t. $Ax \geq b$, then $\mathbb{E}_{(i_1,\ldots,i_B)\sim d} \left[\min_{i\in\{i_1,\ldots,i_B\}} A_i x - b_i\right] \geq \mathbb{E}_{(i_1,\ldots,i_B)\sim d} \left[\min_{i\in\{i_1,\ldots,i_B\}} 0\right] = 0$ and so the oracle problem would be solvable. □