A Appendix: Proofs and Algorithms

A.1 Proofs of results in Section 4

Proof of Proposition 4.1. Plug $\mathbb{B}_s^q$ in (3) into (4), and apply the minimax theorem, the original problem $[\mathfrak{T}(\boldsymbol{v})]_s,\ \forall s \in \mathcal{S}$ is given by:

$$\begin{array}{rcll} & \max\limits_{\boldsymbol{\pi}_s \in \Delta_A} & \min\limits_{\boldsymbol{p}_s^1, \cdots, \boldsymbol{p}_s^N} \sum\limits_{a \in \mathcal{A}} \pi_{sa} \left(\dfrac{1}{N} \sum\limits_{i=1}^N \boldsymbol{p}_{sa}^i \right)^\top (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}) \\ & & \text{s.t.} & \dfrac{1}{N} \sum\limits_{i=1}^N \left\| \boldsymbol{p}_s^i - \hat{\boldsymbol{p}}_s^i \right\|_q^q \leq \theta^q, \\ & & & \boldsymbol{p}_{sa}^i \in \Delta_S,\ \forall i \in [N],\ \forall a \in \mathcal{A} \\ = & \min & \max\limits_{\boldsymbol{\pi}_s \in \Delta_A} \dfrac{1}{N} \sum\limits_{a \in \mathcal{A}} \pi_{sa} \sum\limits_{i=1}^N (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v})^\top \boldsymbol{p}_{sa}^i \\ & & \text{s.t.} & \dfrac{1}{N} \sum\limits_{i=1}^N \left\| \boldsymbol{p}_s^i - \hat{\boldsymbol{p}}_s^i \right\|_q^q \leq \theta^q, \\ & & & \boldsymbol{p}_{sa}^i \in \Delta_S,\ \forall i \in [N],\ \forall a \in \mathcal{A} \\ = & \min & \max\limits_{a \in \mathcal{A}} \dfrac{1}{N} \sum\limits_{i=1}^N (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v})^\top \boldsymbol{p}_{sa}^i \\ & & \text{s.t.} & \dfrac{1}{N} \sum\limits_{i=1}^N \sum\limits_{a \in \mathcal{A}} \left\| \boldsymbol{p}_s^i - \hat{\boldsymbol{p}}_s^i \right\|_q^q \leq \theta^q, \\ & & & \boldsymbol{p}_{sa}^i \in \Delta_S,\ \forall i \in [N],\ \forall a \in \mathcal{A}. \end{array}$$

Hence (5) is a direct consequence of above formulation by introducing the following epigraph variable γ which satisfies $\gamma \geq \max_{a \in \mathcal{A}} \dfrac{1}{N} \sum\limits_{i=1}^N (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v})^\top \boldsymbol{p}_{sa}^i$. □

Proof of Proposition 4.2. To prove the upper bound of $\gamma^\star$, we consider $\bar{\gamma} = \max_{a \in \mathcal{A}} \frac{1}{N} \sum_{i=1}^N \boldsymbol{b}_{sa}^\top \hat{\boldsymbol{p}}_{sa}^i$, which equals to $\bar{\gamma} \geq \frac{1}{N} \sum_{i=1}^N \boldsymbol{b}_{sa}^\top \hat{\boldsymbol{p}}_{sa}^i,\ \forall a \in \mathcal{A}$. This implies $\{\hat{\boldsymbol{p}}_{sa}^i\}_{i=1}^N$ satisfies every constraint in problem (6) with the lowest possible objective value 0, for every $a \in \mathcal{A}$. Therefore, $\sum_{a \in \mathcal{A}} \mathfrak{P}\left(\{\hat{\boldsymbol{p}}_{sa}^i\}_{i=1}^N; \boldsymbol{r}_{sa} + \lambda \boldsymbol{v}, \bar{\gamma}\right) = 0 < \theta^q$. At this time, $\bar{\gamma}$ is feasible for (5). Hence, we provide an upper bound $\bar{\gamma} = \max_{a \in \mathcal{A}} \frac{1}{N} \sum_{i=1}^N \boldsymbol{b}_{sa}^\top \hat{\boldsymbol{p}}_{sa}^i$.
To prove the lower bound of $\gamma^\star$, we assume the contrary $\max_{a \in \mathcal{A}} \{\min\{\boldsymbol{b}_{sa}\}\} > \gamma^\star$. So there exists $\hat{a} \in \mathcal{A}$ with $\gamma^\star < \min\{\boldsymbol{b}_{s\hat{a}}\}$. Then there is no $\boldsymbol{p}_{s\hat{a}}^i$ satisfies the first constraint in (5) for $a = \hat{a}$. This contradiction verifies the lower bound $\underline{\gamma} = \max_{a \in \mathcal{A}} \{\min\{\boldsymbol{b}_{sa}\}\}$.

□

Proof of Proposition 4.3 . By definition, for any fixed $s \in \mathcal{S}$ and $a \in \mathcal{A}$, we have

$$\begin{array}{rcll} \mathfrak{P}\left(\{\hat{\boldsymbol{p}}_{sa}^i\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma\right) & = & \min & \dfrac{1}{N} \sum\limits_{i=1}^N \left\| \boldsymbol{p}_{sa}^i - \hat{\boldsymbol{p}}_{sa}^i \right\|_q^q \\ & & \text{s.t.} & \dfrac{1}{N} \sum\limits_{i=1}^N \boldsymbol{b}_{sa}^\top \boldsymbol{p}_{sa}^i \leq \gamma, \\ & & & \boldsymbol{p}_{sa}^i \in \Delta_S,\ \forall i \in [N]\,. \end{array}$$

Then, we introduce the dual variable $\alpha \geq 0$ for the constraint $\frac{1}{N} \sum_{i=1}^N \boldsymbol{b}_{sa}^\top \boldsymbol{p}_{sa}^i \leq \gamma$ and obtain

$$\begin{array}{rcll} \mathfrak{P}\left(\{\hat{\boldsymbol{p}}_{sa}^i\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma\right) & = & \max\limits_{\alpha \geq 0} & \left\{ \min\limits_{\boldsymbol{p}_{sa}^i \in \Delta_S,\ \forall i \in [N]} \dfrac{1}{N} \sum\limits_{i=1}^N \left\| \boldsymbol{p}_{sa}^i - \hat{\boldsymbol{p}}_{sa}^i \right\|_q^q + \alpha \left(\dfrac{1}{N} \sum\limits_{i=1}^N \boldsymbol{b}_{sa}^\top \boldsymbol{p}_{sa}^i - \gamma \right) \right\} \\ & = & \max\limits_{\alpha \geq 0} & -\alpha\gamma + \dfrac{1}{N} \sum\limits_{i=1}^N \mathfrak{D}_q(\hat{\boldsymbol{p}}_{sa}^i, \boldsymbol{b}_{sa}, \alpha), \end{array}$$

where $\mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \alpha)$ is defined in the Proposition 4.3.

