Appendix. Online Planning in Non-stationary Environments

The appendix is organized as follows. Appendix A presents the proofs of our technical results. In Appendix B, we provide the implementation details of our O2O framework and the OMD algorithm. The cases of correlated distributions and imperfect information are discussed in Appendix C. Finally, in Appendix D, we provide supplementary numerical experiments to validate the effectiveness of our proposed O2O framework.

A. Proofs

The three key theorems in this work are proved in Appendices A.3, A.4, and A.5, respectively.

A.1. Proof of Proposition 1

PROPOSITION 1. *Consider* $K=2$ *and* $\phi(\boldsymbol{w})=\min\{w_1,w_2\}$. *For any online algorithm that does not have any distributional information on* $\Xi_{1:\Gamma}$, *there exists an instance such that* $\mathbb{E}[\mathrm{opt}(\mathrm{DP}_1(\phi,\mathbb{R}^2))-\phi(\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma)/\Gamma)]\geq 1/4$, *and* $\Xi_1,\ldots,\Xi_\Gamma$ *changes only once within the horizon* $1,\ldots,\Gamma$.

PROOF. Without loss of generality, let $\Gamma>0$ be even. We construct two instances $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$ that share the same Ω, $\{\mathcal{X}(\boldsymbol{\omega})\}_{\boldsymbol{\omega}\in\Omega}$, $\boldsymbol{f}$, but have different scenario probability distributions, denoted as $\Xi_{1:\Gamma}^{(1)}$ and $\Xi_{1:\Gamma}^{(2)}$, respectively. We set $\Omega=\{a,b,c\}$, and $\mathcal{X}(\boldsymbol{\omega})=\{\boldsymbol{y},\boldsymbol{z}\}$ for all $\boldsymbol{\omega}\in\Omega$, and set

$$\begin{aligned}&\boldsymbol{f}(\boldsymbol{y},a)=(1,0),\quad \boldsymbol{f}(\boldsymbol{y},b)=(1,0),\quad \boldsymbol{f}(\boldsymbol{y},c)=(0,1),\\ &\boldsymbol{f}(\boldsymbol{z},a)=(0,1),\quad \boldsymbol{f}(\boldsymbol{z},b)=(1,0),\quad \boldsymbol{f}(\boldsymbol{z},c)=(0,1).\end{aligned}$$

The two different scenario probability distributions $\Xi_{1:\Gamma}^{(1)}$ and $\Xi_{1:\Gamma}^{(2)}$ are defined as

$$\begin{aligned}&\Pr_{\boldsymbol{\omega}\sim\Xi_\gamma^{(i)}}(\boldsymbol{\omega}=a)=1 \quad \text{for both } i=1,2, \text{ and } \gamma\in\{1,\ldots,\Gamma/2\},\\ &\Pr_{\boldsymbol{\omega}\sim\Xi_\gamma^{(1)}}(\boldsymbol{\omega}=b)=1 \quad \text{for } \gamma\in\{\Gamma/2+1,\ldots,\Gamma\}, \qquad \Pr_{\boldsymbol{\omega}\sim\Xi_\gamma^{(2)}}(\boldsymbol{\omega}=c)=1 \quad \text{for } \gamma\in\{\Gamma/2+1,\ldots,\Gamma\},\end{aligned}$$

which display exactly one change within the horizon $1,\ldots,\Gamma$. Recall that $\phi(\boldsymbol{w})=\min\{w_1,w_2\}$. For each of the instances $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$, it is clear that $\mathrm{opt}(\mathrm{DP}_1(\phi,\mathbb{R}^2))=1/2$ with certainty. Despite the common optimal value, the optimum is achieved differently between $\mathcal{I}^{(1)}$ and $\mathcal{I}^{(2)}$. An optimal policy must choose $\boldsymbol{x}_1^*=\ldots=\boldsymbol{x}_{\Gamma/2}^*=\boldsymbol{z}$ in $\mathcal{I}^{(1)}$, but $\boldsymbol{x}_1^*=\ldots=\boldsymbol{x}_{\Gamma/2}^*=\boldsymbol{y}$ in $\mathcal{I}^{(2)}$.

Consider an arbitrary but fixed online policy that has no prior distributional information on $\Xi_{1:\Gamma}$. During periods $1,\ldots,\Gamma/2$, the policy cannot distinguish $\mathcal{I}^{(1)}$ from $\mathcal{I}^{(2)}$, since the policy observes the same scenario $\boldsymbol{\omega}_\gamma=a$ but has no other auxiliary information. Consequently, for $\gamma\in\{1,\ldots,\Gamma/2\}$, the action $\boldsymbol{x}_\gamma^{(1)}$ chosen in $\mathcal{I}^{(1)}$ and the action $\boldsymbol{x}_\gamma^{(2)}$ chosen in $\mathcal{I}^{(2)}$ are the same (i.e., same action when the policy is deterministic, and the same probability on $\mathcal{X}(a)=\{\boldsymbol{y},\boldsymbol{z}\}$ when the policy is randomized). We denote $\boldsymbol{x}_\gamma=\boldsymbol{x}_\gamma^{(1)}=\boldsymbol{x}_\gamma^{(2)}$ as the common value. The quantity $N(\boldsymbol{x})=\sum_{\gamma=1}^{\Gamma/2}\mathbf{1}(\boldsymbol{x}_\gamma=\boldsymbol{x})$ for $\boldsymbol{x}\in\{\boldsymbol{y},\boldsymbol{z}\}$

is the same in both instances. Next, note that $0 \le N(\boldsymbol{y}), N(\boldsymbol{z}) \le \Gamma/2$, and $N(\boldsymbol{y}) + N(\boldsymbol{z}) = \Gamma/2$. A routine calculation shows that the policy's objective value is

$$\begin{cases} \min\left\{\dfrac{1-N(\boldsymbol{z})}{\Gamma}, \dfrac{N(\boldsymbol{z})}{\Gamma}\right\} = \dfrac{N(\boldsymbol{z})}{\Gamma} = \dfrac{1}{2} - \dfrac{N(\boldsymbol{y})}{\Gamma} & \text{in } \mathcal{I}^{(1)}, \\ \min\left\{\dfrac{N(\boldsymbol{y})}{\Gamma}, 1 - \dfrac{N(\boldsymbol{y})}{\Gamma}\right\} = \dfrac{N(\boldsymbol{y})}{\Gamma} & \text{in } \mathcal{I}^{(2)}. \end{cases}$$

Finally, to establish the Proposition, denote $\mathrm{Reg}_1^{(i)}$ as the regret of the online policy in instance $\mathcal{I}^{(i)}$. From the objective values above, it is straightforward to check that

$$\mathrm{Reg}_1^{(1)} + \mathrm{Reg}_1^{(2)} = \left[\frac{1}{2} - \left(\frac{1}{2} - \frac{N(\boldsymbol{y})}{\Gamma}\right)\right] + \left[\frac{1}{2} - \frac{N(\boldsymbol{y})}{\Gamma}\right] = \frac{1}{2}. \tag{13}$$

Hence, $\max\{\mathrm{Reg}_1^{(1)}, \mathrm{Reg}_1^{(2)}\} = 1/4$, and the proposition is proved. ■

A.2. Proof of Proposition 2

PROPOSITION 2. *Under Assumption 1, it holds that* $\mathrm{opt}(\mathrm{UB}(\phi, S)) \ge \mathbb{E}[\mathrm{opt}(\mathrm{DP}_1(\phi, S))]$.

PROOF. We start by noting that a feasible non-anticipatory policy can be expressed as a sequence of functions $h_1, \ldots, h_\Gamma$, where h_γ maps the input $\{(\boldsymbol{\omega}_s, \boldsymbol{x}_s)\}_{s=1}^{\gamma-1} \cup \{U_\gamma\} \cup \{\boldsymbol{\omega}_\gamma\}$ to a decision $\boldsymbol{x}_\gamma \in \mathcal{X}(\boldsymbol{\omega}_\gamma)$. That is, $h_\gamma(\{(\boldsymbol{\omega}_s, \boldsymbol{x}_s)\}_{s=1}^{\gamma-1}, U_\gamma, \boldsymbol{\omega}_\gamma) \in \mathcal{X}(\boldsymbol{\omega}_\gamma)$ for all $\{(\boldsymbol{\omega}_s, \boldsymbol{x}_s)\}_{s=1}^{\gamma-1}, U_\gamma, \boldsymbol{\omega}_\gamma$. The existence of a feasible non-anticipatory policy is ensured by Assumption 1. Now, let $h_1^*, \ldots, h_\Gamma^*$ be a function sequence that corresponds to an optimal non-anticipatory policy, and define

$$\begin{aligned} \boldsymbol{x}_\gamma^* &= h_\gamma^*(\{(\boldsymbol{\omega}_s, \boldsymbol{x}_s^*)\}_{s=1}^{\gamma-1}, U_\gamma, \boldsymbol{\omega}_\gamma) \\ \boldsymbol{H}_\gamma^*(\boldsymbol{\omega}_\gamma) &= \mathbb{E}_{\{(\boldsymbol{\omega}_s, \boldsymbol{x}_s^*)\}_{s=1}^{\gamma-1} \cup \{U_\gamma\}}\left[\boldsymbol{f}(x_\gamma^*, \boldsymbol{\omega}_\gamma) \,|\, \boldsymbol{\omega}_\gamma\right] \qquad (14) \\ \boldsymbol{f}_\gamma^* &= \mathbb{E}_{\boldsymbol{\omega}_\gamma \sim \Xi_\gamma}[\boldsymbol{H}_\gamma^*(\boldsymbol{\omega}_\gamma)]. \qquad (15) \end{aligned}$$

We claim that $(\boldsymbol{f}^*, \boldsymbol{H}^*)$ is feasible to $\mathrm{UB}(\phi, S)$. By definition, we have $\boldsymbol{H}_\gamma^*(\boldsymbol{\omega}_\gamma) \in \mathrm{Conv}(\boldsymbol{\omega}_\gamma)$. Hence, it suffices to check constraint (2). By the convexity of $d(\cdot, S)$, we have

$$d\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}_\gamma^*, S\right) \le \mathbb{E}\left[d\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma^*, \boldsymbol{\omega}_\gamma), S\right)\right] \le 0,$$

where the expectation $\mathbb{E}$ is over $\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_\Gamma$ and $U_1, \ldots, U_\Gamma$. Finally, the proposition follows by

$$\begin{aligned} \mathbb{E}[\mathrm{opt}(\mathrm{DP}(S, \phi))] &= \mathbb{E}\left[\phi\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma^*, \boldsymbol{\omega}_\gamma)\right)\right] \\ &\le \phi\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \mathbb{E}\left[\boldsymbol{f}(\boldsymbol{x}_\gamma^*, \boldsymbol{\omega}_\gamma)\right]\right) \qquad (16) \\ &= \phi\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}_\gamma^*\right) \le \mathrm{opt}(\mathrm{UB}(\phi, S)). \end{aligned}$$

Step (16) is by the concavity of ϕ. ■

A.3. Proof of Theorem 1

Before we state the proof of Theorem 1, we need to introduce the Azuma-Hoeffding inequality to enable our analysis. The Azuma-Hoeffding inequality is summarized as follows:

PROPOSITION 3. *Let $\{X_t\}_{t=1}^T \in [0,1]^T$ be a martingale difference sequence adapted to a filtration $\{\mathcal{F}_t\}_{t=0}^T$. That is, X_t is $\mathcal{F}_t$-measurable, and $\mathbb{E}[X_t|\mathcal{F}_{t-1}] = 0$. For any $\delta \in (0,1)$, we have*

$$\mathbb{P}\left[\left|\frac{1}{T}\sum_{t=1}^T X_t\right| \le \sqrt{\frac{2\log(2/\delta)}{T}}\right] \ge 1-\delta.$$

In particular, for any independent but not necessarily identically distributed random variables $Y_1, \ldots, Y_T$, it holds that

$$\mathbb{P}\left[\left|\frac{1}{T}\sum_{t=1}^T Y_t - \frac{1}{T}\sum_{t=1}^T \mathbb{E}[Y_t]\right| \le \sqrt{\frac{2\log(2/\delta)}{T}}\right] \ge 1-\delta. \tag{17}$$

Recall that we apply the Online Mirror Descent (OMD) algorithm to produce the set of weight vectors Θ in the offline algorithm. Next, we state the notion of strong convexity, which is often required for the mirror map Λ used in OMD:

DEFINITION 1. Let $\alpha \ge 0$. A function $\Lambda : D \to \mathbb{R}$ is α-strongly convex over the domain D w.r.t. the norm $\|\cdot\|_*$, if we have

$$\Lambda(\boldsymbol{u}) \ge \Lambda(\boldsymbol{w}) + \boldsymbol{z}^\top(\boldsymbol{u}-\boldsymbol{w}) + \frac{\alpha}{2}\|\boldsymbol{u}-\boldsymbol{w}\|_*^2,$$

for any $\boldsymbol{u}, \boldsymbol{w} \in D$ and $\boldsymbol{z} \in \underline{\partial}\Lambda(\boldsymbol{w})$.

Our analysis is further facilitated by the performance guarantee of the OMD algorithm, which is stated in the following proposition.

