Contents of Main Article and Appendix

A. Technical Lemmas

A.1. Good Event and High-Confidence Argument

We define the good event as

$$\mathcal{E} = \left\{ \forall i \in [M],\ \forall t \in \mathbb{N} \colon |n_i(t)\mu^\star - U_i(t) - \mathsf{Reg}_i(t)| \le c\sqrt{n_i(t)\,\ln\frac{M\ln n_i(t)}{\delta}} \right\}. \tag{6}$$

Lemma 5. *There is an absolute constant c such that the event $\mathcal{E}$ has probability at least $1-\delta$*

Proof. Consider a fixed $i \in [M]$ and write the LHS in the event definition as

$$\begin{aligned} & n_i(t)\mu^\star - U_i(t) - \mathsf{Reg}_i(t) && (7)\\ &= \sum_{k\in T_i(t)} \left(\mu^\star - r_k - \max_{\pi'\in\Pi} \mathbb{E}[r_k|\pi', x_k] + \mathbb{E}[r_k|\pi_k, x_k] \right) \\ &= \sum_{k\in T_i(t)} \left(\mu^\star - \max_{\pi'\in\Pi} \mathbb{E}[r_k|\pi', x_k] \right) + \sum_{k\in T_i(t)} \left(\mathbb{E}[r_k|\pi_k, x_k] - r_k \right). && (8) \end{aligned}$$

Consider the first sum and let $\mathcal{F}_t$ be the sigma-field induced by all variables up to round t, i.e., $(\mathcal{I}_k, x_k, i_k, a_k, r_k)_{k\le t}$. Note that i_{t+1}, the learner chosen at $t+1$ is $\mathcal{F}_t$-measurable. Hence, $X_k = \mathbf{1}\{i_k = i\}(\mu^\star - \max_{\pi'\in\Pi}\mathbb{E}[r_k|\pi', x_k]) \in [-1, +1]$ is a martingale-difference sequence w.r.t. $\mathcal{F}_k$. We will now apply a Hoeffding-style uniform concentration bound from Howard et al. (2021). Using the terminology and definition in this article, by case Hoeffding I in Table 4, the process $S_k = \sum_{j=1}^k X_k$ is sub-ψ_N with variance process $V_k = \sum_{j=1}^k \mathbf{1}\{i_j = i\}/4$. Thus by using the boundary choice in Equation (11) of Howard et al. (2021), we get

$$\begin{aligned} S_k &\le 1.7\sqrt{V_k\,(\ln\ln(2V_k) + 0.72\ln(5.2/\delta))} \\ &= 0.85\sqrt{n_i(k)\,(\ln\ln(n_i(k)/2) + 0.72\ln(5.2/\delta))} \end{aligned}$$

for all k where $V_k \ge 1$ with probability at least $1-\delta$. Applying the same argument to $-S_k$ gives that

$$\left| \sum_{k\in T_i(t)} \left(\mu^\star - \max_{\pi'\in\Pi} \mathbb{E}[r_k|\pi', x_k] \right) \right| \le 3 \vee 0.85\sqrt{n_i(k)\,(\ln\ln(n_i(k)/2) + 0.72\ln(10.4/\delta))}$$

holds with probability at least $1-\delta$ for all t.

Consider now the second term in (8) and let $\mathcal{F}_t$ now be the sigma-field induced by all variables up to the reward at round $t+1$, i.e., $\sigma((\mathcal{I}_k, x_k, i_k, a_k, r_k)_{k\le t}, \mathcal{I}_{t+1}, x_{t+1}, i_{t+1}, a_{t+1})$. Then $X_k = \mathbf{1}\{i_k = i\}(\mathbb{E}[r_k|\pi_k, x_k] - r_k) \in [-1, +1]$ is a martingale-difference sequence w.r.t. $\mathcal{F}_k$ and we can apply the same concentration argument as for the first term to get with probability at least $1-\delta$ for all t

$$\left| \sum_{k\in T_i(t)} \left(\mathbb{E}[r_k|\pi_k, x_k] - r_k \right) \right| \le 3 \vee 0.85\sqrt{n_i(k)\,(\ln\ln(n_i(k)/2) + 0.72\ln(10.4/\delta))}\,.$$

We now take a union bound over both concentration results and $i \in [M]$ and rebind $\delta \to \delta/M$. Then picking the absolute constant c sufficiently large gives the desired statement. □

A.2. Ancillary Technical Lemmas

Lemma 6. *Let $(x_i)_{i\in[n]}$ be a sequence of non-negative numbers and let $\beta \in (0, 1]$. Then*

$$\sum_{i=1}^n \frac{x_i}{\left(\sum_{j=1}^i x_j\right)^{1-\beta}} \le \frac{1}{\beta}\left(\sum_{i=1}^n x_i\right)^\beta$$

Proof. We proceed by induction on n. The $n = 1$ case is clear. Suppose the statement is true for some n. Since $\beta \leq 1$, the function $f(z) = z^\beta$ is concave. Therefore

$$f\left(\sum_{i=1}^{n+1} x_i\right) \geq f\left(\sum_{i=1}^{n} x_i\right) + x_{n+1} f'\left(\sum_{i=1}^{n+1} x_i\right)$$
$$\left(\sum_{i=1}^{n+1} x_i\right)^\beta \geq \left(\sum_{i=1}^{n} x_i\right)^\beta + \frac{\beta x_{n+1}}{\left(\sum_{i=1}^{n+1} x_i\right)^{1-\beta}}$$

apply the induction hypothesis:

$$\geq \sum_{i=1}^{n+1} \frac{\beta x_i}{\left(\sum_{j=1}^{i} x_i\right)^{1-\beta}}$$

□

Lemma 7 (Elliptical potential)**.** *Let $x_1, \ldots, x_n \in \mathbb{R}^d$ and $V_t = V_0 + \sum_{i=1}^t x_i x_i^\top$ and $b > 0$ then*

$$\sum_{t=1}^{n} b \wedge \|x_t\|^2_{V_{t-1}^{-1}} \leq \frac{b}{\ln(b+1)} \ln \frac{\det V_n}{\det V_0} \leq (1+b) \ln \frac{\det V_n}{\det V_0}.$$

Proof Sketch. The proof is identical to the usual elliptical potential lemma (Lattimore & Szepesvári, 2020, Lemma 19.4) where $b = 1$ except that we need to argue that for any $b > 0$

$$b \wedge u \leq c \ln(u+1)$$

holds whenever $c \geq \frac{b}{\ln(1+b)}$. Since $\ln(1+\cdot)$ is strictly concave and strictly monotonically increasing, it is sufficient for us to check that this inequality holds at the critical point $u = b$ which is the case. □

Lemma 8 (Randomized elliptical potential)**.** *Let $x_1, x_2, \cdots \in \mathbb{R}^d$ and $I_1, I_2, \cdots \in \{0, 1\}$ and $V_0 \in \mathbb{R}^{d \times d}$ be random variables so that $\mathbb{E}[I_k | x_1, I_1, \ldots, x_{k-1}, I_{k-1}, x_k, V_0] = p$ for all $k \in \mathbb{N}$. Further, let $V_t = V_0 + \sum_{i=1}^t I_i x_i x_i^\top$. Then*

$$\sum_{t=1}^{n} b \wedge \|x_t\|^2_{V_{t-1}^{-1}} \leq 1 \vee 2.9 \frac{b}{p} \left(1.4 \ln \ln (2bn \vee 2) + \ln \frac{5.2}{\delta}\right) + \frac{2}{p}(1+b) \ln \frac{\det V_n}{\det V_0}$$
$$= \frac{4}{p}(1+b) \ln \frac{\ln(2bn \vee 2) 5.2 \det V_n}{\delta \det V_0}$$

holds with probability at least $1 - \delta$ for all n simultaneously.

Proof. We decompose the sum of squares as

$$\sum_{t=1}^{n} b \wedge \|x_t\|^2_{V_{t-1}^{-1}} = \frac{1}{p} \sum_{t=1}^{n} (bI_t \wedge \|I_t x_t\|^2_{V_{t-1}^{-1}}) + \frac{1}{p} \sum_{t=1}^{n} (p - I_t)(b \wedge \|x_t\|^2_{V_{t-1}^{-1}}) \tag{9}$$

The first term can be controlled using the standard elliptical potential lemma in Lemma 7 as

$$\frac{1}{p} \sum_{t=1}^{n} (bI_t \wedge \|I_t x_t\|^2_{V_{t-1}^{-1}}) \leq \frac{1}{p} \sum_{t=1}^{n} (b \wedge \|I_t x_t\|^2_{V_{t-1}^{-1}}) \leq \frac{1}{p}(1+b) \ln \frac{\det V_n}{\det V_0}.$$

For the second term, we apply an empirical variance uniform concentration bound. Let $\mathcal{F}_{i-1} = \sigma(V_0, x_1, I_1, \ldots, x_{i-1}, I_{i-1}, x_i)$ be the sigma-field up to before the i-th indicator. Let $Y_i = \frac{1}{p}(p - I_i) \left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)$

which is a martingale difference sequence because $\mathbb{E}[Y_i|\mathcal{F}_{i-1}] = 0$ and consider the process $S_t = \sum_{i=1}^t Y_i$ with variance process

$$W_t = \sum_{i=1}^t \mathbb{E}[Y_i^2|\mathcal{F}_{i-1}] = \sum_{i=1}^t \frac{1}{p^2}\left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)^2 \mathbb{E}[(p - I_i)^2|\mathcal{F}_{i-1}]$$
$$= \frac{1-p}{p}\sum_{i=1}^t \left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)^2 \le \frac{b}{p}\sum_{i=1}^t \left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right) \le \frac{tb^2}{p}.$$

Note that $Y_t \le b$ and therefore, S_t satisfies with variance process W_t the sub-ψ_P condition of Howard et al. (2021) with constant $c = b$ (see Bennett case in Table 3 of Howard et al. (2021)). By Lemma 9 below, the bound

$$S_t \le 1.44\sqrt{(W_t \vee m)\left(1.4\ln\ln\left(2(W_t/m \vee 1)\right) + \ln\frac{5.2}{\delta}\right)}$$
$$+ 0.41b\left(1.4\ln\ln\left(2(W_t/m \vee 1)\right) + \ln\frac{5.2}{\delta}\right)$$

holds for all $t \in \mathbb{N}$ with probability at least $1 - \delta$. We set $m = \frac{b}{p}$ and upper-bound the RHS further as

$$1.44\sqrt{\frac{b}{p}\left(1 \vee \sum_{i=1}^t \left(b \wedge \|x_i\|^2_{V_{i-1}^{-1}}\right)\right)\left(1.4\ln\ln\left(2bt \vee 2\right) + \ln\frac{5.2}{\delta}\right)}$$
$$+ 0.41b\left(1.4\ln\ln\left(2bt \vee 2\right) + \ln\frac{5.2}{\delta}\right)$$
$$\le \frac{1}{2}\left(1 \vee \sum_{i=1}^t \left(b \wedge \|x_i\|^2_{V_{i-1}^{-1}}\right)\right) + 1.45\frac{b}{p}\left(1.4\ln\ln\left(2bt \vee 2\right) + \ln\frac{5.2}{\delta}\right),$$

where the inequality is an application of the AM-GM inequality. Thus, we have shown that with probability at least $1 - \delta$, for all n, the second term in (9) is bounded as

$$\frac{1}{p}\sum_{t=1}^n (p - I_t)(b \wedge \|x_t\|^2_{V_{t-1}^{-1}}) \le \frac{1}{2}\left(1 \vee \sum_{i=1}^n \left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right)\right) + Z.$$

where $Z = 1.45\frac{b}{p}\left(1.4\ln\ln\left(2bn \vee 2\right) + \ln\frac{5.2}{\delta}\right)$. And when combining all bounds on the sum of squares term in (9), we get that either $\sum_{i=1}^n \left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right) \le 1$ or

$$\sum_{i=1}^n \left(\|x_i\|^2_{V_{i-1}^{-1}} \wedge b\right) \le 2Z + \frac{2}{p}(1+b)\ln\frac{\det V_n}{\det V_0}$$
$$\le \frac{4}{p}(1+b)\ln\frac{\ln(2bn \vee 2)5.2\det V_n}{\delta \det V_0}$$

which gives the desired statement. □

Lemma 9 (Uniform empirical Bernstein bound)**.** *In the terminology of Howard et al. (2021), let $S_t = \sum_{i=1}^t Y_i$ be a sub-ψ_P process with parameter $c > 0$ and variance process W_t. Then with probability at least $1 - \delta$ for all $t \in \mathbb{N}$*

$$S_t \le 1.44\sqrt{(W_t \vee m)\left(1.4\ln\ln\left(2\left(\frac{W_t}{m} \vee 1\right)\right) + \ln\frac{5.2}{\delta}\right)}$$
$$+ 0.41c\left(1.4\ln\ln\left(2\left(\frac{W_t}{m} \vee 1\right)\right) + \ln\frac{5.2}{\delta}\right)$$

where $m > 0$ is arbitrary but fixed.

Proof. Setting $s = 1.4$ and $\eta = 2$ in the polynomial stitched boundary in Equation (10) of Howard et al. (2021) shows that $u_{c,\delta}(v)$ is a sub-ψ_G boundary for constant c and level δ where

$$u_{c,\delta}(v) = 1.44\sqrt{(v \vee 1)\left(1.4 \ln \ln (2(v \vee 1)) + \ln \frac{5.2}{\delta}\right)} + 1.21c\left(1.4 \ln \ln (2(v \vee 1)) + \ln \frac{5.2}{\delta}\right).$$

By the boundary conversions in Table 1 in Howard et al. (2021) $u_{c/3,\delta}$ is also a sub-ψ_P boundary for constant c and level δ. The desired bound then follows from Theorem 1 by Howard et al. (2021). □

B. Regret Analysis for Simplified Dynamic Balancing Algorithm

In this section, we provide analysis for a simplified version of Algorithm 1. This method (Algorithm 2) does not employ the balancing coefficients v_i or the biases $b_i(t)$ (i.e. $v_i = 1$, $b_i(t) = 0$) and modifies the activation condition in a subtle way. The new activation condition is:

$$\eta_i(t) + \gamma_i(t) + \frac{R_i(n_i(t))}{n_i(t)} \geq \max_{j \in [M]} \eta_j(t) - \gamma_j(t)$$

The change is that $\gamma_j(t)$ is now subtracted on the RHS rather than added. Although this loosens some bounds, it can yield simpler and more intuitive analysis. In particular, since $b_i(t) = 0$ for all i and t, we are able to argue that a well-specified learner will *never* become inactive. This makes the balancing condition becomes more powerful as the optimal well-specified learner is always being compared against when choosing which learner to play at any given iteration. This background helps inform the general analysis of Algorithm 1 in which we combine all desired properties into a single setting of the parameters in Section C

B.1. Well-Specified Learners Remain Active

Lemma 10 (Well-specified learners always active when bias is zero)**.** *When Algorithm 2 is used without biases ($b_i(t) = 0$), then all well-specified learners are active in all rounds in event $\mathcal{E}$.*

Proof. Without biases, the condition for learner i to be active in round $t + 1$ evaluates to

$$\frac{U_i(t)}{n_i(t)} + \frac{R_i(n_i(t))}{n_i(t)} + c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} \geq \max_{j \in [M]} \frac{U_j(t)}{n_j(t)} - c\sqrt{\frac{\ln(M \ln n_j(t)/\delta)}{n_j(t)}}. \tag{10}$$

Now, by the definition of event $\mathcal{E}$, we have for every learner $j \in [M]$,

$$\mu^\star - \frac{U_j(t)}{n_j(t)} + c\sqrt{\frac{\ln(M \ln n_j(t)/\delta)}{n_j(t)}} \geq \frac{\mathsf{Reg}_i(t)}{n_j(t)} \geq 0,$$

which implies that the RHS of Equation 10 is upper-bounded by the optimal expected reward μ^*. Similarly, we turn to the LHS and bound

$$\frac{U_i(t)}{n_i(t)} + \frac{R_i(n_i(t))}{n_i(t)} + c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} - \mu^* \geq \frac{U_i(t)}{n_i(t)} + \frac{\mathsf{Reg}_i(t)}{n_i(t)} + c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} - \mu^* \geq 0,$$

where we first used the fact that $i \in \mathcal{G}$ is well-specified and then applied the definition of $\mathcal{E}$. This implies that the LHS of Equation 10 is lower-bounded by μ^*. Thus, $i \in \mathcal{G}$ has to be active in $t + 1$. Since this holds for all t and $\mathcal{I}_1 = [M]$, we have shown the desired statement. □

Algorithm 2: Simplified Dynamic Balancing Algorithm

input : M base learners
Candidate regret bound R_i for each learner
Confidence parameter $\delta \in (0,1)$

$U_i(0) = n_i(0) = 0$ for all $i \in [M]$
Active set: $\mathcal{I}_1 \leftarrow [M]$
for *round* $t = 1, 2, \ldots$ **do**
  Select learner from active set as
    $i_t \in \operatorname{argmin}_{i \in \mathcal{I}_t} R_i(n_i(t-1))$
  Play action a_t of learner i_t and receive reward r_t
  Update learner i_t with r_t
  Update $n_i(\cdot)$ and $U_i(\cdot)$:
    $U_{i_t}(t) \leftarrow U_{i_t}(t-1) + r_t$
    $n_{i_t}(t) \leftarrow n_{i_t}(t-1) + 1$
  foreach *learner* $i \in [M]$ **do**
    Compute adjusted avg. reward:
      $\eta_i(t) \leftarrow \frac{U_i(t)}{n_i(t)}$
    Compute confidence band:
      $\gamma_i(t) \leftarrow c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}}$
  Set active learners $\mathcal{I}_{t+1}$ as all $i \in [M]$ that satisfy
    $\eta_i(t) + \gamma_i(t) + \frac{R_i(n_i(t))}{n_i(t)} \geq \max_{j \in [M]} \eta_j(t) - \gamma_j(t)$; `// Note the sign compared to Algorithm 1`

B.2. Regret Contribution of Any Active Learner

Lemma 11. *In all rounds t of Algorithm 2, the regret of any learner $i \in \mathcal{I}_{t+1}$ that is active in the next round can be bounded in event $\mathcal{E}$ as*

$$\begin{aligned}\mathsf{Reg}_i(t) \leq R_i(n_i(t)) + \frac{n_i(t)}{n_j(t)}(\mathsf{Reg}_j(t) - \mathbf{1}\{j \notin \mathcal{I}_{t+1}\} R_j(n_j(t))) \\ + 2c\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}} \left(1 + \mathbf{1}\{j \in \mathcal{I}_{t+1}\} \sqrt{\frac{n_i(t)}{n_j(t)}}\right) ,\end{aligned}$$

where c is a universal constant and $j \in [M]$ is any learner.

Proof. Since $i \in \mathcal{I}_{t+1}$ is active, it satisfies

$$\frac{U_i(t)}{n_i(t)} + c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} + \frac{R_i(n_i(t))}{n_i(t)} \geq \max_{h \in [M]} \frac{U_h(t)}{n_h(t)} - c\sqrt{\frac{\ln(M \ln n_h(t)/\delta)}{n_h(t)}} .$$

Let $j \in [M]$ be an arbitrary base learner. If $j \notin \mathcal{I}_{t+1}$ is inactive, then

$$\frac{U_j(t)}{n_j(t)} + c\sqrt{\frac{\ln(M \ln n_j(t)/\delta)}{n_j(t)}} + \frac{R_j(n_j(t))}{n_j(t)} \leq \max_{h \in [M]} \frac{U_h(t)}{n_h(t)} - c\sqrt{\frac{\ln(M \ln n_h(t)/\delta)}{n_h(t)}} ,$$

and otherwise

$$\frac{U_j(t)}{n_j(t)} - c\sqrt{\frac{\ln(M \ln n_j(t)/\delta)}{n_j(t)}} \leq \max_{h \in [M]} \frac{U_h(t)}{n_h(t)} - c\sqrt{\frac{\ln(M \ln n_h(t)/\delta)}{n_h(t)}} ,$$

Combining all inequalities above yields

$$\frac{U_i(t)}{n_i(t)} + c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} + \frac{R_i(n_i(t))}{n_i(t)} \\ \geq \frac{U_j(t)}{n_j(t)} - (\mathbf{1}\{j \in \mathcal{I}_{t+1}\} - \mathbf{1}\{j \notin \mathcal{I}_{t+1}\})c\sqrt{\frac{\ln(M \ln n_j(t)/\delta)}{n_j(t)}} + \mathbf{1}\{j \notin \mathcal{I}_{t+1}\}\frac{R_j(n_j(t))}{n_j(t)} .$$

Subtracting $\mu^\star$ from both sides and rearranging terms gives

$$\mu^\star - \frac{U_i(t)}{n_i(t)} - c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} - \frac{R_i(n_i(t))}{n_i(t)} \\ \leq \mu^\star - \frac{U_j(t)}{n_j(t)} + (\mathbf{1}\{j \in \mathcal{I}_{t+1}\} - \mathbf{1}\{j \notin \mathcal{I}_{t+1}\})c\sqrt{\frac{\ln(M \ln n_j(t)/\delta)}{n_j(t)}} - \mathbf{1}\{j \notin \mathcal{I}_{t+1}\}\frac{R_j(n_j(t))}{n_j(t)} .$$

Applying the definition of $\mathcal{E}$, we obtain an inequality in terms of pseudo-regrets:

$$\frac{\mathsf{Reg}_i(t)}{n_i(t)} - 2c\sqrt{\frac{\ln(M \ln n_i(t)/\delta)}{n_i(t)}} - \frac{R_i(n_i(t))}{n_i(t)} \\ \leq \frac{\mathsf{Reg}_j(t)}{n_j(t)} + (1 + \mathbf{1}\{j \in \mathcal{I}_{t+1}\} - \mathbf{1}\{j \notin \mathcal{I}_{t+1}\})c\sqrt{\frac{\ln(M \ln n_j(t)/\delta)}{n_j(t)}} - \mathbf{1}\{j \notin \mathcal{I}_{t+1}\}\frac{R_j(n_j(t))}{n_j(t)} .$$

Multiplying both sides by $n_i(t)$ and rearranging terms gives

$$\mathsf{Reg}_i(t) \leq R_i(n_i(t)) + \frac{n_i(t)}{n_j(t)}(\mathsf{Reg}_j(t) - \mathbf{1}\{j \notin \mathcal{I}_{t+1}\} R_j(n_j(t))) \\ + 2c\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}} \left(1 + \mathbf{1}\{j \in \mathcal{I}_{t+1}\}\sqrt{\frac{n_i(t)}{n_j(t)}}\right) .$$

□

Corollary 12. *In all rounds t of Algorithm 2, the regret of any learner $i \in \mathcal{I}_{t+1}$ that is active in the next round can be bounded in event $\mathcal{E}$ as*

$$\mathsf{Reg}_i(t) \leq R_i(n_i(t)) + \mathbf{1}\{\star \in \mathcal{I}_{t+1}\}\frac{n_i(t)}{n_\star(t)}\mathsf{Reg}_\star(t) \\ + 2c\sqrt{n_i(t) \ln \frac{M \ln t}{\delta}} \left(1 + \mathbf{1}\{\star \in \mathcal{I}_{t+1}\}\sqrt{\frac{n_i(t)}{n_\star(t)}}\right) ,$$

where c is a universal constant and $\star \in \mathcal{G}$ is any well-specified learner.

Proof. This statement follows immediately from Lemma 11 by noting that since $\star$ is well-specified, it satisfies $\mathsf{Reg}_\star(t) \leq R_\star(n_\star(t))$. □

B.3. Regret Contribution with Regret Bounds of the form $\epsilon_i C_1 n + C_2\sqrt{n}$

While for most of this paper, we consider regret bounds of the form $Cd_i n^\beta$, we now provide a sketch for the case where learners have regret bounds of the form

$$R_i(n) = \epsilon_i C_1 n + C_2\sqrt{n},$$

which naturally occur in approximately linear bandits and MDPs. We decompose the regret of Algorithm 2 as

$$\mathsf{Reg}(T) = \sum_{i \in \mathcal{G}} \mathsf{Reg}_i(t_i) + \sum_{i \in \mathcal{B}} \mathsf{Reg}_i(t_i) \leq M + \sum_{i \in \mathcal{G}} \mathsf{Reg}_i(\tilde{t}_i) + \sum_{i \in \mathcal{B}} \mathsf{Reg}_i(\tilde{t}_i),$$

where t_i is the last time up to T that learner i was played and $\tilde{t}_i = t_i - 1$. Now by Lemma 10, the learner $\star \in \mathcal{G}$ is active in all rounds and thus, by the learner selection criterion

$$\sum_{i \in \mathcal{G}} \mathsf{Reg}_i(\tilde{t}_i) \le \sum_{i \in \mathcal{G}} R_i(n_i(\tilde{t}_i)) \le \sum_{i \in \mathcal{G}} R_\star(n_\star(\tilde{t}_i)) \le W R_\star(T) \,.$$

Further, for misspecified learners, we can apply Corollary 12 from above to bound

$$\begin{aligned}
\mathsf{Reg}_i(\tilde{t}_i) &\le R_i(n_i(\tilde{t}_i)) + \frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} \mathsf{Reg}_\star(\tilde{t}_i) + 2c\sqrt{n_i(\tilde{t}_i) \ln \frac{M \ln T}{\delta}} \left(1 + \sqrt{\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)}}\right) \\
&\le R_\star(n_\star(\tilde{t}_i)) + \frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} R_\star(n_\star(\tilde{t}_i)) + 2c\sqrt{n_i(\tilde{t}_i) \ln \frac{M \ln T}{\delta}} \left(1 + \sqrt{\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)}}\right)
\end{aligned}$$

Also note that we can distinguish two regimes for each regret bound:

$$\begin{aligned}
&\epsilon_i C_1 n \le R_i(n) \le 2\epsilon_i C_1 n && \text{when} && \sqrt{n} \ge \frac{C_2}{C_1 \epsilon_i} \\
&C_2 \sqrt{n} \le R_i(n) \le 2 C_2 \sqrt{n} && \text{when} && \sqrt{n} \le \frac{C_2}{C_1 \epsilon_i}
\end{aligned}$$

We now go through different cases distinguishing which regime $R_\star(n_\star(\tilde{t}_i))$ and $R_i(n_i(\tilde{t}_i))$ are in.