To show $\bar{\alpha}$ defined in the proposition is indeed an upper bound of the optimal $\alpha^\star$, we denote

$$f(\alpha) = -\alpha\gamma + \frac{1}{N}\sum_{i=1}^{N} \mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \alpha).$$

Notice that $\forall \alpha \geq \max_{i\in[N]} \frac{\left\|\boldsymbol{e}_j - \hat{\boldsymbol{p}}^i_{sa}\right\|_q^q}{\gamma - \min\{\boldsymbol{b}_{sa}\}}$, where $j \in \arg\min_{s'\in\mathcal{S}} b_{sas'}$:

$$\begin{aligned} f(\alpha) &\leq -\alpha\gamma + \frac{1}{N}\sum_{i=1}^{N} \left\|\boldsymbol{e}_j - \hat{\boldsymbol{p}}^i_{sa}\right\|_q^q + \alpha \boldsymbol{b}_{sa}^\top \boldsymbol{e}_j \\ &= \frac{1}{N}\sum_{i=1}^{N} \left\|\boldsymbol{e}_j - \hat{\boldsymbol{p}}^i_{sa}\right\|_q^q + (\min\{\boldsymbol{b}_{sa}\} - \gamma)\,\alpha \\ &\leq 0 = f(0), \end{aligned}$$

where the first inequality is from the definition of $\mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \alpha)$, the second equality is due to the selection of j, and the last inequality is due to the selection of α. Hence $\alpha^\star \in [0, \bar{\alpha}]$ by $f(\alpha)$ is concave w.r.t. α (otherwise $\exists \alpha^\star > \bar{\alpha}$ such that $f(\alpha^\star) > f(0) \geq f(\bar{\alpha})$, where nonconcavity of f follows), and (8) is true. One can further compute the subdifferential of $f(\alpha)$ by Danskin's theorem (Bertsekas, 1999). Typically,

$$-\gamma + \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{b}_{sa}^\top \boldsymbol{p}^{i,\star}_{sa} \in \partial f(\alpha),$$

where $\boldsymbol{p}^{i,\star}_{sa}$ is any minimizer of the inner minimization problem corresponding with given $\alpha \geq 0$. $\square$

Proof of Theorem 4.4. For simplicity of notations, throughout this proof we use

$$\delta \triangleq \frac{A\epsilon_2}{2}\left(\max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\} + \bar{\gamma}\right) \qquad \text{and} \qquad f(\gamma) \triangleq \sum_{a\in\mathcal{A}} \mathfrak{P}\left(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma\right), \quad \forall \gamma \in [\underline{\gamma}, \bar{\gamma}].$$

By (6), we can get that $\mathfrak{P}\left(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma\right)$ and $f(\gamma)$ are non-increasing in $[\underline{\gamma}, \bar{\gamma}]$. For each given $\gamma \in [\underline{\gamma}, \bar{\gamma}]$ and $a \in \mathcal{A}$, we denote $\hat{\mathfrak{P}}\left(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma\right)$ the corresponding value calculated by the Algorithm 1. Furthermore, we call

$$\hat{f}(\gamma) \triangleq \sum_{a\in\mathcal{A}} \hat{\mathfrak{P}}\left(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma\right), \quad \forall \gamma \in [\underline{\gamma}, \bar{\gamma}].$$

We can show that

$$|f(\gamma) - \hat{f}(\gamma)| \leq \delta, \quad \forall \gamma \in [\underline{\gamma}, \bar{\gamma}]. \tag{14}$$

We fix $\gamma \in [\underline{\gamma}, \bar{\gamma}]$, and consider (8). Algorithm 1 provides the optimal solution for (8) with tolerance $\epsilon_2/2$, so we have $|\alpha^\star_{sa,\gamma} - \hat{\alpha}_{sa,\gamma}| \leq \epsilon_2/2$, where $\alpha^\star_{sa,\gamma}$ is the true optimal solution of (8) and $\hat{\alpha}_{sa,\gamma}$ is the solution computed in the inner bisection in Algorithm 1. Thus we get

$$\begin{aligned} &\left|\mathfrak{P}\left(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma\right) - \hat{\mathfrak{P}}\left(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma\right)\right| \\ =& -\alpha^\star_{sa,\gamma}\gamma + \frac{1}{N}\sum_{i=1}^{N} \mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \alpha^\star_{sa,\gamma}) - \left(-\hat{\alpha}_{sa,\gamma}\gamma + \frac{1}{N}\sum_{i=1}^{N} \mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \hat{\alpha}_{sa,\gamma})\right) \\ \leq& \left|\alpha^\star_{sa,\gamma} - \hat{\alpha}_{sa,\gamma}\right|\gamma + \frac{1}{N}\sum_{i=1}^{N} \left|\mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \alpha^\star_{sa,\gamma}) - \mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \hat{\alpha}_{sa,\gamma})\right| \\ \leq& \frac{\epsilon_2}{2}\left(\bar{\gamma} + \max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\}\right), \end{aligned}$$

where the last step is due to that

$$\begin{aligned}
&\mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \alpha^\star_{sa,\gamma}) - \mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \hat{\alpha}_{sa,\gamma}) \\
=&\min_{\boldsymbol{p}^i_{sa}\in\Delta_S} \left\|\boldsymbol{p}^i_{sa} - \hat{\boldsymbol{p}}^i_{sa}\right\|_q^q + \alpha^\star_{sa,\gamma}\cdot \boldsymbol{b}_{sa}^\top \boldsymbol{p}^i_{sa} - \left(\min_{\boldsymbol{p}^i_{sa}\in\Delta_S} \left\|\boldsymbol{p}^i_{sa} - \hat{\boldsymbol{p}}^i_{sa}\right\|_q^q + \hat{\alpha}_{sa,\gamma}\cdot \boldsymbol{b}_{sa}^\top \boldsymbol{p}^i_{sa}\right) \\
\leq& \left(\alpha^\star_{sa,\gamma} - \hat{\alpha}_{sa,\gamma}\right) \boldsymbol{b}_{sa}^\top \boldsymbol{p}^{i,\hat{\alpha}}_{sa} \\
\leq& \frac{\epsilon_2}{2} \max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\},
\end{aligned}$$

and

$$\begin{aligned}
&\mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \hat{\alpha}_{sa,\gamma}) - \mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \alpha^\star_{sa,\gamma}) \\
=&\min_{\boldsymbol{p}^i_{sa}\in\Delta_S} \left\|\boldsymbol{p}^i_{sa} - \hat{\boldsymbol{p}}^i_{sa}\right\|_q^q + \hat{\alpha}_{sa,\gamma}\cdot \boldsymbol{b}_{sa}^\top \boldsymbol{p}^i_{sa} - \left(\min_{\boldsymbol{p}^i_{sa}\in\Delta_S} \left\|\boldsymbol{p}^i_{sa} - \hat{\boldsymbol{p}}^i_{sa}\right\|_q^q + \alpha^\star\cdot \boldsymbol{b}_{sa}^\top \boldsymbol{p}^i_{sa}\right) \\
\leq& \left(\hat{\alpha}_{sa,\gamma} - \alpha^\star_{sa,\gamma}\right) \boldsymbol{b}_{sa}^\top \boldsymbol{p}^{i,\alpha^\star}_{sa} \\
\leq& \frac{\epsilon_2}{2} \max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\},
\end{aligned}$$

here $\boldsymbol{p}^{i,\hat{\alpha}}_{sa}$ and $\boldsymbol{p}^{i,\alpha^\star}_{sa}$ are the optimal solutions to (9) for $\alpha = \hat{\alpha}_{sa,\gamma}$ and $\alpha = \alpha^\star_{sa,\gamma}$ respectively.

Hence, (14) is the direct consequence of above estimation, together with the definitions of f, $\hat{f}$ and δ. And we have $\gamma^\star = \inf\{\gamma \in [\underline{\gamma}, \bar{\gamma}] : f(\gamma) \leq \theta^q\}$. We further define $\hat{\gamma} \triangleq \inf\{\gamma \in [\underline{\gamma}, \bar{\gamma}] : \hat{f}(\gamma) \leq \theta^q\}$. The outer bisection of Algorithm 1 implies that $|\gamma' - \hat{\gamma}| \leq \frac{\epsilon_1}{2}$, so to prove the claimed result in the theorem, it suffices to show that

$$|\gamma^\star - \hat{\gamma}| \leq \frac{2\delta\left(\max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)}{\theta^q}. \tag{15}$$

Notice that $\hat{\mathfrak{P}}\left(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma\right) \leq \mathfrak{P}\left(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma\right), \forall \gamma \in [\underline{\gamma}, \bar{\gamma}]$ by definitions, we get $\hat{f}(\gamma) \leq f(\gamma)$, so $\{\gamma \in [\underline{\gamma}, \bar{\gamma}] : f(\gamma) \leq \theta^q\} \subseteq \{\gamma \in [\underline{\gamma}, \bar{\gamma}] : \hat{f}(\gamma) \leq \theta^q\}$, hence we get $\hat{\gamma} \leq \gamma^\star$.
We assume that $\gamma^\star - \underline{\gamma} \geq \frac{2\delta\left(\max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)}{\theta^q}$, otherwise (15) is trivially satisfied. To achieve (15), we claim the following statement:

$$f\left(\gamma^\star - \frac{2\delta\left(\max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)}{\theta^q}\right) > \theta^q + \delta. \tag{16}$$