PROPOSITION 4 **(Nemirovskij and Yudin (1983), Shalev-Shwartz (2012))**. *Let $g^1, \ldots, g^T$ be a sequence of convex and R-Lipschitz function w.r.t. the norm $\|\cdot\|_*$ on domain D. In addition, let Λ be a 1-strongly convex function over the domain D w.r.t. $\|\cdot\|_*$. Consider the OMD algorithm with learning rate $\eta = \lambda/(R\sqrt{T})$, where $\lambda^2 := \max_{\boldsymbol{\theta}\in D}\{\Lambda(\boldsymbol{\theta})\} - \min_{\boldsymbol{\theta}\in D}\{\Lambda(\boldsymbol{\theta})\}$. For each $t = 1, \ldots, T$, perform*

$$\boldsymbol{\theta}^t \leftarrow \mathit{argmin}_{\boldsymbol{\theta}\in D}\left\{\eta \cdot \left[\sum_{q=1}^{t-1} \boldsymbol{\theta}^\top \boldsymbol{z}^q\right] + \Lambda(\boldsymbol{\theta})\right\}, \tag{18}$$

where $\boldsymbol{z}^q \in \underline{\partial} g^q(\boldsymbol{\theta}^q)$. The following inequality holds:

$$\frac{1}{T}\sum_{t=1}^T g^t(\boldsymbol{\theta}^t) - \min_{\boldsymbol{\theta}\in D}\left\{\frac{1}{T}\sum_{t=1}^T g^t(\boldsymbol{\theta})\right\} \le \frac{2\lambda R}{\sqrt{T}}.$$

Now, we are ready to present the proof of Theorem 1.

THEOREM 1. *Consider the unconstrained problem* $\mathrm{DP}_1(\phi, \mathbb{R}^K)$*, and implement the O2O algorithm framework with Algorithm 1 as the offline algorithm and Algorithm 2 as the online algorithm. (a) In the* SIMU *setting, with probability at least* $1-\delta$*, we have*

$$\mathrm{Reg}_1 \le \mathrm{Loss}^{(1)}_{\mathrm{SIMU}}(L) + \mathrm{Loss}^{(2)}_{\mathrm{SIMU}}(L) = \tilde{O}\left(L\,\|\mathbf{1}_K\|\left[\frac{1}{\sqrt{T}} + \frac{\lambda}{\sqrt{T}} + \frac{1}{\sqrt{\Gamma}}\right]\right), \tag{7}$$

where

$$\mathrm{Loss}^{(1)}_{\mathrm{SIMU}}(L) = L\,\|\mathbf{1}_K\|\left[\sqrt{\frac{2\log(6K/\delta)}{T}} + \sqrt{\frac{2\log(6K/\delta)}{\Gamma}}\right],$$
$$\mathrm{Loss}^{(2)}_{\mathrm{SIMU}}(L) = L\|\mathbf{1}_K\|\left[\frac{4\lambda + \sqrt{2\log(6K/\delta)}}{\sqrt{T}}\right].$$

(b) Consider the SAMP *setting, where Assumption 3 holds. Using* M *sample trajectories and a rounding parameter* ϵ *in round*$(\cdot,\epsilon)$*, with probability at least* $1-2\delta$*, we have*

$$\mathrm{Reg}_1 \le \mathrm{Loss}^{(1)}_{\mathrm{SAMP}}(L) + \mathrm{Loss}^{(2)}_{\mathrm{SAMP}}(L) = \tilde{O}\left(L\,\|\mathbf{1}_K\|\left[\frac{1}{\sqrt{T}} + \frac{\lambda}{\sqrt{T}} + \frac{1}{\sqrt{\Gamma}} + \sqrt{\frac{K}{M\Gamma}}\right] + c'K\epsilon\right), \tag{8}$$

where

$$\mathrm{Loss}^{(1)}_{\mathrm{SAMP}}(L) = L\,\|\mathbf{1}_K\|\left[\sqrt{\frac{2\log(6K/\delta)}{T}} + \sqrt{\frac{2\log(6K/\delta)}{\Gamma}} + \sqrt{\frac{2\log(6K/\delta) + 2K\log(2Lc/\epsilon)}{M\Gamma}}\right],$$
$$\mathrm{Loss}^{(2)}_{\mathrm{SAMP}}(L) = L\|\mathbf{1}_K\|\left[\frac{4\lambda + \sqrt{2\log(6K/\delta)}}{\sqrt{T}} + \sqrt{\frac{2\log(6K/\delta) + 2K\log(2Lc/\epsilon)}{M\Gamma}}\right] + 2c'K\epsilon.$$

We provide the proofs for the SIMU and SAMP settings in Sections A.3.1 and A.3.2, respectively.

A.3.1. Proof of Theorem 1 in SIMU setting. The proof involves two major steps. In Step 1, we show that

$$\phi\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma)\right) \ge \phi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\boldsymbol{\omega}^t)\right) - L\,\|\mathbf{1}_K\|\left[\sqrt{\frac{2\log(6K/\delta)}{T}} + \sqrt{\frac{2\log(6K/\delta)}{\Gamma}}\right] \tag{19}$$

holds with probability $\ge 1-(2\delta/3)$. In step 2, we show that

$$\phi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\boldsymbol{\omega}^t)\right) \ge \mathrm{opt}(\mathrm{UB}(\phi,\mathbb{R}^K)) - L\|\mathbf{1}_K\|\left[\frac{4\lambda + \sqrt{2\log(6K/\delta)}}{\sqrt{T}}\right] \tag{20}$$

holds with probability $\ge 1-\delta/3$. Altogether, the theorem is proved. More specifically, Step 1 bounds the regret due to the difference between the online and the offline phases, while Step 2 bounds the regret between the offline estimation and the benchmark.

Step 1: Proving (19). To start, for each $1 \le \gamma \le \Gamma$, we define

$$\boldsymbol{F}_\gamma := \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\omega}_\gamma\sim\Xi_\gamma}\left[\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma)\middle|\boldsymbol{\theta}^t\right].$$

In the conditional expectation above, the vector $\boldsymbol{\theta}^t$ is held deterministic, and the expectation is only over the randomness of $\boldsymbol{\omega}_\gamma \sim \Xi_\gamma$. We lower bound the reward gained in the online phase:

$$\begin{aligned}
&\phi\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma)\right) && (21)\\
\geq&\phi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\boldsymbol{\omega}^t)\right) - L\left\|\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma) - \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\boldsymbol{\omega}^t)\right\| \\
\geq&\phi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\boldsymbol{\omega}^t)\right) - \underbrace{L\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\boldsymbol{\omega}^t) - \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{F}_\gamma\right\|}_{(\ddagger)} - \underbrace{L\left\|\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}[\boldsymbol{F}_\gamma - \boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma)]\right\|}_{(\dagger)} \\
\geq&\phi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\boldsymbol{\omega}^t)\right) - L\left\|\mathbf{1}_K\right\|\left[\sqrt{\frac{2\log(6K/\delta)}{T}} + \sqrt{\frac{2\log(6K/\delta)}{\Gamma}}\right] \text{ w.p. } \geq 1-2\delta/3. && (22)
\end{aligned}$$

Step (22) is the essential step that helps us bound the generalization error incurred by employing the sampling oracle during the offline phase in the SIMU setting. We bound the terms $(\dagger,\ddagger)$ by applying the Hoeffding inequality provided in Proposition 3. Bounding the terms $(\dagger),(\ddagger)$ requires analyzing the randomness in the online and the offline algorithms, respectively.

Proving that $\Pr\left[(\dagger) \leq \|\mathbf{1}_K\|\sqrt{2\log(6K/\delta)/\Gamma}\right] \geq 1-\delta/3$**.** Consider an execution of the online algorithm (Algorithm 2), conditioned on the output $\Theta = \{\boldsymbol{\theta}^t\}_{t=1}^T$ by the offline algorithm (Algorithm 1), where Θ is fed to the online algorithm as the input. Observe that $\mathbb{E}[\boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma)|\Theta] = \boldsymbol{F}_\gamma$, where the expectation is taken over $\boldsymbol{x}_\gamma$ and $\boldsymbol{\omega}_\gamma$. Indeed, Line 5 in Algorithm 2 asserts that $\boldsymbol{x}_\gamma = \mathcal{O}(-\boldsymbol{\theta}_\gamma,\boldsymbol{\omega}_\gamma)$, and $\Pr[\boldsymbol{\theta}_\gamma = \boldsymbol{\theta}^t|\Theta] = 1/T$. Consequently,

$$\begin{aligned}
&\Pr\left[(\dagger) \leq \|\mathbf{1}_K\|\sqrt{\frac{2\log(6K/\delta)}{\Gamma}}\middle|\Theta\right] \\
\geq&\Pr\left[\left|\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}f_k(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma) - \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}F_{\gamma,k}\right| \leq \sqrt{\frac{2\log(6K/\delta)}{\Gamma}} \text{for each } 1\leq k\leq K \;\middle|\; \Theta\right] && (23)\\
\geq&1-\delta/3. && (24)
\end{aligned}$$

Step (23) is by the fact that, for each k, we have $f_k(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma) - F_{\gamma,k} \in [-1,1]$. Step (24) is by Proposition 3 and a union bound over $k\in\{1,\ldots,K\}$. Finally, by taking the expectation over Θ, we establish the bound for $(\dagger)$.

Proving that $\Pr\left[(\ddagger) \leq \|\mathbf{1}_K\|\sqrt{2\log(6K/\delta)/T}\right] \geq 1-\delta/3$**.** Consider an execution of the offline algorithm. For each $t\in\{1,\ldots,T\}$, set

$$\begin{aligned}
\mathcal{F}^{t-1} &:= \sigma(\{\boldsymbol{\omega}^s,\boldsymbol{\theta}^s\}_{s=1}^{t-1}\cup\{\boldsymbol{\theta}^t\}) \\
\boldsymbol{F}^t &:= \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\mathbb{E}_{\boldsymbol{\omega}_\gamma^t\sim\Xi_\gamma}\left[\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\boldsymbol{\omega}_\gamma^t),\boldsymbol{\omega}_\gamma^t)\mid\boldsymbol{\theta}^t\right] = \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\mathbb{E}_{\boldsymbol{\omega}_\gamma^t\sim\Xi_\gamma}\left[\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\boldsymbol{\omega}_\gamma^t),\boldsymbol{\omega}_\gamma^t)\mid\mathcal{F}^{t-1}\right], \\
\boldsymbol{Y}^t &:= \boldsymbol{f}(\boldsymbol{x}^t,\boldsymbol{\omega}^t) - \boldsymbol{F}^t.
\end{aligned}$$

The filtration $\mathcal{F}^{t-1}$ represents the information available to the DM at the end of time step $t-1$. To this end, we remark that $\boldsymbol{x}^s, \boldsymbol{f}(\boldsymbol{x}^s, \boldsymbol{\omega}^s)$ are $\mathcal{F}^{t-1}$-measurable for each $s \in \{1, \ldots, t-1\}$, since $\boldsymbol{x}^s = \mathcal{O}(-\boldsymbol{\theta}^s, \boldsymbol{\omega}^s)$. In fact, $\boldsymbol{\theta}^t$ is $\sigma(\{\boldsymbol{\omega}^s, \boldsymbol{\theta}^s\}_{s=1}^{t-1})$-measurable, but we include $\boldsymbol{\theta}^t$ in $\mathcal{F}^{t-1}$ explicitly to signal that $\boldsymbol{\theta}^t$ is held deterministic in the conditional expectation $\mathbb{E}[\cdot \mid \mathcal{F}^{t-1}]$.

Now, for each t, the random variable $\boldsymbol{Y}^t$ is $\mathcal{F}^t$-measurable, and $\mathbb{E}[\boldsymbol{Y}_t | \mathcal{F}^{t-1}] = \boldsymbol{0}$. Consequently, by applying the Hoeffding inequality with $X_t = Y_k^t$ for every $k \in \{1, \ldots, K\}$ and the filtration process $\{\mathcal{F}^t\}_{t=1}^T$, we know that $\Pr\left[|\sum_{t=1}^T Y_k^t/T| \le \sqrt{2\log(2K/\delta)/T}\right] \ge 1 - \delta/K$. Applying a union over $k \in \{1, \ldots, K\}$, we obtain the inequality $\Pr\left[\|\sum_{t=1}^T \boldsymbol{Y}^t/T\| \le \|\boldsymbol{1}_K\|\sqrt{2\log(6K/\delta)/T}\right] \ge 1 - \delta/3$. Finally, the desired bound on (‡) is shown by the fact that $\frac{1}{T}\sum_{t=1}^T \boldsymbol{F}^t = \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{F}_\gamma$.