Case $\star$ in linear regime: Here, we have

$$\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} R_\star(n_\star(\tilde{t}_i)) \le 2\epsilon_\star C_1 n_i(\tilde{t}_i) \qquad \text{and} \qquad \frac{n_i(\tilde{t}_i)}{\sqrt{n_\star(\tilde{t}_i)}} \le \frac{C_1 \epsilon_\star}{C_2} n_i(\tilde{t}_i)$$

because $\frac{1}{\sqrt{n_\star(\tilde{t}_i)}} \le \frac{C_1 \epsilon_\star}{C_2}$. This gives

$$\mathsf{Reg}_i(\tilde{t}_i) \le \epsilon_\star C_1 n_\star(\tilde{t}_i) + 2\epsilon_\star C_1 n_i(\tilde{t}_i) + 2c\sqrt{n_i(\tilde{t}_i) \ln \frac{M \ln T}{\delta}} + 2c\sqrt{\ln \frac{M \ln T}{\delta}} \frac{C_1 \epsilon_\star}{C_2} n_i(\tilde{t}_i) \,.$$

Case $\star$ in square-root regime and i in linear regime: Here, we have $\sqrt{n_\star(\tilde{t}_i)} \le \frac{C_2}{C_1 \epsilon_\star}$ and $\sqrt{n_i(\tilde{t}_i)} \ge \frac{C_2}{C_1 \epsilon_i} \ge \frac{C_2}{C_1 \epsilon_\star} \ge \sqrt{n_\star(\tilde{t}_i)}$ where we used that if i is misspecified then $\epsilon_i \le \epsilon_\star$. The balancing condition implies

$$C_1 \epsilon_i n_i(\tilde{t}_i) \le R_i(n_i(\tilde{t}_i)) \le R_\star(n_\star(\tilde{t}_i)) \le 2C_2 \sqrt{n_\star(\tilde{t}_i)}$$

and thus

$$\frac{n_i(\tilde{t}_i)}{\sqrt{n_\star(\tilde{t}_i)}} \le \frac{2C_2}{C_1 \epsilon_i} \le 2\sqrt{n_i(\tilde{t}_i)} \,.$$

This yields

$$\mathsf{Reg}_i(\tilde{t}_i) \le 6C_2 \sqrt{n_i(\tilde{t}_i)} + 6c\sqrt{n_i(\tilde{t}_i) \ln \frac{M \ln T}{\delta}} \,.$$

Case $\star$ and i in square-root regime: Here we have by the balancing condition

$$C_2 \sqrt{n_i(\tilde{t}_i)} \le R_i(n_i(\tilde{t}_i)) \le R_\star(n_\star(\tilde{t}_i)) \le 2C_2 \sqrt{n_\star(\tilde{t}_i)} \,.$$

and thus $\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} \leq 4$.

$$\mathsf{Reg}_i(\tilde{t}_i) \leq 5C_2\sqrt{n_i(\tilde{t}_i)} + 6c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln T}{\delta}}\,.$$

Combining the bounds in all cases yields

$$\mathsf{Reg}_i(\tilde{t}_i) \leq \epsilon_\star C_1 n_\star(\tilde{t}_i) + 2\epsilon_\star C_1 n_i(\tilde{t}_i) + 6C_2\sqrt{n_i(\tilde{t}_i)} + 6c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln T}{\delta}} + 2c\sqrt{\ln\frac{M\ln T}{\delta}}\frac{C_1\epsilon_\star}{C_2}n_i(\tilde{t}_i)$$

and summing over all $i \in \mathcal{B}$

$$\begin{aligned}\sum_{i\in\mathcal{B}}\mathsf{Reg}_i(\tilde{t}_i) \leq{}& \epsilon_\star C_1\sum_{i\in\mathcal{B}} n_\star(\tilde{t}_i) + 2\epsilon_\star C_1\sum_{i\in\mathcal{B}} n_i(\tilde{t}_i) + 6C_2\sum_{i\in\mathcal{B}}\sqrt{n_i(\tilde{t}_i)}\\ &+ 6c\sum_{i\in\mathcal{B}}\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln T}{\delta}} + 2c\sum_{i\in\mathcal{B}}\sqrt{\ln\frac{M\ln T}{\delta}}\frac{C_1\epsilon_\star}{C_2}n_i(\tilde{t}_i)\\ \leq{}&\epsilon_\star C_1 T\left(B + 2 + \frac{2c}{C_2}\sqrt{\ln\frac{M\ln T}{\delta}}\right) + 6C_2\sqrt{BT} + 6c\sqrt{BT\ln\frac{M\ln T}{\delta}}\,.\end{aligned}$$

Therefore, the total regret of Algorithm 2 is bounded as

$$\mathsf{Reg}(T) \leq M + \epsilon_\star C_1 T\left(M + 2 + \frac{2c}{C_2}\sqrt{\ln\frac{M\ln T}{\delta}}\right) + 6MC_2\sqrt{T} + 6c\sqrt{BT\ln\frac{M\ln T}{\delta}}$$

B.4. Regret Contribution of Misspecified Learners

We now show that the regret contribution of a learner that is significantly misspecified can be bounded by a gap-dependent quantity with logarithmic dependency on T. We first carry out the argument for the special case of $\sqrt{n}$ candidate regret learners and subsequently generalize the result to n^β in Lemma 14.

Lemma 13. *Assume candidate regret bounds of the form $R_i(n) = Cd_i\sqrt{n}$ and let $\Delta = \frac{\mathsf{Reg}_i(\tilde{t}_i)}{n_i(\tilde{t}_i)}$ where $\tilde{t}_i = t_i - 1$ with t_i being the last round where learner i was played. Then the total regret contributed by learner i in T rounds of Algorithm 2 is bounded by*

$$\mathsf{Reg}_i(T) \leq 1 + \frac{5C^2d_i^2 + 20c^2\ln\frac{M\ln T}{\delta}}{\Delta} + \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\}\frac{d_\star^2}{d_i^2}\frac{5C^2d_\star^2 + 20c^2\ln\frac{M\ln T}{\delta}}{\Delta}$$

in event $\mathcal{E}$ where c is a universal constant and $\star \in [M]$ is any well-specified learner..

Proof. Let t_i be the last time learner i was played up to round T and $\tilde{t}_i = t_i - 1$. By Corollary 12, the regret of i is bounded as

$$\begin{aligned}\mathsf{Reg}_i(\tilde{t}_i) \leq{}& Cd_i\sqrt{n_i(\tilde{t}_i)} + \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\}\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)}Cd_\star\sqrt{n_\star(\tilde{t}_i)}\\ &+ 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}\left(1 + \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\}\sqrt{\frac{n_i(\tilde{t}_i)}{n_\star(t_i)}}\right)\,,\end{aligned}$$

where we used the form of the regret bounds $R_i(n_i(\tilde{t}_i)) \leq Cd_i\sqrt{n_i(\tilde{t}_i)}$. If $\star$ was active in round t_i then the selection criterion implies $R_i(n_i(t_i - 1)) \leq R_\star(n_\star(t_i - 1))$ because i was played in t_i. Plugging in the form of both regret bounds and rearranging terms then gives $\sqrt{n_i(\tilde{t}_i)/n_\star(\tilde{t}_i)} \leq d_\star/d_i$. Applying this bound to the regret bound above gives

$$\mathsf{Reg}_i(\tilde{t}_i) \leq Cd_i\sqrt{n_i(\tilde{t}_i)} + \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\}\sqrt{n_i(\tilde{t}_i)}C\frac{d_\star^2}{d_i}$$

$$+ 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}\left(1+\mathbf{1}\left\{\star\in\mathcal{I}_{t_i}\right\}\frac{d_\star}{d_i}\right)$$
$$= \sqrt{n_i(\tilde{t}_i)}\left(Cd_i + 2c\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}} + \mathbf{1}\left\{\star\in\mathcal{I}_{t_i}\right\}\left(C\frac{d_\star^2}{d_i}+\frac{d_\star}{d_i}2c\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}\right)\right).$$

The inequality has the form $\mathsf{Reg}_i(\tilde{t}_i) \le \sqrt{n_i(\tilde{t}_i)}D$. Since i has linear regret, we further have $\Delta n_i(\tilde{t}_i) \le \mathsf{Reg}_i(\tilde{t}_i)$ which implies that $\sqrt{n_i(\tilde{t}_i)} \le D/\Delta$ and $\mathsf{Reg}_i(\tilde{t}_i) \le \frac{D^2}{\Delta}$. When we write out the full expression for D, we have

$$\mathsf{Reg}_i(\tilde{t}_i) \le \frac{1}{\Delta}\left(Cd_i + 2c\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}} + \mathbf{1}\left\{\star\in\mathcal{I}_{t_i}\right\}\left(C\frac{d_\star^2}{d_i}+\frac{d_\star}{d_i}2c\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}\right)\right)^2$$
$$\le \frac{5C^2d_i^2 + 20c^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{\Delta} + \mathbf{1}\left\{\star\in\mathcal{I}_{t_i}\right\}\frac{d_\star^2}{d_i^2}\frac{5C^2d_\star^2+20c^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{\Delta}$$

where the last inequality follows from $\left(\sum_{i=1}^n a_i\right)^2 \le n\sum_{i=1}^n a_i^2$, an application of Cauchy-Schwarz inequality. Finally, since t_i is the last round where i was played, $\mathsf{Reg}_i(T) = \mathsf{Reg}_i(t_i) \le \mathsf{Reg}_i(\tilde{t}_i) + 1$, which finishes the proof. □

Lemma 14. *Assume candidate regret bounds of the form $R_i(n) = Cd_in^\beta$ and let $\Delta = \frac{\mathsf{Reg}_i(\tilde{t}_i)}{n_i(\tilde{t}_i)}$ where $\tilde{t}_i = t_i - 1$ with t_i being the last round where learner i was played. Then the total regret contributed by learner i in T rounds of Algorithm 2 is bounded by*

$$\mathsf{Reg}_i(T) \le 2(1-\beta)\left(Cd_i\right)^{\frac{1}{1-\beta}}\left(\frac{12\beta}{\Delta}\right)^{\frac{\beta}{1-\beta}} + \mathbf{1}\left\{\star\in\mathcal{I}_{t_i}\right\}2(1-\beta)\left(Cd_\star\right)^{\frac{1}{1-\beta}}\left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}}\left(\frac{12\beta}{\Delta}\right)^{\frac{\beta}{1-\beta}}$$
$$+ \frac{16c^2}{\Delta}\ln\frac{M\ln T}{\delta} + \mathbf{1}\left\{\star\in\mathcal{I}_{t_i}\right\}\frac{16c^2}{\Delta}\ln\frac{M\ln T}{\delta}\left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}} + 1$$

in event $\mathcal{E}$ where c is a universal constant and $\star\in[M]$ is any well-specified learner.

Proof. Let t_i be the last time learner i was played up to round T and $\tilde{t}_i = t_i - 1$. By Corollary 12, the regret of i at round $\tilde{t}_i$ is bounded as

$$\mathsf{Reg}_i(\tilde{t}_i) \le Cd_in_i(\tilde{t}_i)^\beta + \mathbf{1}\left\{\star\in\mathcal{I}_{t_i}\right\}\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)}Cd_\star n_\star(\tilde{t}_i)^\beta$$
$$+ 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}\left(1+\mathbf{1}\left\{\star\in\mathcal{I}_{t_i}\right\}\sqrt{\frac{n_i(\tilde{t}_i)}{n_\star(t_i)}}\right),$$

where we used the form of the regret bounds $R_i(n_i(\tilde{t}_i)) \le Cd_i\sqrt{n_i(\tilde{t}_i)}$. If $\star$ was active in round t_i then the selection criterion implies $R_i(n_i(t_i-1)) \le R_\star(n_\star(t_i-1))$ because i was played in t_i. Plugging in the form of both regret bounds and rearranging terms then gives $\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} \le \left(\frac{d_\star}{d_i}\right)^{1/\beta}$. Applying this bound to the regret bound above gives

$$\mathsf{Reg}_i(\tilde{t}_i) \le Cd_in_i(\tilde{t}_i)^\beta + n_i(\tilde{t}_i)^\beta\mathbf{1}\left\{\star\in\mathcal{I}_{t_i}\right\}Cd_\star\left(\frac{d_\star}{d_i}\right)^{\frac{1-\beta}{\beta}}$$
$$+ 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}\left(1+\mathbf{1}\left\{\star\in\mathcal{I}_{t_i}\right\}\left(\frac{d_\star}{d_i}\right)^{\frac{1}{2\beta}}\right)$$
$$= n_i(\tilde{t}_i)^\beta\underbrace{\left(Cd_i + \mathbf{1}\left\{\star\in\mathcal{I}_{t_i}\right\}Cd_\star\left(\frac{d_\star}{d_i}\right)^{\frac{1-\beta}{\beta}}\right)}_{D_1}$$

$$
\begin{aligned}
&+ \sqrt{n_i(\tilde{t}_i)} 2c \sqrt{\ln \frac{M \ln \tilde{t}_i}{\delta}} \left(1 + \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\} \left(\frac{d_\star}{d_i}\right)^{\frac{1}{2\beta}}\right) \\
&= n_i(\tilde{t}_i)^\beta \underbrace{\left(C d_i + \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\} C d_\star \left(\frac{d_\star}{d_i}\right)^{\frac{1-\beta}{\beta}}\right)}_{D_1'} \\
&\quad + \sqrt{n_i(\tilde{t}_i)} \underbrace{2c \sqrt{\ln \frac{M \ln \tilde{t}_i}{\delta}} \left(1 + \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\} \left(\frac{d_\star}{d_i}\right)^{\frac{1}{2\beta}}\right)}_{D_2} .
\end{aligned}
$$

We now use the assumption that learner i has linear regret and thus $\Delta n_i(\tilde{t}_i) \le \mathsf{Reg}_i(\tilde{t}_i)$. We can therefore subtract $\frac{\Delta}{2} n_i(\tilde{t}_i)$ from both sides of the inequality above and get

$$
\begin{aligned}
\frac{1}{2}\mathsf{Reg}_i(\tilde{t}_i) &\le \left(n_i(\tilde{t}_i)^\beta D_1 - \frac{\Delta}{4} n_i(\tilde{t}_i)\right) + \left(\sqrt{n_i(\tilde{t}_i)} D_2 - \frac{\Delta}{4} n_i(\tilde{t}_i)\right) \\
&= \left(n_i(\tilde{t}_i)^\beta D_1' - \frac{\Delta}{4} n_i(\tilde{t}_i)\right) + \left(\sqrt{n_i(\tilde{t}_i)} D_2 - \frac{\Delta}{4} n_i(\tilde{t}_i)\right) .
\end{aligned}
$$

Let us first consider the second term above, which we bound as follows

$$
\sqrt{n_i(\tilde{t}_i)} D_2 - \frac{\Delta}{4} n_i(\tilde{t}_i) \le \max_{x \ge 0} \left(\sqrt{x} D_2 - \frac{\Delta}{4} x\right) = \frac{D_2^2}{\Delta}
$$

since the maximum is attained at $\sqrt{x} = \frac{2D_2}{\Delta}$. Similarly, we can bound the first term as

$$
n_i(\tilde{t}_i)^\beta D_1 - \frac{\Delta}{4} n_i(\tilde{t}_i) \le \max_{x \ge 0} \left(x^\beta D_1 - \frac{\Delta}{4} x\right) = \max_{x \ge 0} x^\beta \left(D_1 - \frac{\Delta}{4} x^{1-\beta}\right) = \left(\frac{4\beta D_1}{\Delta}\right)^{\frac{\beta}{1-\beta}} (1-\beta) D_1
$$

where the maximum is attained at $x^{1-\beta} = \frac{4\beta D_1}{\Delta}$. This bound holds both for D_1 and D_1'. Combining the bounds for both terms yields

$$
\begin{aligned}
\mathsf{Reg}_i(\tilde{t}_i) &\le 2(1-\beta) D_1^{\frac{1}{1-\beta}} \left(\frac{4\beta}{\Delta}\right)^{\frac{\beta}{1-\beta}} + 2\frac{D_2^2}{\Delta} \qquad \text{and} \\
\mathsf{Reg}_i(\tilde{t}_i) &\le 2(1-\beta) {D_1'}^{\frac{1}{1-\beta}} \left(\frac{4\beta}{\Delta}\right)^{\frac{\beta}{1-\beta}} + 2\frac{D_2^2}{\Delta} .
\end{aligned}
$$

Now, by Hölder's inequality

$$
\begin{aligned}
D_2^2 &\le 8c^2 \ln \frac{M \ln \tilde{t}_i}{\delta} + 8c^2 \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\} \ln \frac{M \ln \tilde{t}_i}{\delta} \left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}} , \\
(D_1')^{\frac{1}{1-\beta}} &\le \left(2^\beta C d_i\right)^{\frac{1}{1-\beta}} + \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\} \left(2^\beta C d_\star\right)^{\frac{1}{1-\beta}} \left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}} , \\
(D_1)^{\frac{1}{1-\beta}} &\le \left(3^\beta C d_i\right)^{\frac{1}{1-\beta}} + \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\} \left(3^\beta C d_\star\right)^{\frac{1}{1-\beta}} \left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}} .
\end{aligned}
$$

Finally, combining all expressions yields the desired bound

$$
\begin{aligned}
\mathsf{Reg}_i(\tilde{t}_i) \le{}& 2(1-\beta) \left(C d_i\right)^{\frac{1}{1-\beta}} \left(\frac{12\beta}{\Delta}\right)^{\frac{\beta}{1-\beta}} + \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\} 2(1-\beta) \left(C d_\star\right)^{\frac{1}{1-\beta}} \left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}} \left(\frac{12\beta}{\Delta}\right)^{\frac{\beta}{1-\beta}} \\
&+ \frac{16c^2}{\Delta} \ln \frac{M \ln T}{\delta} + \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\} \frac{16c^2}{\Delta} \ln \frac{M \ln T}{\delta} \left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}}
\end{aligned}
$$

□

B.5. Main Regret Bounds Without Biases

B.5.1. Worst-Case Bound

This appendix contains the proofs our main regret bounds for Algorithm 2. Our main bound is:

Corollary 15 (Worst-Case Regret Bound without Biases)**.** *Assume that $R_i(n) = Cd_i n^\beta$ with $d_i \geq 1$ for all learners $i \in [M]$. Then the regret of Algorithm 2 is bounded with probability at least $1-\delta$ for all rounds $T \in \mathbb{N}$ as*

$$\mathsf{Reg}(T) \leq \tilde{O}\left(\left(M + B^{1-\beta} d_\star^{\frac{1}{\beta}-1}\right) R_\star(T) + 4c d_\star^{\frac{1}{2\beta}} \sqrt{BT}\right),$$

where $\star \in \mathcal{G}$ is any well-specified learner.

Instead of proving Corollary 15 directly, we show the following version that can be sharper in some cases:

Theorem 16 (Worst-Case Regret Bound without Biases)**.** *Assume that $R_i(n) = Cd_i n^\beta$ with $d_i \geq 1$ and $b_i(t) = 0$ for all learners $i \in [M]$. Then the regret of Algorithm 2 is bounded with probability at least $1-\delta$ for all rounds $T \in \mathbb{N}$ as*

$$\mathsf{Reg}(T) \leq \left(M + \left(\sum_{i \in \mathcal{B}} \frac{d_\star^{1/\beta}}{d_i^{1/\beta}}\right)^{1-\beta}\right) C d_\star T^\beta + 2c\sqrt{BT \ln \frac{M \ln T}{\delta}} + 2c\sqrt{\sum_{i \in \mathcal{B}} \left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}}} \sqrt{T \ln \frac{M \ln T}{\delta}} + M\,.$$

where $\star \in \mathcal{G}$ is any well-specified learner.

Proof. We decompose the regret of Algorithm 2 as

$$\mathsf{Reg}(T) = \sum_{i \in \mathcal{G}} \mathsf{Reg}_i(t_i) + \sum_{i \in \mathcal{B}} \mathsf{Reg}_i(t_i) \leq M + \sum_{i \in \mathcal{G}} \mathsf{Reg}_i(\tilde{t}_i) + \sum_{i \in \mathcal{B}} \mathsf{Reg}_i(\tilde{t}_i).$$

We now bound the regret of well-specified learners as

$$\sum_{i \in \mathcal{G}} \mathsf{Reg}_i(\tilde{t}_i) \leq \sum_{i \in \mathcal{G}} R_i(n_i(\tilde{t}_i)) = \sum_{i \in \mathcal{G}} R_i(n_i(t_i - 1)) \overset{(i)}{\leq} \sum_{i \in \mathcal{G}} R_\star(n_\star(t_i - 1)) \leq W C d_\star T^\beta,$$

where step (i) uses the fact that i was played in round t_i and $\star \in \mathcal{G}$ was active (by Lemma 10). For misspecified learners, we apply Corollary 12 to bound their regret contribution as

$$\begin{aligned}
\sum_{i \in \mathcal{B}} \mathsf{Reg}_i(\tilde{t}_i) &\leq \sum_{i \in \mathcal{B}} R_i(n_i(\tilde{t}_i)) + \sum_{i \in \mathcal{B}} \frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} \mathsf{Reg}_\star(\tilde{t}_i) + 2c \sum_{i \in \mathcal{B}} \sqrt{n_i(\tilde{t}_i) \ln \frac{M \ln \tilde{t}_i}{\delta}} \left(1 + \sqrt{\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)}}\right) \\
&\leq \underbrace{\sum_{i \in \mathcal{B}} R_i(n_i(\tilde{t}_i))}_{(A)} + C d_\star \underbrace{\sum_{i \in \mathcal{B}} \frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} n_\star(\tilde{t}_i)^\beta}_{(B)} + 2c\sqrt{\ln \frac{M \ln T}{\delta}} \underbrace{\sum_{i \in \mathcal{B}} \sqrt{n_i(\tilde{t}_i)} \left(1 + \sqrt{\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)}}\right)}_{(C)}.
\end{aligned}$$

Before we bound each term (A), (B) and (C) individually, we first derive a bound on $\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)}$. Note that learner i was played in round $\tilde{t}_i + 1$ and $\star$ is well-specified and thus active in round $\tilde{t}_i + 1$ (by Lemma 10). Therefore, by the learner selection criterion $C d_i n_i(\tilde{t}_i)^\beta = R_i(n_i(\tilde{t}_i)) \leq R_\star(n_\star(\tilde{t}_i)) = C d_\star n_\star(\tilde{t}_i)^\beta$. Rearranging this condition yields

$$\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} \leq \left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}}.$$

We now bound each term as

$$(A) \leq \sum_{i \in \mathcal{B}} R_\star(n_\star(\tilde{t}_i)) \leq B R_\star(T) = B C d_\star T^\beta\,,$$

$$
\begin{aligned}
(B) &= \sum_{i\in\mathcal{B}} \left(\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)}\right)^{1-\beta} n_i(\tilde{t}_i)^\beta \le \sum_{i\in\mathcal{B}} \left(\frac{d_\star}{d_i}\right)^{\frac{1-\beta}{\beta}} n_i(\tilde{t}_i)^\beta \le \sum_{i\in\mathcal{B}} \left(\frac{d_\star}{d_i}\right)^{\frac{1-\beta}{\beta}} n_i(\tilde{t}_i)^\beta \\
&\le \left(\sum_{i\in\mathcal{B}} \frac{d_\star^{1/\beta}}{d_i^{1/\beta}}\right)^{1-\beta} \left(\sum_{i\in\mathcal{B}} n_i(\tilde{t}_i)\right)^\beta \le d_\star^{\frac{1}{\beta}-1} \left(\sum_{i\in\mathcal{B}} \frac{1}{d_i^{1/\beta}}\right)^{1-\beta} T^\beta \le d_\star^{\frac{1}{\beta}-1} B^{1-\beta} T^\beta \,, \\
(C) &\le \sum_{i\in\mathcal{B}} \sqrt{n_i(\tilde{t}_i)} \left(1 + \left(\frac{d_\star}{d_i}\right)^{\frac{1}{2\beta}}\right) \le \sqrt{BT} + \sqrt{T}\sqrt{\sum_{i\in\mathcal{B}} \left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}}} \le \sqrt{BT} + d_\star^{\frac{1}{2\beta}} \sqrt{T}\sqrt{\sum_{i\in\mathcal{B}} \left(\frac{1}{d_i}\right)^{\frac{1}{\beta}}} \\
&\le \sqrt{BT} + d_\star^{\frac{1}{2\beta}} \sqrt{BT} \,.
\end{aligned}
$$

Combining all terms above bounds the total regret contribution of misspecified learners as

$$
\sum_{i\in\mathcal{B}} \mathsf{Reg}_i(\tilde{t}_i) \le d_\star^{\frac{1}{\beta}} \left(\sum_{i\in\mathcal{B}} \frac{1}{d_i^{1/\beta}}\right)^{1-\beta} C T^\beta + BC d_\star T^\beta + 2c\sqrt{BT \ln \frac{M \ln T}{\delta}} + 2c\sqrt{\sum_{i\in\mathcal{B}} \left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}}} \sqrt{T \ln \frac{M \ln T}{\delta}} \,,
$$

which yields a total regret bound of

$$
\mathsf{Reg}(T) \le \left(M + \left(\sum_{i\in\mathcal{B}} \frac{d_\star^{1/\beta}}{d_i^{1/\beta}}\right)^{1-\beta}\right) C d_\star T^\beta + 2c\sqrt{BT \ln \frac{M \ln T}{\delta}} + 2c\sqrt{\sum_{i\in\mathcal{B}} \left(\frac{d_\star}{d_i}\right)^{\frac{1}{\beta}}} \sqrt{T \ln \frac{M \ln T}{\delta}} + M \,.
$$

□

B.5.2. GAP-DEPENDENT BOUND

Corollary 17 (Gap-Dependent Regret Bound without Biases)**.** *Assume that $R_i(n) = C d_i n^\beta$ with $d_i \ge 1$ for all learners $i \in [M]$. Let us further assume that all misspecified learners $i \in \mathcal{B}$ have regret $\mathsf{Reg}_i(t)$ bounded from below as $\mathsf{Reg}_i(t) \ge \Delta_i n_i(t)$, for some constants $\Delta_i > 0$. Then the regret of Algorithm 2 with $b_i(t) = 0$ for all $i \in [M]$ and $t \in [T]$ is bounded with probability at least $1-\delta$ for all rounds $T \in \mathbb{N}$ as*

$$
\mathsf{Reg}(T) \le \tilde{O}\left(\min\left\{W R_\star(T), \sum_{i\in\mathcal{G}} \mathsf{Reg}_i(T)\right\} + \sum_{i\in\mathcal{B}} \left(\frac{C(d_i + d_\star^{1/\beta} d_i^{1-1/\beta})}{\Delta_i^\beta}\right)^{\frac{1}{1-\beta}} + \sum_{i\in\mathcal{B}} \frac{d_\star^{1/\beta}}{\Delta_i}\right)
$$

where c is a universal constant, $\star \in \mathcal{G}$ is any well-specified learner, and $W = |\mathcal{G}| \le M$ is the number of well-specified learners.

