A. Appendix

A.1. Bootstrap Convergence

In this section we provide a high-level discussion of the bootstrap procedure and its asymptotic validity. We refer the readers to the works by (Cao, 1993; Hall, 1990) for a more fine-grained analysis and convergence rates when estimating MSE using statistical bootstrap. Individual treatment of bias (Efron, 1990; Efron and Tibshirani, 1994; Hong, 1999; Shi, 2012; Mikusheva, 2013) and variance (Chen, 2017b; Gamero et al., 1998; Shao, 1990; Ghosh et al., 1984; Li and Maddala, 1999) can also be found.

In the following, we will discuss the consistency of $\hat{A}$ estimated using bootstrap,

$$\hat{A}_{i.j} - A_{i,j} \xrightarrow{a.s.} 0.$$

Towards this goal, we will consider the following conditions imposed on the set of the base estimators $\{\hat{\theta}_i\}_{i=1}^k$,

- $\forall i, \quad \hat{\theta}_i$ is uniformly bounded.
- $\forall i, \quad \hat{\theta}_i \xrightarrow{a.s.} c_i$.
- $\forall i, \quad \hat{\theta}_i$ is smooth with respect to data distribution.
- $\exists \hat{\theta}_k : \hat{\theta}_k \xrightarrow{a.s.} c_k = \theta^*$.

Recall from (2),

$$\begin{aligned}
A_{i,j} &= \mathbb{E}\Big[\Big(\hat{\theta}_i - \theta^*\Big)\Big(\hat{\theta}_j - \theta^*\Big)\Big] \\
&= \mathbb{E}\Big[\Big(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i] + \mathbb{E}[\hat{\theta}_i] - \theta^*\Big) \\
&\qquad \Big(\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j] + \mathbb{E}[\hat{\theta}_j] - \theta^*\Big)\Big] \\
&= \mathbb{E}\Big[\Big(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i]\Big)\Big(\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j]\Big)\Big] \\
&\quad + \mathbb{E}\Big[\Big(\mathbb{E}[\hat{\theta}_i] - \theta^*\Big)\Big(\mathbb{E}[\hat{\theta}_j] - \theta^*\Big)\Big].
\end{aligned}$$

Let $X_n := \Big(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i]\Big)$ and $Y_n := \Big(\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j]\Big)$. As $\hat{\theta}_i \xrightarrow{a.s.} c_i$ and $\hat{\theta}_i$ is uniformly bounded, using (Thomas and Brunskill, 2016, Lemma 2), we have $\mathbb{E}[\hat{\theta}_i] \xrightarrow{a.s.} c_i$. Similarly, we have $\mathbb{E}[\hat{\theta}_j] \xrightarrow{a.s.} c_j$ as $\hat{\theta}_j \xrightarrow{a.s.} c_j$. Then using continuous mapping theorem,

$$X_n Y_n \xrightarrow{a.s.} (c_i - c_i)(c_j - c_j) = 0.$$

Now using (Thomas and Brunskill, 2016, Lemma 2),

$$\mathbb{E}\Big[\Big(\hat{\theta}_i - \mathbb{E}[\hat{\theta}_i]\Big)\Big(\hat{\theta}_j - \mathbb{E}[\hat{\theta}_j]\Big)\Big] = \mathbb{E}[X_n Y_n] \xrightarrow{a.s.} 0. \quad (9)$$

Similarly,

$$\Big(\mathbb{E}[\hat{\theta}_i] - \theta^*\Big)\Big(\mathbb{E}[\hat{\theta}_j] - \theta^*\Big) \xrightarrow{a.s.} (c_i - \theta^*)(c_j - \theta^*) \quad (10)$$

Therefore, using (9) and (10),

$$A_{i,j} \xrightarrow{a.s.} 0 + (c_i - \theta^*)(c_j - \theta^*). \quad (11)$$

Now we consider the asymptotic property of the bootstrap estimate $\hat{A}$ of A.

$$\hat{A}_{i,j} = \mathbb{E}_{D^*_{n_1}|D_n}\Big[\Big(\hat{\theta}_i(D^*_{n_1}) - \hat{\theta}_k\Big)\Big(\hat{\theta}_j(D^*_{n_1}) - \hat{\theta}_k\Big)\Big] \quad (12)$$

where $\hat{\theta}_k$ is known to be a consistent estimator, i.e., $\hat{\theta}_k \xrightarrow{a.s.} \theta^*$. Here, $\hat{\theta}_k$ could be the WIS or IS or doubly-robust estimators that are known to provide consistent estimates of $\theta^* = J(\pi)$. For brevity, we drop the conditional notation on the subscript, and write (12) as,

$$\hat{A}_{i,j} = \mathbb{E}_{D^*_{n_1}}\Big[\Big(\hat{\theta}_i(D^*_{n_1}) - \hat{\theta}_k\Big)\Big(\hat{\theta}_j(D^*_{n_1}) - \hat{\theta}_k\Big)\Big] \quad (13)$$