Step 2: Proving (20). We focus on $\phi(\sum_{t=1}^T \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)/T)$, which is compared against the offline benchmark via the analytical tools from existing OMD results. To this end, recall $\lambda^2 = \max_{\boldsymbol{\theta} \in B_*(L)}\{\Lambda(\boldsymbol{\theta})\} - \min_{\boldsymbol{\theta} \in B_*(L)}\{\Lambda(\boldsymbol{\theta})\}$. We have

$$\begin{aligned}
\phi\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)\right) &= \min_{\boldsymbol{\theta} \in B_*(L)} \left\{\frac{1}{T}\sum_{t=1}^T \phi^*(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)\right\} \\
&\ge \frac{1}{T}\sum_{t=1}^T \left\{\phi^*(\boldsymbol{\theta}^t) - {\boldsymbol{\theta}^t}^\top \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)\right\} - \frac{4\lambda L\|\boldsymbol{1}_K\|}{\sqrt{T}} \qquad (25) \\
&= \frac{1}{T}\sum_{t=1}^T \left\{\phi^*(\boldsymbol{\theta}^t) - {\boldsymbol{\theta}^t}^\top \boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t, \boldsymbol{\omega}^t), \boldsymbol{\omega}^t)\right\} - \frac{4\lambda L\|\boldsymbol{1}_K\|}{\sqrt{T}}.
\end{aligned}$$

Next, we show that with probability at least $1 - \delta/3$, we have

$$\frac{1}{T}\sum_{t=1}^T \left\{\phi^*(\boldsymbol{\theta}^t) - {\boldsymbol{\theta}^t}^\top \boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t, \boldsymbol{\omega}^t), \boldsymbol{\omega}^t)\right\} \ge \mathrm{opt}(\mathrm{UB}(\phi, \mathbb{R}^K)) - L\|\boldsymbol{1}_K\|\frac{\sqrt{2\log(6K/\delta)}}{\sqrt{T}}. \qquad (26)$$

The inequality (26) is shown as follows:

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^T \left\{\phi^*(\boldsymbol{\theta}^t) - {\boldsymbol{\theta}^t}^\top \boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t, \boldsymbol{\omega}^t), \boldsymbol{\omega}^t)\right\} \\
\ge &\frac{1}{T}\sum_{t=1}^T \left\{\phi^*(\boldsymbol{\theta}^t) + \mathbb{E}_{\boldsymbol{\omega}_{1:\Gamma}^t \sim \Xi_{1:\Gamma}}\left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}(-\boldsymbol{\theta}^t)^\top \boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t, \boldsymbol{\omega}_\gamma^t), \boldsymbol{\omega}_\gamma^t) \mid \mathcal{F}^{t-1}\right]\right\} - L\|\boldsymbol{1}_K\|\frac{\sqrt{2\log(6K/\delta)}}{\sqrt{T}} \\
&(\text{w.p.} \ge 1 - \delta/3) \qquad (27) \\
\ge &\frac{1}{T}\sum_{t=1}^T \left\{\mathbb{E}_{\boldsymbol{\omega}_{1:\Gamma}^t \sim \Xi_{1:\Gamma}}\left[\phi^*(\boldsymbol{\theta}^t) + (-\boldsymbol{\theta}^t)^\top \left\{\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{H}_\gamma^*(\boldsymbol{\omega}_\gamma^t)\right\} \mid \mathcal{F}^{t-1}\right]\right\} - L\|\boldsymbol{1}_K\|\frac{\sqrt{2\log(6K/\delta)}}{\sqrt{T}} \qquad (28) \\
\ge &\frac{1}{T}\sum_{t=1}^T \min_{\boldsymbol{\theta}^* \in B_*(L)} \left\{\phi^*(\boldsymbol{\theta}^*) - {\boldsymbol{\theta}^*}^\top \left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}_\gamma^*\right)\right\} - L\|\boldsymbol{1}_K\|\frac{\sqrt{2\log(6K/\delta)}}{\sqrt{T}} \qquad (29) \\
= &\mathrm{opt}(\mathrm{UB}(\phi, \mathbb{R}^K)) - L\|\boldsymbol{1}_K\|\frac{\sqrt{2\log(6K/\delta)}}{\sqrt{T}}. \qquad (30)
\end{aligned}$$

Step (25) is by applying Proposition 4 on the series of functions $\{g_t\}_{t=1}^T$, defined as $g_t(\boldsymbol{\theta}) = \phi^*(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)$, with the mirror map Λ as stated in Algorithm 1. For every t, the function g_t is $2\|\mathbf{1}_K\|$-Lipschitz continuous w.r.t. $\|\cdot\|_*$. Step (27) is by applying the Hoeffding inequality on the random variables, similar to the analysis of (‡). In steps (28, 29), we recall the definitions of $(f^*, \boldsymbol{H}^*)$ from (15, 14) in the proof of Proposition 2. The inequality in (29) holds by the definition of the optimization oracle $\mathcal{O}(-\boldsymbol{\theta}^t, \boldsymbol{\omega}_\gamma^t) \in \operatorname{argmax}_{\boldsymbol{x}\in\mathcal{X}(\boldsymbol{\omega}_\gamma^t)}(-\boldsymbol{\theta}^t)^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\omega}_\gamma^t)$, and the fact that $\max_{\boldsymbol{x}\in\mathcal{X}(\boldsymbol{\omega}_\gamma^t)}(-\boldsymbol{\theta}^t)^\top \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\omega}_\gamma^t) = \max_{\boldsymbol{f}\in\operatorname{Conv}(\mathcal{F}(\omega_\gamma^t))}(-\boldsymbol{\theta}^t)^\top \boldsymbol{f}$.

Altogether, combining (22) with (26), Theorem 1 in the SIMU setting is proved. ■

A.3.2. Proof of Theorem 1 in SAMP setting. The proof of the SAMP shares some similarities with that of SIMU, but they differ in the analysis of the resampling procedure (Line 5 in Algorithm 1) and the rounding of $-\boldsymbol{\theta}^t$. The proof consists of two major steps. In step 1, we show that

$$\phi\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma)\right) \geq \phi\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)\right) - \mathrm{Loss}^{(1)}_{\mathrm{SAMP}}(L) \tag{31}$$

holds with probability at least $1-(2\delta/3)$. In step 2, we show that

$$\phi\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)\right) \geq \mathrm{opt}(\mathrm{UB}(\phi, \mathbb{R}^K)) - \mathrm{Loss}^{(2)}_{\mathrm{SAMP}}(L) \tag{32}$$

holds with probability at least $1-(4\delta/3)$.

Step 1: Proving (31). To start, for each $1 \leq \gamma \leq \Gamma$, we define

$$\boldsymbol{F}_\gamma := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\omega}_\gamma\sim\Xi_\gamma}\left[\boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t, \boldsymbol{\omega}_\gamma), \boldsymbol{\omega}_\gamma)\middle| - \tilde{\boldsymbol{\theta}}^t\right],$$
$$\hat{\boldsymbol{F}}_\gamma := \frac{1}{M\cdot T}\sum_{t=1}^{T}\sum_{m=1}^{M} \boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t, \boldsymbol{\omega}_\gamma^{(m)}), \boldsymbol{\omega}_\gamma^{(m)}),$$

where $-\tilde{\boldsymbol{\theta}}_t = \mathrm{round}(-\boldsymbol{\theta}_t, \epsilon)$ is the rounded version of the gradient vector $-\boldsymbol{\theta}_t$. We bound the reward gained in the online problem $\mathrm{DP}_1(\phi, \mathbb{R}^K))$ as follows:

$$\begin{aligned}
&\phi\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma)\right)\\
\geq&\phi\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)\right) - L\left\|\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma) - \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)\right\|\\
\geq&\phi\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)\right) - \underbrace{L\left\|\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t) - \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \hat{\boldsymbol{F}}_\gamma\right\|}_{(\ddagger')} - \underbrace{L\left\|\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \hat{\boldsymbol{F}}_\gamma - \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{F}_\gamma\right\|}_{(\P)} - \underbrace{L\left\|\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} [\boldsymbol{F}_\gamma - \boldsymbol{f}(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma)]\right\|}_{(\dagger')}\\
\geq&\phi\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)\right) - L\|\mathbf{1}_K\|\left[\sqrt{\frac{2\log(6K/\delta)}{T}} + \sqrt{\frac{2\log(6K/\delta) + 2K\log(2Lc/\epsilon)}{M\Gamma}} + \sqrt{\frac{2\log(6K/\delta)}{\Gamma}}\right].\\
&\text{(w.p. } 1-2\delta/3)
\end{aligned} \tag{33}$$

To establish (33), we provide high probability upper bounds for (†′), (¶), and (‡′), which are established as follows.

Proving $\Pr\left[(\dagger') \le \|\mathbf{1}_K\|\sqrt{2\log(6K/\delta)/\Gamma}\right] \ge 1-\delta/3$**.** The proof is by the same argument for bounding (†) in the Simu setting in Section A.3.1, using the Hoeffding inequality.

Proving $\Pr\left[(\P) \le \|\mathbf{1}_K\|\sqrt{\frac{2\log(6K/\delta)+2K\log(2Lc/\epsilon)}{M\Gamma}}\right] \ge 1-\delta/3$**.** To analyze (¶), we start by observing that

$$\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{F}_\gamma - \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\hat{\boldsymbol{F}}_\gamma = \frac{1}{T}\sum_{t=1}^{T}\mathsf{Diff}(-\tilde{\boldsymbol{\theta}}^t),$$

where

$$\mathsf{Diff}(-\tilde{\boldsymbol{\theta}}) = \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\mathbb{E}_{\boldsymbol{\omega}_\gamma\sim\Xi_\gamma}\left[\boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}},\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma)\,|\,-\tilde{\boldsymbol{\theta}}\right] - \frac{1}{M\Gamma}\sum_{m=1}^{M}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}},\boldsymbol{\omega}_\gamma^{(m)}),\boldsymbol{\omega}_\gamma^{(m)}). \tag{34}$$

By the rounding procedure and Assumption 3, we know that for all t, $-\tilde{\boldsymbol{\theta}}^t \in \mathsf{Grid}_\epsilon(L)$, where

$$\mathsf{Grid}_\epsilon(L) = \{\lceil -cL/\epsilon\rceil\epsilon, (\lceil -cL/\epsilon\rceil+1)\epsilon, \ldots -\epsilon, 0, \epsilon, \ldots, \lceil cL/\epsilon\rceil\epsilon\}^K$$

is a set of discrete grid points in the cube $[-cL, cL+\epsilon)^K$, positioned with a distance of ϵ apart in each of the K canonical directions in $\mathbb{R}^K$. Consequently, by applying the Chernoff inequality (17), with a union bound over all $-\tilde{\boldsymbol{\theta}} \in \mathsf{Grid}_\epsilon(L)$ as well as all K coordinates, we derive that

$$\begin{aligned}
&\Pr\left(\left\|\mathsf{Diff}(-\tilde{\boldsymbol{\theta}})\right\| \le \|\mathbf{1}_K\|\sqrt{\frac{2\log(2K(2Lc/\epsilon)^K/\delta)}{M\Gamma}} \text{ for all } -\tilde{\boldsymbol{\theta}} \in \mathsf{Grid}_\epsilon(L)\right)\\
\ge& 1 - \sum_{\tilde{\boldsymbol{\theta}}\in\mathsf{Grid}_\epsilon(L)}\frac{\delta}{(2Lc/\epsilon)^K} \ge 1-\delta,
\end{aligned} \tag{35}$$

since we have $|\mathsf{Grid}_\epsilon(L)| \le (2Lc/\epsilon)^K$. Consequently, we have

$$\begin{aligned}
&\Pr\left((\P) \le \|\mathbf{1}_K\|\sqrt{\frac{2\log(6K/\delta)+2K\log(2Lc/\epsilon)}{M\Gamma}}\right)\\
\ge&\Pr\left(\frac{1}{T}\sum_{t=1}^{T}\|\mathsf{Diff}(-\tilde{\boldsymbol{\theta}}^t)\| \le \|\mathbf{1}_K\|\sqrt{\frac{2\log(6K(2Lc/\epsilon)^K/\delta)}{M\Gamma}}\right)\\
\ge&\Pr\left(\left\|\mathsf{Diff}(-\tilde{\boldsymbol{\theta}})\right\| \le \|\mathbf{1}_K\|\sqrt{\frac{2\log(6K(2Lc/\epsilon)^K/\delta)}{M\Gamma}} \text{ for all } -\tilde{\boldsymbol{\theta}} \in \mathsf{Grid}_\epsilon(L)\right) \ge 1-\delta/3.
\end{aligned}$$

Proving $\Pr\left[(\ddagger') \le \|\mathbf{1}_K\|\sqrt{2\log(6K/\delta)/T}\right] \ge 1-\delta/3$**.** While the analysis shares some similarities with (‡) in Simu, the considered filtration here is different due to the bootstrapping in Samp that differs from Simu. Consider an execution of the offline algorithm. We define $\hat{\Xi}^{(M)}$ as the uniform distribution over the $M\Gamma$ elements in the (multi-)set $\{\boldsymbol{\omega}_{1:\Gamma}^{(m)}\}_{m=1}^M = \{\boldsymbol{\omega}_1^{(1)}, \ldots, \boldsymbol{\omega}_\Gamma^{(1)}, \ldots, \boldsymbol{\omega}_1^{(M)}, \ldots, \boldsymbol{\omega}_\Gamma^{(M)}\}$, which consists of the $M\Gamma$ realized scenarios in the M samples $\{\boldsymbol{\omega}_{1:\Gamma}^{(m)}\}_{m=1}^M$, endowed in the Samp

setting. By the algorithm design, we know that $\boldsymbol{\omega}^1, \ldots, \boldsymbol{\omega}^T$ are i.i.d. with the common distribution $\hat{\Xi}^{(M)}$ under SAMP. Now, for each $t \in \{1, \ldots, T\}$, consider

$$\begin{aligned}
\hat{\mathcal{F}}^{t-1} &:= \sigma(\{\boldsymbol{\omega}_{1:\Gamma}^{(m)}\}_{m=1}^{M} \cup \{\boldsymbol{\omega}^s, \boldsymbol{\theta}^s\}_{s=1}^{t-1} \cup \{\boldsymbol{\theta}^t\}), \\
\hat{\boldsymbol{F}}^t &:= \mathbb{E}_{\boldsymbol{\omega}^t \sim \hat{\Xi}^{(M)}} \left[\boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t, \boldsymbol{\omega}^t), \boldsymbol{\omega}^t) \mid -\boldsymbol{\theta}^t \right] \\
&= \mathbb{E}_{\boldsymbol{\omega}^t \sim \hat{\Xi}^{(M)}} \left[\boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t, \boldsymbol{\omega}^t), \boldsymbol{\omega}^t) \mid \hat{\mathcal{F}}^{t-1} \right] = \frac{1}{M \cdot \Gamma} \sum_{\gamma=1}^{\Gamma} \sum_{m=1}^{M} \boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t, \boldsymbol{\omega}_\gamma^{(m)}), \boldsymbol{\omega}_\gamma^{(m)}), \\
\hat{\boldsymbol{Y}}^t &:= \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t) - \hat{\boldsymbol{F}}^t.
\end{aligned}$$

Recall that $\boldsymbol{x}^t = \mathcal{O}(-\tilde{\boldsymbol{\theta}}^t, \boldsymbol{\omega}^t)$, where $-\tilde{\boldsymbol{\theta}}^t = \text{round}(-\boldsymbol{\theta}^t, \epsilon)$ The filtration $\hat{\mathcal{F}}^{t-1}$ represents the information available to the DM at the end of time step $t-1$ under SAMP, which consists of the M samples and the observations received during the offline algorithm's time steps $1, \ldots, t-1$. Similar to the case in SIMU, $\boldsymbol{\theta}^t$ is $\sigma(\{\boldsymbol{\omega}_{1:\Gamma}^{(m)}\}_{m=1}^{M} \cup \{\boldsymbol{\omega}^s, \boldsymbol{\theta}^s\}_{s=1}^{t-1})$-measurable. For each $t \in \{1, \ldots, T\}$, the random variable $\hat{\boldsymbol{Y}}^t$ is $\hat{\mathcal{F}}^t$-measurable, and $\mathbb{E}_{\boldsymbol{\omega}^t \sim \hat{\Xi}^{(M)}}[\hat{\boldsymbol{Y}}_t | \hat{\mathcal{F}}^{t-1}] = \mathbf{0}$. By applying the Hoeffding inequality with $X_t = \hat{Y}_k^t$ for every $k \in \{1, \ldots, K\}$ and the filtration process $\{\hat{\mathcal{F}}^t\}_{t=1}^T$, we know that $\Pr\left[|\sum_{t=1}^T \hat{Y}_k^t / T| \le \sqrt{2\log(6K/\delta)/T} \right] \ge 1 - \delta/(3K)$. Applying a union over $k \in \{1, \ldots, K\}$, we obtain $\Pr[\|\sum_{t=1}^T \hat{\boldsymbol{Y}}^t / T \le \|\mathbf{1}_K\| \sqrt{2\log(6K/\delta)/T}] \ge 1 - \delta/3$. Finally, the required bound on (‡′) is obtained by noting that $\frac{1}{T}\sum_{t=1}^T \hat{\boldsymbol{F}}^t = \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \hat{\boldsymbol{F}}_\gamma$.