If the statement is true, we have $\forall \gamma \in [\underline{\gamma}, \gamma^\star - \left(\max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)(2\delta/\theta^q)]$:

$$\hat{f}(\gamma) \geq f(\gamma) - \delta \geq f\left(\gamma^\star - \frac{2\delta\left(\max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)}{\theta^q}\right) - \delta > \theta^q,$$

where the first inequality is from (14) and $\hat{f}(\gamma) \leq f(\gamma)$, the second inequality is due to that $f(\gamma)$ is non-increasing, and the third inequality is from the above statement.
So $\{\gamma \in [\underline{\gamma}, \bar{\gamma}] : \hat{f}(\gamma) \leq \theta^q\} \subseteq (\gamma^\star - \left(\max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)(2\delta/\theta^q), \bar{\gamma}]$, thus

$$\gamma^\star - \frac{2\delta\left(\max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)}{\theta^q} \leq \hat{\gamma} \leq \gamma^\star,$$

which implies the desired (15).
For the proof of the statement (16). We argue that

$$\forall \gamma \in [\gamma^\star - \frac{2\delta\left(\max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)}{\theta^q}, \gamma^\star) : \sum_{a\in\mathcal{A}} \alpha^\star_{sa,\gamma} \geq \frac{\theta^q}{\left(\max_{a\in\mathcal{A}} \max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)}, \tag{17}$$

by contradiction. Assume there is some $\gamma'' \in [\gamma^\star - \frac{2\delta\left(\max_{a\in\mathcal{A}}\max\{\boldsymbol{b}_{sa}\}-\underline{\gamma}\right)}{\theta^q}, \gamma^\star)$ with that $\sum_{a\in\mathcal{A}}\alpha^\star_{sa,\gamma''} < \frac{\theta^q}{\left(\max_{a\in\mathcal{A}}\max\{\boldsymbol{b}_{sa}\}-\underline{\gamma}\right)}$. Clearly $f(\gamma'') > \theta^q$ since $\gamma'' < \gamma^\star$. Then

$$
\begin{aligned}
\theta^q &> \sum_{a\in\mathcal{A}} \alpha^\star_{sa,\gamma''}\left(\max_{a\in\mathcal{A}}\max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)\\
&\geq \sum_{a\in\mathcal{A}} \alpha^\star_{sa,\gamma''}\left(\frac{1}{N}\sum_{i=1}^N \left(\boldsymbol{b}_{sa}^\top \hat{\boldsymbol{p}}^i_{sa} - \gamma\right)\right)\\
&= \sum_{a\in\mathcal{A}} \frac{1}{N}\sum_{i=1}^N \alpha^\star_{sa,\gamma''}\left(\boldsymbol{b}_{sa}^\top \hat{\boldsymbol{p}}^i_{sa} - \gamma\right)\\
&\geq \sum_{a\in\mathcal{A}} \frac{1}{N}\sum_{i=1}^N \min_{\boldsymbol{p}^i_{sa}\in\Delta_S} \left\|\boldsymbol{p}^i_{sa} - \hat{\boldsymbol{p}}^i_{sa}\right\|_q^q + \alpha^\star_{sa,\gamma''}\cdot\left(\boldsymbol{b}_{sa}^\top \boldsymbol{p}^i_{sa} - \gamma\right)\\
&= \sum_{a\in\mathcal{A}} -\alpha^\star_{sa,\gamma''}\gamma'' + \frac{1}{N}\sum_{i=1}^N \mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \alpha^\star_{sa,\gamma''})\\
&= \sum_{a\in\mathcal{A}} \mathfrak{P}\left(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma''\right)\\
&= f(\gamma''),
\end{aligned}
$$

which is contradicted with $f(\gamma'') > \theta^q$. This contradiction implies that (17) is true.

For simplicity of notations, we call $\gamma''' = \gamma^\star - \frac{2\delta\left(\max_{a\in\mathcal{A}}\max\{\boldsymbol{b}_{sa}\}-\underline{\gamma}\right)}{\theta^q}$, and we fix any $\gamma^\ell \in (\gamma''', \gamma^\star)$, then

$$
\begin{aligned}
&\theta^q - f\left(\gamma^\star - \frac{2\delta\left(\max_{a\in\mathcal{A}}\max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)}{\theta^q}\right)\\
<&f(\gamma^\ell) - f\left(\gamma'''\right)\\
=&\sum_{a\in\mathcal{A}}\left(\mathfrak{P}\left(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma^\ell\right) - \mathfrak{P}\left(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma'''\right)\right)\\
\leq&\sum_{a\in\mathcal{A}} -\alpha^\star_{sa,\gamma^\ell}\gamma^\ell + \frac{1}{N}\sum_{i=1}^N \mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \alpha^\star_{sa,\gamma^\ell}) - \left(-\alpha^\star_{sa,\gamma^\ell}\gamma''' + \frac{1}{N}\sum_{i=1}^N \mathfrak{D}_q(\hat{\boldsymbol{p}}^i_{sa}, \boldsymbol{b}_{sa}, \alpha^\star_{sa,\gamma^\ell})\right)\\
=&\sum_{a\in\mathcal{A}} \alpha^\star_{sa,\gamma^\ell}(\gamma''' - \gamma^\ell)\\
\leq&(\gamma''' - \gamma^\ell)\frac{\theta^q}{\left(\max_{a\in\mathcal{A}}\max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)},
\end{aligned}
$$

where the first step is due to $\gamma^\ell < \gamma^\star$ and definition of γ''', the second step is from the definition of function f, the third step is from the definition of $\mathfrak{P}(\{\hat{\boldsymbol{p}}^i_{sa}\}_{i=1}^N; \boldsymbol{b}_{sa}, \gamma)$ and $\alpha^\star_{sa,\gamma}$, and the last step is due to (17) and $\gamma^\ell \in (\gamma''', \gamma^\star)$. Notice that above inequality is true for all $\gamma^\ell \in (\gamma''', \gamma^\star)$, we could let $\gamma^\ell \to \gamma^\star$, which leads to

$$
\theta^q - f\left(\gamma^\star - \frac{2\delta\left(\max_{a\in\mathcal{A}}\max\{\boldsymbol{b}_{sa}\} - \underline{\gamma}\right)}{\theta^q}\right) \leq -2\delta < -\delta.
$$

Then (16) is the direct consequence of above inequality. This finishes the proof of statement, hence finishes the proof of the theorem. □

Proof of Theorem 4.5 . The Algorithm 1 is the direct consequence of the procedures in the content, except computing the slope, which has been explained at the end of the proof of Proposition 4.3.
For the time complexity, we can see that the bisection method on γ and α uses complexity $\mathcal{O}(\log\epsilon_1^{-1}\log\epsilon_2^{-1})$. For the subproblem (9), which costs time complexity $h_q(S)$, we need to solve

Algorithm 2: Fast algorithm to solve (9) with $q = 1$

Input: Sorted $\boldsymbol{b}_{sa}$ with $b_{san_1} \geq b_{san_2} \geq \cdots \geq b_{san_S}$.
Initialization: $r \leftarrow \boldsymbol{b}_{sa}^\top \hat{\boldsymbol{p}}_{sa}^i$ and $\boldsymbol{p}_{sa}^i \leftarrow \hat{\boldsymbol{p}}_{sa}^i$.
for $k = 1, \ldots, S-1$ **do**
  if $2\hat{p}_{san_k}^i + \alpha(b_{san_S} - b_{san_k})\hat{p}_{san_k}^i \geq 0$ **then** break
  else
    $r+ = 2\hat{p}_{san_k}^i + \alpha(b_{san_S} - b_{san_k})\hat{p}_{sas'}^i$.
    $p_{san_S}^i + = p_{san_k}^i$ and $p_{san_k}^i = 0$.
end
Result: Optimal objective value r and optimal solution $\boldsymbol{p}_{sa}^i$ of (9) with $q = 1$.

it NA times. Besides, computing the upper bound claimed in Proposition 4.2 requires finding $\min\{\boldsymbol{b}_{sa}\}$ for each $a \in \mathcal{A}$, which is in time complexity $\mathcal{O}(AS)$. So we get the time complexity of Algorithm 1 is $\mathcal{O}\left(h_q(S)NA \log \epsilon_1^{-1} \log \epsilon_2^{-1} + AS\right)$.