This corrollary is a simplified version of this stronger bound:

Theorem 18 (Gap-Dependent Regret Bound without Biases)**.** *Assume that $R_i(n) = C d_i n^\beta$ with $d_i \ge 1$ for all learners $i \in [M]$. Let us further assume that all misspecified learners $i \in \mathcal{B}$ have regret $\mathsf{Reg}_i(t)$ bounded from below as $\mathsf{Reg}_i(t) \ge \Delta_i n_i(t)$, for some constants $\Delta_i > 0$. Then the regret of Algorithm 2 is bounded with probability at least $1-\delta$ for all rounds $T \in \mathbb{N}$ as*

$$
\begin{aligned}
\mathsf{Reg}(T) &\le \min\left\{W R_\star(T), \sum_{i\in\mathcal{G}} \mathsf{Reg}_i(T)\right\} \\
&\quad + 2(1-\beta)\, C^{\frac{1}{1-\beta}} \sum_{i\in\mathcal{B}} \left(\frac{12\beta}{\Delta_i}\right)^{\frac{\beta}{1-\beta}} \left(d_i^{\frac{1}{1-\beta}} + \frac{d_\star^{\frac{1}{\beta(1-\beta)}}}{d_i^{\frac{1}{\beta}}}\right) + \ln\frac{M \ln T}{\delta} \sum_{i\in\mathcal{B}} \frac{16c^2}{\Delta_i} \left(1 + \frac{d_\star^{1/\beta}}{d_i^{1/\beta}}\right) + M \,,
\end{aligned}
$$

where c is a universal constant, $\star \in \mathcal{G}$ is any well-specified learner, and $W = |\mathcal{G}| \le M$ is the number of well-specified learners.

Proof. We decompose the regret of Algorithm 2 as

$$\mathsf{Reg}(T) = \sum_{i\in\mathcal{G}} \mathsf{Reg}_i(t_i) + \sum_{i\in\mathcal{B}} \mathsf{Reg}_i(t_i) \le W + \sum_{i\in\mathcal{G}} \mathsf{Reg}_i(\tilde{t}_i) + \sum_{i\in\mathcal{B}} \mathsf{Reg}_i(T).$$

We now bound the regret of well-specified learners as

$$\sum_{i\in\mathcal{G}} \mathsf{Reg}_i(\tilde{t}_i) \le \sum_{i\in\mathcal{G}} R_i(n_i(\tilde{t}_i)) = \sum_{i\in\mathcal{G}} R_i(n_i(t_i-1)) \overset{(i)}{\le} \sum_{i\in\mathcal{G}} R_\star(n_\star(t_i-1)) \le WCd_\star T^\beta,$$

where step (i) uses the fact that i was played in round t_i and $\star \in \mathcal{G}$ was active (by Lemma 10). For misspecified learners, we apply Lemma 14 to bound their regret contribution as

$$\sum_{i\in\mathcal{B}} \mathsf{Reg}_i(T) \le B + 2(1-\beta)\sum_{i\in\mathcal{B}} \left(\frac{12\beta}{\Delta_i}\right)^{\frac{\beta}{1-\beta}} C^{\frac{1}{1-\beta}} \left(d_i^{\frac{1}{1-\beta}} + \frac{d_\star^{\frac{1}{1-\beta}+\frac{1}{\beta}}}{d_i^{\frac{1}{\beta}}}\right) + \ln\frac{M\ln T}{\delta}\sum_{i\in\mathcal{B}}\frac{16c^2}{\Delta_i}\left(1+\frac{d_\star^{1/\beta}}{d_i^{1/\beta}}\right).$$

Combining all terms above bounds the total regret as

$$\begin{aligned}\mathsf{Reg}(T) \le \min\left\{WCd_\star T^\beta, \sum_{i\in\mathcal{G}}\mathsf{Reg}_i(T)\right\} \\ + 2(1-\beta)\sum_{i\in\mathcal{B}} \left(\frac{12\beta}{\Delta_i}\right)^{\frac{\beta}{1-\beta}} C^{\frac{1}{1-\beta}} \left(d_i^{\frac{1}{1-\beta}} + \frac{d_\star^{\frac{1}{1-\beta}+\frac{1}{\beta}}}{d_i^{\frac{1}{\beta}}}\right) + \ln\frac{M\ln T}{\delta}\sum_{i\in\mathcal{B}}\frac{16c^2}{\Delta_i}\left(1+\frac{d_\star^{1/\beta}}{d_i^{1/\beta}}\right) + M.\end{aligned}$$

□

C. Regret Analysis for General Dynamic Balancing Algorithm (Algorithm 1)

In this section we provide the regret analysis of Algorithm 1. In particular, this analysis will prove Theorem 1 in Corollary 23 and 30 respectively. Note that the Theorems presented in the main text are simplified versions of the results presented here.

Further, all the results in this section assume that the value c is the same absolute constant present in Lemma 5 and Equation 6 and all results hold on event $\mathcal{E}$ (which holds with probability at least $1-\delta$).

The procedure is described in Algorithm 1. Note first of all that $\mathcal{I}_t$ is never empty so long as $R_i(n_i(t)) \ge 0$ for all i, as it will always be the case that the learner which maximizes $\eta_i(t)+\gamma_i(t)$ will be active.

C.1. Regret Contribution of Individual Base Learners

Now we proceed to analyze the algorithm, by first providing a refinement of Lemma 11, which applied only to the simplified version of the method, Algorithm 2.

Lemma 19 (Regret contribution of any active learner)**.** *In all rounds t of Algorithm 1, the regret of any learner $i \in \mathcal{I}_{t+1}$ that is active in the next round can be bounded in event $\mathcal{E}$ as*

$$\mathsf{Reg}_i(t) \le R_i(n_i(t)) + \frac{n_i(t)}{n_j(t)}(\mathsf{Reg}_j(t) - \mathbf{1}\{j\notin\mathcal{I}_{t+1}\}R_j(n_j(t))) + n_i(t)(b_j(t)-b_i(t)) + 2c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}},$$

where c is a universal constant and $j\in[M]$ is any learner.

Proof. Since $i\in\mathcal{I}_{t+1}$ is active, it satisfies

$$\frac{U_i(t)}{n_i(t)} + \gamma_i(t) + \frac{R_i(n_i(t))}{n_i(t)} - b_i(t) \ge \max_{h\in[M]}\frac{U_h(t)}{n_h(t)} + \gamma_h(t) - b_h(t).$$

Let $j\in[M]$ be an arbitrary base learner. If $j\notin\mathcal{I}_{t+1}$ is inactive, then

$$\frac{U_j(t)}{n_j(t)} + \gamma_j(t) + \frac{R_j(n_j(t))}{n_j(t)} - b_j(t) \le \max_{h\in[M]}\frac{U_h(t)}{n_h(t)} + \gamma_h(t) - b_h(t),$$

and otherwise we still have

$$\frac{U_j(t)}{n_j(t)} + (1-2\rho)\gamma_j(t) - b_j(t) \le \max_{h\in[M]} \frac{U_h(t)}{n_h(t)} + \gamma_h(t) - b_h(t)\,,$$

Combining all inequalities above yields

$$\frac{U_i(t)}{n_i(t)} + \gamma_i(t) + \frac{R_i(n_i(t))}{n_i(t)} - b_i(t) \ge \frac{U_j(t)}{n_j(t)} + \gamma_j(t) + \mathbf{1}\left\{j \notin \mathcal{I}_{t+1}\right\} \frac{R_j(n_j(t))}{n_j(t)} - b_j(t)\,.$$

Adding $\mu^\star$ from both sides and rearranging terms gives

$$\begin{aligned}\mu^\star - \frac{U_i(t)}{n_i(t)} - \gamma_i(t) - \frac{R_i(n_i(t))}{n_i(t)} + b_i(t)& \\ \le \mu^\star - \frac{U_j(t)}{n_j(t)} - \gamma_j(t) - \mathbf{1}\left\{j \notin \mathcal{I}_{t+1}\right\} \frac{R_j(n_j(t))}{n_j(t)}\,.&\end{aligned}$$

Applying the definition of $\mathcal{E}$, we obtain an inequality in terms of pseudo-regrets:

$$\begin{aligned}\frac{\mathsf{Reg}_i(t)}{n_i(t)} - 2\gamma_i(t) - \frac{R_i(n_i(t))}{n_i(t)} + b_i(t)& \\ \le \frac{\mathsf{Reg}_j(t)}{n_j(t)} - \mathbf{1}\left\{j \notin \mathcal{I}_{t+1}\right\} \frac{R_j(n_j(t))}{n_j(t)} + b_j(t)\,.&\end{aligned}$$

Multiplying both sides by $n_i(t)$ and rearranging terms gives

$$\begin{aligned}\mathsf{Reg}_i(t) \le R_i(n_i(t)) + \frac{n_i(t)}{n_j(t)}(\mathsf{Reg}_j(t) - \mathbf{1}\left\{j \notin \mathcal{I}_{t+1}\right\} R_j(n_j(t))) + n_i(t)(b_j(t) - b_i(t))& \\ + 2c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}}.&\end{aligned}$$

□

Corollary 20. *In all rounds t of Algorithm 1, the regret of any learner $i \in \mathcal{I}_{t+1}$ that is active in the next round can be bounded in event $\mathcal{E}$ as*

$$\begin{aligned}\mathsf{Reg}_i(t) \le R_i(n_i(t)) + \mathbf{1}\left\{\star \in \mathcal{I}_{t+1}\right\} \frac{n_i(t)}{n_\star(t)}\mathsf{Reg}_\star(t) + n_i(t)(b_\star(t) - b_i(t))& \\ + 2c\sqrt{n_i(t)\ln\frac{M\ln t}{\delta}}\,,&\end{aligned}$$

where c is a universal constant and $\star \in \mathcal{G}$ is any well-specified learner.

Proof. This statement follows immediately from Lemma 19 by noting that since $\star$ is well-specified, it satisfies $\mathsf{Reg}_\star(t) \le R_\star(n_\star(t))$. □

Lemma 21. *Assume that Algorithm 1 is used with candidate regret bounds of the form $R_i(n) = Cd_i n^\beta$ and positive biases $b_i(t) > 0$. Then in event $\mathcal{E}$, the regret contribution of any subset $\mathcal{D} \subseteq [M]$ of learners after T total rounds can be bounded as*

$$\begin{aligned}\sum_{i\in\mathcal{D}} \mathsf{Reg}_i(T) \le |\mathcal{D}| + \sum_{i\in\mathcal{D}}\left(n_i(\tilde{t}_i)b_\star(\tilde{t}_i) + \left(\frac{2Cd_i}{b_i(\tilde{t}_i)^\beta}\right)^{\frac{1}{1-\beta}}\right) + 8c^2\sum_{i\in\mathcal{D}}\frac{\ln\frac{M\ln\tilde{t}_i}{\delta}}{b_i(\tilde{t}_i)}& \\ + Cv_\star^{\frac{1-\beta}{\beta}} d_\star^{1/\beta}\left(\sum_{i\in\mathcal{D}:\,\star\in\mathcal{I}_{t_i}} \frac{1}{(v_i d_i)^{1/\beta}}\right)^{1-\beta} T^\beta\,.&\end{aligned}$$

where $\tilde{t}_i = t_i - 1$ and t_i is the last round where learner i was played, $\star \in \mathcal{G}$ a well-specified learner, and c is a universal positive constant.

Proof. First, note that since t_i was the last time i was played and the maximum instantaneous regret is 1, we have $\mathsf{Reg}_i(T) = \mathsf{Reg}_i(t_i) \le 1 + \mathsf{Reg}_i(\tilde{t}_i)$. Now applying Corollary 20 to learner i in round $\tilde{t}_i$, we have

$$\begin{aligned}
\sum_{i\in\mathcal{D}} \mathsf{Reg}_i(T) &= \sum_{i\in\mathcal{D}} \mathsf{Reg}_i(t_i) \le |\mathcal{D}| + \sum_{i\in\mathcal{D}} \mathsf{Reg}_i(\tilde{t}_i) \\
&\le |\mathcal{D}| + \sum_{i\in\mathcal{D}} R_i(n_i(\tilde{t}_i)) + \sum_{i\in\mathcal{D}} \mathbf{1}\left\{\star \in \mathcal{I}_{t_i}\right\} \frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} \mathsf{Reg}_\star(\tilde{t}_i) + \sum_{i\in\mathcal{D}} n_i(\tilde{t}_i)(b_\star(\tilde{t}_i) - b_i(\tilde{t}_i)) \\
&\quad + 2c \sum_{i\in\mathcal{D}} \sqrt{n_i(\tilde{t}_i) \ln \frac{M \ln \tilde{t}_i}{\delta}} \\
&= |\mathcal{D}| + \sum_{i\in\mathcal{D}} \underbrace{\left(R_i(n_i(\tilde{t}_i)) - \frac{1}{2} n_i(\tilde{t}_i) b_i(\tilde{t}_i)\right)}_{(A)} + \sum_{i\in\mathcal{D}} n_i(\tilde{t}_i) b_\star(\tilde{t}_i) + \underbrace{\sum_{i\in\mathcal{D}} \left(2c\sqrt{n_i(\tilde{t}_i) \ln \frac{M \ln \tilde{t}_i}{\delta}} - \frac{1}{2} n_i(\tilde{t}_i) b_i(\tilde{t}_i)\right)}_{(B)} \\
&\quad + \underbrace{\sum_{i\in\mathcal{D}:\, \star\in\mathcal{I}_{t_i}} \frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} \mathsf{Reg}_\star(\tilde{t}_i)}_{(C)} .
\end{aligned}$$

We will now treat each of the three terms, (A), (B), and (C) separately. First, we consider (A), which we bound for each $i \in \mathcal{D}$ as

$$(A) = C d_i n_i(\tilde{t}_i)^\beta - \frac{1}{2} n_i(\tilde{t}_i) b_i(\tilde{t}_i) \le \sup_{x\ge 0} \left\{ C d_i x^\beta - \frac{x}{2} b_i(\tilde{t}_i) \right\} \le \frac{2^{\frac{\beta}{1-\beta}} (\beta^\beta C d_i)^{\frac{1}{1-\beta}}}{b_i(\tilde{t}_i)^{\frac{\beta}{1-\beta}}} \le \left(\frac{2^\beta C d_i}{b_i(\tilde{t}_i)^\beta}\right)^{\frac{1}{1-\beta}}$$

which holds because for $b_i(t) > 0$ and $\beta < 1$, the supremum is attained at $x^{\beta-1} = \frac{b_i(\tilde{t}_i)}{2\beta d_i C}$.

Now, we can handle (B) by again considering each $i \in \mathcal{D}$, and defining $K = 2c\sqrt{\ln \frac{M \ln \tilde{t}_i}{\delta}}$

$$(B) = K\sqrt{n_i(\tilde{t}_i)} - \frac{1}{2} n_i(\tilde{t}_i) b_i(\tilde{t}_i) \le \sup_{x\ge 0} \left\{ K\sqrt{x} - \frac{x}{2} b_i(\tilde{t}_i) \right\} \le \frac{2K^2}{b_i(\tilde{t}_i)}$$

Where the supremum is computed exactly as in the argument for bounding part (A).

To handle term (C), we derive a bound on $\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)}$. Since learner i is chosen in round t_i, then whenever $\star \in \mathcal{I}_{t_i}$ we have:

$$\begin{aligned}
v_i R_i(n_i(t_i - 1)) &\le v_\star R_\star(n_\star(t_i - 1)) \\
C v_i d_i n_i(\tilde{t}_i)^\beta &\le C v_\star d_\star n_\star(\tilde{t}_i)^\beta \\
\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} &\le \left(\frac{v_\star d_\star}{v_i d_i}\right)^{1/\beta} .
\end{aligned}$$

Equipped with this bound on the ratio of plays of $\star$ and i, we can bound (C) as

$$\begin{aligned}
(C) &\le \sum_{i\in\mathcal{D}:\, \star\in\mathcal{I}_{t_i}} \frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} R_\star(n_\star(\tilde{t}_i)) = C d_\star \sum_{i\in\mathcal{D}:\, \star\in\mathcal{I}_{t_i}} \frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} n_\star(\tilde{t}_i)^\beta = C d_\star \sum_{i\in\mathcal{D}:\, \star\in\mathcal{I}_{t_i}} \left(\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)}\right)^{1-\beta} n_i(\tilde{t}_i)^\beta \\
&\le C d_\star \sum_{i\in\mathcal{D}:\, \star\in\mathcal{I}_{t_i}} \left(\frac{v_\star d_\star}{v_i d_i}\right)^{\frac{1}{\beta}-1} n_i(\tilde{t}_i)^\beta \le C v_\star^{\frac{1-\beta}{\beta}} d_\star^{1/\beta} \left(\sum_{i\in\mathcal{D}:\, \star\in\mathcal{I}_{t_i}} \frac{1}{(v_i d_i)^{1/\beta}}\right)^{1-\beta} T^\beta ,
\end{aligned}$$

where the last step is an application of Hölder's inequality. Combining the bounds for each all terms, we have shown

$$\sum_{i\in\mathcal{D}}\mathsf{Reg}_i(T) \le |\mathcal{D}| + \sum_{i\in\mathcal{D}}\left(\frac{2^\beta C d_i}{b_i(\tilde{t}_i)^\beta}\right)^{\frac{1}{1-\beta}} + \sum_{i\in\mathcal{D}} n_i(\tilde{t}_i)b_\star(\tilde{t}_i) + 8c^2\ln\frac{M\ln T}{\delta}\sum_{i\in\mathcal{D}}\frac{1}{b_i(\tilde{t}_i)}$$
$$+ C v_\star^{\frac{1-\beta}{\beta}} d_\star^{1/\beta}\left(\sum_{i\in\mathcal{D}:\ \star\in\mathcal{I}_{t_i}}\frac{1}{(v_i d_i)^{1/\beta}}\right)^{1-\beta} T^\beta\,.$$

□

C.2. Worst-Case Regret Bound

Now, we are finally able to prove our general bound on the regret. This bound provides a setting for the parameters $b_i(t)$ and v_i that are themselves based on arbitrary user-specified parameters $Z_1,\dots,Z_M$ and $W_1,\dots,W_M$. Although in most of our applications, we will consider the case $W_i = \sqrt{M}$ for all i, it will be useful to allow for arbitrary W_i when matching the Pareto frontier for multi-armed bandits in Corollary 25.

Intuitively, the value for $Z_\star$ represents the multiplier that will be applied to the regret of of the optimal bound $Cd_\star T^\beta$: the regret will be roughly $Z_\star^{\frac{1-\beta}{\beta}} Cd_\star T^\beta$. However, there is a tradeoff term in the regret of $O\left(T^\beta\sum_{i\neq\star}\frac{d_i}{Z_i}\right)$ that prevents one from simply setting $Z_i = 0$ for all i.