Step 2: Proving (32). We continue by focusing on $\phi(\sum_{t=1}^T \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)/T)$. Similar to before, the analysis also involves bounding the error due to the limitation to M samples. Recall $\lambda^2 = \max_{\boldsymbol{\theta} \in B_*(L)}\{\Lambda(\boldsymbol{\theta})\} - \min_{\boldsymbol{\theta} \in B_*(L)}\{\Lambda(\boldsymbol{\theta})\}$, and recall the shorthand $-\tilde{\boldsymbol{\theta}}^t = \text{round}(-\boldsymbol{\theta}^t, \epsilon)$. We have

$$\begin{aligned}
&\phi\left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t) \right) \\
=& \min_{\boldsymbol{\theta} \in B_*(L)} \left\{ \frac{1}{T} \sum_{t=1}^{T} \phi^*(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t) \right\} \ge \frac{1}{T} \sum_{t=1}^{T} \left\{ \phi^*(\boldsymbol{\theta}^t) - \boldsymbol{\theta}^{t\top} \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t) \right\} - \frac{4\lambda L \|\mathbf{1}_K\|}{\sqrt{T}} \qquad (36) \\
=& \frac{1}{T} \sum_{t=1}^{T} \left\{ \phi^*(\boldsymbol{\theta}^t) - \boldsymbol{\theta}^{t\top} \boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t, \boldsymbol{\omega}^t), \boldsymbol{\omega}^t) \right\} - \frac{4\lambda L \|\mathbf{1}_K\|}{\sqrt{T}} \\
\ge& \frac{1}{T} \sum_{t=1}^{T} \left\{ \phi^*(\boldsymbol{\theta}^t) + (-\tilde{\boldsymbol{\theta}}^t)^\top \boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t, \boldsymbol{\omega}^t), \boldsymbol{\omega}^t) \right\} - c'\epsilon \|\mathbf{1}_K\| \|\mathbf{1}_K\|_* - \frac{4\lambda L \|\mathbf{1}_K\|}{\sqrt{T}} \qquad (37) \\
=& \frac{1}{T} \sum_{t=1}^{T} \left\{ \phi^*(\boldsymbol{\theta}^t) + (-\tilde{\boldsymbol{\theta}}^t)^\top \boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t, \boldsymbol{\omega}^t), \boldsymbol{\omega}^t) \right\} - c' K \epsilon - \frac{4\lambda L \|\mathbf{1}_K\|}{\sqrt{T}} \qquad (38)
\end{aligned}$$

Step (36) is by Proposition 4, which provides regret bounds on the AMD algorithm. Step (37) is by Assumption 3, which ensures that $\| - \boldsymbol{\theta}^t - (-\tilde{\boldsymbol{\theta}}^t)\|_* \le c'\epsilon \|\mathbf{1}_K\|_*$, and the model assumption that $\|\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\omega})\| \le \|\mathbf{1}_K\|$ for all $\boldsymbol{x}, \boldsymbol{\omega}$. To proceed from (38), we first observe that

$$\frac{1}{T} \sum_{t=1}^{T} (-\tilde{\boldsymbol{\theta}}^t)^\top \boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t, \boldsymbol{\omega}^t), \boldsymbol{\omega}^t) \ge \frac{1}{T} \sum_{t=1}^{T} \left[\frac{1}{M \cdot \Gamma} \sum_{\gamma=1}^{\Gamma} \sum_{m=1}^{M} (-\tilde{\boldsymbol{\theta}}^t)^\top \boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t, \boldsymbol{\omega}_\gamma^{(m)}), \boldsymbol{\omega}_\gamma^{(m)}) \right]$$

$$- L\|\mathbf{1}_K\|\sqrt{\frac{2\log(6K/\delta)}{T}} \text{ with probability at least } 1-\delta/3. \tag{39}$$

The inequality (39) follows from a similar logic to (‡′), where the former is shown by applying the Hoeffding inequality on $\{(-\tilde{\boldsymbol{\theta}}^t)^\top \hat{\boldsymbol{Y}}^t\}_{t=1}^T$, which is a martingale difference sequence w.r.t. the filtration $\{\mathcal{F}^t\}_{t=1}^T$, and $|(-\tilde{\boldsymbol{\theta}}^t)^\top \hat{\boldsymbol{Y}}^t| \le L\|\mathbf{1}_K\|$ with certainty. Next, we observe that, with certainty for each $t \in \{1,\ldots,T\}$, it holds that

$$\frac{1}{M\cdot\Gamma}\sum_{\gamma=1}^{\Gamma}\sum_{m=1}^{M} \boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t,\boldsymbol{\omega}_\gamma^{(m)}),\boldsymbol{\omega}_\gamma^{(m)}) = \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\mathbb{E}_{\boldsymbol{\omega}_\gamma\sim\Xi_\gamma}\left[\boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma)\middle| -\tilde{\boldsymbol{\theta}}^t\right] - \mathsf{Diff}(-\tilde{\boldsymbol{\theta}}^t),$$

where $\mathsf{Diff}(-\tilde{\boldsymbol{\theta}})$ is defined in (34). By invoking the concentration bound (35) on $\mathsf{Diff}(-\tilde{\boldsymbol{\theta}})$ that holds simultaneously for all $-\tilde{\boldsymbol{\theta}} \in \mathsf{Grid}_\epsilon(L)$, we know that

$$\begin{aligned}
&\frac{1}{M\cdot\Gamma}\sum_{\gamma=1}^{\Gamma}\sum_{m=1}^{M} \boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t,\boldsymbol{\omega}_\gamma^{(m)}),\boldsymbol{\omega}_\gamma^{(m)}) \\
\ge&\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\mathbb{E}_{\boldsymbol{\omega}_\gamma\sim\Xi_\gamma}\left[\boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma)\middle| -\tilde{\boldsymbol{\theta}}^t\right] - L\|\mathbf{1}_K\|\sqrt{\frac{2\log(6K/\delta)+2K\log(2Lc/\epsilon)}{M\Gamma}}
\end{aligned} \tag{40}$$

for all $t \in \{1,\ldots,T\}$ holds with probability at least $1-\delta$. Combining (39, 40) provides us with a key inequality, which holds with probability at least $1-(4\delta/3)$, to proceed from (38):

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}(-\tilde{\boldsymbol{\theta}}^t)^\top \boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t,\boldsymbol{\omega}^t),\boldsymbol{\omega}^t) \ge \frac{1}{T}\sum_{t=1}^{T}(-\tilde{\boldsymbol{\theta}}^t)^\top\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\mathbb{E}_{\boldsymbol{\omega}_\gamma\sim\Xi_\gamma}\left[\boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma)\middle| -\tilde{\boldsymbol{\theta}}^t\right]\right) \\
&- L\|\mathbf{1}_K\|\left\{\sqrt{\frac{2\log(6K/\delta)+2K\log(2Lc/\epsilon)}{M\Gamma}}+\sqrt{\frac{2\log(6K/\delta)}{T}}\right\}.
\end{aligned} \tag{41}$$

Further combining (38, 41), so far we have

$$\begin{aligned}
&\phi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\boldsymbol{\omega}^t)\right) \ge \frac{1}{T}\sum_{t=1}^{T}\left\{\phi^*(\boldsymbol{\theta}^t)+\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\mathbb{E}_{\boldsymbol{\omega}_\gamma\sim\Xi_\gamma}\left[(-\tilde{\boldsymbol{\theta}}^t)^\top\boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma)\middle| -\tilde{\boldsymbol{\theta}}^t\right]\right)\right\} \\
&- c'K\epsilon - L\|\mathbf{1}_K\|\left\{\sqrt{\frac{2\log(6K/\delta)+2K\log(2Lc/\epsilon)}{M\Gamma}}+\sqrt{\frac{2\log(6K/\delta)}{T}}+\frac{4\lambda}{\sqrt{T}}\right\}
\end{aligned} \tag{42}$$

with probability at least $1-(4\delta/3)$. Now, for any t and γ, and any realization of $-\tilde{\boldsymbol{\theta}}^t$ and $\boldsymbol{\omega}_\gamma$, we have the following with certainty

$$\begin{aligned}
(-\tilde{\boldsymbol{\theta}}^t)^\top\boldsymbol{f}(\mathcal{O}(-\tilde{\boldsymbol{\theta}}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma) &\ge (-\tilde{\boldsymbol{\theta}}^t)^\top\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma) \\
&\ge (-\boldsymbol{\theta}^t)^\top\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma) - \|-\tilde{\boldsymbol{\theta}}^t-(-\boldsymbol{\theta}^t)\|_*\cdot\|\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma)\| \\
&\ge (-\boldsymbol{\theta}^t)^\top\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma) - c'K\epsilon,
\end{aligned} \tag{43}$$

where the last step (43) follows the same reasoning as (37). Plugging (43) into (42) gives

$$\phi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\boldsymbol{\omega}^t)\right) \geq \frac{1}{T}\sum_{t=1}^{T}\left\{\phi^*(\boldsymbol{\theta}^t)+\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\mathbb{E}_{\boldsymbol{\omega}_\gamma\sim\Xi_\gamma}\left[(-\boldsymbol{\theta}^t)^\top\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma)\middle|-\boldsymbol{\theta}^t\right]\right)\right\}$$
$$-2c'K\epsilon - L\|\mathbf{1}_K\|\left[\sqrt{\frac{2\log(6K/\delta)+2K\log(2Lc/\epsilon)}{M\Gamma}}+\sqrt{\frac{2\log(6K/\delta)}{T}}+\frac{4\lambda}{\sqrt{T}}\right] \tag{44}$$

with probability at least $1-(4\delta/3)$. Lastly, by repeating the arguments with steps (27-30) in the analysis for SIMU, we derive that

$$\frac{1}{T}\sum_{t=1}^{T}\left\{\phi^*(\boldsymbol{\theta}^t)+\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\mathbb{E}_{\boldsymbol{\omega}_\gamma\sim\Xi_\gamma}\left[(-\boldsymbol{\theta}^t)^\top\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma)\middle|-\boldsymbol{\theta}^t\right]\right)\right\} \geq \mathrm{opt}(\mathrm{UB}(\phi,\mathbb{R}^K)) \tag{45}$$

with certainty. Altogether, combining (45) with (44), **Step 2** is established, and Theorem 1 for the SAMP setting is proved. ■

A.4. Proof of Theorem 2

THEOREM 2. *Consider the general problem* $\mathrm{DP}_1(\phi,S)$*, and implement the O2O algorithm framework with Algorithm 3 as the offline algorithm and Algorithm 2 as the online algorithm. It holds that* $\mathrm{LCB}\leq\mathrm{opt}(\mathrm{UB}(\phi,S))\leq\mathrm{UCB}$*. (a) In the* SIMU *setting, with probability* $\geq 1-2\delta$*, we have*

$$\mathrm{Reg}_1 = \tilde{O}\left(L\|\mathbf{1}_K\|\left[\frac{\lambda}{\sqrt{T}}+\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{\Gamma}}\right]\right),\quad \mathrm{Reg}_2 = \tilde{O}\left(\|\mathbf{1}_K\|\left[\frac{\lambda}{\sqrt{T}}+\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{\Gamma}}\right]\right).$$