□

Proof of Proposition 4.6. Plug $\mathbb{B}_s^\infty$ in (3) into (4), we get the Bellman update is given by

$$\begin{aligned}
& \frac{1}{N} \max_{\boldsymbol{\pi}_s \in \Delta_A} \left\{ \min \sum_{a \in \mathcal{A}} \sum_{i=1}^N \pi_{sa} {\boldsymbol{p}_{sa}^i}^\top (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}) \ : \ \left\| \boldsymbol{p}_{sa}^i - \hat{\boldsymbol{p}}_{sa}^i \right\|_\infty \leq \theta,\ \boldsymbol{p}_{sa}^i \in \Delta_S,\ \forall i \in [N],\ \forall a \in \mathcal{A} \right\} \\
= \ & \frac{1}{N} \max_{\boldsymbol{\pi}_s \in \Delta_A} \sum_{a \in \mathcal{A}} \pi_{sa} \sum_{i=1}^N \min \left\{ {\boldsymbol{p}_{sa}^i}^\top (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v}) \ : \ \left\| \boldsymbol{p}_{sa}^i - \hat{\boldsymbol{p}}_{sa}^i \right\|_\infty \leq \theta,\ \boldsymbol{p}_{sa}^i \in \Delta_S \right\} \\
= \ & \frac{1}{N} \max_{a \in \mathcal{A}} \sum_{i=1}^N \min_{\boldsymbol{p}_{sa}^i \in \Delta_S} \left\{ \boldsymbol{b}_{sa}^\top \boldsymbol{p}_{sa}^i \ : \ \left\| \boldsymbol{p}_{sa}^i - \hat{\boldsymbol{p}}_{sa}^i \right\|_\infty \leq \theta \right\}.
\end{aligned}$$

The first equality is due to that the decision varaibles $\boldsymbol{p}_{sa}^i,\ \forall i \in [N]$ and $\forall a \in \mathcal{A}$, are independent from each other. Then we can divide the orignal problem into NA subproblems. The second equality is from the fact that the objective function is affine w.r.t. π_{sa}, and the maximum is simply the greatest coefficient of π_{sa}. □

A.2 Proofs of results in Section 5

A.2.1 Proof of results in Section 5.1

Proof of Proposition 5.1. We introduce the variable $v \geq \|\boldsymbol{p}_{sa}^i - \hat{\boldsymbol{p}}_{sa}^i\|_1$, so

$$\begin{aligned}
\mathfrak{D}_1(\hat{\boldsymbol{p}}_{sa}^i, \boldsymbol{b}_{sa}, \alpha) &= \min_{\boldsymbol{p}_{sa}^i \in \Delta_S} \left\| \boldsymbol{p}_{sa}^i - \hat{\boldsymbol{p}}_{sa}^i \right\|_1 + \alpha \cdot \boldsymbol{b}_{sa}^\top \boldsymbol{p}_{sa}^i \\
&= \min_{v \geq 0} \min_{\boldsymbol{p}_{sa}^i \in \Delta_S} \left\{ v + \alpha \boldsymbol{b}_{sa}^\top \boldsymbol{p}_{sa}^i \ : \ \|\boldsymbol{p}_{sa}^i - \hat{\boldsymbol{p}}_{sa}^i\|_1 \leq v \right\} \\
&= \min_{v \geq 0} \left\{ v + \alpha \min_{\boldsymbol{p}_{sa}^i \in \Delta_S} \boldsymbol{b}_{sa}^\top \boldsymbol{p}_{sa}^i \ : \ \|\boldsymbol{p}_{sa}^i - \hat{\boldsymbol{p}}_{sa}^i\|_1 \leq v \right\}.
\end{aligned} \tag{18}$$

□

Proof of Theorem 5.2. We denote the objective function in (11) as

$$F(v) \ = \ v + \alpha \left\{ \min_{\boldsymbol{p}_{sa}^i \in \Delta_S} \boldsymbol{b}_{sa}^\top \boldsymbol{p}_{sa}^i \ : \ \|\boldsymbol{p}_{sa}^i - \hat{\boldsymbol{p}}_{sa}^i\|_1 \leq v \right\}.$$

W.L.O.G., we assume that $b_{sa1} > b_{sa2} > \cdots > b_{saS}$, and $\hat{\boldsymbol{p}}^i_{sa} > \mathbf{0}$. We claim that

$$F(v) = \begin{cases} v + \alpha\left(-\dfrac{v\,(b_{sa1} - b_{saS})}{2} + \boldsymbol{b}_{sa}^\top \hat{\boldsymbol{p}}^i_{sa}\right) & \text{if } v \in \left[0, 2\hat{p}^i_{sa1}\right), \\ v + \alpha \displaystyle\sum_{k=K+1}^{S} r_k b_{sak} & \text{if } v \in \left[2\displaystyle\sum_{k=1}^{K} \hat{p}^i_{sak}, 2\sum_{k=1}^{K+1} \hat{p}^i_{sak}\right), \text{ for some } K \in [S-2], \\ v + \alpha b_{saS} & \text{if } v \in \left[2\displaystyle\sum_{k=1}^{S-1} \hat{p}^i_{sak}, +\infty\right), \end{cases}$$

where

$$r_k = \begin{cases} \hat{p}^i_{sa(K+1)} - \left(v - 2\displaystyle\sum_{k=1}^{K} \hat{p}^i_{sak}\right) \Big/ 2 & \text{if } k = K+1, \\ \hat{p}^i_{sak} & \text{if } K+2 \le k \le S-1, \\ \hat{p}^i_{saS} + \dfrac{v}{2} & \text{if } k = S. \end{cases}$$

To prove the first case of the claim, it suffices to show that the optimal solution for the minimization problem in $F(v)$ is given by $\boldsymbol{p}^\star$, whose components are $p^\star_1 = \hat{p}^i_{sa1} - \frac{v}{2}$, $p^\star_k = \hat{p}^i_{sak},\ \forall 2 \le k \le S-1$ and $p^\star_S = \hat{p}^i_{saS} + \frac{v}{2}$. It can be easily verified that $\boldsymbol{p}^\star$ defined in this way satisfies the constraints of the minimization problem in $F(v)$.
To see the optimality, we consider any optimal solution $\bar{\boldsymbol{p}}$. We first notice that $\bar{p}_1 > 0$ and $\bar{p}_S < 1$. Actually, Let

$$\begin{aligned} \mathcal{N} &= \{k \in [S] \,:\, \bar{p}_k < \hat{p}^i_{sak}\}, \\ \mathcal{P} &= \{k \in [S] \,:\, \bar{p}_k > \hat{p}^i_{sak}\}, \\ \mathcal{E} &= \{k \in [S] \,:\, \bar{p}_k = \hat{p}^i_{sak}\}. \end{aligned}$$

Then by $\boldsymbol{e}^\top \bar{\boldsymbol{p}} = \boldsymbol{e}^\top \hat{\boldsymbol{p}}^i_{sa} = 1$ and $\|\bar{\boldsymbol{p}} - \hat{\boldsymbol{p}}^i_{sa}\|_1 \le v$, we get

$$\begin{aligned} \sum_{k\in\mathcal{N}} \hat{p}^i_{sak} - \bar{p}_k &= \sum_{k\in\mathcal{P}} \bar{p}_k - \hat{p}^i_{sak} \\ \sum_{k\in\mathcal{N}} (\hat{p}^i_{sak} - \bar{p}_k) + \sum_{k\in\mathcal{P}} (\bar{p}_k - \hat{p}^i_{sak}) &\le v \end{aligned}$$