Theorem 22. *Suppose $R_i(n) = Cd_i n^\beta$, and let $Z_1,\dots,Z_M$ and $W_1,\dots,W_M$ be arbitrary positive real numbers. Let*

$$b_i(t) = \max\left[2CZ_i^{\frac{1-\beta}{\beta}} d_i t^{\beta-1},\ \frac{cW_i\sqrt{\ln\frac{M\ln t}{\delta}}}{\sqrt{t}}\right]$$

and set

$$v_i = \frac{Z_i^\beta}{d_i^{\beta+1}}$$

Then on event $\mathcal{E}$, the total regret of Algorithm 1 is bounded for all rounds $T\in\mathbb{N}$ as follows:

$$\mathsf{Reg}(T) \le \mathsf{Reg}_\star(T) + 2cW_\star\sqrt{T\ln\frac{M\ln T}{\delta}} + 8c\sqrt{T\ln\frac{M\ln T}{\delta}}\sum_{i\neq\star}\frac{1}{W_i} + M$$
$$+ (2\beta^{-1}+\beta)CZ_\star^{\frac{1-\beta}{\beta}} d_\star T^\beta + 3CT^\beta\sum_{i\neq\star}\frac{d_i}{Z_i}\,.$$

Proof. Since $\mathsf{Reg}(T) = \mathsf{Reg}_\star(T) + \sum_{i\neq\star}\mathsf{Reg}_i(T)$, applying Lemma 21 yields:

$$\mathsf{Reg}(T) \le \mathsf{Reg}_\star(T) + M + \sum_{i\neq\star}\left(n_i(\tilde{t}_i)b_\star(\tilde{t}_i) + \left(\frac{2Cd_i}{b_i(\tilde{t}_i)^\beta}\right)^{\frac{1}{1-\beta}}\right) + 8c^2\sum_{i\neq\star}\frac{\ln\frac{M\ln\tilde{t}_i}{\delta}}{b_i(\tilde{t}_i)}$$
$$+ C v_\star^{\frac{1-\beta}{\beta}} d_\star^{1/\beta}\left(\sum_{i\neq\star:\ \star\in\mathcal{I}_{t_i}}\frac{1}{(v_i d_i)^{1/\beta}}\right)^{1-\beta} T^\beta\,.$$

Without loss of generality, order the learners such that $\tilde{t}_1\le\tilde{t}_2\le\cdots\le\tilde{t}_M$. Then if $j\le i$ we have $n_j(\tilde{t}_i)\ge n_j(\tilde{t}_j)$ so that

$$\tilde{t}_i = \sum_{j=1}^M n_j(\tilde{t}_i) \ge \sum_{j\le i} n_j(\tilde{t}_i) \ge \sum_{j\le i} n_j(\tilde{t}_j)\,.$$

Let S be the set of indices $i \neq \star$ such that $b_\star(\tilde{t}_i) = CZ_i^{\frac{1-\beta}{\beta}} d_i \tilde{t}_i^{\beta-1}$. Then using the above relation, we obtain

$$\sum_{i\in S} n_i(\tilde{t}_i) b_\star(\tilde{t}_i) = \sum_{i\in S} 2CZ_\star^{\frac{1-\beta}{\beta}} d_\star \frac{n_i(\tilde{t}_i)}{\tilde{t}_i^{1-\beta}} \leq \sum_{i=1}^{M} 2CZ_\star^{\frac{1-\beta}{\beta}} d_\star \frac{n_i(\tilde{t}_i)}{\left(\sum_{j\leq i} n_j(\tilde{t}_j)\right)^{1-\beta}} \leq \frac{2CZ_\star^{\frac{1-\beta}{\beta}} d_\star}{\beta} T^\beta \,,$$

where the final inequality is an application of Lemma 6. Next, let S' be the set of indices $i \neq \star$ such that $i \notin S$. Then

$$\sum_{i\in S'} n_i(\tilde{t}_i) b_\star(\tilde{t}_i) \leq c\sqrt{\ln \frac{M\ln T}{\delta}} \sum_{i\in S'} W_\star \frac{n_i(\tilde{t}_i)}{\sqrt{\tilde{t}_i}} \leq c\sqrt{\ln \frac{M\ln T}{\delta}} \sum_{i=1}^{M} W_\star \frac{n_i(\tilde{t}_i)}{\sqrt{\sum_{j\leq i} n_j(\tilde{t}_j)}} \leq 2cW_\star \sqrt{Tln \frac{M\ln T}{\delta}} \,,$$

Thus, overall

$$\sum_{i\neq\star} n_i(\tilde{t}_i) b_\star(\tilde{t}_i) \leq \frac{2CZ_\star^{\frac{1-\beta}{\beta}} d_\star}{\beta} T^\beta + 2cW_\star \sqrt{Tln \frac{M\ln T}{\delta}} \,.$$

Next, we bound:

$$\sum_{i\neq\star} \left(\frac{2Cd_i}{b_i(\tilde{t}_i)^\beta}\right)^{\frac{1}{1-\beta}} \leq \sum_{i\neq\star} \left(\frac{2Cd_i\tilde{t}_i^\beta}{Z_i}\right) \leq 2CT^\beta \sum_{i\neq\star} \frac{d_i}{Z_i}$$

And similarly,

$$8c^2 \sum_{i\neq\star} \frac{\ln \frac{M\ln \tilde{t}_i}{\delta}}{b_i(\tilde{t}_i)} \leq 8c \sum_{i\neq\star} \frac{\sqrt{T\ln \frac{M\ln \tilde{t}_i}{\delta}}}{W_i}$$

Finally, by Young's inequality $xy \leq \frac{x^p}{p} + \frac{y^q}{q}$ with $p = 1/\beta$ and $q = \frac{1}{1-\beta}$:

$$\begin{aligned} Cv_\star^{\frac{1-\beta}{\beta}} d_\star^{1/\beta} \left(\sum_{i\neq\star:\ \star\in\mathcal{I}_{t_i}} \frac{1}{(v_i d_i)^{1/\beta}}\right)^{1-\beta} T^\beta &\leq \beta C v_\star^{\frac{1-\beta}{\beta^2}} d_\star^{1/\beta^2} T^\beta + (1-\beta)CT^\beta \sum_{i\neq\star} \frac{1}{(v_i d_i)^{1/\beta}} \\ &= \beta CZ_\star^{\frac{1-\beta}{\beta}} d_\star T^\beta + (1-\beta)CT^\beta \sum_{i\neq\star} \frac{d_i}{Z_i} \end{aligned}$$

Combining all, we have shown

$$\begin{aligned} \mathsf{Reg}(T) \leq \mathsf{Reg}_\star(T) &+ 2cW_\star \sqrt{T\ln \frac{M\ln T}{\delta}} + 8c\sqrt{T\ln \frac{M\ln T}{\delta}} \sum_{i\neq\star} \frac{1}{W_i} + M \\ &+ (2\beta^{-1} + \beta) CZ_\star^{\frac{1-\beta}{\beta}} d_\star T^\beta + (3-\beta)CT^\beta \sum_{i\neq\star} \frac{d_i}{Z_i} \end{aligned}$$

which implies the desired result. □

With this theorem in hand, we can proceed to prove the fully-detailed version of Theorem 1 from the main text.

The Corollary only holds when assuming event $\mathcal{E}$ (as do all results in this section). However, recall the $\mathcal{E}$ occurs with probability at least $1-\delta$, as shown in Lemma 5.

Corollary 23. *Suppose $R_i(n) = Cd_i n^\beta$, and suppose that $d_1 \leq \cdots \leq d_M$. Let $W_i = W = \sqrt{M}$ for all i, and $Z_i = \frac{d_i}{d_1} i^\beta$. Set*

$$b_i(t) = \max\left[2CZ_i^{\frac{1-\beta}{\beta}} d_i t^{\beta-1},\ \frac{cW\sqrt{\ln \frac{M\ln t}{\delta}}}{\sqrt{t}}\right]$$

and set

$$v_i = \frac{Z_i^\beta}{d_i^{\beta+1}}$$

Then if $B = |\mathcal{B}|$, on event $\mathcal{E}$, the total regret of Algorithm 1 is bounded for all rounds $T \in \mathbb{N}$ as follows:

$$\mathsf{Reg}(T) \leq \mathsf{Reg}_\star(T) + 10c\sqrt{MT\ln\frac{M\ln T}{\delta}} + M + (2\beta^{-1}+\beta)C\left(\frac{B}{d_1^{1/\beta}}\right)^{1-\beta} d_\star^{1/\beta}T^\beta + \frac{3Cd_1T^\beta M^{1-\beta}}{1-\beta}$$

where $\star = \operatorname{argmin}_{i\notin\mathcal{B}} d_i$ is the index of the well-specified learner with minimum value of $d_\star$.

Proof. Notice that $Z_\star \leq \frac{d_i}{d_1}B^\beta$ since any learning with $d_i \leq d_\star$ must be misspecified by definition of $\star$. Further, by Lemma 6,

$$\sum_{i\neq\star}\frac{d_i}{Z_i} = \sum_{i\neq\star}\frac{d_1}{i^\beta} \leq \frac{d_1}{1-\beta}M^{1-\beta}$$

Now, apply Theorem 22:

$$\mathsf{Reg}(T) \leq \mathsf{Reg}_\star(T) + 2cW\sqrt{T\ln\frac{M\ln T}{\delta}} + 8c\sqrt{T\ln\frac{M\ln T}{\delta}}\sum_{i\neq\star}\frac{1}{W} + M + (2\beta^{-1}+\beta)CZ_\star^{\frac{1-\beta}{\beta}}d_\star T^\beta + 3CT^\beta\sum_{i\neq\star}\frac{d_i}{Z_i}$$

and the result now follows. □

C.3. Worst-Case Regret Bound Recovering Corral Guarantees

Using a different setting of the parameters, we are able to recover the same bounds as available in CORRAL and stochastic CORRAL (Agarwal et al., 2017; Pacchiano et al., 2020b):

Corollary 24. *Suppose $R_i(n) = Cd_in^\beta$, and suppose that $d_1 \leq \cdots \leq d_M$. Let η be some arbitrary parameter. Set $W_i = W = \sqrt{M}$ for all i and $Z_i = \eta T^\beta Cd_i$. Set*

$$b_i(t) = \max\left[2CZ_i^{\frac{1-\beta}{\beta}}d_it^{\beta-1},\ \frac{cW\sqrt{\ln\frac{M\ln t}{\delta}}}{\sqrt{t}}\right]$$

and set

$$v_i = \frac{Z_i^\beta}{d_i^{\beta+1}}$$

Then on event $\mathcal{E}$, the total regret of Algorithm 1 is bounded for all rounds $T \in \mathbb{N}$ as follows:

$$\mathsf{Reg}(T) \leq \mathsf{Reg}_\star(T) + 10c\sqrt{MT\ln\frac{M\ln T}{\delta}} + M + (2\beta^{-1}+\beta)\eta^{\frac{1-\beta}{\beta}}(Cd_\star)^{1/\beta}T + \frac{3M}{\eta}$$

where $\star = \operatorname{argmin}_{i\notin\mathcal{B}} d_i$ is the index of the well-specified learner with minimum value of $d_\star$.

Proof. Observe that

$$CZ_\star^{\frac{1-\beta}{\beta}}d_\star T^\beta = \eta^{\frac{1-\beta}{\beta}}(Cd_\star)^{1/\beta}T$$
$$CT^\beta\sum_{i\neq\star}\frac{d_i}{Z_i} = \sum_{i\neq\star}\frac{1}{\eta}$$

Now the result follows from Theorem 22. □

C.4. Worst-Case Regret Bound Recovering Pareto Frontier for Multi-Armed Bandits

We can also recover (up to log factors) the Pareto frontier for bandits (Lattimore, 2015):

Corollary 25. *Consider an M-armed bandit problem for which we have M learners, each dedicated to only playing one specific arm. Let $\star$ be the optimal arm. Set $R_i(n) = \sqrt{n}$ for all i. Let $P_1, \dots, P_M$ be numbers satisfying for all i.*

$$P_i \geq \min\left(T,\ \sum_{j \neq i} \frac{T}{P_j}\right)$$

Set

$$b_i(t) = \frac{2cP_i\sqrt{\ln \frac{M \ln t}{\delta}}}{\sqrt{t}\sqrt{T}}$$

and set

$$v_i = \sqrt{P_i/\sqrt{T}}$$

Then on event $\mathcal{E}$, the total regret of Algorithm 1 is bounded as

$$\mathsf{Reg}(T) \leq M + \left(10c\sqrt{\ln \frac{M \ln T}{\delta}} + 8\right) P_\star$$

Proof. Notice that the settings in the statement correspond to setting $Z_i = W_i = Z = W = \frac{P_i}{\sqrt{T}}$ for all i, $\beta = 1/2$ and $C = d_i = 1$ in Theorem 22. Thus:

$$\begin{aligned}\mathsf{Reg}(T) &\leq \mathsf{Reg}_\star(T) + 2cW\sqrt{T \ln \frac{M \ln T}{\delta}} + 8c\sqrt{T \ln \frac{M \ln T}{\delta}} \sum_{i \neq \star} \frac{1}{W} + M + 5CZd_\star\sqrt{T} + 3C\sqrt{T}\sum_{i \neq \star} \frac{d_i}{Z} \\ &= 10c\sqrt{\ln \frac{M \ln T}{\delta}} \sum_{i \neq \star} \frac{T}{P_i} + M + 5P_\star + 3\sum_{i \neq \star} \frac{T}{P_i}\,.\end{aligned}$$

The result then follows after observing that by definition of $P_\star$,

$$\sum_{i \neq \star} \frac{T}{P_i} \leq P_\star\,.$$

This concludes the proof. □

Finally, we illustrate our ability to "bias" the dynamic balancing routine in favor of a particular learner. Specifically, if we suspect that some learner j will be the best learner, we can bias the algorithm to have very low overhead when $j = \star$, at the expense of always suffering $O(Cd_jT^\beta)$ regret:

Corollary 26. *Suppose $R_i(n) = Cd_in^\beta$ and $d_1 \leq \cdots \leq d_M$. Let j be some arbitrary learner. set $W_i = W = \sqrt{M}$ for all i. Set $Z_i = \frac{d_i}{d_1}/i^\beta$ for $i \neq j$, and $Z_j = 1$. Set*

$$b_i(t) = \max\left[2CZ_i^{\frac{1-\beta}{\beta}} d_i t^{\beta-1},\ \frac{cW\sqrt{\ln \frac{M \ln t}{\delta}}}{\sqrt{t}}\right]$$

and set

$$v_i = \frac{Z_i^\beta}{d_i^{\beta+1}}$$

Then if $B = |\mathcal{B}|$, on event $\mathcal{E}$, the total regret of Algorithm 1 is bounded for all rounds $T \in \mathbb{N}$ as follows:

$$\mathsf{Reg}(T) \le \mathsf{Reg}_\star(T) + 8c\sqrt{M \ln \frac{M \ln T}{\delta}} + M + (2\beta^{-1} + \beta)C \left(\frac{B}{d_1^{1/\beta}}\right)^{1-\beta} d_\star^{1/\beta} T^\beta + 3C d_j T^\beta + \frac{3C d_1 T^\beta M^{1-\beta}}{1-\beta}$$

where $\star = \operatorname{argmin}_{i \notin \mathcal{B}} d_i$. Further, if j is well-specified,

$$\mathsf{Reg}(T) \le \mathsf{Reg}_\star(T) + 8c\sqrt{M \ln \frac{M \ln T}{\delta}} + M + (2\beta^{-1} + \beta)C d_j T^\beta + \frac{3C d_1 T^\beta M^{1-\beta}}{1-\beta}$$

Proof. We have:

$$\sum_{i \neq j} \frac{d_i}{Z_i} = \sum_{i \neq j} \frac{d_1}{i^\beta} \le \frac{d_1 M^{1-\beta}}{1-\beta} .$$

And for $\star \neq j$:

$$\sum_{i \neq \star} \frac{d_i}{Z_i} = d_j + \frac{d_1 M^{1-\beta}}{1-\beta} .$$

Regardless, of whether $j = \star$, we have

$$\sum_i \frac{1}{W_i} \le \sqrt{M} .$$

Now, Apply Theorem 22:

$$\mathsf{Reg}(T) \le \mathsf{Reg}_\star(T) + 2cW_\star\sqrt{T \ln \frac{M \ln T}{\delta}} + 8c\sqrt{T \ln \frac{M \ln T}{\delta}} \sum_{i \neq \star} \frac{1}{W_i} + M + (2\beta^{-1} + \beta)C Z_\star^{\frac{1-\beta}{\beta}} d_\star T^\beta + 3CT^\beta \sum_{i \neq \star} \frac{d_i}{Z_i}$$

and the result now follows from a case analysis of whether $j = \star$. □

C.5. Gap-Dependent Regret Bounds

To show a gap-dependent bound for Algorithm 1, we first show that any algorithm that suffers a large gap between its performance and that of the optimal action (e.g., has linear regret), must be played essentially only a constant number of times (Lemma 27). This means that the major source of regret will be the well-specified algorithms that do not experience such a gap in their regret. In order to bound the regret contributed by these algorithms, we leverage the particular settings of b_i and v_i described in Corollary 23. These provide two main properties:

1. If i is a learner whose regret bound is larger than $\star$, then either $\star$ is inactive when i is last played, or i is played fewer times than $\star$. This property is a consequence of the setting for v.

2. If i is a learner whose regret bound is larger than $\star$, then we use the fact that $b_i \ge b_\star$, and that if R_i is a constant factor larger than $R_\star$, then b_i is also a constant factor larger than $b_\star$ bound the contributions to the regret from the $b_\star$ as well as R_i.

In order to show a gap-dependent bound, we need an analog of Lemma 14:

Lemma 27. *Define Δ_i by $\mathsf{Reg}_i(\tilde{t}_i) = n_i(\tilde{t}_i)\Delta_i$ where $\tilde{t}_i = t_i - 1$ and t_i is the last round where learner i was played. Suppose $R_i(n) = Cd_i n^\beta$, $\star$ is well-specified, and $b_\star(t) = \max\left[2CZ_\star^{1-\beta}\beta d_\star t^{\beta-1}, \frac{cW_\star\sqrt{\ln\frac{M\ln t}{\delta}}}{\sqrt{t}}\right]$ for some $Z_\star$ and $W_\star$. Then Algorithm 1 guarantees:*

$$\mathsf{Reg}_i(\tilde{t}_i) \le \max\left\{\frac{(Cd_\star)^{\frac{1}{1-\beta}}Z_\star^{1/\beta}}{\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6Cd_i)^{\frac{1}{1-\beta}}}{\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6C)^{\frac{1}{1-\beta}}(v_\star d_\star)^{1/\beta}}{(v_i d_i)^{1/\beta}\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{12c^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{\Delta_i}, \min\left[\frac{n_i(\tilde{t}_i)cW_\star\sqrt{\ln\frac{M\ln T}{\delta}}}{2\sqrt{\tilde{t}_i}}, \frac{c^2W_\star^2\ln\frac{M\ln T}{\delta}}{4\Delta_i}\right]\right\}.$$

Proof. First, consider the case in which $\Delta_i \le \frac{1}{2}b_\star(\tilde{t}_i)$. In this situation, we have either

$$\Delta_i \le CZ_\star^{1-\beta}\beta d_\star \tilde{t}_i^{\beta-1}$$

or

$$\Delta_i \le \frac{cW_\star\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}}{2\sqrt{\tilde{t}_i}}.$$

In the former case, it must hold that:

$$n_i(\tilde{t}_i) \le \tilde{t}_i \le \frac{(Cd_\star)^{\frac{1}{1-\beta}}Z_\star^{1/\beta}}{\Delta_i^{\frac{1}{1-\beta}}}.$$

And in the latter case, instead we have:

$$\mathsf{Reg}_i(\tilde{t}_i) = n_i(\tilde{t}_i)\Delta_i \le \frac{n_i(\tilde{t}_i)cW_\star\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}}{2\sqrt{\tilde{t}_i}}.$$

Also, in the latter case we can say:

$$n_i(\tilde{t}_i) \le \tilde{t}_i \le \frac{c^2W_\star^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{4\Delta_i^2}$$
$$\mathsf{Reg}_i(\tilde{t}_i) \le \frac{c^2W_\star^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{4\Delta_i},$$

so that when $\Delta_i \le \frac{W_\star}{2\sqrt{\tilde{t}_i}}$, we have

$$\mathsf{Reg}_i(\tilde{t}_i) \le \min\left[\frac{n_i(\tilde{t}_i)cW_\star\sqrt{\ln\frac{M\ln T}{\delta}}}{2\sqrt{\tilde{t}_i}}, \frac{c^2W_\star^2\ln\frac{M\ln T}{\delta}}{4\Delta_i}\right].$$

Next, let us suppose $\Delta_i > \frac{1}{2}b_\star(\tilde{t}_i)$. Then from Corollary 20, we have

$$\mathsf{Reg}_i(\tilde{t}_i) \le R_i(n_i(\tilde{t}_i)) + \mathbf{1}\left\{\star \in \mathcal{I}_{\tilde{t}_i+1}\right\}\frac{n_i(t)}{n_\star(\tilde{t}_i)}\mathsf{Reg}_\star(\tilde{t}_i) + n_i(\tilde{t}_i)(b_\star(\tilde{t}_i) - b_i(\tilde{t}_i)) + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}},$$

Note that if $\star \in \mathcal{I}_{\tilde{t}_i+1}$, we must have

$$v_i d_i n_i(\tilde{t}_i)^\beta \le v_\star d_\star n_\star(\tilde{t}_i)^\beta$$

that is

$$\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} \leq \left(\frac{v_\star d_\star}{v_i d_i}\right)^{1/\beta}.$$

Therefore, since $\star$ is well-specified:

$$\begin{aligned}
n_i(\tilde{t}_i)\Delta_i &\leq C d_i n_i(\tilde{t}_i)^\beta + \mathbf{1}\left\{\star \in \mathcal{I}_{\tilde{t}_i+1}\right\} \frac{n_i(t)}{n_\star(\tilde{t}_i)} C d_\star n_\star(\tilde{t}_i)^\beta + n_i(\tilde{t}_i)(b_\star(\tilde{t}_i) - b_i(\tilde{t}_i)) + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}} \\
&\leq C d_i n_i(\tilde{t}_i)^\beta + \mathbf{1}\left\{\star \in \mathcal{I}_{\tilde{t}_i+1}\right\} \left(\frac{n_i(t)}{n_\star(\tilde{t}_i)}\right)^{1-\beta} C d_\star n_i(\tilde{t}_i)^\beta + n_i(\tilde{t}_i)(b_\star(\tilde{t}_i) - b_i(\tilde{t}_i)) + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}} \\
&\leq C d_i n_i(\tilde{t}_i)^\beta + C d_\star \left(\frac{v_\star d_\star}{v_i d_i}\right)^{\frac{1-\beta}{\beta}} n_i(\tilde{t}_i)^\beta + n_i(\tilde{t}_i)(b_\star(\tilde{t}_i) - b_i(\tilde{t}_i)) + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}} \\
&\leq C d_i n_i(\tilde{t}_i)^\beta + C d_\star \left(\frac{v_\star d_\star}{v_i d_i}\right)^{\frac{1-\beta}{\beta}} n_i(\tilde{t}_i)^\beta + \frac{n_i(\tilde{t}_i)\Delta_i}{2} + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}
\end{aligned}$$

Thus, subtracting $\frac{n_i(\tilde{t}_i)\Delta_i}{2}$ from both sides:

$$\frac{n_i(\tilde{t}_i)\Delta_i}{2} \leq C d_i n_i(\tilde{t}_i)^\beta + C d_\star \left(\frac{v_\star d_\star}{v_i d_i}\right)^{\frac{1-\beta}{\beta}} n_i(\tilde{t}_i)^\beta + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}$$

Now, clearing the denominator, and using the fact that $x + y + z \leq 3\max(x, y, z)$:

$$n_i(\tilde{t}_i)\Delta_i \leq 6\max\left[C d_i n_i(\tilde{t}_i)^\beta,\ C d_\star \left(\frac{v_\star d_\star}{v_i d_i}\right)^{\frac{1-\beta}{\beta}} n_i(\tilde{t}_i)^\beta,\ 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}\right]$$

Now, working through the three cases yields:

$$n_i(\tilde{i}) \leq \max\left[\frac{(6Cd_i)^{\frac{1}{1-\beta}}}{\Delta_i^{\frac{1}{1-\beta}}},\ \frac{(6C)^{\frac{1}{1-\beta}}(v_\star d_\star)^{1/\beta}}{(v_i d_i)^{1/\beta}\Delta_i^{\frac{1}{1-\beta}}},\ \frac{12c^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{\Delta_i^2}\right]$$

Putting together this with the two cases at the start of the proof, we have:

$$\begin{aligned}
\mathsf{Reg}_i(\tilde{t}_i) &= n_i(\tilde{t}_i)\Delta_i \\
&\leq \max\Bigg\{\frac{(Cd_\star)^{\frac{1}{1-\beta}} Z_\star^{1/\beta}}{\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{(6Cd_i)^{\frac{1}{1-\beta}}}{\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{(6C)^{\frac{1}{1-\beta}}(v_\star d_\star)^{1/\beta}}{(v_i d_i)^{1/\beta}\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{12c^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{\Delta_i}, \\
&\qquad \min\left[\frac{n_i(\tilde{t}_i)cW_\star\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}}{2\sqrt{\tilde{t}_i}},\ \frac{c^2W_\star^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{4\Delta_i}\right]\Bigg\}.
\end{aligned}$$

□

Next, we need an observation about the particular form of the balancing parameters used in Corollary 23:

Lemma 28. *Suppose $R_i(n) = C d_i n^\beta$, and suppose that $d_1 \leq \cdots \leq d_M$. Let $W_i = W = \sqrt{M}$ and $Z_i = \frac{d_i}{d_1} i^\beta$ for all i. Set*

$$b_i(t) = \max\left[2CZ_i^{\frac{1-\beta}{\beta}} d_i t^{\beta-1},\ \frac{cW\sqrt{\ln\frac{M\ln t}{\delta}}}{\sqrt{t}}\right]$$

and set

$$v_i = \frac{Z_i^\beta}{d_i^{\beta+1}} .$$

Then if learners i and $\star$ satisfy $d_i \geq d_\star$, and $\tilde{t}_i = t_i - 1$ where t_i is the last time at which learner i was played, Algorithm 1 guarantees for all t:

$$\mathbf{1}\left\{\star \in \mathcal{I}_{\tilde{t}_i+1}\right\} \frac{n_i(\tilde{t}_i)}{n_\star(t)} \leq 1 .$$

Proof. Clearly the statement holds if $\star \notin \mathcal{I}_{\tilde{t}_i+1}$. Let us consider then the case $\star \in \mathcal{I}_{\tilde{t}_i+1}$. Then, since learner i was played at time t_i, we must have

$$\begin{aligned}
v_i R_i(n_i(\tilde{t}_i)) &\leq v_\star R_\star(n_\star(\tilde{t}_i)) \\
v_i C d_i n_i(\tilde{t}_i)^\beta &\leq v_\star C d_\star n_\star(\tilde{t}_i)^\beta \\
\frac{n_i(\tilde{t}_i)}{n_\star(\tilde{t}_i)} &\leq \left(\frac{v_\star d_\star}{v_i d_i}\right)^{1/\beta} \\
&= \left(\frac{(Z_\star/d_\star)^\beta}{(Z_i/d_i)^\beta}\right)^{1/\beta} \\
&= \frac{Z_\star d_i}{d_\star Z_i} \\
&= \frac{i_\star^\beta}{i^\beta} \\
&\leq 1
\end{aligned}$$

where the last line holds since $d_\star \leq d_i$. □

Next, we need a special-case version of Lemma 21 that takes advantage of Lemma 28:

Lemma 29. *Suppose $R_i(n) = C d_i n^\beta$, and suppose that $d_1 \leq \cdots \leq d_M$. Let $W_i = W = \sqrt{M}$ and $Z_i = \frac{d_i}{d_1} i^\beta$ for all i. Set*

$$b_i(t) = \max\left[2CZ_i^{\frac{1-\beta}{\beta}} d_i t^{\beta-1},\ \frac{cW\sqrt{\ln\frac{M\ln t}{\delta}}}{\sqrt{t}}\right]$$

and set

$$v_i = \frac{Z_i^\beta}{d_i^{\beta+1}} .$$

Let $d_i \geq d_\star$ and let $\tilde{t}_i = t_i - 1$ where t_i is the last time at which learner i is played. Then in event $\mathcal{E}$:

$$\mathsf{Reg}_i(\tilde{t}_i) \leq \frac{C d_1 \tilde{t}_i^\beta}{i^\beta} + 2R_\star(n_i(\tilde{t}_i)) + \mathsf{Reg}_\star(\tilde{t}_i) + \frac{n_i(\tilde{t}_i) c W_\star \sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}}{\sqrt{\tilde{t}_i}} + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}} .$$

Proof. From Corollary 20, we have

$$\mathsf{Reg}_i(\tilde{t}_i) \leq R_i(n_i(\tilde{t}_i)) + \mathbf{1}\left\{\star \in \mathcal{I}_{\tilde{t}_i+1}\right\}\frac{n_i(\tilde{t}_i)}{n_\star(t)}\mathsf{Reg}_\star(\tilde{t}_i) + n_i(\tilde{t}_i)(b_\star(\tilde{t}_i) - b_i(\tilde{t}_i)) + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}} ,$$

while from Lemma 28:

$$\leq R_i(n_i(\tilde{t}_i)) + \mathsf{Reg}_\star(\tilde{t}_i) + n_i(\tilde{t}_i)(b_\star(\tilde{t}_i) - b_i(\tilde{t}_i)) + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}\,.$$

Now, we consider two cases, either $d_i \leq 2d_\star$ or not.