More precisely, we have $\mathrm{Reg}_1\leq\mathrm{Loss}^{(1)}_{\mathrm{SIMU}}(L)+\mathrm{Loss}^{(2)}_{\mathrm{SIMU}}(L+1)$ *and* $\mathrm{Reg}_2\leq\mathrm{Loss}^{(1)}_{\mathrm{SIMU}}(1)+\mathrm{Loss}^{(2)}_{\mathrm{SIMU}}(1)$*. (b) In the* SAMP *setting, with probability* $\geq 1-2\delta$*, we have*

$$\mathrm{Reg}_1 = \tilde{O}\left(L\|\mathbf{1}_K\|\left[\frac{\lambda}{\sqrt{T}}+\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{\Gamma}}+\sqrt{\frac{K}{M\Gamma}}\right]+c'K\epsilon\right),$$
$$\mathrm{Reg}_2 = \tilde{O}\left(\|\mathbf{1}_K\|\left[\frac{\lambda}{\sqrt{T}}+\frac{1}{\sqrt{T}}+\frac{1}{\sqrt{\Gamma}}+\sqrt{\frac{K}{M\Gamma}}\right]+c'K\epsilon\right).$$

More precisely, $\mathrm{Reg}_1\leq L\cdot\mathrm{Loss}^{(1)}_{\mathrm{SAMP}}(1)+(L+1)\cdot\mathrm{Loss}^{(2)}_{\mathrm{SAMP}}(1)$*, and* $\mathrm{Reg}_2\leq\mathrm{Loss}^{(1)}_{\mathrm{SAMP}}(1)+\mathrm{Loss}^{(2)}_{\mathrm{SAMP}}(1)$*.*

PROOF. The proof is based on the claim that the **For** loop breaks for some $n\in\{0,\ldots,N\}$. We first provide some set-up to state the claim in more details. By the assumption on LCB and UCB, there exists $n^*\in\{0,\ldots,N\}$ such that

$$\mathrm{UCB}-(n^*+1)\varepsilon\leq\mathrm{opt}(\mathrm{UB}(\phi,S))\leq\mathrm{UCB}-n^*\varepsilon.$$

We claim that when $n=n_\epsilon+1$, the **For** loop breaks with probability at least $1-(4\delta/3)$. Now, from (20, 32) in our analysis of the O2O framework, we first observe that, for any fixed n, we have

$$\breve{\phi}_n\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\boldsymbol{\omega}^t)\right)\geq\mathrm{opt}(\mathrm{UB}(\breve{\phi}_n,\mathbb{R}^K))-\varepsilon\ \text{w.p.}\geq 1-(4\delta/3) \tag{46}$$

in each of the SIMU or SAMP setting. Next, we observe that $\text{opt}(\text{UB}(\breve{\phi}_{n^*+1}, \mathbb{R}^K)) = 0$. Recall from the proof of Proposition 2 for the optimal solution $(\boldsymbol{f}_\gamma^*)_{\gamma=1}^\Gamma$ to $\text{UB}(\phi, S)$. Our observation is immediate by observing that $\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}_\gamma^* \in \breve{S}_{n^*+1}$, since

$$\phi\left(\frac{1}{\Gamma}\sum_{\gamma=1}^\Gamma \boldsymbol{f}_\gamma^*\right) = \text{opt}(\text{UB}(\phi, S)) \geq \text{UCB} - (n^*+1)\varepsilon, \quad d\left(\frac{1}{\Gamma}\sum_{t=1}^\Gamma \boldsymbol{f}_\gamma^*, S\right) = 0,$$

by the feasibility of $(\boldsymbol{f}_\gamma^*)_{\gamma=1}^\Gamma$ to $\text{UB}(\phi, S)$. Combining $\text{opt}(\text{UB}(\breve{\phi}_{n^*+1}, \mathbb{R}^K)) = 0$ and (46), we have

$$\begin{aligned}\phi\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t)\right) &\geq \text{UCB} - (n^*+1)\varepsilon - L\varepsilon \geq \text{opt}(\text{UB}(\phi, S)) - (L+1)\varepsilon \\ &= \begin{cases} \text{opt}(\text{UB}(\phi, S)) - \text{Loss}_{\text{Simu}}^{(2)}(L+1) & \text{in SIMU} \\ \text{opt}(\text{UB}(\phi, S)) - (L+1)\cdot \text{Loss}_{\text{Samp}}^{(2)}(1) & \text{in SAMP} \end{cases}, \end{aligned} \tag{47}$$

$$d\left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t), S\right) \leq \varepsilon = \begin{cases} \text{Loss}_{\text{Simu}}^{(2)}(1) & \text{in SIMU} \\ \text{Loss}_{\text{Samp}}^{(2)}(1) & \text{in SAMP} \end{cases}, \tag{48}$$

with probability at least $1-(4\delta/3)$ in each case. Lastly, we recall the upper bounds

$$\left\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t) - \frac{1}{\Gamma}\sum_{\gamma=1}^\Gamma \boldsymbol{f}(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma)\right\| \leq \text{Loss}_{\text{SIMU}}^{(1)}(1)$$

in SIMU via (24), and

$$\left\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{f}(\boldsymbol{x}^t, \boldsymbol{\omega}^t) - \frac{1}{\Gamma}\sum_{\gamma=1}^\Gamma \boldsymbol{f}(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma)\right\| \leq \text{Loss}_{\text{SAMP}}^{(1)}(1)$$

in SAMP via (33). In each of the two cases of SIMU and SAMP, the respective inequality above holds with probability at least $1-(2\delta/3)$. In each of the two cases, L is set to be 1 since $\breve{\phi}_n$ is 1-Lipschitz continuous w.r.t. $\|\cdot\|$. Altogether, the theorem is proved. Lastly, we verify the statement on the default setting of UCB and LCB. By the L-Lipschitz continuity of ϕ, we know that

$$\phi\left(\frac{1}{\Gamma}\sum_{\gamma=1}^\Gamma \boldsymbol{f}_\gamma^*\right) \geq \phi(\boldsymbol{0}_K) - L\left\|\frac{1}{\Gamma}\sum_{\gamma=1}^\Gamma \boldsymbol{f}_\gamma^*\right\| \geq \phi(\boldsymbol{0}_K) - L\|\boldsymbol{1}_K\| = \text{LCB},$$

and that

$$\phi\left(\frac{1}{\Gamma}\sum_{\gamma=1}^\Gamma \boldsymbol{f}_\gamma^*\right) \leq \phi(\boldsymbol{0}_K) + L\left\|\frac{1}{\Gamma}\sum_{\gamma=1}^\Gamma \boldsymbol{f}_\gamma^*\right\| \leq \phi(\boldsymbol{0}_K) + L\|\boldsymbol{1}_K\| = \text{UCB}.$$

■

A.5. Proof of Theorem 3

THEOREM 3. *Suppose Assumption 4 holds. Algorithm 5 satisfies all $K-1$ resource constraints with certainty, and with probability at least $1-3\delta$, it achieves*

$$\sum_{\gamma=1}^\Gamma r(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma) \geq \left(1 - \frac{\xi + \sqrt{(1/\Gamma)\log(K/\delta)}}{\xi + c_{min}} - \frac{\xi + \sqrt{(1/\Gamma)\log(K/\delta)}}{\text{opt}(\text{DP}_1(\phi, S))}\right)\Gamma \cdot \text{opt}(\text{DP}_1(\phi, S)) \tag{10}$$

$$\geq [1 - O(\xi)] \times \textsf{optimum} \geq \textsf{optimum} - O(\xi \cdot \Gamma) \tag{11}$$

where ξ in (10) is set differently according to the cases of Simu *or* Samp *as specified in Line 3 of Algorithm 4, while ϕ, S in (10) are as defined in (9). The quantity* optimum *in (11) is the expected optimum of the non-stationary online resource allocation problem.*

Proof. We show that with probability $\geq 1-3\delta$, the breaking of **for** loop by Line 10 in Algorithm 4 does not occurs, and the stated inequalities in the theorem hold.

Firstly, according to the performance guarantee of our O2O framework in Section 3.4.1 on $\mathrm{DP}_1(\phi, S)$, the online performance, disregarding the thinning by $\{B(\gamma)\}_{\gamma=1}^{\Gamma}$ and the breaking of **for** loop by Line 10 in Algorithm 4, satisfies

$$\sum_{\gamma=1}^{\Gamma} r(\bar{\boldsymbol{x}}_\gamma, \boldsymbol{\theta}_\gamma) \geq \Gamma\left(\mathrm{opt}(\mathrm{DP}_1(\phi, S)) - \xi\right), \text{ and } \sum_{\gamma=1}^{\Gamma} a_k(\bar{\boldsymbol{x}}_\gamma, \boldsymbol{\theta}_\gamma) \leq \Gamma\left(c_k + \xi\right) \text{ for all } k \in \{1, \ldots, K-1\}$$

with probability $\geq 1-2\delta$, where ξ aligns with Line 3 in Algorithm 4. Considering the thinning conducted in the online phase, Algorithm 4 without the breaking of **for** loop by Line 10 earns a reward of $\sum_{\gamma=1}^{\Gamma} B(\gamma) r(\bar{\boldsymbol{x}}_\gamma, \boldsymbol{\theta}_\gamma)$, while consumes $\sum_{\gamma=1}^{\Gamma} B(\gamma) a_k(\bar{\boldsymbol{x}}_\gamma, \boldsymbol{\theta}_\gamma)$ for each k during the online horizon. By a union bound, with probability $\geq 1-3\delta$, the resource consumption on resource k satisfies

$$\sum_{\gamma=1}^{\Gamma} B(\gamma) a_k(\bar{\boldsymbol{x}}_\gamma, \boldsymbol{\theta}_\gamma) \leq p\Gamma(c_k + \xi) + \sqrt{\Gamma \log \frac{K}{\delta}} \leq c_k \Gamma,$$

where the first inequality is by the Chernoff bound, and the second is by the definition of p, which falls within the range $[0,1]$ by Assumption 4. The above inequality implies that, with probability $\geq 1-3\delta$, the breaking of **for** loop by Line 10 in Algorithm 4 is never invoked, and that

$$\sum_{\gamma=1}^{\Gamma} a_k(\boldsymbol{x}_\gamma, \boldsymbol{\theta}_\gamma) = \sum_{\gamma=1}^{\Gamma} B(\gamma) a_k(\bar{\boldsymbol{x}}_\gamma, \boldsymbol{\theta}_\gamma) \text{ for all } k \ , \ \sum_{\gamma=1}^{\Gamma} r(\boldsymbol{x}_\gamma, \boldsymbol{\theta}_\gamma) = \sum_{\gamma=1}^{\Gamma} B(\gamma) r(\bar{\boldsymbol{x}}_\gamma, \boldsymbol{\theta}_\gamma).$$

Moreover, the total reward obtained is at least

$$\sum_{\gamma=1}^{\Gamma} B(\gamma) r(\bar{\boldsymbol{x}}_\gamma, \boldsymbol{\theta}_\gamma) \geq p\Gamma(\mathrm{opt}(\mathrm{DP}_1(\phi, S)) - \xi) - \sqrt{\Gamma \log \frac{K}{\delta}},$$

which is equal to the bound stated in the theorem. Altogether, we complete the proof. ■

B. Supplementary Details for the Implementability of the O2O Framework

In this section, we provide supplementary details on the implementability of our O2O framework.

Proof of (4) for Line 7 in Algorithm 1. The set inclusion above follows from $\max_{\boldsymbol{w} \in [0,1]^K} \{(\boldsymbol{\theta}^t)^\top \boldsymbol{w} - d(\boldsymbol{w}, S)\} = \max_{\boldsymbol{w} \in S} \{(\boldsymbol{\theta}^t)^\top \boldsymbol{w}\}$. The equality can be seen by the following. Let $\boldsymbol{w}^*$ be an optimal solution to $\max_{\boldsymbol{w} \in [0,1]^K} \{(\boldsymbol{\theta}^t)^\top \boldsymbol{w} - d(\boldsymbol{w}, S)\}$, and denote $\pi_S(\boldsymbol{w})$ as the projection of $\boldsymbol{w}$ to S, so that $d(\boldsymbol{w}, S) = \|\boldsymbol{w} - \pi_S(\boldsymbol{w})\|$. Now, we have $(\boldsymbol{\theta}^t)^\top \pi_S(\boldsymbol{w}^*) \geq (\boldsymbol{\theta}^t)^\top \boldsymbol{w}^* - |(\boldsymbol{\theta}^t)^\top (\boldsymbol{w}^* - \pi_S(\boldsymbol{w}^*))|$, and $|(\boldsymbol{\theta}^t)^\top (\boldsymbol{w}^* - \pi_S(\boldsymbol{w}^*))| \leq \|\boldsymbol{\theta}^t\|_* \cdot \|\boldsymbol{w}^* - \pi_S(\boldsymbol{w}^*)\| \leq d(\boldsymbol{w}, S)$, since $\|\boldsymbol{\theta}^t\|_* \leq$

1 by the 1-Lipschitz continuity of $\phi(\boldsymbol{w}) = -d(\boldsymbol{w}, S)$. Consequently, we have $\max_{\boldsymbol{w}\in S}\{(\boldsymbol{\theta}^t)^\top \boldsymbol{w}\} \geq (\boldsymbol{\theta}^t)^\top \pi_S(\boldsymbol{w}^*) \geq \max_{\boldsymbol{w}\in[0,1]^K}\{(\boldsymbol{\theta}^t)^\top \boldsymbol{w} - d(\boldsymbol{w}, S)\}$, and then the desired equality holds since clearly we have $\max_{\boldsymbol{w}\in[0,1]^K}\{(\boldsymbol{\theta}^t)^\top \boldsymbol{w} - d(\boldsymbol{w}, S)\} \geq \max_{\boldsymbol{w}\in S}\{(\boldsymbol{\theta}^t)^\top \boldsymbol{w}\}$.