Hence $\sum_{k\in\mathcal{N}} \hat{p}^i_{sak} - \bar{p}_k = \sum_{k\in\mathcal{P}} \bar{p}_k - \hat{p}^i_{sak} \le \frac{v}{2} < \hat{p}^i_{sa1}$, so we get $\bar{p}_1 > 0$ and $\bar{p}_S < 1$.
Next we show that $\bar{p}_k = \hat{p}^i_{sak},\ \forall 2 \le k \le S-1$. Otherwise we have some $2 \le \hat{k} \le S-1$ such that $|\bar{p}_{\hat{k}} - \hat{p}^i_{sa\hat{k}}| > 0$. If $\bar{p}_{\hat{k}} > \hat{p}^i_{sa\hat{k}}$, we define $\tilde{\boldsymbol{p}}$ with that $\tilde{p}_{\hat{k}} = \bar{p}_{\hat{k}} - \varepsilon$ and $\tilde{p}_S = \bar{p}_S + \varepsilon$, here $0 < \varepsilon < \min\{\frac{|\bar{p}_{\hat{k}} - \hat{p}^i_{sa\hat{k}}|}{2}, \frac{1-\bar{p}_S}{2}\}$, while keeping the other components of $\tilde{\boldsymbol{p}}$ same as $\bar{\boldsymbol{p}}$. We can see that $\tilde{\boldsymbol{p}}$ achieves smaller objective value than $\bar{\boldsymbol{p}}$ does due to $b_{sa\hat{k}} > b_{saS}$, which contradicts the optimality of $\bar{\boldsymbol{p}}$. If $\bar{p}_{\hat{k}} < \hat{p}^i_{sa\hat{k}}$, then we define $\tilde{\boldsymbol{p}}$ with that $\tilde{p}_{\hat{k}} = \bar{p}_{\hat{k}} + \varepsilon$ and $\tilde{p}_1 = \bar{p}_1 - \varepsilon$, here $0 < \varepsilon < \min\{\frac{|\bar{p}_{\hat{k}} - \hat{p}^i_{sa\hat{k}}|}{2}, \frac{\bar{p}_1}{2}\}$, while keeping the other components of $\tilde{\boldsymbol{p}}$ same as $\bar{\boldsymbol{p}}$. Similarly, $\tilde{\boldsymbol{p}}$ achieves smaller objective value and this implies the contradiction.
Finally, we verify the rest two components. We introduce the variables $d_1 = \bar{p}_1 - \hat{p}^i_{sa1}$ and $d_S = \bar{p}_S - \hat{p}^i_{saS}$, then we get the equivalent reformulation of the inner minimization in $F(v)$:

$$\begin{aligned} \min_{d_1, d_S} \quad & b_{sa1} d_1 + b_{saS} d_S \\ \text{s.t.} \quad & d_1 + d_S = 0, |d_1| + |d_S| \le v. \end{aligned}$$

The optimal d_1 and d_S are given by $-v/2$ and $v/2$ respectively. Hence we proved $\boldsymbol{p}^\star$ is indeed an optimal solution.

To prove the second case of the claim, the optimal solution for the minimization problem in $F(v)$ is given by $\boldsymbol{p}^\star$, whose components are $p^\star_k = 0,\ \forall k \in [K]$, and $p^\star_k = r_k,\ \forall K+1 \le k \le S$. The decision variables $\boldsymbol{p}^\star$ defined in this way satisfies the constraints of the minimization problem in $F(v)$.
To see the optimality, we consider any optimal solution $\bar{\boldsymbol{p}}$. We first notice that $\bar{p}_{K+1} > 0$ and $\bar{p}_S < 1$.

To argue this by contradiction, we suppose the contrary $\bar{p}_{K+1} = 0$. Define the notations $\mathcal{N}, \mathcal{P}$ and $\mathcal{E}$ same as before. Then there exists $\hat{k} \leq K$ with $\bar{p}_{\hat{k}} > 0$, otherwise we assume that $\bar{p}_k = 0,\ \forall k \in [K]$, which implies the contradiction as follows.

$$\sum_{k=1}^{K+1} \hat{p}^i_{sak} > \frac{v}{2} \geq \sum_{k\in\mathcal{N}} \hat{p}^i_{sak} - \bar{p}_k \geq \sum_{k=1}^{K+1} \hat{p}^i_{sak}.$$

Here the first inequality is due to the selection of v, the second inequality has been deduced in the first case and the third inequality is from $\bar{p}_k = 0,\ \forall k \in [K+1]$. So we are able to find $\hat{k} \leq K$ with $\bar{p}_{\hat{k}} > 0$. By moving probability $\varepsilon = \min\{\frac{\bar{p}_{\hat{k}}}{2}, \frac{\hat{p}^i_{sa(K+1)}}{2}\}$ from $\bar{p}_{\hat{k}}$ to $\bar{p}_{K+1}$, we can achieve smaller objective value while keeping the feasibility, hence we get the contradiction, which implies that $\bar{p}_{K+1} > 0$ and $\bar{p}_S < 1$. By applying the similar procedures in the first case (moving some probability from $\bar{p}_{K+1}$ or to $\bar{p}_S$), we can show that $\bar{p}_k = \hat{p}^i_{sak} = r_k$ for $K+2 \leq k \leq S-1$.
Next we prove that $\bar{p}_k = r_k = 0$ for $k \in [K]$. Suppose the contrary is true; that is, $\bar{p}_{\hat{k}} > 0$ for some $\hat{k} \leq K$. Then we are able to apply the same procedures as before which illustrate that $\bar{p}_k = \hat{p}^i_{sak},\ \forall \hat{k} < k < S$. Provided this, one can verified that the optimal strategy is putting all the rest probability $1 - \sum_{k=\hat{k}+1}^{S-1} \hat{p}^i_{sak}$ to $\bar{p}_S$ since $b_{sa1} > b_{sa2} > \cdots > b_{saS}$, and $v \geq 2\sum_{k=1}^{K} \hat{p}^i_{sak} \geq 2\sum_{k=1}^{\hat{k}} \hat{p}^i_{sak}$, which implies $\bar{p}_{sa\hat{k}} = 0$ and we get a contradiction. Hence $\bar{p}_k = r_k = 0,\ \forall k \in [K]$.
Finally, we verify the rest two components. We introduce the variables $d_{K+1} = \bar{p}_{K+1} - \hat{p}^i_{sa(K+1)}$ and $d_S = \bar{p}_S - \hat{p}^i_{saS}$, then we get the equivalent reformulation of the inner minimization in $F(v)$:

$$\begin{aligned} \min_{d_{K+1}, d_S} \quad & b_{sa(K+1)} d_{K+1} + b_{saS} d_S \\ \text{s.t.} \quad & d_{K+1} + d_S = \sum_{k=1}^{K} \hat{p}^i_{sak}, |d_{K+1}| + |d_S| \leq v - \sum_{k=1}^{K} \hat{p}^i_{sak}. \end{aligned}$$

The optimal d_{K+1} and d_S are given by $-\frac{v - 2\sum_{k=1}^{K} \hat{p}^i_{sak}}{2}$ and $\frac{v}{2}$ respectively. Hence we proved $\boldsymbol{p}^\star$ is an optimal solution.

To prove the third case of the claim, we notice that the optimal solution $\boldsymbol{e}_S$ to $\min_{\boldsymbol{p}^i_{sa} \in \Delta_S} \boldsymbol{b}_{sa}^\top \boldsymbol{p}^i_{sa}$ is also feasible for the inner minimization problem in $F(v)$, hence it becomes the optimal solution we desire, which implies $F(v) = v + \alpha b_{saS}$ at this time. This finishes the proof of our claim.