Case $d_i \leq 2d_\star$: In this case, we have $R_i(n_i(\tilde{t}_i)) \leq 2R_\star(n_i(\tilde{t}_i))$. Also, since $d_i \geq d_\star$, by our expression for b_i we have $b_i(\tilde{t}_i) \geq b_\star(\tilde{t}_i)$. Then we have:

$$\mathsf{Reg}_i(\tilde{t}_i) \leq 2R_\star(n_i(\tilde{t}_i)) + \mathsf{Reg}_\star(\tilde{t}_i) + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}\,.$$

Case $d_i > 2d_\star$: In this case, we also must have $b_i(\tilde{t}_i) \geq 2b_\star(\tilde{t}_i) - 2\frac{cW_\star\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}}{\sqrt{\tilde{t}_i}}$ as well. Therefore:

$$\begin{aligned} b_\star(\tilde{t}_i) - b_i(\tilde{t}_i) &= b_\star(\tilde{t}_i) - \frac{1}{2}b_i(\tilde{t}_i) - \frac{1}{2}b_i(\tilde{t}_i) \\ &\leq \frac{cW_\star\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}}{\sqrt{\tilde{t}_i}} - \frac{1}{2}b_i(\tilde{t}_i)\,. \end{aligned}$$

Using this, we bound the regret:

$$\mathsf{Reg}_i(\tilde{t}_i) \leq R_i(n_i(\tilde{t}_i)) - \frac{n_i(\tilde{t}_i)b_i(\tilde{t}_i)}{2} + \mathsf{Reg}_\star(\tilde{t}_i) + \frac{n_i(\tilde{t}_i)cW_i\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}}{\sqrt{\tilde{t}_i}} + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}\,.$$

Now, we bound $R_i(n_i(\tilde{t}_i)) - \frac{n_i(\tilde{t}_i)b_i(\tilde{t}_i)}{2}$ just as in Lemma 21:

$$\begin{aligned} R_i(n_i(\tilde{t}_i)) - \frac{n_i(\tilde{t}_i)b_i(\tilde{t}_i)}{2} &= Cd_i n_i(\tilde{t}_i)^\beta - \frac{1}{2}n_i(\tilde{t}_i)b_i(\tilde{t}_i) \\ &\leq \sup_{x\geq 0}\left\{Cd_i x^\beta - \frac{x}{2}b_i(\tilde{t}_i)\right\} \\ &\leq \frac{2^{\frac{\beta}{1-\beta}}(\beta^\beta Cd_i)^{\frac{1}{1-\beta}}}{b_i(\tilde{t}_i)^{\frac{\beta}{1-\beta}}} \\ &\leq \left(\frac{2^\beta Cd_i}{b_i(\tilde{t}_i)^\beta}\right)^{\frac{1}{1-\beta}}\,. \end{aligned}$$

Now, using the fact that $b_i(\tilde{t}_i) \geq 2C\frac{d_i^{1/\beta}i^{1-\beta}}{d_1^{\frac{1-\beta}{\beta}}\tilde{t}_i^{1-\beta}}$ we can see that the last expression is upper bounded by

$$\frac{Cd_1\tilde{t}_i^\beta}{i^\beta}\,.$$

Combining all the bounds proves the Lemma □

Now, we are ready to provide an alternative *gap-dependent* bound on the regret of Algorithm 1. This result (Theorem 30) is the full version of the second part of Theorem 1 from the main text.

Now, again, the full result specifies all values for the parameters (which are *the same* as the specifications in the non-gap-dependent result Corollary 23), and provides a bound that holds on event $\mathcal{E}$, which itself occurs with probability at least $1-\delta$.

Theorem 30. *Suppose $R_i(n) = Cd_i n^\beta$, and suppose that $d_1 \le \cdots \le d_M$. Set $W_i = W = \sqrt{M}$ and $Z_i = \frac{d_i}{d_1} i^\beta$ for all i. Set*

$$b_i(t) = \max\left[2CZ_i^{\frac{1-\beta}{\beta}} d_i t^{\beta-1},\ \frac{cW_i\sqrt{\ln\frac{M\ln t}{\delta}}}{\sqrt{t}}\right]$$

and set

$$v_i = \frac{Z_i^\beta}{d_i^{\beta+1}}\,.$$

Let $\tilde{t}_i = t_i - 1$ where t_i is the last time at which learner i is played. Then on event $\mathcal{E}$, the total regret of Algorithm 1 is bounded for all rounds $T \in \mathbb{N}$ as follows:

$$\begin{aligned}\mathsf{Reg}(T) \le \sum_{i<\star} \max\left[\frac{C^{\frac{1}{1-\beta}} d_\star^{\frac{1}{\beta-\beta^2}} i_*}{d_1^{1/\beta}\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{(6Cd_i)^{\frac{1}{1-\beta}}}{\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{(6C)^{\frac{1}{1-\beta}} i_*}{i\,\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{12c^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{\Delta_i}\right] \\ + \mathsf{Reg}_\star(T) + 3MR_\star(T) + \frac{Cd_1T^\beta M^{1-\beta}}{1-\beta} + 6c\sqrt{MT\ln\frac{M\ln T}{\delta}}\,,\end{aligned}$$

where $\star$ (or i_*) is the index of the well-specified learner with minimum value of $d_\star$, and $\Delta_i = \frac{\mathsf{Reg}_i(\tilde{t}_i)}{n_i(\tilde{t}_i)}$.

Proof. As usual, we start by bounding the regret:

$$\mathsf{Reg}(T) = \mathsf{Reg}_\star(T) + \sum_{i\neq\star}\mathsf{Reg}_i(t_i) \le \mathsf{Reg}_\star(T)M + \sum_{i\neq\star}\mathsf{Reg}_i(\tilde{t}_i)\,.$$

Now, however, we split the last sum into two parts: when $i < \star$ and when $i > \star$. For $i < \star$, we define $\Delta_i = \frac{\mathsf{Reg}_i(\tilde{t}_i)}{n_i(\tilde{t}_i)}$. Then by Lemma 27, we have

$$\begin{aligned}\mathsf{Reg}_i(\tilde{t}_i) &\le \max\left[\frac{(Cd_\star)^{\frac{1}{1-\beta}} Z_\star^{1/\beta}}{\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{(6Cd_i)^{\frac{1}{1-\beta}}}{\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{(6C)^{\frac{1}{1-\beta}}(v_\star d_\star)^{1/\beta}}{(v_i d_i)^{1/\beta}\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{12c^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{\Delta_i},\ \frac{n_i(\tilde{t}_i)cW_\star\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}}{2\sqrt{\tilde{t}_i}}\right] \\ &\le \max\left[\frac{C^{\frac{1}{1-\beta}} d_\star^{\frac{1}{\beta-\beta^2}} i_*}{d_1^{1/\beta}\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{(6Cd_i)^{\frac{1}{1-\beta}}}{\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{(6C)^{\frac{1}{1-\beta}} i_*}{i\,\Delta_i^{\frac{\beta}{1-\beta}}},\ \frac{12c^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{\Delta_i}\right] + \frac{n_i(\tilde{t}_i)cW_\star\sqrt{\ln\frac{M\ln t}{\delta}}}{2\sqrt{\tilde{t}_i}}\,.\end{aligned}$$

Next, let's consider $i > \star$. In this case we have $d_i \ge d_\star$, so by Lemma 29:

$$\mathsf{Reg}_i(\tilde{t}_i) \le \frac{Cd_1\tilde{t}_i^\beta}{i^\beta} + 2R_\star(n_i(\tilde{t}_i)) + \mathsf{Reg}_\star(\tilde{t}_i) + \frac{n_i(\tilde{t}_i)cW_\star\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}}{\sqrt{\tilde{t}_i}} + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}\,.$$

Summing over these indices i and using $R_\star(n_i(\tilde{t}_i)) \le R_\star(T)$ and $R_\star(T) \ge \mathsf{Reg}_\star(\tilde{t}_i)$ yields:

$$\sum_{i>\star}\mathsf{Reg}_i(\tilde{t}_i) \le 3MR_\star(T) + \sum_{i>\star}\left[\frac{Cd_1\tilde{t}_i^\beta}{i^\beta} + \frac{n_i(\tilde{t}_i)cW_\star\sqrt{M\ln\frac{M\ln T}{\delta}}}{\sqrt{\tilde{t}_i}} + 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}}\right]\,.$$

Now, we bound the sums of the first two terms inside the summation using identical arguments as in Theorem 22:

$$\sum_{i>\star}\frac{Cd_1\tilde{t}_i^\beta}{i^\beta} \le CT^\beta\sum_{i\neq\star}\frac{d_i}{Z_i} \le \frac{Cd_iT^\beta M^{1-\beta}}{1-\beta}$$

$$\sum_{i>\star} 2c\sqrt{n_i(\tilde{t}_i)\ln\frac{M\ln\tilde{t}_i}{\delta}} \le 2c\sqrt{MT\ln\frac{M\ln T}{\delta}}\,.$$

So all together, we have

$$\sum_{i>\star}\mathsf{Reg}_i(\tilde{t}_i) \le 3MR_\star(T) + \frac{Cd_iT^\beta M^{1-\beta}}{1-\beta} + 6c\sqrt{MT\ln\frac{M\ln T}{\delta}} + \sum_{i>\star}\frac{n_i(\tilde{t}_i)cW_\star\sqrt{\ln\frac{M\ln\tilde{t}_i}{\delta}}}{\sqrt{\tilde{t}_i}}\,.$$

Combining the two cases $i<\star$ and $i>\star$ we have:

$$\begin{aligned}
\mathsf{Reg}(T) &\le \mathsf{Reg}_\star(T) + \sum_{i<\star}\max\left[\frac{C^{\frac{1}{1-\beta}}d_\star^{\frac{1}{\beta-\beta^2}}i_*}{d_1^{1/\beta}\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6Cd_i)^{\frac{1}{1-\beta}}}{\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6C)^{\frac{1}{1-\beta}}i_*}{i\,\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{12c^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{\Delta_i}\right] + \frac{n_i(\tilde{t}_i)cW_\star\sqrt{\ln\frac{M\ln T}{\delta}}}{2\sqrt{\tilde{t}_i}} \\
&\quad + 3MR_\star(T) + \frac{Cd_iT^\beta M^{1-\beta}}{1-\beta} + 6c\sqrt{MT\ln\frac{M\ln T}{\delta}} + \sum_{i>\star}\frac{n_i(\tilde{t}_i)cW_\star\sqrt{\ln\frac{M\ln T}{\delta}}}{\sqrt{\tilde{t}_i}} \\
&\le \mathsf{Reg}_\star(T) + \sum_{i<\star}\max\left[\frac{C^{\frac{1}{1-\beta}}d_\star^{\frac{1}{\beta-\beta^2}}i_*}{d_1^{1/\beta}\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6Cd_i)^{\frac{1}{1-\beta}}}{\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6C)^{\frac{1}{1-\beta}}i_*}{i\,\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{12c^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{\Delta_i}\right] \\
&\quad + 3MR_\star(T) + \frac{Cd_iT^\beta M^{1-\beta}}{1-\beta} + 6c\sqrt{MT\ln\frac{M\ln T}{\delta}} + \sum_{i\neq\star}\frac{n_i(\tilde{t}_i)cW_\star\sqrt{\ln\frac{M\ln T}{\delta}}}{\sqrt{\tilde{t}_i}}\,.
\end{aligned}$$

Let's focus on the last summation: $\sum_{i\neq\star}\frac{n_i(\tilde{t}_i)W_\star}{\sqrt{\tilde{t}_i}}$. This can again be bounded using the same argument as in the proof of Theorem 22:

$$\sum_{i=1}^{M}\frac{n_i(\tilde{t}_i)W_\star}{\sqrt{\tilde{t}_i}} \le 2W_\star\sqrt{T} \le 2\sqrt{MT}\,.$$

Thus combining all the sums yields the claimed bound. □

Theorem 30 does not provide any gains in the case that $\beta < 1/2$. For this scenario, we will instead show the following bound, which demonstrates that Algorithm 1 is able to successfully eliminate all misspecified algorithms to obtain regret $O(T^\beta)$, although the model selection guarantee degrades significantly. The key idea is to use the fact that the bound in Lemma 27 actually provides a bound on the total regret of any misspecified learner that has a significant gap in the regret, and then rely on the well-specified-ness of the remaining learners to bound their regret.

Theorem 31. *Suppose $R_i(n) = Cd_in^\beta$, and suppose that $d_1 \le \cdots \le d_M$. Set $W_i = W = \sqrt{M}$ and $Z_i = \frac{d_i}{d_1}i^\beta$ for all i. Set*

$$b_i(t) = \max\left[2CZ_i^{\frac{1-\beta}{\beta}}d_it^{\beta-1}, \frac{cW\sqrt{\ln\frac{M\ln t}{\delta}}}{\sqrt{t}}\right]$$

and set

$$v_i = \frac{Z_i^\beta}{d_i^{\beta+1}}\,.$$

Let $\tilde{t}_i = t_i - 1$ where t_i is the last time at which learner i is played. Then on event $\mathcal{E}$, the total regret of Algorithm 1 is bounded for all rounds $T \in \mathbb{N}$ as follows:

$$\mathsf{Reg}(T) \le \min\left[\sum_{i\notin\mathcal{B}}\mathsf{Reg}_i(T),\ CT^\beta\left(\sum_{i\notin\mathcal{B}}d_i^{\frac{1}{1-\beta}}\right)^{1-\beta},\ \mathsf{Reg}_\star(T) + 3MR_\star(T) + \frac{Cd_1T^\beta M^{1-\beta}}{1-\beta} + 6c\sqrt{MT\ln\frac{M\ln T}{\delta}}\right]$$

$$+ M + \sum_{i \in \mathcal{B}} \max \left[\frac{C^{\frac{1}{1-\beta}} d_\star^{\frac{1}{\beta-\beta^2}} i_*}{d_1^{1/\beta} \Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6Cd_i)^{\frac{1}{1-\beta}}}{\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6C)^{\frac{1}{1-\beta}} i_*}{i\, \Delta_i^{\frac{\beta}{1-\beta}}}, \frac{12c^2 \ln \frac{M \ln \tilde{t}_i}{\delta}}{\Delta_i}, \frac{c^2 W^2 \ln \frac{M \ln \tilde{t}_i}{\delta}}{4\Delta_i} \right]$$

where $\star$ (or i_*) is the index of the well-specified learner with minimum value of $d_\star$, and $\Delta_i = \frac{\mathsf{Reg}_i(\tilde{t}_i)}{n_i(\tilde{t}_i)}$.

Proof. We start splitting the regret into the sum over the misspecified learners, and the sum over the well-specified learners:

$$\begin{aligned}
\mathsf{Reg}(T) &\le \sum_{i \notin \mathcal{B}} \mathsf{Reg}_i(T) + \sum_{i \in \mathcal{B}} \mathsf{Reg}_i(T) \\
&\le \sum_{i \notin \mathcal{B}} \mathsf{Reg}_i(t_i) + M + \sum_{i \in \mathcal{B}} \mathsf{Reg}_i(\tilde{t}_i)
\end{aligned}$$

by well-specified-ness:

$$\le \min \left(\sum_{i \notin \mathcal{B}} \mathsf{Reg}_i(T), \sum_{i \notin \mathcal{B}} C d_i n_i(t_i)^\beta \right) + M + \sum_{i \in \mathcal{B}} \mathsf{Reg}_i(\tilde{t}_i)$$

from Hölder's inequality

$$\le \min \left(\sum_{i \notin \mathcal{B}} \mathsf{Reg}_i(T), C \left(\sum_{i \notin . \mathcal{B}} d_i^{\frac{1}{1-\beta}} \right)^{1-\beta} T^\beta \right) + M + \sum_{i \in \mathcal{B}} \mathsf{Reg}_i(\tilde{t}_i)\,.$$

Further, the sum $\sum_{i \notin \mathcal{B}} \mathsf{Reg}_i(T)$ can also be bounded using exactly the same argument as in 30 to yield:

$$\sum_{i \notin \mathcal{B}} \mathsf{Reg}_i(T) \le \mathsf{Reg}_\star(T) + 3M R_\star(T) + \frac{C d_1 T^\beta M^{1-\beta}}{1-\beta} + 6c\sqrt{MT \ln \frac{M \ln T}{\delta}}\,.$$

Now, let us bound the last sum using Lemma 27:

$$\sum_{i \in \mathcal{B}} \mathsf{Reg}_i(\tilde{t}_i) \le \sum_{i \in \mathcal{B}} \max \left[\frac{C^{\frac{1}{1-\beta}} d_\star^{\frac{1}{\beta-\beta^2}} i_*}{d_1^{1/\beta} \Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6Cd_i)^{\frac{1}{1-\beta}}}{\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6C)^{\frac{1}{1-\beta}} i_*}{i\, \Delta_i^{\frac{\beta}{1-\beta}}}, \frac{12c^2 \ln \frac{M \ln \tilde{t}_i}{\delta}}{\Delta_i}, \frac{c^2 W_\star^2 \ln \frac{M \ln T}{\delta}}{4\Delta_i} \right],$$

as claimed. □

C.6. Anytime Results with Logarithmic Factor Bounds

In this section, we briefly sketch how to extend our gap-dependent result Theorem 30 to the case that the regret bounds are not polynomial, but instead take the form $Cd_i \log(n)^\alpha n^\beta$. The argument is nearly identical, so we do not go into detail.

First, we have the following analog of Lemma 21:

Lemma 32. *Assume that Algorithm 1 is used with candidate regret bounds of the form $R_i(n) = Cd_i(1 + \log(n))^\alpha n^\beta$ for some $\alpha > 0$ and positive biases $b_i(t) > 0$. Then in event $\mathcal{E}$, the regret contribution of any subset $\mathcal{D} \subseteq [M]$ of learners after T total rounds can be bounded as*

$$\begin{aligned}
\sum_{i \in \mathcal{D}} \mathsf{Reg}_i(T) \le (1 + \log(T))^{\frac{\alpha}{1-\beta} + \frac{\alpha}{\beta}} \Bigg[&\sum_{i \in \mathcal{D}} \left(n_i(\tilde{t}_i) b_\star(\tilde{t}_i) + \left(\frac{2Cd_i}{b_i(\tilde{t}_i)^\beta} \right)^{\frac{1}{1-\beta}} \right) + 8c^2 \ln \frac{M \ln T}{\delta} \sum_{i \in \mathcal{D}} \frac{1}{b_i(\tilde{t}_i)} + |\mathcal{D}| \\
&+ C v_\star^{\frac{1-\beta}{\beta}} d_\star^{1/\beta} \left(\sum_{i \in \mathcal{D} :\, \star \in \mathcal{I}_{t_i}} \frac{1}{(v_i d_i)^{1/\beta}} \right)^{1-\beta} T^\beta \Bigg],
\end{aligned}$$

where $\tilde{t}_i = t_i - 1$ and t_i is the last round where learner i was played, $\star \in \mathcal{G}$ a well-specified learner, and c is a universal positive constant.

Proof. The proof is nearly identical to that of Lemma 21. The only difficulty is to keep track of the logarithmic terms. This is easily extracted by adding a $\log(T)^\alpha$ to every instance of d_i or $d_\star$ that appears in the numerator of the final bound, under-approximating the $\log(t)^\alpha$ terms that appear along with d_is in the denominator by 1. □

Similarly, we can produce a direct analog of our main regret bounds. We provide the analog of Theorem 30 below for concreteness:

Theorem 33. *Suppose $R_i(n) = Cd_i(1+\log(n)))^\alpha n^\beta$ for some $\alpha > 0$, and suppose that $d_1 \le \cdots \le d_M$. Set $W_i = W = \sqrt{M}$ and $Z_i = \frac{d_i}{d_1} i^\beta$ for all i. Set*

$$b_i(t) = \max\left[2CZ_i^{\frac{1-\beta}{\beta}} d_i t^{\beta-1}, \frac{cW_i\sqrt{\ln\frac{M\ln T}{\delta}}}{\sqrt{t}}\right]$$

and set

$$v_i = \frac{Z_i^\beta}{d_i^{\beta+1}}\,.$$

Let $\tilde{t}_i = t_i - 1$ where t_i is the last time at which learner i is played. Then on event $\mathcal{E}$, the total regret of Algorithm 1 is bounded for all rounds $T \in \mathbb{N}$ as follows:

$$\mathsf{Reg}(T) \le (1+\log(T))^{\frac{\alpha}{\beta-\beta^2}} \sum_{i<\star} \max\left[\frac{C^{\frac{1}{1-\beta}} d_\star^{\frac{1}{\beta-\beta^2}}(\star)}{d_1^{1/\beta}\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6Cd_i)^{\frac{1}{1-\beta}}}{\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{(6C)^{\frac{1}{1-\beta}}(\star)}{(i)\Delta_i^{\frac{\beta}{1-\beta}}}, \frac{12c^2\ln\frac{M\ln\tilde{t}_i}{\delta}}{\Delta_i}\right]$$
$$+ \mathsf{Reg}_\star(T) + 3MR_\star(T) + \frac{Cd_1\log(T)^\alpha T^\beta M^{1-\beta}}{1-\beta} + 6c\sqrt{MT\ln\frac{M\ln T}{\delta}}\,,$$

where $\star$ is the index of the well-specified learning with minimum value of $d_\star$, and $\Delta_i = \frac{\mathsf{Reg}_i(\tilde{t}_i)}{n_i(\tilde{t}_i)}$.

Proof. The proof is again identical to that of Theorem 30, but now we under-approximate each instance of $\log(t)^\alpha$ for any t in the denominators by 1, and over-approximate each instance of $\log(t)^\alpha$ in the numerators by $\log(T)^\alpha$. □

D. Example Applications

One important application of the method we presented in Section 3 is the setting of contextual linear bandits. Since this setting is often tackled using the OFUL Algorithm, we focus on instances of this algorithm as base learners but our results apply more generally. In the following, we first briefly review in Appendix D.1 the linear bandit setting and the OFUL Algorithm. We then present in Appendix D.2 a modification of OFUL which is either well-specified or suffers linear regret asymptotically. This version is particularly suited as base learner to achieve strong gap-dependent regret guarantee for Dynamic Balancing. Finally, we discuss two additional applications in Appendix D.3 and Appendix D.2.

D.1. Brief Review of Contextual Linear Bandits and the OFUL Algorithm

To keep consistency with previous sections, we shall assume here that contexts are drawn i.i.d. from some distribution over context space $\mathcal{X}$. Yet, the algorithmic solutions we present (specifically, the OFUL algorithm) actually work unchanged even in the more general fixed design or adaptive design scenarios.