Details for Line 8 in Example 1. Recall that this is the case with the ϕ being L-Lipschitz continuous w.r.t. $\|\cdot\| = \|\cdot\|_2$. In coherence with Proposition 4, we incorporate a mirror map Λ that is 1-strongly convex w.r.t. the dual norm $\|\cdot\|_* = \|\cdot\|_2$ over the domain $D = B_*(L)$, the Euclidean ball of radius L. An eligible candidate is $\Lambda_2(\boldsymbol{\theta}) := \boldsymbol{\theta}^\top\boldsymbol{\theta}/2 + I_{B(L,\|\cdot\|_*)}(\boldsymbol{\theta})$. Next, in order to achieve the convergence postulated in Proposition 4, we set the learning rate $\eta_2 = \lambda_2/(R_2\sqrt{T})$, where $R_2 = 2\|\mathbf{1}_K\|_2 = 2\sqrt{K}$, and $\lambda_2 = L/\sqrt{2}$. Consequently, we have $\eta_2 = L/(\sqrt{8KT})$. Altogether, under the specified learning rate η_2, mirror map Λ_2 and domain $D = B_*(L)$, the gradient update rule in Line 8 of Algorithm 1 is:

$$\begin{aligned}
\boldsymbol{\theta}^{t+1} =& \operatorname{argmin}_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|_2\leq L}\left\{\frac{L}{\sqrt{8KT}}\left[\sum_{q=1}^{t}(\boldsymbol{z}^q)^\top\boldsymbol{\theta}\right] + \frac{1}{2}\boldsymbol{\theta}^\top\boldsymbol{\theta}\right\} \\
=& \operatorname{argmin}_{\boldsymbol{\theta}:\|\boldsymbol{\theta}\|_2\leq L}\left\{\left\|\boldsymbol{\theta} + \frac{L}{\sqrt{8KT}}\sum_{q=1}^{t}\boldsymbol{z}^q\right\|_2^2\right\} \\
=& -\frac{L\sum_{q=1}^{t}\boldsymbol{z}^q}{\max\left\{\sqrt{8KT}, \|\sum_{q=1}^{t}\boldsymbol{z}^q\|_2\right\}},
\end{aligned}$$

which establishes the closed form update rule (5).

We apply this update rule to solve the resource allocation problem with fill rate constraints discussed in Appendix D. By employing the optimization oracle $\mathcal{O}$, it is straightforward to see that the optimal allocation rule follows a priority rule according to the value of $\sum_{q=1}^{t}\boldsymbol{z}^q$. This reduces to the debt-based allocation rule proposed by Zhong et al. (2018). We also implement this update rule in the two applications discussed in Section 5.

Details on the Multiplicative Weight Update Rule under $\|\cdot\|_\infty$-Lipschitz Continuity for Line 8. In the stated set-up, it suffices to define a mirror map Λ_∞ that is 1-strongly convex w.r.t. to $\|\cdot\|_1$ over the domain $D = S_{\geq 0}(L, \|\cdot\|_1)$. An eligible candidate is the negative entropy function:

$$\Lambda_\infty(\boldsymbol{\theta}) = L\sum_{k=1}^{K}\theta_k\log\theta_k + I_{S_{\geq 0}(L,\|\cdot\|_1)}(\boldsymbol{\theta}),$$

where $0\log 0 := 0$. To achieve the convergence rate postulated in Proposition 4, we set the learning rate $\eta_\infty = \lambda_\infty/(R_\infty\sqrt{T})$, where $\lambda_\infty = L\sqrt{\log K}$ and $R_\infty = 2$. Consequently, we have $\eta_\infty = L\sqrt{\log K}/(2\sqrt{T})$. Altogether, the gradient update rule in Line 8 of Algorithm 1 has the following incarnation:

$$\boldsymbol{\theta}^{t+1} = \frac{(Le^{w_{t,1}}, \ldots, Le^{w_{t,K}})}{\sum_{k=1}^{K}e^{w_{t,k}}}, \text{ where } w_{t,k} := -\sqrt{\frac{\log K}{4T}}\sum_{q=1}^{t}z_k^q.$$

We apply this update rule to solve the resource allocation problem with budget constraints discussed in Section 4.3. By employing the optimization oracle $\mathcal{O}$, the optimal decision can be computed in closed form, eliminating the need for commercial solvers.

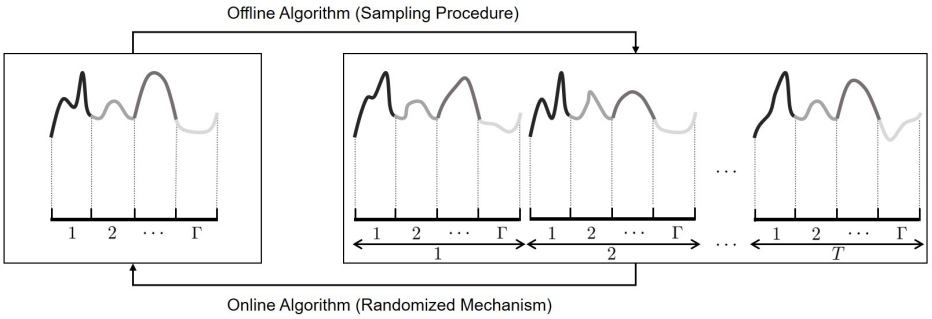
C. Supplementary Details for Model Extensions

In this section, we extend the O2O algorithm framework to address two additional online planning scenarios, including the case with correlated scenario distributions and the case with imperfect information.

C.1. Online Planning with Correlated Scenario Distributions

We extend the O2O algorithm framework to address the online planning problem with potentially correlated scenario distributions $\Xi_1, \ldots, \Xi_\Gamma$ in the SIMU setting. For ease of exposition, we focus on the non-constrained problem $\mathrm{DP}_1(\phi, \mathbb{R}^K)$, while the general problem $\mathrm{DP}_1(\phi, S)$ could be solved through a similar generalization as described in Section 3.4. The refined O2O framework, which is depicted in Figure 8, consists of Algorithm 6 as the offline algorithm, and Algorithm 2 (specific to the SIMU case) as the online algorithm.

Figure 8 Schematic drawing for the (refined) O2O framework.



Note. Under the sampling scheme, the stochastic processes over Γ periods are i.i.d. re-generated for T samples in the offline problem.

With a slight abuse of notation, we denote $\boldsymbol{\omega}_\gamma^t \sim \Xi_\gamma$ as the generated scenario for period γ at sample t, and $\boldsymbol{x}_\gamma^t$ as the corresponding solution. The offline algorithm is run before the DM encounters the online problem $\mathrm{DP}_1(\phi, \mathbb{R}^K)$. As described in Algorithm 6, the offline algorithm involves a refined sampling procedure (Line 3-10) to capture the correlations among $\Xi_1, \ldots, \Xi_\Gamma$ during the entire planning horizon, and produces a collection of weight vectors Θ, which are crucial for solving $\mathrm{DP}_1(\phi, \mathbb{R}^K)$. Through the refined sampling procedure, we also translate the non-stationary stochastic problem into a "multi-sample" optimization problem (cf. Figure 8). Notably, since the scenarios could be correlated, we cannot sample the scenario independently for each single period as described in Algorithm

1. To capture the correlations, we treat the entire planning problem over Γ periods as a "single" period problem and sample the scenarios $\boldsymbol{\omega}_{1:\Gamma}^t \sim \Xi_{1:\Gamma}$ for each $t = 1, \ldots, T$ (Line 4). Furthermore, from Line 5 to 7, we call the oracle $\mathcal{O}(-\boldsymbol{\theta}^t, \cdot)$ to solve the planning problem using the same $\boldsymbol{\theta}^t$ at each period $\gamma = 1, \ldots, \Gamma$. In Line 9, we also apply the OMD algorithm to solve this multi-sample problem. Although we change the way to sample the scenarios in the offline algorithm, we highlight that the online algorithm remains the same as in the case of independent distributions under Simu.

Algorithm 6 Extended offline algorithm for the case with correlated distributions.

1: **Input:** Number of iterations T; OMD parameters $\eta > 0$, $D \subseteq B_*(L)$, and $\Lambda : D \to \mathbb{R}$.
2: **Initialize:** Initial weight vector $\boldsymbol{\theta}^1 = \min_{\boldsymbol{\theta} \in D} \Lambda(\boldsymbol{\theta})$.
3: **for** $t = 1, \ldots, T$ **do**
4: Sample a sequence of random scenarios $\boldsymbol{\omega}_\gamma^t \sim \Xi_\gamma$ for $\gamma = 1, \ldots, \Gamma$.
5: **for** $\gamma = 1, \ldots, \Gamma$ **do**
6: Compute decision $\boldsymbol{x}_\gamma^t = \mathcal{O}(-\boldsymbol{\theta}^t, \boldsymbol{\omega}_\gamma^t)$, by calling the oracle $\mathcal{O}$ at each period.
7: **end for**
8: Compute gradient $\boldsymbol{z}^t = \nabla_{\boldsymbol{\theta}}[g^t(\boldsymbol{\theta}^t)]$, where $g^t(\boldsymbol{\theta}) := \phi^*(\boldsymbol{\theta}) + (-\boldsymbol{\theta})^\top \left[\frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma^t, \boldsymbol{\omega}_\gamma^t)\right]$.
9: Compute weight vector $\boldsymbol{\theta}^{t+1} = \operatorname{argmin}_{\boldsymbol{\theta} \in D} \left\{\eta \cdot \left[\sum_{q=1}^{t} \boldsymbol{z}^{q\top} \boldsymbol{\theta}\right] + \Lambda(\boldsymbol{\theta})\right\}$.
10: **end for**
11: **Output:** The collection of weight vectors $\Theta = \{\boldsymbol{\theta}^t\}_{t=1}^T$.

Next, we demonstrate the performance guarantee of this refined O2O algorithm on $\mathrm{DP}_1(\phi, \mathbb{R}^K)$ with correlated scenario distributions, in the form of regret bound. Different from previous analysis, we investigate the performance of our refined O2O algorithm in terms of the expected average value over the entire planning horizon:

$$\mathrm{Reg}_3 := \mathrm{opt}(\mathrm{UB}(\phi, S)) - \phi\left(\mathbb{E}\left[\frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma)\right]\right),$$

which captures the difference between the offline benchmark and the objective of the algorithm. Given the same algorithm, we note that $\phi(\mathbb{E}[\frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma)]) \geq \mathbb{E}[\phi(\frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma))]$, by the Jensen's inequality. Therefore, we have $\mathrm{Reg}_3 \leq \mathbb{E}[\mathrm{Reg}_1]$. In this sense, we remark that the performance guarantee using Reg_3 is relatively "weaker" compared to using Reg_1.

While the scenarios across different periods (within each sample t) could be correlated, the sequences $\boldsymbol{\omega}_{1:\Gamma}^t$ are independently generated for different samples. Therefore, we apply the concentration equality on the average reward vector $\frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma)$, instead of the vector $\boldsymbol{f}(\boldsymbol{x}_\gamma, \boldsymbol{\omega}_\gamma)$ at a single period. This modification is crucial to derive the performance guarantee as follows.

THEOREM 4. *Consider the unconstrained problem* $\mathrm{DP}_1(\phi, \mathbb{R}^K)$ *with correlated scenario distributions, and implement the refined O2O framework with Algorithm 6 as the offline algorithm and Algorithm 2 as the online algorithm. In the* SIMU *setting, with probability at least* $1-\delta$*, we have*

$$\mathrm{Reg}_3 = \tilde{O}\left(L\left\|\mathbf{1}_K\right\|\left(\frac{\lambda+1}{\sqrt{T}}\right)\right).$$

PROOF. We bound the expected average reward gained in the online problem $\mathrm{DP}_1(\phi, \mathbb{R}^K)$) as follows:

$$\begin{aligned}
&\phi\left(\mathbb{E}\left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma)\right]\right)\\
\geq&\phi\left(\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma^t,\boldsymbol{\omega}_\gamma^t)\right]\right) - L\underbrace{\left\|\mathbb{E}\left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma)\right] - \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{\Gamma}\sum_{t=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma^t,\boldsymbol{\omega}_\gamma^t)\right]\right\|}_{(\S)}\\
\geq&\phi\left(\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma^t,\boldsymbol{\omega}_\gamma^t)\right]\right) - L\left\|\mathbf{1}_K\right\|\left[\sqrt{\frac{2\log(4K/\delta)}{T}}\right] \text{ w.p. } 1-\delta/2. \qquad (49)
\end{aligned}$$

We prove that $\Pr\left[(\S) \leq \|\mathbf{1}_K\|\sqrt{2\log(4K/\delta)/\Gamma}\right] \geq 1-\delta/2$. Consider an execution of the online algorithm, conditioned on the output $\Theta = \{\boldsymbol{\theta}^t\}_{t=1}^T$ by the offline algorithm, where Θ is fed to the online algorithm as the input. We claim that

$$\mathbb{E}\left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma) \;\middle|\; \Theta\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma^t,\boldsymbol{\omega}_\gamma^t)\right],$$

where the expectation on the left hand-side is taken over all $\boldsymbol{x}_\gamma$ and $\boldsymbol{\omega}_\gamma$. The claim is readily justified by Line 5 in Algorithm 2, which asserts that $\boldsymbol{x}_\gamma = \mathcal{O}(-\boldsymbol{\theta}_\gamma, \boldsymbol{\omega}_\gamma)$ with $\Pr[\boldsymbol{\theta}_\gamma = \boldsymbol{\theta}^t|\Theta] = 1/T$. Using a similar argument in the proof of Theorem 1, we have

$$\begin{aligned}
&\Pr\left[(\S) \leq \|\mathbf{1}_K\|\sqrt{\frac{2\log(4K/\delta)}{T}} \;\middle|\; \Theta\right]\\
\geq&\Pr\left[\left|\frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}f_k(\boldsymbol{x}_\gamma^t,\boldsymbol{\omega}_\gamma^t) - \frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\mathbb{E}\left[f_k(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma)\right]\right)\right| \leq \sqrt{\frac{2\log(4K/\delta)}{T}} \text{for each } 1\leq k\leq K \;\middle|\; \Theta\right]\\
\geq&1-\delta/2. \qquad (50)
\end{aligned}$$

Step (50) is by Proposition 3 and a union bound over $k \in \{1,\ldots,K\}$. Finally, by taking the expectation over Θ, we establish the bound for (§).