Our claim directly illustrates that $F(v)$ is a piecewise-linear function in v with breakpoints $\{0\} \cup \{2\sum_{k=1}^{K} \hat{p}^i_{sak} : \forall K \in [S-1]\}$. Furthermore, based on the provided formulation of $F(v)$, we can compute the difference of value for $F(\cdot)$ between any two adjacent breakpoints, given by

$$F(v_K) - F(v_{K-1}) = 2\hat{p}^i_{saK} + \alpha(b_{saS} - b_{saK})\hat{p}^i_{saK} \quad \forall K \in [S-1],$$

where $v_K = 2\sum_{k=1}^{K} \hat{p}^i_{sak},\ \forall K \in [S-1]$ and $v_0 = 0$.
Hence we provide Algorithm 2 to compute (11), whose time complexity is $\mathcal{O}(S \log S)$ generally and can be reduced to $\mathcal{O}(S)$ if the sorted $\boldsymbol{b}_{sa}$ is provided.

□

Proof of Corollary 5.2.1. We can see that Algorithm 2 is in time complexity $\mathcal{O}(S)$ if $\cup_{a\in\mathcal{A}} \boldsymbol{b}_{sa}$ are sorted, which can be done at the Initialization step in Algorithm 1 with time complexity $\mathcal{O}(AS \log S)$. So by Theorem 4.5, we get the overall complexity is $\mathcal{O}\left(NAS \log \epsilon_1^{-1} \log \epsilon_2^{-1} + AS \log S\right)$.

□

Algorithm 3: Fast algorithm to solve the inner minimization problem in (10)

Input: $\boldsymbol{v}$, $\boldsymbol{r}_{sa}$, and $\hat{\boldsymbol{p}}^i_{sa} \in \Delta_S$.
Initialization: $\boldsymbol{p}^{i,\star}_{sa} = \mathbf{0}$.
Sort $\boldsymbol{b}_{sa}$ as $b_{san_1} \leq \cdots \leq b_{san_S}$.
Find the smallest k such that $\sum_{j=1}^{k} \left(\hat{p}^i_{san_j} + \theta\right) \geq 1$.
Set $p^{i,\star}_{san_j} = \hat{p}^i_{san_j} + \theta$ for $j \leq k-1$ and $p^{i,\star}_{san_k} = 1 - \sum_{j=1}^{k-1} \left(\hat{p}^i_{san_j} + \theta\right)$.
$r = \boldsymbol{b}^\top_{sa} \boldsymbol{p}^{i,\star}_{sa}$.
Result: Optimal objective value r and optimal solution $\boldsymbol{p}^\star_{sa}$ of the inner minimization problem in (10).

A.2.2 Proof of results in Section 5.2

Proof of Theorem 5.3. From problem (9), we can get the minimization problem for $q = 2$

$$\begin{aligned}
& \min_{\boldsymbol{p}^i_{sa} \in \Delta_S} \left\|\boldsymbol{p}^i_{sa} - \hat{\boldsymbol{p}}^i_{sa}\right\|_2^2 + \alpha \boldsymbol{b}^\top_{sa} \boldsymbol{p}^i_{sa} \\
= & \min_{\boldsymbol{p}^i_{sa} \in \Delta_S} \left\|\boldsymbol{p}^i_{sa}\right\|_2^2 - 2{\boldsymbol{p}^i_{sa}}^\top \hat{\boldsymbol{p}}^i_{sa} + \left\|\hat{\boldsymbol{p}}^i_{sa}\right\|_2^2 + \alpha \boldsymbol{b}^\top_{sa} \boldsymbol{p}^i_{sa} \\
= & \min_{\boldsymbol{p}^i_{sa} \in \Delta_S} \left\|\boldsymbol{p}^i_{sa}\right\|_2^2 - \left(2\hat{\boldsymbol{p}}^i_{sa} - \alpha \boldsymbol{b}_{sa}\right)^\top \boldsymbol{p}^i_{sa} + \left\|\hat{\boldsymbol{p}}^i_{sa}\right\|_2^2 \\
= & -\frac{\alpha^2 \left\|\boldsymbol{b}_{sa}\right\|_2^2}{4} + \alpha \boldsymbol{b}^\top_{sa} \hat{\boldsymbol{p}}^i_{sa} + \min_{\boldsymbol{p}^i_{sa} \in \Delta_S} \left\|\boldsymbol{p}^i_{sa} - \frac{2\hat{\boldsymbol{p}}^i_{sa} - \alpha \boldsymbol{b}_{sa}}{2}\right\|_2^2 .
\end{aligned}$$

So it suffices to solve $\min_{\boldsymbol{p}^i_{sa} \in \Delta_S} \left\|\boldsymbol{p}^i_{sa} - \frac{2\hat{\boldsymbol{p}}^i_{sa} - \alpha \boldsymbol{b}_{sa}}{2}\right\|_2^2$, which can be done by Euclidean projection algorithm (Wang and Carreira-Perpinán, 2013) with time complexity $\mathcal{O}(S \log S)$. □

Proof of Corollary 5.3.1. The result is the direct consequence of Theorem 4.5 with $h_2(S)$ is $\mathcal{O}(S \log S)$, provided by Theorem 5.3.

□

A.2.3 Proof of results in Section 5.3

Proof of Theorem 5.4. We claim that the Algorithm 3 solves problem (13) with time complexity $\mathcal{O}(S \log S)$. By expanding the ∞-norm, we formulate (13) as the following box constraints problem.

$$\begin{array}{ll}
\min & \boldsymbol{b}^\top_{sa} \boldsymbol{p}^i_{sa} \\
\text{s.t.} & \max\left\{\mathbf{0}, \hat{\boldsymbol{p}}^i_{sa} - \theta \boldsymbol{e}\right\} \leq \boldsymbol{p}^i_{sa} \leq \min\left\{\boldsymbol{e}, \hat{\boldsymbol{p}}^i_{sa} + \theta \boldsymbol{e}\right\}, \\
& \boldsymbol{e}^\top \boldsymbol{p}^i_{sa} = 1, \\
& \boldsymbol{p}^i_{sa} \in \mathbb{R}^S.
\end{array}$$

To get the optimal solution, we put the probability on the index where $\boldsymbol{b}_{sa}$ is small as much as possible. Specifically, we assume $b_{sa1} < b_{sa2} < \cdots < b_{saS}$ w.l.o.g., and assume k is the smallest index such that $\sum_{j=1}^{k} \left(\hat{p}^i_{saj} + \theta\right) \geq 1$. We claim that

$$p^{i,\star}_{saj} = \begin{cases}
\hat{p}^i_{saj} + \theta & \text{if } 1 \leq j \leq k-1 \\
1 - \sum_{\ell=1}^{k-1} \left(\hat{p}^i_{sa\ell} + \theta\right) & \text{if } j = k \\
0 & \text{otherwise .}
\end{cases}$$

is the optimal solution to the above formulation. To see this, suppose $\bar{\boldsymbol{p}}$ is optimal and there is some $\hat{j} \in [k-1]$ with $\bar{p}_{\hat{j}} \neq \hat{p}^i_{saj} + \theta$. By above box constraint, we get $\bar{p}_{\hat{j}} < \hat{p}^i_{sa\hat{j}} + \theta$, so there exists $\bar{j} \geq k$ such that $\bar{p}_{\bar{j}} > p^{i,\star}_{sa\bar{j}}$. By moving the probability from $\bar{p}_{\bar{j}}$ to $\bar{p}_{\hat{j}}$, we can achieve strictly smaller objective value, which is a contradiction with that $\bar{\boldsymbol{p}}$ is optimal. This gives us an optimal solution with the first $k-1$ components coincides $p^{i,\star}_{sa}$. Then we can get $\boldsymbol{p}^{i,\star}_{sa}$ is indeed optimal by putting the

extra $1-\sum_{\ell=1}^{k-1}\left(\hat{p}^i_{sa\ell}+\theta\right)$ probability on $p^{i,\star}_{sak}$, since $b_{sak}\le\cdots\le b_{saS}$.
The major time complexity of Algorithm 3 is sorting the vector $\boldsymbol{b}_{sa}\in\mathbb{R}^S$, which is $\mathcal{O}(S\log S)$.