In the contextual bandit setting, context x_t determines the set of actions $\mathcal{A}_t \subseteq \mathcal{A}$ that can be played at time t. When the bandit setting is *linear* the policies we consider are of the form $\pi_\theta(x_t) = \arg\max_{a\in\mathcal{A}_t}\langle a,\theta\rangle$, for some $\theta \in \mathbb{R}^d$, and the class of policies Π can then be thought of as a class of d-dimensional vectors $\Pi \subseteq \mathbb{R}^d$. Moreover, rewards are generated according to a noisy linear function, that is, $r_t = \langle a_t, \theta_\star\rangle + \xi_t$, where $\theta_\star \in \Pi$ is unknown, and ξ_t is a conditionally zero mean $\sigma-$subgaussian random variable. We denote the time-t optimal action as $a_t^\star = \mathrm{argmax}_{a\in\mathcal{A}_t}\langle a,\theta_\star\rangle$. The learner's objective is to control its pseudo-regret:

$$\mathsf{Reg}(T) = \sum_{t=1}^T \langle a_t^\star,\theta_\star\rangle - \langle a_t,\theta_\star\rangle\,.$$

Algorithm 3: OFUL (Abbasi-Yadkori et al., 2011)

Input: regularization parameter $\lambda > 0$, confidence scaling $\beta_1, \beta_2, \ldots$
for *round* $t = 1, 2, \ldots$ **do**
  Update regularized least-squares estimator $\hat{\theta}_t$ and covariance matrix Σ_t
  Receive context x_t/action space $\mathcal{A}_t$
  Play optimistic action:

$$a_t \in \operatorname*{argmax}_{a \in \mathcal{A}_t} \max_{\theta \in \mathcal{C}_t} \langle a, \theta \rangle = \operatorname*{argmax}_{a \in \mathcal{A}_t} \langle \hat{\theta}_t, a \rangle + \beta_t \|a\|_{\Sigma_t^{-1}}$$

  Receive reward $r_t = \langle a_t, \theta_\star \rangle + \xi_t$.

OFUL Algorithm. We now recall the relevant components of the OFUL algorithm (Abbasi-Yadkori et al., 2011) shown in Algorithm 3. Instances of this algorithm will play the role of base learners in subsequent sections. The OFUL algorithm proceeds by computing a regularized least-squares (RLS) estimator $\hat{\theta}_t$ of the true parameter $\theta_\star$ using the data collected so far:

$$\hat{\theta}_t := \Sigma_t^{-1} \left(\sum_{\ell=1}^{t-1} a_\ell \, r_\ell \right) \quad \text{where} \quad \Sigma_t = \lambda \mathbb{I} + \sum_{\ell=1}^{t-1} a_\ell a_\ell^\top \, . \tag{11}$$

Here, Σ_t is the regularized covariance matrix of the played actions up to the beginning of round t with regularization parameter λ, and $\mathbb{I}$ denotes the $d \times d$ identity matrix. Using $\hat{\theta}_t$ and Σ_t, OFUL proceeds by computing a confidence ellipsoid

$$\mathcal{C}_t := \{\theta \, : \, \|\theta - \hat{\theta}_t\|_{\Sigma_t} \leq \beta_t\} \tag{12}$$

that should contain the optimal parameter $\theta_\star$. We will discuss a choice of the (possibly data-dependent) scaling factor $\beta_t \in \mathbb{R}_+$ below that ensures that this happens in all rounds with high probability. Algorithm 3 now plays any action that achieves highest expected return in what we refer to as the optimistic model

$$\tilde{\theta}_t = \operatorname*{argmax}_{\theta \in \mathcal{C}_t} \max_{a \in \mathcal{A}_t} \langle a, \theta \rangle \, . \tag{13}$$

This choice of action is equivalent to picking $a_t \in \operatorname{argmax}_{a \in \mathcal{A}_t} \langle \hat{\theta}_t, a \rangle + \beta_t \|a\|_{\Sigma_t^{-1}}$. We define the event that the above-mentioned ellipsoidal confidence set $\mathcal{C}_t$ contains θ^* at all times $t \in \mathbb{N}$ as

$$\mathcal{E}' = \{\theta_* \in \mathcal{C}_t, \quad \forall t \in \mathbb{N}\} \, . \tag{14}$$

In this event $\mathcal{E}'$, the optimistic model $\tilde{\theta}_t$ indeed gives rise to an optimistic estimate of the expected reward in each round

$$\langle a_t, \tilde{\theta}_t \rangle \geq \max_{a \in \mathcal{A}_t} \langle a, \theta_\star \rangle = \langle a_t^\star, \theta_\star \rangle \, . \tag{15}$$

Abbasi-Yadkori et al. (2011) show that the following choice for β_t is sufficient to make $\mathcal{E}'$ happen with high probability:

Lemma 34 (Theorem 2 in (Abbasi-Yadkori et al., 2011))**.** *For any $\delta \in (0, 1)$, let the confidence scaling be*

$$\beta_t := \sqrt{2\sigma^2 \ln \left(\frac{\det(\Sigma_t)^{1/2} \det(\lambda I)^{-1/2}}{\delta} \right)} + \sqrt{\lambda} S \leq \sqrt{\sigma^2 d \ln \left(\frac{1 + tL^2/\lambda}{\delta} \right)} + \sqrt{\lambda} S \tag{16}$$

where S is a known bound on the parameter norm $\max_{\theta \in \Pi} \|\theta\|_2$ and L is a known bound on the action norm in all rounds, i.e., $\max_{a \in \mathcal{A}_t} \|a\|_2 \leq L$ for all t. Then $\theta_\star$ is contained in the confidence ellipsoid with high probability, i.e., $\mathbb{P}(\mathcal{E}') \geq 1 - \delta$.

In event $\mathcal{E}'$, one can show that the regret of Algorithm 3 is bounded for all $t \in [T]$ as

$$\mathsf{Reg}(t) \leq \sum_{t=1}^{T} 2\beta_t \|a_t\|_{\Sigma_t^{-1}} \leq 2\beta_{\max} \sqrt{dt \left(1 + \frac{L^2}{\lambda}\right) \ln \frac{d\lambda + tL}{d\lambda}} \, ,$$

where $\beta_{\max} = \max_{k \in [t]} \beta_k$. We reproduce a slightly more general version of the standard proof for this regret bound in the following in Lemma 35. The right side of the above inequality will play the role of our presumed regret bound $R(n_i(t))$ when OFUL is used as a base learner.

Lemma 35 (Regret Bound for OFUL)**.** *Assume* OFUL *(Algorithm 3) uses regularization parameter $\lambda > 0$ and chooses the each action as*

$$a_t \in \operatorname*{argmax}_{a \in \mathcal{A}_t} \langle \widehat{\theta}_t, a \rangle + \beta_t \|a\|_{V_t^{-1}},$$

where $\widehat{\theta}_t$ is a parameter estimate, $\beta_t \in \mathbb{R}$ is a confidence width and $V_t \succcurlyeq \lambda I + \sum_{l=1}^{t-1} a_l a_l^\top$ is a covariance matrix. In the event that the true parameter $\theta_\star$ was contained at all times in the confidence ellipsoid, that is, $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$ for all $t \in [T]$, the (pseudo-)regret is bounded as

$$\mathsf{Reg}(t) \leq 2\beta_{\max} \sqrt{dt \left(1 + \frac{L^2}{\lambda}\right) \ln \frac{d\lambda + tL^2}{d\lambda}},$$

where $\beta_{\max} = \max_{t \in [T]} \beta_t$ is the largest confidence width during all rounds and $L = \max_{a \in \bigcup_t \mathcal{A}_t} \|a\|_2$ be a bound on the action norms.

Remark 2. *This regret bound for OFUL holds for any, possibly random, sequence of confidence widths as long as the true parameter is contained in the confidence ellipsoid. It does not assume any specific form or monotonicity or $\beta_t \geq 1$. It also does not prescribe that the covariance matrix exactly matches $\lambda I + \sum_{l=1}^{t-1} a_l a_l^\top$. This makes this regret bounds applicable to the case where $\hat{\theta}_t$ includes additional observations besides the ones from previous rounds played by the algorithm.*

Proof. The immediate regret at time t (defined as the difference of the expected reward of the optimal action choice $a_t^\star \in \operatorname{argmax}_{a \mathcal{A}_t} \langle \theta_\star, a \rangle$ and the action a_t taken by the algorithm) is bounded as

$$\begin{aligned}
\langle \theta_\star, a_t^\star - a_t \rangle &\overset{(i)}{\leq} \langle \widehat{\theta}_t, a_t^\star \rangle + \beta_t \|a_t^\star\|_{V_t^{-1}} - \langle \theta_\star, a_t \rangle \\
&\overset{(ii)}{\leq} \langle \widehat{\theta}_t, a_t \rangle + \beta_t \|a_t\|_{V_t^{-1}} - \langle \theta_\star, a_t \rangle \\
&\overset{(iii)}{\leq} 2\beta_t \|a_t\|_{V_t^{-1}} \overset{(iv)}{\leq} 2\beta_t \|a_t\|_{\Sigma_t^{-1}},
\end{aligned}$$

where $\Sigma_t = \lambda I + \sum_{l=1}^{t-1} a_l a_l^\top$. Step (i) follows from $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$, step (ii) from the algorithm's action choice and step (iii) again from the confidence ellipsoid $\|\theta_\star - \hat{\theta}_t\|_{V_t} \leq \beta_t$. Finally, step (iv) follows from the assumption that $V_t \succcurlyeq \lambda I + \sum_{l=1}^{t-1} a_l a_l^\top = \Sigma_t$.

Since L is a bound of the action norm and $\Sigma_t \succcurlyeq \lambda I$, we have $\|a_t\|_{\Sigma_t^{-1}} = \|\Sigma_t^{-1/2} a_t\|_2 \leq \frac{L}{\sqrt{\lambda}}$. Thus, we can bound the regret as

$$\begin{aligned}
\mathsf{Reg}(T) &\leq 2 \sum_{t=1}^{T} \beta_t \|a_t\|_{\Sigma_t^{-1}} \\
&\leq 2 \sqrt{\sum_{t=1}^{T} \beta_t^2} \sqrt{\sum_{t=1}^{T} \|a_t\|_{\Sigma_t^{-1}}^2} && \text{(Cauchy-Schwarz)} \\
&\leq 2\beta_{\max} \sqrt{T \sum_{i=1}^{T} \frac{L^2}{\lambda} \wedge \|a_t\|_{\Sigma_t^{-1}}^2} \\
&\leq 2\beta_{\max} \sqrt{T \left(1 + \frac{L^2}{\lambda}\right) \ln \frac{\det \Sigma_{T+1}}{\det \Sigma_1}} && \text{(Lemma 7)} \\
&\leq 2\beta_{\max} \sqrt{dT \left(1 + \frac{L^2}{\lambda}\right) \ln \frac{d\lambda + TL^2}{d\lambda}}.
\end{aligned}$$

□

D.2. OFUL with misspecification test

In this section, we describe a variant of the OFUL algorithm for stochastic linear bandits that that incorporates a misspecification test. This will cause the regret to satisfy two possibilities. Either the regret bound will be essentially the same as the ordinary OFUL algorithm's regret bound, or it will be asymptotically linear in T. Moreover, if the algorithm is well-specified in the sense that the rewards are indeed a linear function of the contexts, then the regret is guaranteed to match the OFUL regret.

We again compute the regularized least-squares estimator $\hat{\theta}_t$ of the true parameter $\theta_\star$ using the data collected so far:

$$\hat{\theta}_t := \Sigma_t^{-1} \left(\sum_{\ell=1}^{t-1} a_\ell \, r_\ell \right) \quad \text{where} \quad \Sigma_t = \lambda \mathbb{I} + \sum_{\ell=1}^{t-1} a_\ell a_\ell^\top \, . \tag{17}$$

However, in a slight deviation from the standard OFUL algorithm, we define the confidence sets:

$$\mathcal{C}'_t := \bigcap_{i=1}^{t} \mathcal{C}_i = \{\theta \, : \, \|\theta - \hat{\theta}_i\|_{\Sigma_i} \le \beta_i \; \forall \, i \le t\} \tag{18}$$

along with the corresponding event:

$$\mathcal{E}' = \{\theta_* \in \mathcal{C}'_t \;\; \forall t \in \mathbb{N}\} \, , \tag{19}$$

where β_t is defined as in Lemma 34. Our algorithm again computes

$$\tilde{\theta}_t = \operatorname*{argmax}_{\theta \in \mathcal{C}'_t} \max_{a \in \mathcal{A}_t} \langle a, \theta \rangle \, . \tag{20}$$

and takes the action $a_t = \operatorname{argmax}_{a \in \mathcal{A}_t} \langle a, \tilde{\theta}_t \rangle$. We add one small twist: we specify some values of $\gamma_1, \gamma_2, \ldots$ and if at any time t it does not hold that

$$\left| \sum_{i=1}^{t} r_i - \operatorname{clip}\left(\langle \hat{\theta}_i, a_i \rangle \right) \right| \le \gamma_t, \tag{21}$$

then we conclude that the rewards are not linear in the contexts and terminate the algorithm (or play actions uniformly at random which has linear regret). Here, $\operatorname{clip}(x) = \min\{1, \max\{0, x\}\}$ clips the value of x to the range $[0, 1]$.

We are now ready to state the main guarantee for this modified OFUL algorithm:

Theorem 36. *Suppose that the action norms satisfy* $\max_{a \in \mathcal{A}_t} \|a\| \le L$ *for all* t *and set*

$$\gamma_t = c\sqrt{t \, \ln \frac{\ln t}{\delta}} + \beta_t \sqrt{dt \left(1 + \frac{L^2}{\lambda}\right) \ln\left(\frac{d\lambda + tL^2}{d\lambda}\right)}$$

where c *is the same constant as in the definition of* $\widetilde{\mathcal{E}}$*. Let the modified version of* OFUL *described above be run with* γ_t *on any sequential decision process with stochastic contexts that may or may not be a linear bandit. Then the following statements hold with probability at least* $1 - 2\delta$*:*

- *Whenever the regret violates the bound* $\mathsf{Reg}(\tau) \le 4\gamma_\tau \approx \tilde{O}(\beta_\tau \sqrt{d\tau})$ *for some* $\tau \in \mathbb{N}$*, the algorithm either stops or suffers linear regret for large enough* T*. More precisely, it has regret* $\mathsf{Reg}(T) \ge \frac{\gamma_\tau}{2\tau} T$ *for all* T *large enough (in this case* $T \gtrapprox 36\tau$ *is sufficient).*
- *If rewards are generated by a linear bandit instance and the confidence ellipsoids contain the true parameter vector, then the misspecification test never triggers and* $\mathsf{Reg}(T) \le 4\gamma_T$*.*

Proof. We first show the second part. As long as the misspecification test does not trigger, this OFUL version behaves identical to the standard OFUL algorithm except that confidence sets can never increase. We can follow the standard analysis outlined in Lemma 35 to show that, despite this small change, OFUL still satisfies the the regret bound in Lemma 35 which is upper-bounded by γ_T. It therefore only remains to show that regret bound test never triggers when rewards are indeed

generated by a linear bandit: Here, $r_t = \langle \theta_\star, a_t \rangle + \xi_t$ for all t for some $\theta_\star \in \mathbb{R}^d$ with $\|\theta_\star\| \le S$ and sub-Gaussian noise ξ_t. Now, we can write the quantity in the LHS of the misspecification in Equation 21 as

$$\sum_{i=1}^{t} r_i - \text{clip}\left(\langle \hat{\theta}_i, a_i \rangle\right) = \sum_{i=1}^{t} \left(r_i - \langle \theta_\star, a_i \rangle\right) + \sum_{i=1}^{t} \left(\langle \theta_\star, a_i \rangle - \text{clip}\left(\langle \hat{\theta}_i, a_i \rangle\right)\right) .$$

Using standard concentration arguments, the magnitude of the first term is at most $c\sqrt{t \ln \frac{\ln t}{\delta}}$ with probability $1-\delta$ for all t. The magnitude of the second term can be bounded on event $\mathcal{E}'$ (where the confidence ellipsoids contain $\theta_\star$) as

$$\begin{aligned}\left|\sum_{i=1}^{t} \langle \theta_*, a_i \rangle - \text{clip}\left(\langle \hat{\theta}_i, a_i \rangle\right)\right| &\le \left|\sum_{i=1}^{t} \langle \theta_* - \hat{\theta}_i, a_i \rangle\right| \le \sum_{i=1}^{t} \|\theta_* - \hat{\theta}_i\|_{\Sigma_i} \|a_i\|_{\Sigma_i^{-1}} \le \beta_t \sum_{i=1}^{t} \|a_i\|_{\Sigma_i^{-1}} \\ &\le \beta_t \sqrt{dt\left(1 + \frac{L^2}{\lambda}\right) \ln\left(\frac{d\lambda + tL^2}{d\lambda}\right)} ,\end{aligned}$$

where the first step uses the fact that expected rewards are in $[0,1]$ and the last step follows the same logic as the proof of Lemma 35. Thus from our definition of γ_t, the misspecification test will never trigger.

Let us now move on to show the first part of the theorem. We here look at

$$\sum_{t=1}^{T} \mathbb{E}\left[\mu_t^\star - \text{clip}\left(\langle \tilde{\theta}_t, a_t \rangle\right) \mid \mathcal{F}_{t-1}\right],$$

the total amount that the algorithm underestimates the optimal reward per round in expectation over the current context x_t. Using this quantity, we define the stopping time

$$\tau = \min\left\{T \in \mathbb{N}\colon \sum_{t=1}^{T} \mathbb{E}\left[\mu_t^\star - \text{clip}\left(\langle \tilde{\theta}_t, a_t \rangle\right) \mid \mathcal{F}_{t-1}\right] > \gamma_T \text{ or } \left|\sum_{t=1}^{T} r_t - \text{clip}\left(\langle \hat{\theta}_t, a_t \rangle\right)\right| > \gamma_T\right\}$$

as the first time the misspecification test triggers or the total underestimation amount surpasses γ_T. Here $\mathcal{F}_{t-1}$ is the sigma-field of everything up to round $t-1$ (before the context x_t is revealed).

Regret in rounds $T < \tau$: We now show that up to time τ, the regret of the algorithm is well-behaved. To that end, we decompose the regret for any round $T \le \tau$ as

$$\begin{aligned}\mathsf{Reg}(T) &= \sum_{t=1}^{T} \left(\mu_t^\star - \mathbb{E}[r_t | x_t, a_t]\right) \\ &= \underbrace{\sum_{t=1}^{T} \left(\mu_t^\star - \text{clip}\left(\langle \tilde{\theta}_t, a_t \rangle\right)\right)}_{(A)} + \underbrace{\sum_{t=1}^{T} [\text{clip}\left(\langle \tilde{\theta}_t, a_t \rangle\right) - \text{clip}\left(\langle \hat{\theta}_t, a_t \rangle\right)]}_{(B)} \qquad (22) \\ &\quad + \underbrace{\sum_{t=1}^{T} \left(\text{clip}\left(\langle \hat{\theta}_t, a_t \rangle\right) - r_t\right)}_{(C)} + \underbrace{\sum_{t=1}^{T} \left(r_t - \mathbb{E}[r_t | \mathcal{F}_{t-1}, a_t]\right)}_{(D)}, \qquad (23)\end{aligned}$$

and bound each term individually. Starting with (A), we can apply the definition of τ and a concentration argument to get (with probability at least $1-\delta$ for all $T \le \tau$)

$$(A) = c\sqrt{T \ln \frac{\ln T}{\delta}} + \sum_{t=1}^{T} \mathbb{E}\left[\mu_t^\star - \text{clip}\left(\langle \tilde{\theta}_t, a_t \rangle\right) \mid \mathcal{F}_{t-1}\right] \le c\sqrt{T \ln \frac{\ln T}{\delta}} + \gamma_T .$$

Moving on to the second term, we can apply the standard machinery of the OFUL analysis (see Lemma 35) to get

$$(B) \leq \sum_{t=1}^{T} \langle \tilde{\theta}_t - \hat{\theta}_t, a_t \rangle \leq \sum_{t=1}^{T} \|\tilde{\theta}_t - \hat{\theta}_t\|_{\Sigma_t} \|a_t\|_{\Sigma_t^{-1}} \leq \sum_{t=1}^{T} \beta_t \|a_t\|_{\Sigma_t^{-1}} \leq \beta_T \sqrt{T \sum_{t=1}^{T} \|a_t\|^2_{\Sigma_t^{-1}}}$$
$$\leq \beta_T \sqrt{dT \left(1 + \frac{L^2}{\lambda}\right) \ln\left(\frac{d\lambda + TL^2}{d\lambda}\right)}.$$

Note that this bound follows directly from the construction of a_t and $\tilde{\theta}_t$ in OFUL and does not require the ellipsoid confidence sets to be valid. Again, by the definition of τ, term $(C) \leq \gamma_T$. Finally, term (D) can be bounded using standard concentration arguments as $(D) \leq c\sqrt{T \ln \frac{\ln T}{\delta}}$ for all $T \in \mathbb{N}$ with probability $1 - \delta$. Combining all terms, we have shown that with probability at least $1 - 2\delta$, the regret for all rounds $T \leq \tau$ is bounded as

$$\mathsf{Reg}(T) \leq 2\gamma_T + 2c\sqrt{T \ln \frac{\ln T}{\delta}} + \beta_T \sqrt{dT \left(1 + \frac{L^2}{\lambda}\right) \ln\left(\frac{d\lambda + TL^2}{d\lambda}\right)} \leq 4\gamma_T \,.$$

Regret in rounds $T \geq \tau$ where misspecification test has not triggered yet: Consider now rounds $T \geq \tau$ after τ but where the misspecification test has not triggered yet. In this case, it must hold that $\sum_{t=1}^{\tau} \mathbb{E}\left[\mu_t^\star - \text{clip}\left(\langle \tilde{\theta}_t, a_t \rangle\right) \mid \mathcal{F}_{t-1}\right] > \gamma_\tau$ and thus, there must be a round $t \leq \tau$ which satisfies $\mathbb{E}\left[\mu_t^\star - \text{clip}\left(\langle \tilde{\theta}_t, a_t \rangle\right) \mid \mathcal{F}_{t-1}\right] > \frac{\gamma_\tau}{\tau}$. We now show that this has to also hold for all future rounds. Let $T \geq t$. Then

$$\begin{aligned}
\mathbb{E}\left[\mu_T^\star - \text{clip}\left(\langle \tilde{\theta}_T, a_T \rangle\right) \mid \mathcal{F}_{T-1}\right] &= \mathbb{E}\left[\mu_T^\star \mid \mathcal{F}_{T-1}\right] - \mathbb{E}\left[\text{clip}\left(\max_{a \in \mathcal{A}_T} \sup_{\theta \in \mathcal{C}'_T} \langle \theta, a \rangle\right) \mid \mathcal{F}_{T-1}\right] \\
&\geq \mathbb{E}\left[\mu_T^\star \mid \mathcal{F}_{T-1}\right] - \mathbb{E}\left[\text{clip}\left(\max_{a \in \mathcal{A}_T} \sup_{\theta \in \mathcal{C}'_t} \langle \theta, a \rangle\right) \mid \mathcal{F}_{T-1}\right] \\
&= \mathbb{E}\left[\mu_t^\star \mid \mathcal{F}_{t-1}\right] - \mathbb{E}\left[\text{clip}\left(\max_{a \in \mathcal{A}_t} \sup_{\theta \in \mathcal{C}'_t} \langle \theta, a \rangle\right) \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\mu_t^\star - \text{clip}\left(\langle \tilde{\theta}_t, a_t \rangle\right) \mid \mathcal{F}_{t-1}\right] \geq \frac{\gamma_\tau}{\tau}
\end{aligned}$$

where we first used that the confidence sets satisfy $\mathcal{C}'_1 \supseteq \mathcal{C}'_2 \supseteq \mathcal{C}'_3 \ldots$ and that the distribution of the context (here $\mathcal{A}_t$) is the same across all rounds t. We now return to the regret decomposition in Equation 23 but lower-bound each term instead of upper-bounding them. Starting with term (A), we have

$$(A) \geq \sum_{t=1}^{T} \mathbb{E}\left[\mu_t^\star - \text{clip}\left(\langle \tilde{\theta}_t, a_t \rangle\right) \mid \mathcal{F}_{t-1}\right] - c\sqrt{T \ln \frac{\ln T}{\delta}} \geq \gamma_\tau + (T - \tau)\frac{\gamma_\tau}{\tau} - c\sqrt{T \ln \frac{\ln T}{\delta}} = \frac{T\gamma_\tau}{\tau} - c\sqrt{T \ln \frac{\ln T}{\delta}} \,,$$

which holds for all T with probability at least $1 - \delta$, as above. For term (B), we have

$$(B) = \sum_{t=1}^{T} [\text{clip}\left(\langle \tilde{\theta}_t, a_t \rangle\right) - \text{clip}\left(\langle \hat{\theta}_t, a_t \rangle\right)] \geq 0$$

since $\langle \tilde{\theta}_t, a_t \rangle \geq \langle \hat{\theta}_t, a_t \rangle$ by construction of $\tilde{\theta}_t$. We also have $(C) \geq -\gamma_T$ because the algorithm has not stopped yet. Finally, term (D) is again bounded by $(D) \geq -c\sqrt{T \ln \frac{\ln T}{\delta}}$ by standard concentration (for all T with probability at least $1 - \delta$). Combining the individual terms yields

$$\mathsf{Reg}(T) \geq \frac{T\gamma_\tau}{\tau} - 2c\sqrt{T \ln \frac{\ln T}{\delta}} - \gamma_T \geq \frac{T\gamma_\tau}{\tau} - 3\gamma_T \,.$$

Therefore, for T large enough so that $\frac{\gamma_\tau}{6\tau} \geq \frac{\gamma_T}{T}$, the regret is lower-bounded as $\mathsf{Reg}(T) \geq \frac{\gamma_\tau}{2\tau} T$.