We continue by focusing on $\phi(\frac{1}{T}\sum_{t=1}^{T}(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma^t,\boldsymbol{\omega}_\gamma^t)))$. To proceed, we use the same technique in the proof of Theorem 1 to compare the online solutions and certain offline benchmarks by considering the OMD procedure on the dual of the reward function. Recall the notation that $\lambda^2 = \max_{\boldsymbol{\theta}\in B_*(L)}\{\Lambda(\boldsymbol{\theta})\} - \min_{\boldsymbol{\theta}\in B_*(L)}\{\Lambda(\boldsymbol{\theta})\}$. We have

$$\phi\left(\frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma^t,\boldsymbol{\omega}_\gamma^t)\right]\right)$$

$$
\begin{aligned}
&= \min_{\boldsymbol{\theta}\in B_*(L)} \left\{ \frac{1}{T}\sum_{t=1}^{T} \phi^*(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma^t, \boldsymbol{\omega}_\gamma^t) \right] \right\} \\
&\geq \frac{1}{T}\sum_{t=1}^{T} \left\{ \phi^*(\boldsymbol{\theta}^t) - \boldsymbol{\theta}^{t\top} \left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma^t, \boldsymbol{\omega}_\gamma^t) \right] \right\} - \frac{4\lambda L\|\mathbf{1}_K\|}{\sqrt{T}} \qquad (51) \\
&= \frac{1}{T}\sum_{t=1}^{T} \left\{ \phi^*(\boldsymbol{\theta}^t) - \boldsymbol{\theta}^{t\top} \left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t, \boldsymbol{\omega}_\gamma^t), \boldsymbol{\omega}_\gamma^t) \right] \right\} - \frac{4\lambda L\|\mathbf{1}_K\|}{\sqrt{T}} \\
&\geq \frac{1}{T}\sum_{t=1}^{T} \left\{ \phi^*(\boldsymbol{\theta}^t) + \mathbb{E}_{\boldsymbol{\omega}_{1:\Gamma}^t \sim \Xi_{1:\Gamma}} \left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} (-\boldsymbol{\theta}^t)^\top \boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t, \boldsymbol{\omega}_\gamma^t), \boldsymbol{\omega}_\gamma^t) \right] \right\} - \frac{\left(4\lambda + \sqrt{2\log(4K/\delta)}\right) L\|\mathbf{1}_K\|}{\sqrt{T}} \\
&\quad \text{(w.p. } \geq 1-\delta/2) \qquad (52) \\
&\geq \frac{1}{T}\sum_{t=1}^{T} \left\{ \phi^*(\boldsymbol{\theta}^t) + \mathbb{E}_{\boldsymbol{\omega}_{1:\Gamma}^t \sim \Xi_{1:\Gamma}} \left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} (-\boldsymbol{\theta}^t)^\top \boldsymbol{f}_\gamma^* \right] \right\} - L\|\mathbf{1}_K\| \left[\frac{4\lambda + \sqrt{2\log(4K/\delta)}}{\sqrt{T}} \right] \qquad (53) \\
&\geq \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\omega}_{1:\Gamma}^t \sim \Xi_{1:\Gamma}} \left[\min_{\boldsymbol{\theta}^*\in B_*(L)} \left\{ \phi^*(\boldsymbol{\theta}^*) - \boldsymbol{\theta}^{*\top} \left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}_\gamma^* \right) \right\} \right] - L\|\mathbf{1}_K\| \left[\frac{4\lambda + \sqrt{2\log(4K/\delta)}}{\sqrt{T}} \right] \\
&= \mathrm{opt}(\mathrm{UB}(\phi, \mathbb{R}^K)) - L\|\mathbf{1}_K\| \left[\frac{4\lambda + \sqrt{2\log(4K/\delta)}}{\sqrt{T}} \right]. \qquad (54)
\end{aligned}
$$

Step (51) is by applying Proposition 4 on the series of functions $\{g_t\}_{t=1}^T$, defined as $g_t(\boldsymbol{\theta}) = \phi^*(\boldsymbol{\theta}) - \boldsymbol{\theta}^\top \left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma} \boldsymbol{f}(\boldsymbol{x}_\gamma^t, \boldsymbol{\omega}_\gamma^t) \right]$, with the mirror map Λ as stated in Algorithm 6. For every t, the function g_t is $2\|\mathbf{1}_K\|$-Lipschitz continuous w.r.t. $\|\cdot\|_*$. Step 52 is by an application of the Hoeffding inequality. In step (53), we recall the definitions of $(\boldsymbol{f}^*, \boldsymbol{H}^*)$ from (15, 14) in the proof of Proposition 2. Step (53) is by the assumption of the optimization oracle, and the last step (54) is by the definition of $\mathrm{opt}(\mathrm{UB}(\phi, \mathbb{R}^K))$. Altogether, Theorem 4 is proved. ■

Finally, we remark that a performance guarantee in SAMP can be attained by replacing T with the number of sample trajectories M in Algorithm 6, and use the M sample trajectories as the sequence of random scenarios in Line 4. Nevertheless, the method of bootstrapping proposed in Section 3.3 cannot be applied in this case, since the random sampling with bootstrapping obscures the correlation between scenarios in different periods, and the refined sampling procedure in Line 4 is crucial.

C.2. Online Planning with Imperfect Information

We extend our techniques to address the case of imperfect knowledge on the distributional information for problem $\mathrm{DP}_1(\phi, \mathbb{R}^K)$ under SIMU.

We start with an additional literature review in this line. Vee et al. (2010) studied the online assignment problem based on a set of random samples from future arriving users. They showed that a sampled problem instance suffices to construct near-optimal solution to the full problem. Hardt et al. (2016) studied the performance of stochastic gradient method (SGM) over a finite set of training (offline) samples. They demonstrated the SGM is stable in the sense that the performance under

SGM varies slightly if a single point in the training samples is replaced by a new sample generated from the same distribution. Indeed, our O2O algorithm can also provide the same sub-linear regret guarantee given a finite training samples, as long as the scenarios are sampled from the actual process. However, since the base optimization model studied in the present work is different from the models in Vee et al. (2010) and Hardt et al. (2016), our technical results are established based on novel analyses. Lastly, we remark that Jiang et al. (2025) also studied a case of imperfect prior information for non-stationary online resource allocation with a different metric on the discrepancy measure that quantifies the mismatch in prior information. We highlight that Jiang et al. (2025) is a concurrent research work compared to ours. Notably, the algorithm design and analyses in Jiang et al. (2025) and ours are fundamentally different.

Next, we proceed with extending our O2O framework to handle the case of imperfect prior information. Following Section 6, we only need to revise the Line 5 of Algorithm 1 to implement our O2O framework under Simu. More concretely, we sample a random scenario $\tilde{\boldsymbol{\omega}}^t$ according to the distribution $\widetilde{\Xi}_{\gamma^t}$, instead of the actual distribution Ξ_{γ^t}. We show that, with this type of imperfect information, the optimality gap of the algorithm scales naturally with the discrepancy between actual and forecast models.

Corollary 1. *Assume the access to a forecast model $\widetilde{\Xi}_{1:\Gamma}$ and an optimization oracle (Assumption 2). Consider the application of Algorithm 2 to the online problem $(\mathrm{DP}_1(\phi,\mathbb{R}^K))$ under the actual model $\Xi_{1:\Gamma}$, where the input Θ is generated by Algorithm 1 under the forecast model $\widetilde{\Xi}_{1:\Gamma}$. With probability at least $1-\delta$, we have*

$$Reg_1 = \tilde{O}\left(L\left\|\mathbf{1}_K\right\|\left(\frac{1}{\sqrt{\Gamma}}+\Psi+\frac{\lambda}{\sqrt{T}}+\frac{1}{\sqrt{T}}\right)\right).$$

Proof. The proof largely follows the analysis in Appendix A.3. We sketch the key difference for completeness. For each $1\le\gamma\le\Gamma$, we define

$$\boldsymbol{F}_\gamma := \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\boldsymbol{\omega}_\gamma\sim\Xi_\gamma}\left[\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\boldsymbol{\omega}_\gamma),\boldsymbol{\omega}_\gamma)\middle|\boldsymbol{\theta}^t\right],\quad \widetilde{\boldsymbol{F}}_\gamma := \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}_{\tilde{\boldsymbol{\omega}}_\gamma\sim\widetilde{\Xi}_\gamma}\left[\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\tilde{\boldsymbol{\omega}}_\gamma),\tilde{\boldsymbol{\omega}}_\gamma)\middle|\boldsymbol{\theta}^t\right].$$

We bound the reward gained in the online problem $\mathrm{DP}_1(\phi,\mathbb{R}^K))$ as follows:

$$\begin{aligned}
&\phi\left(\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma)\right)\\
\ge&\phi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\tilde{\boldsymbol{\omega}}^t)\right)-L\left\|\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma)-\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\tilde{\boldsymbol{\omega}}^t)\right\|\\
\ge&\phi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\tilde{\boldsymbol{\omega}}^t)\right)-\underbrace{L\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\tilde{\boldsymbol{\omega}}^t)-\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\widetilde{\boldsymbol{F}}_\gamma\right\|}_{(\ddagger)}-\underbrace{L\left\|\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\left[\widetilde{\boldsymbol{F}}_\gamma-\boldsymbol{F}_\gamma\right]\right\|}_{(\pounds)}-\underbrace{L\left\|\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}\left[\boldsymbol{F}_\gamma-\boldsymbol{f}(\boldsymbol{x}_\gamma,\boldsymbol{\omega}_\gamma)\right]\right\|}_{(\dagger)}\\
\ge&\phi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\tilde{\boldsymbol{\omega}}^t)\right)-L\left\|\mathbf{1}_K\right\|\left[\sqrt{\frac{2\log(6K/\delta)}{\Gamma}}+\Psi+\sqrt{\frac{2\log(6K/\delta)}{T}}\right]\ \text{w.p. } 1-2\delta/3. \qquad (55)
\end{aligned}$$

Similar to Step (22), we bound the terms $(\tilde{\ddagger}, \dagger)$ by applying the Hoeffding inequality, while the term $(\pounds)$ is bounded by invoking the definition of forecast error Ψ defined in Equation (12). Next, we continue by focusing on the term $\phi(\sum_{t=1}^{T} \boldsymbol{f}(\boldsymbol{x}^t, \tilde{\boldsymbol{\omega}}^t)/T)$.

$$
\begin{aligned}
&\phi\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{f}(\boldsymbol{x}^t,\tilde{\boldsymbol{\omega}}^t)\right)\\
=&\min_{\boldsymbol{\theta}\in B_*(L)}\left\{\frac{1}{T}\sum_{t=1}^{T}\phi^*(\boldsymbol{\theta})-\boldsymbol{\theta}^\top\boldsymbol{f}(\boldsymbol{x}^t,\tilde{\boldsymbol{\omega}}^t)\right\}\\
\geq&\frac{1}{T}\sum_{t=1}^{T}\left\{\phi^*(\boldsymbol{\theta}^t)-\boldsymbol{\theta}^{t\top}\boldsymbol{f}(\boldsymbol{x}^t,\tilde{\boldsymbol{\omega}}^t)\right\}-\frac{4\lambda L\|\mathbf{1}_K\|}{\sqrt{T}}\\
=&\frac{1}{T}\sum_{t=1}^{T}\left\{\phi^*(\boldsymbol{\theta}^t)-\boldsymbol{\theta}^{t\top}\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\tilde{\boldsymbol{\omega}}^t),\tilde{\boldsymbol{\omega}}^t)\right\}-\frac{4\lambda L\|\mathbf{1}_K\|}{\sqrt{T}}\\
\geq&\frac{1}{T}\sum_{t=1}^{T}\left\{\phi^*(\boldsymbol{\theta}^t)+\mathbb{E}_{\tilde{\boldsymbol{\omega}}^t_{1:\Gamma}\sim\tilde{\Xi}_{1:\Gamma}}\left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}(-\boldsymbol{\theta}^t)^\top\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\tilde{\boldsymbol{\omega}}^t_\gamma),\tilde{\boldsymbol{\omega}}^t_\gamma)\right]\right\}-\frac{\left(4\lambda+\sqrt{2\log(6K/\delta)}\right)L\|\mathbf{1}_K\|}{\sqrt{T}}\\
&(\text{w.p.}\geq 1-\delta/3)\\
\geq&\frac{1}{T}\sum_{t=1}^{T}\left\{\phi^*(\boldsymbol{\theta}^t)+\mathbb{E}_{\boldsymbol{\omega}^t_{1:\Gamma}\sim\Xi_{1:\Gamma}}\left[\frac{1}{\Gamma}\sum_{\gamma=1}^{\Gamma}(-\boldsymbol{\theta}^t)^\top\boldsymbol{f}(\mathcal{O}(-\boldsymbol{\theta}^t,\boldsymbol{\omega}^t_\gamma),\boldsymbol{\omega}^t_\gamma)\right]\right\}-L\|\mathbf{1}_K\|\left[\frac{4\lambda+\sqrt{2\log(6K/\delta)}}{\sqrt{T}}+\Psi\right] \qquad (56)\\
\geq&\mathrm{opt}(\mathrm{UB}(\phi,\mathbb{R}^K))-L\|\mathbf{1}_K\|\left[\frac{4\lambda+\sqrt{2\log(6K/\delta)}}{\sqrt{T}}+\Psi\right]. \qquad (57)
\end{aligned}
$$

Step (56) is again by the the definition of Ψ. Step (57) follows similar arguments. Altogether, combing (55) with (57), Corollary 1 is proved. ■

D. Supplementary Details for the Numerical Experiments

In this section, we provide additional numerical experiments to showcase the effectiveness of our O2O framework. We address the resource allocation problem in a capacity pooling system while considering the service level requirement. Specifically, we focus on the Type-II service level requirement, also known as the fill rate requirement. The fill rate is defined as the ratio of the fulfilled demand to the demand mean value. Furthermore, this service level requirement is considered as a "soft" constraint since it is required to met on expectation.