□

Proof of Corollary 5.4.1. As for each $a\in\mathcal{A}$, we need to sort $\boldsymbol{b}_{sa}$, which costs $\mathcal{O}(AS\log S)$, then we need to solve NA subproblems, which costs $\mathcal{O}(NAS)$, so the whole problem is computed in time $\mathcal{O}(AS\log S+NAS)$. □

B Appendix: Computational Complexity for General Convex Optimzation

To compare our algorithm with general convex optimization algorithm, we use general convex optimization to compute the problem (5) and problem (8). The time complexities will be discussed in different situations:

- Suppose $q=1$: For general convex optimization problem, problem (5) is equivalent with the following problem:

$$[\mathfrak{T}(\boldsymbol{v})]_s=\left[\begin{array}{ll}\text{minimize} & \gamma\\ \text{subject to} & \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{r}_{sa}+\lambda\boldsymbol{v})^\top\boldsymbol{p}^i_{sa}\le\gamma,\ \forall a\in\mathcal{A}\\ & \frac{1}{N}\sum_{i=1}^{N}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}\left|p^i_{sas'}-\hat{p}^i_{sas'}\right|\le\theta\\ & \gamma\in\mathbb{R},\ \boldsymbol{p}^i_{sa}\in\Delta_S,\ \forall i\in[N],\ \forall a\in\mathcal{A}.\end{array}\right]\quad\forall s\in\mathcal{S}.$$

  By introducing the variables $t^i_{sas'}=\left|p^i_{sas'}-\hat{p}^i_{sas'}\right|,\ \forall i\in[N],\ \forall a\in\mathcal{A},\ \forall s'\in\mathcal{S}$, the above problem is equivalent with

$$\begin{array}{ll}\text{minimize} & \gamma\\ \text{subject to} & \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{r}_{sa}+\lambda\boldsymbol{v})^\top\boldsymbol{p}^i_{sa}\le\gamma,\ \forall a\in\mathcal{A}\\ & \frac{1}{N}\sum_{i=1}^{N}\sum_{a\in\mathcal{A}}\sum_{s'\in\mathcal{S}}t^i_{sas'}\le\theta\\ & t^i_{sas'}\ge p^i_{sas'}-\hat{p}^i_{sas'},\ t^i_{sas'}\ge\hat{p}^i_{sas'}-p^i_{sas'},\ \forall i\in[N],\ \forall a\in\mathcal{A},\ \forall s'\in\mathcal{S}\\ & \gamma\in\mathbb{R},\ \boldsymbol{p}^i_{sa}\in\Delta_S,\ t^i_{sas'}\in\mathbb{R},\ \forall i\in[N],\ \forall a\in\mathcal{A},\ \forall s'\in\mathcal{S}.\end{array}$$

  There are $1+NSA+NSA=\mathcal{O}(NSA)$ decision variables, and the number of bits in the input is $\mathcal{O}(1)+\mathcal{O}(NAS)+\mathcal{O}(NAS)+\mathcal{O}(NAS)+\mathcal{O}(NAS)+\mathcal{O}(NAS)=\mathcal{O}(NAS)$. So by (Karmarkar, 1984), the complexity of solving this LP is $\mathcal{O}(N^{4.5}S^{4.5}A^{4.5})$.

  We utilize our outer bisection, solving (6) directly using convex optimization. Typically, for each fixed $a\in\mathcal{A}$ and γ, (6) is equivalent with the LP that

$$\begin{array}{ll}\text{minimize} & \frac{1}{N}\sum_{i=1}^{N}\sum_{s'\in\mathcal{S}}t^i_{sas'}\\ \text{subject to} & \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{b}^\top_{sa}\boldsymbol{p}^i_{sa}\le\gamma\\ & t^i_{sas'}\ge p^i_{sas'}-\hat{p}^i_{sas'},\ t^i_{sas'}\ge\hat{p}^i_{sas'}-p^i_{sas'},\ \forall i\in[N],\ \forall s'\in\mathcal{S}\\ & \boldsymbol{p}^i_{sa}\in\Delta_S,\ t^i_{sas'}\in\mathbb{R},\ \forall i\in[N],\ \forall s'\in\mathcal{S}.\end{array}$$

  There are $2NS$ decision variables, and the number of bits in the input is $\mathcal{O}(NS+NS+NS+NS)=\mathcal{O}(NS)$. So the complexity of solving this LP is $\mathcal{O}(N^{4.5}S^{4.5})$. Together with the outer bisection, the total complexity for each Bellman update with ϵ_1 tolerance is $\mathcal{O}(N^{4.5}S^{4.5}A\log\epsilon_1^{-1})$.

We utilize our nested bisection scheme, solving (9) using the general convex optimization algorithm. Typically, for each fixed $a \in \mathcal{A}, i \in [N]$ and α, (9) is equivalent with

$$\begin{array}{ll} \text{minimize} & \sum_{s' \in \mathcal{S}} t^i_{sas'} + \alpha \cdot \boldsymbol{b}_{sa}^\top \boldsymbol{p}^i_{sa} \\ \text{subject to} & t^i_{sas'} \geq p^i_{sas'} - \hat{p}^i_{sas'},\ t^i_{sas'} \geq \hat{p}^i_{sas'} - p^i_{sas'},\ \forall s' \in \mathcal{S} \\ & \boldsymbol{p}^i_{sa} \in \Delta_S,\ t^i_{sas'} \in \mathbb{R},\ \forall s' \in \mathcal{S}. \end{array}$$

There are $2S$ decision variables, and the number of bits in the input is $\mathcal{O}(S+S+S) = \mathcal{O}(S)$. So the complexity of solving this LP is $\mathcal{O}(S^{4.5})$. Together with the nested bisection, the total complexity for each Bellman update where the tolerances of bisections are ϵ_1 and ϵ_2, is $\mathcal{O}(NS^{4.5}A \log \epsilon_1^{-1} \log \epsilon_2^{-1})$.

- Suppose $q = 2$: For general convex optimization problem, problem (5) is equivalent with the following SOCP:

$$[\mathfrak{T}(\boldsymbol{v})]_s = \left[\begin{array}{ll} \text{minimize} & \gamma \\ \text{subject to} & \frac{1}{N} \sum_{i=1}^N (\boldsymbol{r}_{sa} + \lambda \boldsymbol{v})^\top \boldsymbol{p}^i_{sa} \leq \gamma,\ \forall a \in \mathcal{A} \\ & \frac{1}{N} \sum_{i=1}^N \sum_{a \in \mathcal{A}} \left\| \boldsymbol{p}^i_{sa} - \hat{\boldsymbol{p}}^i_{sa} \right\|_2^2 \leq \theta^2 \\ & \gamma \in \mathbb{R},\ \boldsymbol{p}^i_{sa} \in \Delta_S,\ \forall i \in [N],\ \forall a \in \mathcal{A}. \end{array} \right] \quad \forall s \in \mathcal{S}.$$

There are $1 + NSA = \mathcal{O}(NSA)$ decision variables, and the number of constraints are $A+1+NA(S+2) = \mathcal{O}(NAS)$. So the complexity of solving the SOCP with ϵ-accuracy is $\mathcal{O}(\sqrt{NAS} \log \epsilon^{-1} \cdot (NSA)^2 (NAS+1+2(A+NA(S+2)))) = \mathcal{O}(N^{3.5} S^{3.5} A^{3.5} \log \epsilon^{-1})$.