□

D.3. Linear Markov Decision Processes with Nested Model Classes

We can instantiate the regret bound in Corollary 2 to the episodic linear MDP setting of Jin et al. (2020), again with nested feature classes of doubling dimension, as in Section 5.1. Here, each round t of Algorithm 1 corresponds to one episode of H time steps in the MDP, and contexts x_t are the initial state of the episode in the MDP. Jin et al. (2020) prove that their LSVI-UCB algorithm achieves regret $O(H^2\sqrt{d^3K}\ln(dK/\delta))$ after K episodes when used with a realizable function class of dimension d. We deploy $M = O(\ln d_{\max})$ instances of LSVI-UCB as base learners with presumed regret bounds

$$R_i(n) = H^2\sqrt{d_i^3 n}\ln\frac{d_{\max}T}{\delta}.$$

Since the total reward per episode (= round) is in $[0, H]$ instead of $[0, 1]$ in this setting, we scale the biases as well as the constant c in Algorithm 1 by H. By Corollary 2 the total regret of Algorithm 1 after T episodes is bounded as

$$\mathsf{Reg}(T) = O\left(\left(d_\star^3\sqrt{\ln d_\star} + \sqrt{d_\star^3} + \sqrt{\ln d_{\max}}\right)H^2\sqrt{T}\ln\frac{d_{\max}T}{\delta}\right)$$

with probability $1 - M\delta$. Similar to the contextual bandit setting above, we can achieve a tighter bound if all misspecified learners suffer linear regret $\mathsf{Reg}_i(t) \geq \Delta n_i(t)$ for some $\Delta > 0$. Then applying Corollary 2 yields

$$\mathsf{Reg}(T) = O\left(H^2\sqrt{d_\star^3 T}\ln(d_{\max})\ln(d_{\max}T/\delta) + \frac{H^4 d_\star^6}{\Delta}\ln(d_{\max}T/\delta)^2\ln(d_\star)\right)$$

which, up to log factors and gap-dependent lower order terms, again coincides with the regret bound of the best base learner in hindsight.

D.4. Linear Bandits and MDPs with Unknown Approximation Error

Zanette et al. (2020) presents an algorithm for learning a good policy in episodic MDPs where the value functions are all close to a linear feature space of dimension d. Their algorithm admits a high-probability regret bound of order[7] $\tilde{O}(Hd\sqrt{T} + H\sqrt{d}\epsilon T)$ for all T when a bound ϵ on the inherent Bellman error is known a-priori. For details of the setting and the exact definition of inherent Bellman error see Zanette et al. (2020). Unfortunately, in most practical applications, one does not know ϵ ahead of time and picking a conservative value (large ϵ) makes the algorithm over-explore and suffer large regret.

We can address this limitation by applying the simplified balancing algorithm in Algorithm 2 with several instances of their algorithm as base-learners, each associated with a certain value of the inherent Bellman error $\epsilon_i = \frac{2^{1-i}}{\sqrt{d}}$ and the putative regret bound $R_i(n) = CHd\sqrt{n} + CH\sqrt{d}\epsilon_i n$ for an appropriate value C that depends at most logarithmically on d, T or H. It is sufficient to use $M = \lceil 1 + \frac{1}{2}\log_2(T/d^2)\rceil$ base learners since the putative regret bound of learner 1 (with $\epsilon_1 = 1/\sqrt{d}$ and $R_1(n) \geq Hn$) always holds, while the putative regret bound of learner M is at most $R_M(T) \leq 2CHd\sqrt{T}$, which is a constant factor worse than the regret when $\epsilon = 0$.

By the bound given in Appendix B.3, the total regret of Algorithm 2 with these base learners is

$$\begin{aligned}\mathsf{Reg}(T) &= O\left(\epsilon_\star CH\sqrt{d}T\left(M + \frac{1}{CHd}\sqrt{\ln\frac{M\ln T}{\delta}}\right) + MCHd\sqrt{T} + \sqrt{BT\ln\frac{M\ln T}{\delta}}\right)\\ &= \tilde{O}\left(MHd\sqrt{T} + MH\sqrt{d}\epsilon_\star T\right)\end{aligned}$$

with probability $1 - M\delta$. Hence, up to at most logarithmic factors (M is logarithmic here), our model-selection framework can recover the best regret bound without requiring knowing the inherent Bellman error ahead of time. Notice also that the special case $H = 1$ recovers the standard linear bandit setting and the algorithm by Zanette et al. (2020) reduces to OFUL with a confidence ellipsoid that accounts for ϵ_i. In this bandit case $\epsilon_\star$ is the absolute approximation error of expected rewards.

Recently, Foster et al. (2020a) have shown that an adaptation to unknown approximation errors $\epsilon_\star$ is possible in contextual bandits, but their model-selection approach requires base learners that work with importance weights, and whose importance-weighted regret admits a favorable dependency on ϵ_i. Here we have shown that a similar result (up to logarithmic factors)

[7] The $\tilde{O}$ notation is similar to the O-notation but hides poly-logarithmic dependencies.

Algorithm 4: EpochBalancing

Input: set of learners $\mathcal{I}$
for *round* $t = 1, 2, \ldots$ **do**
  Receive context x_t
  foreach *learner* $i \in \mathcal{I}$ **do**
    Ask learner i for a lower bound $B_{t,i}$ on the value of its proposed action
  Sample $i_t \sim p \propto \frac{1}{z_i}$ for $i \in \mathcal{I}$ (see Equation (24))
  Play learner i_t and receive reward r_t
  Update base learner i_t with r_t
  Test for misspecification by checking $\sum_{i \in \mathcal{I}} [U_i(t) + R_i(n_i(t))] + c\sqrt{t \ln \frac{\ln(t)}{\delta}} < \max_{i \in \mathcal{I}} \sum_{k=1}^{t} B_{k,i}$
  if *above condition is triggered* **then**
    Return ; // At least one learner must be misspecified

can be achieved with standard optimistic base learners such as OFUL. Our result also matches the regret guarantee by Pacchiano et al. (2020b) but does not require their smoothing procedure for base learners. Importantly, our result proves that an adaptation to unknown approximation errors $\epsilon_\star$ is also possible without any modification to base learners in the MDP setting where base-learners that achieve the importance-weighted regret guarantee required by Foster et al. (2020a) are (still) unavailable. Note also that our framework is not specific to instances of the algorithm by Zanette et al. (2020) as base learners. Our model selection algorithm can, for example, also be used with approximate versions of LSVI-UCB by Jin et al. (2020) and achieve similar regret guarantees in their setting and for their notion of approximation error.

E. Extension to Adversarial Contexts

In this section, we show that the dynamic balancing principle can also be used for model selection when the contexts x_t are generated in an adversarial manner. For the sake of concreteness, we present our extension for the setting from Section 5.1, but our techniques for adversarial contexts can be easily adapted to all other bandit applications discussed in Section 5 and likely to episodic MDP settings with adversarial start states as well. We consider the setting from Section 5.1. Since the entries of the true parameter $\theta_\star$ are 0 for all dimensions above $d_{i_\star}$, where $i_\star \in [M]$ is an unknown index, all learners $i_\star, i_\star + 1, \ldots M$ are well-specified with high probability. We focus our analysis on the event $\mathcal{E}'$ where this is the case. Unlike Section 5.1 where contexts are assumed to be drawn i.i.d., we here consider the setting where contexts x_t (corresponding to the action set $\mathcal{A}_t$ at round t) are generated adversarially. Since each base learner operates only in a lower-dimensional subspace, we allow the bound on the action norm L_i, the bound on the parameter norm S_i, and the range of expected return $R_i^{\max}$ to vary per base learner i (potentially depending on the number of dimension d_i). Yet, for the sake of simplicity, we assume that all learners use regularization parameter $\lambda = 1$.

Algorithm 1, which assumes stochastic contexts, compares upper- and lower-confidence bounds on the optimal reward value $\mu^\star$ obtained from learners that were executed on two disjoint subsets of rounds to determine misspecification. This strategy does not work with adversarial contexts since the optimal policy that an algorithm could have achieved depends on the contexts in the rounds that it was played. One algorithm may only have seen "bad" contexts with low $\mu_t^\star$, while another may only have encountered favorable contexts with high $\mu_t^\star$. A direct comparison is therefore meaningless.

In order to be able to handle adversarial contexts and address this challenge, we modify our dynamic balancing algorithm in two ways: (1) we randomize the learner's choice for regret balancing, and (2) we change the misspecfication test to compare upper and lower confidence bounds on the optimal policy value of *all* rounds played to far. The resulting algorithm is presented in Algorithm 5. The algorithm operates in epochs, where the subroutine in Algorithm 4 is executed. We start by discussing the regret balancing subroutine in the next section before presenting the main algorithm and its regret guarantee afterwards.

E.1. The Epoch Balancing Subroutine

The subroutine in Algorithm 4 takes in input a set of active base learners $\mathcal{I} = \{s, s+1, \dots, M\}$ and ensures by randomized regret bound balancing that its total regret is controlled for all rounds until it terminates.

In addition to the putative bound R_i on its regret, Algorithm 4 requires that each learner i can also provide a lower-confidence bound on $\mathbb{E}[r_t|a_{t,i}, x_t]$, the expected reward of the action it would play in the current context x_t. Since each base learner i is an instance of OFUL, we can choose these bounds at round t as

$$R_i(n_i(t)) = 2 \sum_{k \in T_i(t)} \left(\beta_{k,i} \|a_{k,i}\|_{\Sigma_{k,i}^{-1}} \wedge R_i^{\max} \right)$$

and

$$B_{t,i} = \left(\langle \widehat{\theta}_{t,i}, a_{t,i} \rangle - \beta_{t,i} \|a_{t,i}\|_{\Sigma_{t,i}^{-1}} \right) \vee -R_i^{\max} ,$$

where $R_i^{\max} \in [1, L_i S_i]$ is the range of expected rewards[8] and $L_i \geq \max_t \|a_{t,i}\|$ and $S_i \geq \|\theta^\star\|$ are the norm bounds used by the OFUL base learners. Further, $\widehat{\theta}_{t,i}$, $\Sigma_{t,i}$ and $\beta_{t,i}$ are the parameter estimate (11), the covariance matrix (11), and the ellipsoid radius (13) of base learner i at time t, respectively. In a similar spirit,

$$a_{t,i} \in \operatorname*{argmax}_{a \in \mathcal{A}_t} \langle \widehat{\theta}_{t,i}, a \rangle + \beta_{t,i} \|a_{t,i}\|_{\Sigma_{t,i}^{-1}}$$

denotes the action that base learner i would take at time t. Note that we mean here the truncated action in $\mathbb{R}^{d_i}$ and the covariance matrix in $\mathbb{R}^{d_i \times d_i}$.

At each round t, Algorithm 4 first requests these bounds from each base learner to be later used in the misspecification test. The algorithm then selects one of the base learners in $\mathcal{I}$ by sampling an index $i_t \sim \text{Categorical}(p)$ from a categorical distribution with probabilities

$$p_i = \frac{1/z_i}{\sum_{j \in \mathcal{I}} 1/z_j} , \tag{24}$$

where $z_i = (d_i^2 + d_i S_i^2) \left(R_i^{\max} \wedge L_i^2 \right)$ for $i \in \mathcal{I}$. Since the regret of OFUL scales roughly at a rate of $\sqrt{z_i T}$, this learner selection rule approximately equalizes the regret of all learners in expectation. The algorithm proceeds by playing the action proposed by i_t, gathering the associated reward r_t, and updating i_t's internal state.[9] Finally, Algorithm 4 performs a misspecification test and terminates if this test triggers. We refer to the execution of Algorithm 4 as an epoch.

Unlike the misspecification test in Algorithm 1 which considers the hypothesis that a *specific* learner i is well specified, the misspecification test in Algorithm 4 tests the hypothesis that *all* active learners are well-specified. If all OFUL learners $i \in \mathcal{I}$ are well-specified, in the sense that their ellipsoid confidence sets contain $\theta_\star$ for all rounds t so far, then each $B_{t,i}$ is also a lower-bound on the optimal value in round t, since

$$B_{t,i} \leq \mathbb{E}[r_t|a_{t,i}, x_t] \leq \max_{a \in \mathcal{A}_t} \mathbb{E}[r_t|a, x_t] = \mu_t^\star .$$

Hence, the right-hand side of the misspecification test in Algorithm 4 is a lower-bound on the optimal value of all rounds so far, that is, it satisfies $\max_{j \in \mathcal{I}} \sum_{k=1}^t B_{k,j} \leq \sum_{k=1}^t \mu_t^\star$. Similarly, when all learners are well-specified and satisfy their putative regret bounds, then the left-hand side of the misspecification test is an upper-bound on $\sum_{k=1}^t \mu_k^\star$. We can see this as follows. First, by basic concentration arguments, the realized rewards cannot be much smaller than their conditional expectations with high probability, that is, $\sum_{i \in \mathcal{I}} U_i(t) \geq \sum_{k=1}^t \mathbb{E}[r_t|a_t, x_t] - c\sqrt{t \ln \frac{\ln(t)}{\delta}}$. This implies that

$$\sum_{i \in \mathcal{I}} [U_i(t) + R_i(n_i(t))] + c\sqrt{t \ln \frac{\ln(t)}{\delta}}$$

[8] We specifically assume that $\mathbb{E}[r_t|a_t, x_t] \in [-R_\star^{\max}, +R_\star^{\max}]$ where $\star$ is the smallest base learner whose model class contains the optimal parameter $\theta_\star$.

[9] We may also pass on the observation to all base learners when base learners can accept *off-policy* samples (which do not necessarily come from the proposed action), as is the case for OFUL.

$$
\begin{aligned}
&\geq \sum_{k=1}^{t} \mathbb{E}[r_t|a_t, x_t] + \sum_{i\in\mathcal{I}} R_i(n_i(t)) \\
&= \sum_{i\in\mathcal{I}} \left[\sum_{k\in T_i(t)} \mathbb{E}[r_t|a_t, x_t] + R_i(n_i(t)) \right] \\
&\geq \sum_{i\in\mathcal{I}} \left[\sum_{k\in T_i(t)} \mathbb{E}[r_t|a_t, x_t] + \mathsf{Reg}_i(t) \right] \\
&= \sum_{i\in\mathcal{I}} \sum_{k\in T_i(t)} \mu_k^\star \\
&= \sum_{k=1}^{k} \mu_k^\star,
\end{aligned}
$$

where the last inequality holds because $R_i(n_i(t)) \geq \mathsf{Reg}_i(t)$ when i is well-specified. Thus, if all learners are well-specified, the misspecification test cannot trigger (with high probability). The following theorem formalizes this argument:

Theorem 37. *With probability at least $1-\delta$, Algorithm 4 does not terminate if all base learners are well-specified and their elliptical confidence sets contain $\theta^\star$ at all times.*

Therefore, if the test does trigger, at least one learner in $\mathcal{I}$ has to be misspecified, that is, either their putative regret bound R_i or a lower bound $B_{k,i}$ does not hold. However, until the test triggers, the condition in the test is sufficient to control the regret as the next theorem formalizes.

In this result, we assume that the base learner regret bounds z_i (see text surrounding Eq. (24)) are sufficiently apart, i.e. $2z_i \leq z_{i+1}$ holds for all $i \in \mathcal{I} \setminus \{M\}$. Note that this assumption can always be ensured by first filtering the base learners. This filtering can increase the regret by at most a factor of 2.

Theorem 38. *Assume that Algorithm 4 is run with instances of OFUL as base learners that use different dimensions d_i and norm bounds L_i, S_i with $2z_i \leq z_{i+1}$. All base learners use expected reward range $R_i^{\max} = 1$ and $\lambda = 1$. Denote by $\star$ the smallest index of the base learner so that all base learners $j \in \mathcal{I}$ with $d_j \geq d_\star$ are well-specified and their elliptical confidence sets always contain the true parameter. Then, with probability at least $1-2\delta$, the regret is bounded for all rounds t until termination as*

$$
\mathsf{Reg}(t) = \tilde{O}\left(\left(d_\star + \sqrt{d_\star} S_\star + |\mathcal{I}| \right) (d_\star + \sqrt{d_\star} S_\star) \sqrt{t} \right) .
$$

Here, we highlighted the regret bound of the single best well-specified learner $\star$ in green, and assumed that the range of expected rewards is known and equal to 1. If this is not the case and we have to rely on the expected reward range induced by the vector norms L_i and S_i, then we have an additional lower-order term:

Theorem 39. *Assume that Algorithm 4 is run with instances of OFUL as base learners that use different dimensions d_i and norm bounds L_i, S_i and $R_i^{\max} = L_i S_i$ with $2z_i \leq z_{i+1}$. Denote by $\star$ the smallest index of the base learner so that all base learners $j \in \mathcal{I}$ with $d_j \geq d_\star$ are well-specified and their elliptical confidence sets always contain the true parameter. Then, with probability at least $1-2\delta$, the regret is bounded for all rounds t until termination as*

$$
\mathsf{Reg}(t) = \tilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star} S_\star L_\star + |\mathcal{I}| \right) (d_\star + \sqrt{d_\star} S_\star) L_\star \sqrt{t} + \sum_{i\in\mathcal{I}} L_i S_i \right) .
$$

The proofs of Theorem 39 and Theorem 38 are similar to the proof of Corollary 2 but requires a randomized version of the standard elliptical potential lemma that we prove in Lemma 8.

E.2. Main Algorithm

We now show how to obtain a robust model selection algorithm for adversarial contexts with the help of the Epoch Balancing subroutine from the previous section. Since Theorem 38 guarantees that the regret of Epoch Balancing is controlled in each

Algorithm 5: Regret Bound Balancing and Elimination with Adversarial Contexts

for $s = 1, \dots, M$ **do**
  EpochBalancing $(\{s, s+1 \dots, M\})$ in Algorithm 4

epoch, all that is left it to ensure that the number of epochs is small. When Algorithm 4 terminates, we know that one of the base learners must have been misspecified but we do not know which one. We here use the hierarchy of base learners: It is safe to remove the learner $i_{\min} = \min_{i \in \mathcal{I}} d_i$ with the smallest dimension as its model class is a subset of the model classes of other base learners. Thus, if there is a model class that fails to contain $\theta^\star$, this must also be the case for $i_{\min}$. Therefore, our main algorithm shown in Algorithm 5 calls Epoch Balancing (Algorithm 4) repeatedly, removing the smallest index from the active learner set each time.

Note that once $d_i \geq d_\star$ for all $i \in \mathcal{I} = \{s, s+1, \dots, M\}$, Epoch balancing will not terminate with high probability because all remaining learners are well-specified and their bounds hold (see Theorem 37). Therefore, there can only be $i_\star \leq M$ epochs where $d_{i_\star} = d_\star$ and the total regret $\mathsf{Reg}(T)$ of Algorithm 5 is just the sum of the regret in each epoch up to the total number of T rounds. We denote by $t^{(s)}(T)$ the total number of rounds in the first s epochs after a total of T rounds. Note that $t^{(s)}(T)$ are stopping times. The regret in the s-th epoch is referred to as $\mathsf{Reg}^{(s)}(t^{(s)}(T) - t^{(s-1)}(T))$ where $t^{(s)}(T) - t^{(s-1)}(T)$ is the number of rounds in episode s. Therefore, we can write the total regret as

$$\mathsf{Reg}(T) = \sum_{s=1}^{M} \mathsf{Reg}^{(s)}(t^{(s)}(T) - t^{(s-1)}(T)) \,. \tag{25}$$

The regret incurred within each epoch can be bound using Theorem 38, which yields the main result of this section:

Theorem 40 (Model Selection for Adversarial Contexts in Stochastic Linear Bandits)**.** *Assume that Algorithm 5 is run with instances of OFUL as base learners that use different dimensions d_i and norm bounds L_i, S_i with $2z_i \leq z_{i+1}$ (see text surrounding Eq. (24)). All base learners use regularizer $\lambda = 1$. With probability at least $1 - 3(M+1)\delta$ the total regret of Algorithm 5 is bounded for all rounds $T \in \mathbb{N}$ as*

$$\mathsf{Reg}(T) = \tilde{O}\left(\left(\sqrt{B}d_\star + \sqrt{Bd_\star}S_\star + \sqrt{B}M\right)(d_\star + \sqrt{d_\star}S_\star)\sqrt{T}\right) \,,$$

if base the learners use a common expected reward range $R_i^{\max} = 1$. Here, B is the number of base learners that use a misspecified model that cannot represent $\theta_\star$. If base learners use instead $R_i^{\max} = L_i S_i$, then the regret bound is

$$\mathsf{Reg}(T) = \tilde{O}\left(\left(\sqrt{B}d_\star L_\star + \sqrt{Bd_\star}S_\star L_\star + \sqrt{B}M\right)(d_\star + \sqrt{d_\star}S_\star)L_\star\sqrt{T} + B\sum_{i \in \mathcal{I}} L_i S_i\right) \,.$$

Proof. First, we consider the event where all learners with $d_i \geq d_\star$ are well-specified in the sense that their elliptical confidence intervals contain $\theta_\star$ at all times. This happens with probability at least $1 - M\delta$ by Lemma 34. Further, only consider outcomes where Theorem 38 and Theorem 37 hold in all epochs.[10] By a union bound, all these assumptions hold with probability at least $1 - (3M+1)\delta$. We now consider the decomposition in Eq. (25) and bound

$$\begin{aligned}
\mathsf{Reg}(T) &= \sum_{s=1}^{M} \mathsf{Reg}^{(s)}(t^{(s)}(T) - t^{(s-1)}(T)) \\
&\overset{(i)}{=} \sum_{s=1}^{i_\star} \mathsf{Reg}^{(s)}(t^{(s)}(T) - t^{(s-1)}(T)) \\
&\overset{(ii)}{\leq} \sum_{s=1}^{i_\star} \left[C^{(s)} \sqrt{t^{(s)}(T) - t^{(s-1)}(T)} + 8.12 \sum_{i \in \mathcal{I}^{(s)}} R_i^{\max} \ln \frac{5.2M \ln(2T)}{\delta} \right]
\end{aligned}$$

[10] We note that both theorems hold for arbitrary sequences of contexts and therefore also when the s-th instance of Epoch Balancing is started after a random number of rounds $t^{(s-1)}(T)$.

$$\leq \max_{s\in[i_\star]} C^{(s)} \sqrt{i_\star \sum_{s=1}^{i_\star} (t^{(s)}(T) - t^{(s-1)}(T))} + 8.12\, i_\star \sum_{i\in\mathcal{I}^{(s)}} R_i^{\max} \ln \frac{5.2M \ln(2T)}{\delta}$$
$$= \max_{s\in[i_\star]} C^{(s)} \sqrt{i_\star T} + 8.12\, i_\star \sum_{i\in\mathcal{I}^{(s)}} R_i^{\max} \ln \frac{5.2M \ln(2T)}{\delta} ,$$

where (i) follows from Theorem 37 and (ii) from Theorem 38 with epoch-dependent factor $C^{(s)} \leq \tilde{O}\left(\left(d_\star + \sqrt{d_\star} S_\star + M\right)\left(d_\star + \sqrt{d_\star} S_\star\right)\right)$ or Theorem 39 with epoch-dependent factor $C^{(s)} \leq \tilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star} S_\star L_\star + M\right)\left(d_\star + \sqrt{d_\star} S_\star\right)\right) L_\star$. □

E.3. Epoch Balancing Termination (Proof of Theorem 37)

Theorem 37. *With probability at least $1-\delta$, Algorithm 4 does not terminate if all base learners are well-specified and their elliptical confidence sets contain $\theta^\star$ at all times.*

Proof. Since all base learners are well-specified and their lower-confidence bounds $L_{t,i}$ satisfy $L_{t,i} \leq \mathbb{E}[r_t|a_{t,i}, x_t] \leq \mu_k^\star$, the right-hand side of the misspecification test satisfies

$$\max_{j\in\mathcal{I}} \sum_{k=1}^{t} B_{k,j} \leq \sum_{k=1}^{t} \mu_k^\star$$

for all $t \in \mathbb{N}$. Further, with probability at least $1-\delta$, by Lemma 42, the left-hand side of the misspecification test satisfies for all $t \in \mathbb{N}$

$$\sum_{i\in\mathcal{I}} [U_i(t) + R_i(n_i(t))] + c\sqrt{t \ln \frac{\ln(t)}{\delta}} \geq \sum_{k=1}^{t} \mu_k^\star .$$

Thus, the misspecification test never triggers and Algorithm 4 does not terminate. □

Let's define the event $\mathcal{E}''$ under which the sum of the cumulative rewards of all active algorithms in $\mathcal{I}$ are close to the sum of their pseudorewards. Let $\delta \in (0,1)$. Define:

$$\mathcal{E}'' = \left\{ \forall t \in \mathbb{N}\colon \left| \sum_{i\in\mathcal{I}} U_i(t) - \sum_{k=1}^{t} \mathbb{E}[r_k|a_k, x_k] \right| \leq c\sqrt{t \ln \frac{\ln(t)}{\delta}} \right\} , \tag{26}$$

where $c > 0$ is an absolute constant such that $c\sqrt{t \ln \frac{\ln(t)}{\delta}} \geq 0.85\sqrt{t\left(\ln\ln(4t) + 0.72\ln(10.4/\delta)\right)}$.