Capacity pooling is a common strategy used in practice to serve multiple demand segments with a common pool of resource (e.g., Alptekindğlu et al. 2013, Zhong et al. 2018, Jiang et al. 2023). For instance, a supplier stocking goods for delivery to multiple retailers may face the challenge of meeting the KPIs promised in its service-level agreements with retailers. In this context, the service-level agreement is a commitment by a supplier to achieve a minimum fill rate over a specified time horizon. This problem is often studied in the literature assuming that the demands faced are i.i.d.,

to facilitate the analysis of the fill rates attained. This ignores the more practical challenge in the problem when the demands faced by the retailers are non-stationary across time. We show that our O2O algorithm is able to deal with this challenge, while the state-of-the-art online algorithms developed in the literature fail to deliver satisfactory performance in the non-stationary environment.

To make the discussion concrete, we consider a capacity pooling system in which a retailer supplies a common product to $K = 3$ customers over $\Gamma = 2000$ periods. The retailer serves these customers using a fixed amount of resources c at each period. Customers are faced with non-stationary demands and we assume that the demand $\boldsymbol{\omega}_\gamma$ follows heterogeneous Poisson distributions with mean values $\boldsymbol{\mu}_\gamma := \boldsymbol{\mu}_0 \times e_\gamma$, where the parameter e_γ represents the demand seasonality at period $\gamma = 1, \ldots, \Gamma$ and $\boldsymbol{\mu}_0 = (3, 6, 9)$ denotes the baseline of demand mean values. In the numerical experiments, we consider two cases: (1) we let $e_\gamma = 3$ for $\gamma = 1, \ldots, \lfloor \Gamma/2 \rfloor$, and $e_\gamma = 1$ for $\gamma = \lfloor \Gamma/2 \rfloor + 1, \ldots, \Gamma$ in case 1; (2) we let $e_\gamma = 1$ for $\gamma = 1, \ldots, \lfloor \Gamma/2 \rfloor$, and $e_\gamma = 3$ for $\gamma = \lfloor \Gamma/2 \rfloor + 1, \ldots, \Gamma$ in case 2. In addition, customers require differentiated fill rate targets $\boldsymbol{\beta} = (0.85, 0.90, 0.95)$, i.e., the expected amount of resource allocated to customer k over the entire horizon should be at least $\beta_k \bar{\mu}_k$, where $\bar{\mu}_k := \frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} \mu_{k,\gamma}$ represents the demand mean value across the entire planning horizon. Let $x_{k,\gamma}$ denote the amount of resource allocated to customer k at period γ. The fill rate constraint can be represented as:

$$\frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} x_{k,\gamma} \geq \beta_k \bar{\mu}_k, \ \forall k = 1, \ldots, K.$$

Following the description in Section 3.3, we formulate the fill rate constrained resource allocation problem in the capacity pooling system (CPS) as a "non-constraint" problem:

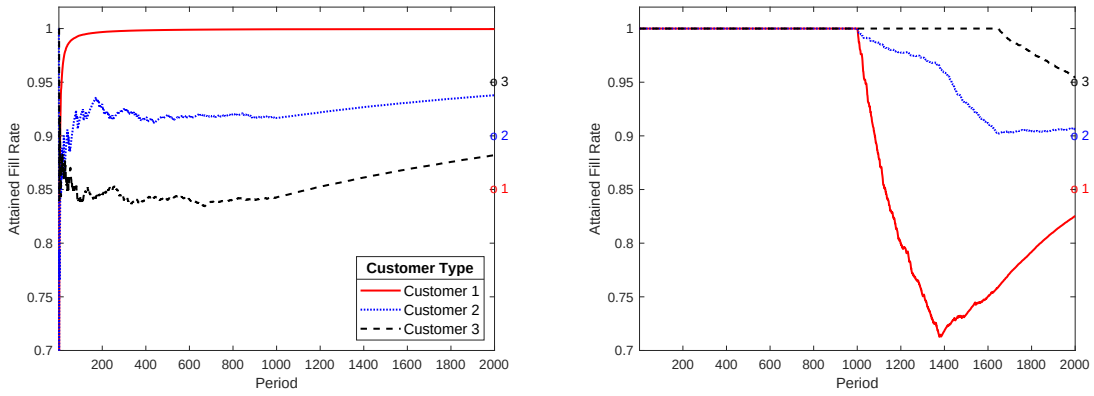
$$\begin{aligned} \text{(CPS)} \quad \min_{\boldsymbol{x}_\gamma} \ & \sum_{k=1}^{K} \left(\beta_k \bar{\mu}_k - \frac{1}{\Gamma} \sum_{\gamma=1}^{\Gamma} x_{k,\gamma} \right)^2 \\ \text{s.t.} \ & \sum_{k=1}^{K} x_{k,\gamma} \leq c, \qquad \forall \gamma = 1, \ldots, \Gamma, \\ & 0 \leq x_{k,\gamma} \leq \omega_{k,\gamma}, \quad \forall \gamma = 1, \ldots, \Gamma, \ k = 1, \ldots, K. \end{aligned}$$

where the objective minimizes the performance gap from the fulfilled demand to the corresponding fill rate target. The first and second sets of constraints require respectively that the non-negative allocation quantity to customer k cannot exceed the realized demand $\omega_{k,\gamma}$ and the total allocation quantity at each period γ cannot exceed the capacity c. This resource allocation problem is "non-constraint" in the sense that there does not exist any linking constraint to tie the decisions across different periods.

Analogous to the setting in Zhong et al. (2018), we consider an identical capacity profile at each period and use sampling average approximation (SAA) method to obtain the minimal capacity level

c^* such that all the fill rate targets are attainable in the offline setting. Note that Zhong et al. (2018) demonstrated that their debt-based online allocation policy, together with the minimal capacity level c^*, are able to attain all the required fill rate targets in a stationary stochastic environment. Nevertheless, Figure 9 plots the glide path of attained fill rate (i.e., $\sum_{\gamma=1}^{s} x_{k,\gamma} / \sum_{\gamma=1}^{s} \omega_{k,\gamma}$ for $s = 1, \ldots, \Gamma$) over the entire planning horizon, and it shows clearly that the debt-based policy cannot meet the fill rate requirements in both cases. We note that the allocation policies developed in Agrawal and Devanur (2015), Zhong et al. (2018), Jiang et al. (2023) provide similar theoretical guarantee and numerical performance for the capacity pooling problem. For the ease of exposition, we take the debt-based allocation policy in Zhong et al. (2018) for comparison.

Figure 9 Attained fill rate over time under the online policy by Zhong et al. (2018).



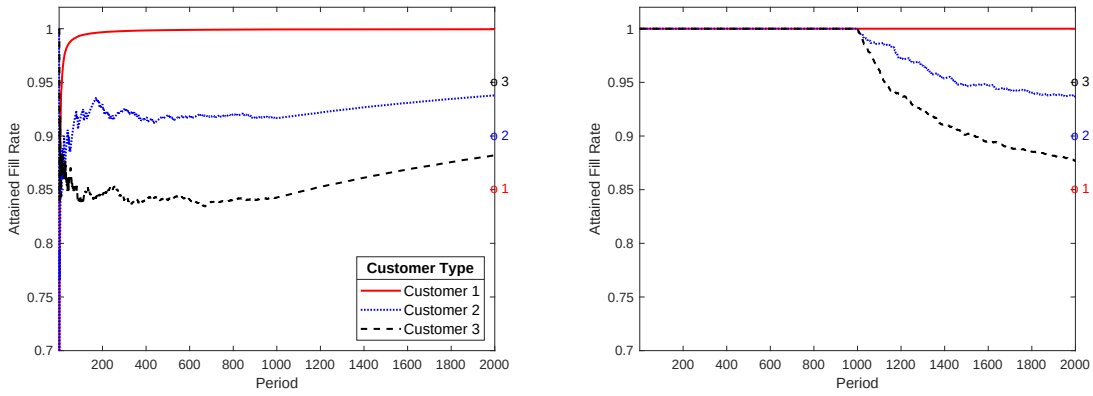
(a) Case 1 (b) Case 2

Note. The fill rate targets are marked on the right-hand-side of each figure.

Due to the drastic scenario change at period $\lfloor \Gamma/2 \rfloor$, it is natural to apply the debt-based online policy on the planning intervals $[1, \lfloor \Gamma/2 \rfloor]$ and $[\lfloor \Gamma/2 \rfloor + 1, \Gamma]$ separately, i.e., we re-start the update of debt vector at time $\lfloor \Gamma/2 \rfloor + 1$ from an initial point. As shown in Figure 10, this re-starting online policy also cannot attain the desired service level target.

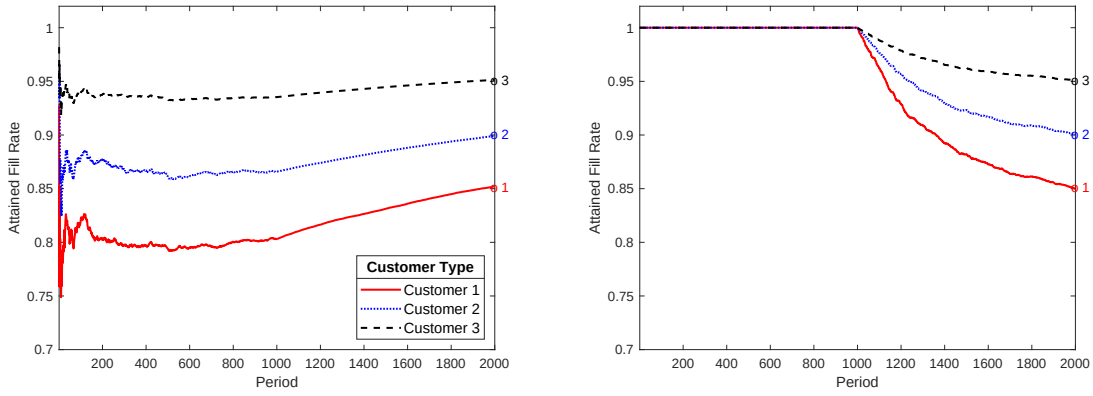
Figure 11 depicts the attained fill rate under our O2O algorithm. It is straightforward to see that all the fill rate requirements are met at the end of planning horizon in both cases. Compared with Figure 9(a), in which customer 1 is served with higher priority due to the (relatively) higher debt, Figure 11(a) shows that customer 1 should be served with lower priority due to the smaller fill rate requirement if future demand samples are incorporated into the allocation policy. In case 2, almost all the demands could be satisfied at the first half of the planning horizon. With the demands increasing above the capacity level, the attained fill rates decline at the second half of the planning horizon. Notably, the fill rate of customer 1 declines more sharply in Figure 9(b), compared with the pattern

Figure 10 Attained fill rate over time under the re-starting online algorithm.



(a) Case 1 (b) Case 2

Figure 11 Attained fill rate over time under our O2O algorithm.



(a) Case 1 (b) Case 2

in Figure 11(b). The reason is that customer 1 accumulates consecutively smaller debt under the priority policy by Zhong et al. (2018) at the first half of the planning horizon, and hence gains lower priority to be served during the remaining time. Differently, the increasing demand trend "guides" the O2O algorithm to balance the resource allocation among three customers under limited capacity so as to fulfill the fill rate requirements. This validates the value of incorporating the offline demand samples into the design of online allocation strategy in non-stationary environments.

References

Agrawal S, Devanur NR (2015) Fast algorithms for online stochastic convex programming. Indyk P, ed. *Proc. 26th Ann. ACM-SIAM Sympos. Discrete Algorithms* (SIAM, Philadelphia), 1405–1424.

Alptekinoğlu A, Banerjee A, Paul MA, Jain N (2013) Inventory pooling to deliver differentiated service. *Manufacturing Service Oper. Management* 15(1):33–44.

Hardt M, Recht B, Singer Y (2016) Train faster, generalize better: Stability of stochastic gradient descent. Balcan M F, Weinberger K L, eds. *Proc. 33rd Internat. Conf. Machine Learn.* (JMLR.org, New York), 1225–1234.

Jiang J, Li X, Zhang J (2025) Online stochastic optimization with Wasserstein based non-stationarity. *Management Sci.*, ePub ahead of print March 3, `https://doi.org/10.1287/mnsc.2020.03850`.

Jiang J, Wang S, Zhang J (2023) Achieving high individual service levels without safety stock? Optimal rationing policy of pooled resources. *Oper. Res.* 71(1):358–377.

Nemirovskij AS, Yudin DB (1983) *Problem Complexity and Method Efficiency in Optimization* (Wiley-Interscience, Hoboken, NJ).

Shalev-Shwartz S (2012) Online learning and online convex optimization. *Foundations Trends Machine Learn.* 4(2):107–194.

Vee E, Vassilvitskii S, Shanmugasundaram J (2010) Optimal online assignment with forecasts. Parkes D C, Dellarocas C, Tennenholtz M, eds. *Proc. 11th ACM Conf. Electronic Commerce* (Association for Computing Machinery (ACM), New York), 109–118.

Zhong Y, Zheng Z, Chou MC, Teo C-P (2018) Resource pooling and allocation policies to deliver differentiated service. *Management Sci.* 64(4):1555–1573.