We utilize our outer bisection, solving (6) directly using convex optimization. Typically, for each fixed $a \in \mathcal{A}$ and γ, (6) is equivalent with the SOCP that

$$\begin{array}{ll} \text{minimize} & \delta \\ \text{subject to} & \frac{1}{N} \sum_{i=1}^N \boldsymbol{b}_{sa}^\top \boldsymbol{p}^i_{sa} \leq \gamma \\ & \frac{1}{N} \sum_{i=1}^N \left\| \boldsymbol{p}^i_{sa} - \hat{\boldsymbol{p}}^i_{sa} \right\|_2^2 \leq \delta \\ & \delta \in \mathbb{R},\ \boldsymbol{p}^i_{sa} \in \Delta_S,\ \forall i \in [N]. \end{array}$$

There are $1 + NS = \mathcal{O}(NS)$ decision variables, and the number of constraints are $\mathcal{O}(NS)$. So the complexity of solving the above SOCP with ϵ-accuracy is $\mathcal{O}(\sqrt{NS} \log \epsilon^{-1} \cdot N^2 S^2 (NS + 1 + 2(1 + N(S+2)))) = \mathcal{O}(N^{3.5} S^{3.5} \log \epsilon^{-1})$, hence the total complexity of the Bellman update is $\mathcal{O}(N^{3.5} S^{3.5} A \log \epsilon^{-1} \log \epsilon_1^{-1})$.

We utilize our nested bisection scheme, solving problem (9) using the general convex optimization algorithm. For each fixed $a \in \mathcal{A}, i \in [N]$ and α, problem (9) is equivalent with

$$\begin{array}{ll} \text{minimize} & \delta + \alpha \cdot \boldsymbol{b}_{sa}^\top \boldsymbol{p}^i_{sa} \\ \text{subject to} & \left\| \boldsymbol{p}^i_{sa} - \hat{\boldsymbol{p}}^i_{sa} \right\|_2^2 \leq \delta \\ & \boldsymbol{p}^i_{sa} \in \Delta_S,\ \delta \in \mathbb{R}. \end{array}$$

There are $1 + S = \mathcal{O}(S)$ decision variables, and the number of constraints are $\mathcal{O}(S)$. So the complexity of solving the above SOCP with ϵ-accuracy is $\mathcal{O}(\sqrt{S} \log \epsilon^{-1} \cdot S^2 (S + 1 + 2(1 + S))) = \mathcal{O}(S^{3.5} \log \epsilon^{-1})$, hence the total complexity of the Bellman update is $\mathcal{O}(NS^{3.5} A \log \epsilon^{-1} \log \epsilon_1^{-1} \log \epsilon_2^{-1})$.

- Suppose $q = \infty$: We can also consider solving the inner problem in (10) using general convex optimization, which is equivalent with

$$\begin{array}{ll} \text{minimize} & \boldsymbol{b}_{sa}^\top \boldsymbol{p}^i_{sa} \\ \text{subject to} & p^i_{sas'} - \hat{p}^i_{sas'} \leq \theta,\ \hat{p}^i_{sas'} - p^i_{sas'} \leq \theta,\ \forall s' \in \mathcal{S} \\ & \boldsymbol{p}^i_{sa} \in \Delta_S, \end{array}$$

Table 1: Comparisons of runtime (average and standard deviation)(second) of Bellman updates for all algorithms in L_1 norm.

Algorithm \ $N = S = A$	10	20	30	40
Fast	0.0253 (0.02)	0.1243 (0.02)	0.2794 (0.02)	0.5112 (0.02)
Gurobi	0.039 (0.02)	0.7084 (0.03)	4.3617 (0.05)	22.6013 (0.92)
FOM(3 its)	0.644 (1.0193)	4.0821 (1.05)	12.6608 (1.00)	29.22 (1.18)

Table 2: Comparisons of runtime (average and standard deviation)(second) of Bellman updates for all algorithms in L_2 norm.

Algorithm \ $N = S = A$	50	60	70	80
Fast	7.3717 (0.11)	10.8293 (0.12)	16.8888 (0.12)	24.6080 (0.07)
Gurobi	5.1669 (0.09)	11.6760 (0.14)	23.6852 (0.34)	44.3305 (0.18)
FOM(3 its)	9.1936 (1.83)	14.5653 (2.01)	21.3651 (1.97)	31.1326 (1.79)

where $i \in [N]$ and $a \in \mathcal{A}$ are fixed.
There are S decision variables, and the number of bits in the input is $\mathcal{O}(S+S+S) = \mathcal{O}(S)$. So the complexity of solving this LP is $\mathcal{O}(S^{4.5})$. Then the total complexity for each Bellman update is $\mathcal{O}(NS^{4.5}A)$.

C Appendix: Details for Numerical Experiments

We compare our fast algorithm with the state-of-the-art solver Gurobi with version v10.0.1rc0 (Gurobi Optimization, LLC, 2023) and the first-order method of (Grand-Clément and Kroer, 2021a). All experiments are implemented in Python 3.8, and they are run on a 2.3 GHz 4-Core Intel Core i7 CPU with 32 GB 3733 MHz DDR4 main memory. We will release our code to ensure reproducibility. `https://github.com/Chill-zd/Fast-Bellman-Updates-DRMDP`

Our algorithms are tested on some random instances generated by the Generalized Average Reward Non-stationary Environment Test-bench (Garnet MDPs) (Archibald et al., 1995; Bhatnagar et al., 2009). The Garnet MDPs are a collection of assessment problems designed to assess the performance of reinforcement learning algorithms in non-stationary environments. It is convenient to construct and implement these problems. We utilize the parameter n_b to regulate the proportion of the next states accessible for each state-action pair (s, a). Following the same setting as (Grand-Clément and Kroer, 2021a), we set n_b to be 0.2 and random uniform rewards to be in $[0, 10]$. The discount factor λ is fixed at 0.8, and parameter $\epsilon = 0.1$. For each MDP instance, we generate the sampled kernels $\boldsymbol{p}^1, \ldots, \boldsymbol{p}^N$, considering N small random (Garnet) perturbations around the nominal kernel $\boldsymbol{p}^0$. We set parameter θ in Proposition 4.1 to be $\sqrt{n_b A}$.

To test the speed of Bellman update, we run the random instances 50 times for all the algorithms, and show the average time of them in tables 1, 2 and 3. We can see that our algorithm performs better than Gurobi and the first-order method. When the states number increases, the running time of our algorithm keeps a small standard deviation.

We also compare the speed of value iteration for all algorithms using the same convergence criteria: $\|\boldsymbol{v} - \boldsymbol{v}^\star\|_\infty \leq 2\lambda\epsilon(1-\lambda)^{-1}$, which follows (Grand-Clément and Kroer, 2021a). The results are shown in tables 4, 5 and 6. The runtimes that exceed 4000s for L_1 and L_∞ case or exceed 10000s for

Table 3: Comparisons of runtime (average and standard deviation) (millisecond) of Bellman updates for all algorithms in L_∞ norm.

Algorithm \ $N = S = A$	10	20	30	40
Fast	0.06 (0.1)	0.21 (0.1)	0.47 (0.2)	0.80 (0.4)
Gurobi	0.92 (0.1)	7.33 (0.8)	26.50 (3.8)	65.22 (7.9)
FOM	144.92 (981.1)	166.19 (966.5)	237.24 (978.2)	366.74 (976.6)

L_2 case will be shown as "$-$". We can see from tables 4, 5 and 6 that our algorithm always performs better than Gurobi and the first-order method. We point out that our algorithm generally becomes better as the state number increases.

Table 4: Runtime (second) of value iteration for all algorithms in L_1 norm.

Algorithm \ $N = S = A$	10	20	30	40
Fast	3.6164	34.0930	116.4698	281.9007
Gurobi	5.7131	202.0723	1967.9552	–
FOM	202.4039	3578.1377	–	–

Table 5: Runtime (second) of value iteration for all algorithms in L_2 norm.

Algorithm \ $N = S = A$	40	50	60	70
Fast	2198.5413	4847.4994	7223.7663	12387.5601
Gurobi	2543.9588	4861.4114	7949.4562	22096.7578
FOM	–	–	–	–

Table 6: Runtime (second) of value iteration for all algorithms in L_∞ norm.

Algorithm \ $N = S = A$	10	20	30	40
Fast	0.0045	0.0348	0.1371	0.3181
Gurobi	0.0907	0.8746	3.9243	11.5589
FOM	16.4344	97.44293	1131.5260	3933.8585