Lemma 41. *Event $\mathcal{E}''$ holds with probability at least $1-\delta$.*

Proof. Let $\mathcal{F}_t = \sigma(x_1, i_1, a_1, r_1, \ldots, x_{t-1}, i_{t-1}, a_{t-1}, r_{t-1}, x_{t-1}, i_{t-1}, a_{t-1})$ be the sigma-field induced by all variables up to the reward at round t. Hence, $X_k = r_k - \mathbb{E}[r_k|a_k, x_k]$ is a martingale-difference sequence w.r.t. $\mathcal{F}_k$. We will now apply a Hoeffding-style uniform concentration bound from Howard et al. (2021). Using the terminology and definition in this article, by case Hoeffding I in Table 4 therein the process $S_k = \sum_{j=1}^{k} X_k$ is sub-ψ_N with variance process $V_k = k/4$. Thus by using the boundary choice in Equation (11) of Howard et al. (2021), we get

$$S_k \leq 1.7\sqrt{V_k\left(\ln\ln(8V_k) + 0.72\ln(5.2/\delta)\right)}$$
$$= 0.85\sqrt{k\left(\ln\ln(4k) + 0.72\ln(5.2/\delta)\right)}$$

for all k with probability at least $1-\delta$. Applying the same argument to $-S_k$ gives that

$$\left| \sum_{k=1}^{t} \left(r_k - \mathbb{E}[r_k|a_k, x_k]\right) \right| \leq 0.85\sqrt{t\left(\ln\ln(4t) + 0.72\ln(10.4/\delta)\right)}$$

holds with probability at least $1-\delta$ for all t. Since $\sum_{i\in\mathcal{I}} U_i(t) = \sum_{k=1}^{t} r_k$, the statement follows. Note that this concentration argument holds for all t uniformly and therefore also when t is random. □

Lemma 42 (Upper-confidence bound on optimal reward)**.** *In event $\mathcal{E}''$ from Equation 26, the following holds. If at time t all learners $i \in \mathcal{I}$ are well-specified, then the left-hand side in the misspecification test of Algorithm 4 is a lower-bound on the optimal rewards, i.e.,*

$$\sum_{i\in\mathcal{I}}[U_i(t) + R_i(n_i(t))] + c\sqrt{t \ln \frac{\ln(t)}{\delta}} \geq \sum_{k=1}^{t} \mu_k^\star \,.$$

Proof. Whenever $\mathcal{E}''$ holds, we have

$$\begin{aligned}
&\sum_{i\in\mathcal{I}}[U_i(t) + R_i(n_i(t))] + c\sqrt{t \ln \frac{\ln(t)}{\delta}} \\
&\geq \sum_{i\in\mathcal{I}} R_i(n_i(t)) + \sum_{k=1}^{t} \mathbb{E}[r_k|a_k, x_k] && \text{(by definition of } \mathcal{E}'') \\
&\geq \sum_{i\in\mathcal{I}} \mathsf{Reg}_i(t) + \sum_{k=1}^{t} \mathbb{E}[r_k|a_k, x_k] && \text{(each learner is well-specified)} \\
&= \sum_{i\in\mathcal{I}} \left[\mathsf{Reg}_i(t) + \sum_{k\in T_i(t)} \mathbb{E}[r_k|a_k, x_k]\right] \\
&= \sum_{i\in\mathcal{I}} \sum_{k\in T_i(t)} \mu_k^\star = \sum_{k=1}^{t} \mu_k^\star && \text{(by definition of regret).}
\end{aligned}$$

□

E.4. Regret Bound for Epoch Balancing (Proof of Theorem 38)

Theorem 38. *Assume that Algorithm 4 is run with instances of OFUL as base learners that use different dimensions d_i and norm bounds L_i, S_i with $2z_i \leq z_{i+1}$. All base learners use expected reward range $R_i^{\max} = 1$ and $\lambda = 1$. Denote by $\star$ the smallest index of the base learner so that all base learners $j \in \mathcal{I}$ with $d_j \geq d_\star$ are well-specified and their elliptical confidence sets always contain the true parameter. Then, with probability at least $1 - 2\delta$, the regret is bounded for all rounds t until termination as*

$$\mathsf{Reg}(t) = \tilde{O}\left(\left(d_\star + \sqrt{d_\star}S_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)\sqrt{t}\right) \,.$$

Proof. We apply Theorem 43 (see below) which immediately yields the desired bound

$$\mathsf{Reg}(t) = \tilde{O}\left(\left(d_\star + \sqrt{d_\star}S_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)\sqrt{t}\right) \,.$$

□

Theorem 39. *Assume that Algorithm 4 is run with instances of OFUL as base learners that use different dimensions d_i and norm bounds L_i, S_i and $R_i^{\max} = L_i S_i$ with $2z_i \leq z_{i+1}$. Denote by $\star$ the smallest index of the base learner so that all base learners $j \in \mathcal{I}$ with $d_j \geq d_\star$ are well-specified and their elliptical confidence sets always contain the true parameter. Then, with probability at least $1 - 2\delta$, the regret is bounded for all rounds t until termination as*

$$\mathsf{Reg}(t) = \tilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star}S_\star L_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)L_\star\sqrt{t} + \sum_{i\in\mathcal{I}} L_i S_i\right) \,.$$

Proof. We apply Theorem 43 (see below) which yields

$$\mathsf{Reg}(t) = \tilde{O}\left(\left(d_\star L_\star + \sqrt{d_\star}S_\star L_\star + |\mathcal{I}|\right)(d_\star + \sqrt{d_\star}S_\star)L_\star\sqrt{t} + \sum_{i\in\mathcal{I}} L_i S_i \ln\ln(t)\right) \,.$$

□

Theorem 43 (General Regret Bound of Epoch Balancing)**.** *Assume that Algorithm 4 is run with instances of OFUL as base learners which use different dimensions $d_i, S_i, L_i, R_i^{\max}$ and regularization parameter $\lambda = 1$. Denote by $\star$ the index of the base learner so that all base learners $j \in \mathcal{I}$ with $d_j \geq d_\star$ are well-specified and their elliptical confidence sets always contain the true parameter. Then, with probability at least $1 - 2\delta$, the regret is bounded for all rounds t as*

$$\mathsf{Reg}(t) \leq \sqrt{(d_\star^2 + d_\star S_\star^2)|\mathcal{I}|}\,(R_\star^{\max} \wedge L_\star)\,\sqrt{t}(2+2c)\,x(t) + (d_\star^2 + d_\star S_\star^2)\,(R_\star^{\max} \wedge L_\star)^2\,\sqrt{\bar{M}t}(2+2c)\,x(t) + 8.12 \sum_{i \in \mathcal{I}} R_i^{\max} \ln \frac{5.2|\mathcal{I}| \ln(2t)}{\delta},$$

where $\bar{M} = |\mathcal{I}|$ for general z_i and $\bar{M} = 2$ when z_i are exponentially increasing (i.e., $2z_i \leq z_{i+1}$ for all $i \in \mathcal{I}$). In the above, $x(t)$ is a short-hand for $O(\ln \frac{tL_{\max}}{\delta} + \ln\ln(R_{\max}^{\max} t \wedge L_{\max} t))$, where $L_{\max} = \max_{i \in [M]} L_i$, and $R_{\max}^{\max} = \max_{i \in [M]} R_i^{\max}$, and c is a universal constant.

Proof. Since learner $i_\star$ is well-specified and its elliptical confidence set contains $\theta^\star$, it holds that

$$\sum_{k=1}^{t} \mu_k^\star \leq \sum_{k=1}^{t} \max_{a \in \mathcal{A}_k} \left[\langle \widehat{\theta}_{k,\star}, a \rangle + \beta_{k,\star} \|a\|_{\Sigma_{k,\star}^{-1}} \right] = \sum_{k=1}^{t} \langle \widehat{\theta}_{k,\star}, a_{k,\star} \rangle + \beta_{k,\star} \|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}.$$

Thus, we can write the total regret up to round t as

$$\begin{aligned}\mathsf{Reg}(t) &= \sum_{k=1}^{t} [\mu_k^\star - \mathbb{E}[r_k | a_k, x_k]] = \sum_{k=1}^{t} \mu_k^\star - \sum_{k=1}^{t} \mathbb{E}[r_k | a_k, x_k] \\ &\leq \sum_{k=1}^{t} \mu_k^\star - \sum_{i \in \mathcal{I}} U_i(n_i(t)) + c\sqrt{t \ln \frac{\ln(t)}{\delta}},\end{aligned}$$

where the inequality holds in event $\mathcal{E}''$. If Algorithm 4 does not stop in iteration t, then the misspecification test does not trigger for any learner, and in particular for learner $i_\star$. This implies that

$$\sum_{i \in \mathcal{I}} [U_i(t) + R_i(n_i(t))] + c\sqrt{t \ln \frac{\ln(t)}{\delta}} \geq \sum_{k=1}^{t} B_{k,\star}\,.$$

Rearranging terms and plugging this inequality back into the regret bound from above yields

$$\mathsf{Reg}(t) \leq \sum_{k=1}^{t} [\mu_k^\star - B_{k,\star}] + \sum_{i \in \mathcal{I}} R_i(n_i(t)) + 2c\sqrt{t \ln \frac{\ln(t)}{\delta}}\,. \tag{27}$$

We bound the first term in Equation 27 as

$$\begin{aligned}&\sum_{k=1}^{t} [\mu_k^\star - B_{k,\star}] \\ &\overset{(i)}{\leq} \sum_{k=1}^{t} \left[R_\star^{\max} \wedge (\langle \widehat{\theta}_{k,\star}, a_{k,\star} \rangle + \beta_{k,\star} \|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}) - (-R_\star^{\max} \vee (\langle \widehat{\theta}_{k,\star}, a_{k,\star} \rangle - \beta_{k,\star} \|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}})) \right] \\ &\leq \sum_{k=1}^{t} \left[2R_\star^{\max} \wedge 2\beta_{k,\star} \|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}} \right] \leq 2\beta_{t,\star} \sum_{k=1}^{t} \left[\frac{R_\star^{\max}}{\beta_{t,\star}} \wedge \|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}} \right] \\ &\overset{(ii)}{\leq} 2\beta_{t,\star} \sqrt{t \sum_{k=1}^{t} \left[\left(\frac{R_\star^{\max}}{\beta_{t,\star}} \right)^2 \wedge \frac{L^2}{\lambda_i} \wedge \|a_{k,\star}\|_{\Sigma_{k,\star}^{-1}}^2 \right]}\end{aligned}$$

where (i) follows from the definition of $B_{k,i}$ and the fact that the ellipsoid confidence set of $\star$ contain the true parameter and (ii) applies the Cauchy-Schwarz inequality. We now apply a randomized version of the elliptical potential lemma which

we prove in Lemma 8. This yields

$$\sum_{k=1}^{t}[\mu_k^\star - B_{\star,k}] \le 4\beta_{t,\star}\sqrt{\frac{t}{p_\star}(1+b_\star^2)\ln\frac{5.2\ln(2b_\star^2 t\vee 2)\det\Sigma_{t,\star}}{\delta\det\Sigma_{0,\star}}}$$
$$\le 4\beta_{t,\star}\sqrt{\frac{td_\star}{p_\star}(1+b_\star^2)\ln\frac{5.2\ln(2b_\star^2 t\vee 2)(d_\star\lambda_\star+tL_\star^2)}{\delta d_\star\lambda_\star}}$$

where $b_\star = \frac{R_\star^{\max}}{\beta_{t,\star}} \wedge \frac{L_\star}{\sqrt{\lambda_\star}}$. For the second term in Equation 27, we apply Lemma 44 (see below) as

$$\sum_{i\in\mathcal{I}} R_i(n_i(t)) \le 8.12\sum_{i\in\mathcal{I}} R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2\sum_{i\in\mathcal{I}}\beta_{t,i}\sqrt{3d_ip_it(1+b_i^2)\ln\frac{d_i\lambda_i+tp_iL_i^2}{d_i\lambda_i}}.$$

Combining the terms for both bounds, we arrive at the regret bound

$$\begin{aligned}\mathsf{Reg}(t) &\le 4\beta_{t,\star}\sqrt{\frac{td_\star}{p_\star}(1+b_\star^2)\ln\frac{5.2\ln(2b_\star^2t\vee 2)(d_\star\lambda_\star+tL_\star^2)}{\delta d_\star\lambda_\star}}\\ &\quad + 2\sum_{i\in\mathcal{I}}\beta_{t,i}\sqrt{3d_ip_it(1+b_i^2)\ln\frac{d_i\lambda_i+tp_iL_i^2}{d_i\lambda_i}}\\ &\quad + 8.12\sum_{i\in\mathcal{I}}R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2c\sqrt{t\ln\frac{\ln(t)}{\delta}}\\ &\le x(t)\sqrt{\frac{z_\star t}{p_\star}} + x\sum_{i\in\mathcal{I}}\sqrt{z_ip_it} + 8.12\sum_{i\in\mathcal{I}}R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2c\sqrt{t\ln\frac{\ln(t)}{\delta}},\end{aligned}$$

where

$$z_i = (\sigma^2 d_i + \lambda_i S_i^2)d_i(1+b_i^2) \le 2(d_i^2+d_iS_i^2)\left(R_i^{\max}\wedge L_i\right)^2$$

and

$$\begin{aligned}x(t) &= 12\max_{i\in\mathcal{I}}\sqrt{\ln\left(\frac{1+tL_i^2/\lambda_i}{\delta}\right)\ln\frac{5.2\ln(2b_i^2t\vee 2)(d_i\lambda_i+tL_i^2)}{\delta d_i\lambda_i}}\\ &\le 12\max_{i\in\mathcal{I}}\sqrt{\ln\left(\frac{1+tL_i^2}{\delta}\right)\ln\frac{10.4\ln(2\left(R_i^{\max}\wedge L_i\right)t)(1+tL_i^2)}{\delta}}\\ &\le 12\ln\frac{10.4(1+tL_{\max}^2)\ln(2\left(R_{\max}^{\max}\wedge L_{\max}\right)t)}{\delta}.\end{aligned}$$

We now use the definition of $p_i \propto \frac{1}{z_i}$ and bound

$$\sum_{i\in\mathcal{I}}\sqrt{z_ip_i} = \sum_{i\in\mathcal{I}}\sqrt{\frac{1}{\sum_{i\in\mathcal{I}}z_i^{-1}}} = \frac{|\mathcal{I}|}{\sqrt{\sum_{i\in\mathcal{I}}z_i^{-1}}} \le \frac{|\mathcal{I}|}{\sqrt{z_\star^{-1}}} = |\mathcal{I}|\sqrt{z_\star}$$

where the inequality uses the fact that $\star\in\mathcal{I}$. Further

$$\sqrt{\frac{z_\star}{p_\star}} = z_\star\sqrt{\sum_{i\in\mathcal{I}}\frac{1}{z_i}} \le z_\star\sqrt{|\mathcal{I}|}$$

holds for any z_i but if we know that $z_1 \le \frac{1}{2}z_2 \le \frac{1}{4}z_3 \cdots \le \frac{1}{2^M z_M}$, then

$$\sqrt{\frac{z_\star}{p_\star}} = z_\star\sqrt{\sum_{i\in\mathcal{I}}\frac{1}{z_i}} \le 2z_\star.$$

Thus, we can bound the total regret as

$$\begin{aligned}\mathsf{Reg}(t) &\le (|\mathcal{I}|\sqrt{z_\star} + z_\star\sqrt{\bar{M}})x(t)\sqrt{t} + 8.12\sum_{i\in\mathcal{I}} R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2c\sqrt{t\ln\frac{\ln(t)}{\delta}} \\ &\le \sqrt{(d_\star^2 + d_\star S_\star^2)|\mathcal{I}|}\left(R_\star^{\max}\wedge L_\star\right)\sqrt{t}(2+2c)x(t) \\ &\quad + (d_\star^2 + d_\star S_\star^2)\left(R_\star^{\max}\wedge L_\star\right)^2\sqrt{\bar{M}t}(2+2c)x(t) \\ &\quad + 8.12\sum_{i\in\mathcal{I}} R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta},\end{aligned}$$

where $\bar{M} = |\mathcal{I}|$ for general z_i and $\bar{M} = 2$ when z_i are exponentially increasing. Note that since this bound holds in the penultimate round of Algorithm 4 and the regret in the final round can be at most 1, this bound holds for all rounds t played by Algorithm 4, including the last one. □

Lemma 44 (Regret bounds are balanced)**.** *Let $\delta \in (0,1)$ be arbitrary but fixed. With probability at least $1-\delta$, the sum of regret bounds satisfy in all iterations t of Algorithm 4 the following upper-bound*

$$\sum_{i\in\mathcal{I}} R_i(n_i(t)) \le 8.12\sum_{i\in\mathcal{I}} R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2\sum_{i\in\mathcal{I}}\beta_{t,i}\sqrt{3d_ip_it\left(1+b_i^2\right)\ln\frac{\lambda_i d_i + 3tp_iL_i^2}{\lambda_i d_i}},$$

where $b_i = \frac{R_i^{\max}}{2\beta_{t,i}}\wedge\frac{L_i}{\sqrt{\lambda_i}}$.

Proof. By the choice of regret bounds we have

$$\begin{aligned}R_i(n_i(t)) &= \sum_{k\in T_i(t)}\left[2\beta_{k,i}\|a_{k,i}\|_{\Sigma_{k,i}^{-1}}\wedge R_i^{\max}\right] \\ &\le R_i^{\max}n_i(t)\wedge 2\beta_{t,i}\sum_{k\in T_i(t)}\left(\|a_{k,i}\|_{\Sigma_{k,i}^{-1}}\wedge\frac{R_i^{\max}}{2\beta_{t,i}}\right) \\ &\le R_i^{\max}n_i(t)\wedge 2\beta_{t,i}\sqrt{n_i(t)\sum_{k\in T_i(t)}\left(\|a_{k,i}\|_{\Sigma_{k,i}^{-1}}^2\wedge\left(\frac{R_i^{\max}}{2\beta_{t,i}}\right)^2\wedge\frac{L_i^2}{\lambda_i}\right)} \\ &\le R_i^{\max}n_i(t)\vee 2\beta_{t,i}\sqrt{d_in_i(t)\left(1+b_i^2\right)\ln\frac{\lambda_i + n_i(t)L_i^2/d_i}{\lambda_i}}\end{aligned}$$

where $b_i = \frac{R_i^{\max}}{2\beta_{t,i}}\wedge\frac{L_i}{\sqrt{\lambda_i}}$ and the last inequality follows from of Lemma 7. To control the the number of times each learner was chosen, we use Lemma 45 (see below). This gives with probability at least $1-\delta$ for all iterations t simultaneously $n_i(t) \le 3tp_i \vee 8.12\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta}$. This yields a regret bound of

$$R_i(n_i(t)) \le 8.12R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} \quad\vee\quad 2\beta_{t,i}\sqrt{3d_ip_it\left(1+b_i^2\right)\ln\frac{\lambda_i + 3tp_iL_i^2/d_i}{\lambda_i}}.$$

Summing over R_i and plugging in $\beta_{t,i}$ yields

$$\sum_{i\in\mathcal{I}} R_i(n_i(t)) \le 8.12\sum_{i\in\mathcal{I}} R_i^{\max}\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} + 2\sum_{i\in\mathcal{I}}\beta_{t,i}\sqrt{3d_ip_it\left(1+b_i^2\right)\ln\frac{\lambda_i + 3tp_iL_i^2/d_i}{\lambda_i}},$$

as claimed. □

Lemma 45. *The number of times each a learner $i\in\mathcal{I}$ has been played in Algorithm 4 after t iterations is bounded with probability at least $1-\delta$ for all $t\in\mathbb{N}$ and $i\in\mathcal{I}$ as*

$$n_i(t) \le \frac{3}{2}tp_i + 4.06\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta} \le 3tp_i \vee 8.12\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta}.$$

Proof. Fix an $i \in \mathcal{I}$ and consider the martingale difference sequence $X_t = \mathbf{1}\{i_t = i\} - p_i$. The process $S_t = \sum_{k=1}^{t} X_k$ with variance process $W_t = tp_i(1-p_i)$ satisfies the sub-ψ_P condition of Howard et al. (2021) with constant $c = 1$ (see Bennett case in Table 3 of Howard et al. (2021)). By Lemma 9, the bound

$$S_t \leq 1.44\sqrt{(W_t \vee m)\left(1.4\ln\ln\left(2(W_t/m \vee 1)\right) + \ln\frac{5.2}{\delta}\right)} + 0.41\left(1.4\ln\ln\left(2(W_t/m \vee 1)\right) + \ln\frac{5.2}{\delta}\right)$$

holds for all $t \in \mathbb{N}$ with probability at least $1-\delta$. We set $m = tp_i$ and upper-bound the RHS further, so as to obtain

$$\begin{aligned} S_t &\leq 1.44\sqrt{tp_i\left(1.4\ln\ln\left(2t\right) + \ln\frac{5.2}{\delta}\right)} + 0.41\left(1.4\ln\ln\left(2t\right) + \ln\frac{5.2}{\delta}\right) \\ &\leq \frac{tp_i}{2} + 1.45\left(1.4\ln\ln\left(2t\right) + \ln\frac{5.2}{\delta}\right), \end{aligned}$$

where used the AM-GM inequality in the final step. We therefore get that with probability at least $1-\delta$, the following upper-bound in the number of times learner i was selected by time t holds for all $i \in \mathcal{I}$ and $t \in \mathbb{N}$:

$$n_i(t) \leq \frac{3}{2}tp_i + 2.9\left(1.4\ln\ln\left(2t\right) + \ln\frac{5.2|\mathcal{I}|}{\delta}\right) \leq \frac{3}{2}tp_i + 4.06\ln\frac{5.2|\mathcal{I}|\ln\left(2t\right)}{\delta}.$$

We can now distinguish between two cases. When $\frac{3}{2}tp_i \leq 4.06\ln\frac{5.2|\mathcal{I}|\ln(2t)}{\delta}$ we have

$$n_i(t) \leq 8.12\ln\frac{5.2|\mathcal{I}|\ln\left(2t\right)}{\delta}\,.$$

In the opposite case, we simply have $n_i(t) \leq 3tp_i$. Putting together concludes the proof. □

F. Experiment Details

Linear bandit instances. We generated the 3 linear bandit instances used in our experiments as follows. All instances have dimensionality $d = 10$ and 100 actions. We drew both the true parameter $\theta \in \mathbb{R}^d$ and each of the 100 actions $a^{(i)} \in \mathbb{R}^d$ independently and uniformly from the d-dimensional unit sphere. The actions are kept fixed throughout all rounds. The reward in each round t was generated as $r_t = \langle a_t, \theta\rangle + \epsilon_t$ where $\epsilon_t \sim \mathcal{N}(0, \sigma^2)$ is independent Gaussian noise and a_t is the action chosen by the algorithm. The three instances are generated in the same fashion but vary in the standard-deviation of the noise: $\sigma = 1.0$, $\sigma = 0.3$ and $\sigma = 0.05$. The problem instances were kept fixed across different runs of the same algorithm and across different algorithms.

The results shown in Figure 1 of the main body of the paper are the averages of 60, 10, 10 independent runs for each of the three instances. We used more seeds for the first instance since the noise level is generally higher but still could not detect a statistically significant difference between Dynamic Balancing and Stochastic Corral.

Dynamic Balancing. In our experiments, we used the simple version of Dynamic Balancing described in Appendix B and shown in Algorithm 2 which does not use reward biases $b_i(t)$ or scaling factors v_i in the balancing condition. Appendix B.5 carefully lays out the regret guarantees of this version of Dynamic Balancing. They are similar to the full version in Algorithm 1 but have additional factors on the number of learners M. For each base learner, we use the following regret bound,

$$R_i(n_i(t)) = \sum_{k \in T_i(t)} \min\left\{1, 2\beta_{k,i}\|a_k\|_{\Sigma_{k,i}^{-1}}\right\}\,,$$

where $\Sigma_{k,i}$ and $\beta_{k,i}$ are the covariance matrix and confidence radius used by the i-th learner in round k. As discussed in Appendix D.1, this is a valid regret bound of OFUL.

Stochastic Corral. While the description of base learner wrapper of Stochastic Corral by Pacchiano et al. (2020b) assumes that the learner executes a previous policy at every second round, this comes with infeasible memory costs. Instead, we only maintained a set of 1000 previous policies using reservoir sampling which ensures that it is a representative sample of the entire history. We do not expect that this has any impact on practical performance.