Simplifying (13),

$$\begin{aligned}
\hat{A}_{i,j} = \mathbb{E}_{D^*_{n_1}}\Bigg[&\Big(\hat{\theta}_i(D^*_{n_1}) - \mathbb{E}_{D^*_{n_1}}\Big[\hat{\theta}_i(D^*_{n_1})\Big] \\
&+ \mathbb{E}_{D^*_{n_1}}\Big[\hat{\theta}_i(D^*_{n_1})\Big] - \hat{\theta}_k\Big) \\
&\Big(\hat{\theta}_j(D^*_{n_1}) - \mathbb{E}_{D^*_{n_1}}\Big[\hat{\theta}_j(D^*_{n_1})\Big] \\
&+ \mathbb{E}_{D^*_{n_1}}\Big[\hat{\theta}_j(D^*_{n_1})\Big] - \hat{\theta}_k\Big)\Bigg] \\
= \mathbb{E}_{D^*_{n_1}}\Bigg[&\Big(\hat{\theta}_i(D^*_{n_1}) - \mathbb{E}_{D^*_{n_1}}\Big[\hat{\theta}_i(D^*_{n_1})\Big]\Big) \\
&\Big(\hat{\theta}_j(D^*_{n_1}) - \mathbb{E}_{D^*_{n_1}}\Big[\hat{\theta}_j(D^*_{n_1})\Big]\Big)\Bigg] \\
&+ \mathbb{E}_{D^*_{n_1}}\Big[\hat{\theta}_i(D^*_{n_1}) - \hat{\theta}_k\Big]\mathbb{E}_{D^*_{n_1}}\Big[\hat{\theta}_j(D^*_{n_1}) - \hat{\theta}_k\Big]
\end{aligned} \quad (14)$$

Let $X_{n_1} := \Big(\hat{\theta}_i(D^*_{n_1}) - \mathbb{E}_{D^*_{n_1}}\Big[\hat{\theta}_i(D^*_{n_1})\Big]\Big)$ and $Y_{n_1} := \Big(\hat{\theta}_j(D^*_{n_1}) - \mathbb{E}_{D^*_{n_1}}\Big[\hat{\theta}_j(D^*_{n_1})\Big]\Big)$. As the empirical distribution $D^*_{n_1}$ converges to the population distribution, i.e., $D_n \xrightarrow{a.s.} D$, the resampled distribution $D^*_{n_1}$ from D_n also converges to the population distribution, i.e., $D^*_{n_1} \xrightarrow{a.s.} D$. Therefore, when the estimator $\hat{\theta}_i(D^*_{n_1})$ is smooth, using the continuous mapping theorem,

$$\forall i, \qquad \lim_{n_1 \to \infty} \hat{\theta}_i(D^*_{n_1}) = \hat{\theta}_i\left(\lim_{n_1 \to \infty} D^*_{n_1}\right) = \hat{\theta}_i(D) = c_i.$$

Therefore, similar to before,

$$X_{n_1} Y_{n_1} \xrightarrow{a.s.} (c_i - c_i)(c_j - c_j) = 0,$$

and subsequently,

$$\mathbb{E}_{D^*_{n_1}}[X_{n_1} Y_{n_1}] \xrightarrow{a.s.} 0. \quad (15)$$

Further, as $\hat{\theta}_k \xrightarrow{a.s.} \theta^*$,

$$\hat{\theta}_i(D^*_{n_1}) - \hat{\theta}_k \xrightarrow{a.s.} c_i - \theta^*.$$

Therefore,

$$\mathbb{E}_{D^*_{n_1}}\left[\hat{\theta}_i(D^*_{n_1}) - \hat{\theta}_k\right]\mathbb{E}_{D^*_{n_1}}\left[\hat{\theta}_j(D^*_{n_1}) - \hat{\theta}_k\right] \xrightarrow{a.s.} (c_i - \theta^*)(c_j - \theta^*). \tag{16}$$

Using (15) and (16) in (14),

$$\hat{A}_{i,j} \xrightarrow{a.s.} 0 + (c_i - \theta^*)(c_j - \theta^*). \tag{17}$$

Finally, combining (11) and (17),

$$\hat{A}_{i.j} - A_{i,j} \xrightarrow{a.s.} 0.$$

which gives the desired result. It is worth highlighting that, theoretically, this result relies upon assumptions that the base estimators satisfy regularity conditions and are consistent. In practice, such assumptions might not hold (for e.g., when using FQE to do policy evaluation if the function approximation is under-parameterized). Nonetheless, in Section 5 we empirically illustrate that even when these assumptions are not directly satisfied, OPERA can be effective.

A.2. Finite Sample Analysis of OPERA

Without loss of generality, let $\forall \pi \in \Pi, |J(\pi)| \leq 1$, such that we can always consider $\forall i, |\hat{\theta}_i| \leq 1$ (this can be trivially achieved by normalizing each estimator's output by $|V_{\max}|$). Let $\bar{\theta}$ be a weighted sum of $\hat{\theta}_i$ with $\alpha^\star$, where the total number of estimators in the ensemble is k.

In the following, we show how the error in estimating the optimal weight coefficients α^* affects the MSE of the resulting estimator $\hat{\bar{\theta}}$. Given $\{\hat{\theta}_i\}_{i=1}^k$, we assume that $\hat{A}$ obtained uisng the bootstrap procedure of OPERA will produce $\hat{\alpha}$ via Equation 8. Whereas, using A would have produced $\alpha^\star$. To provide a finite sample characterization of OPERA's mean squared error, consider the setting where given n samples in dataset D, there exists $\lambda > 0$, such that

$$\forall i, \quad \mathbb{E}_{D_n}[|\hat{\alpha}_i - \alpha^*_i|] \leq n^{-\lambda}, \tag{18}$$

where the expectation is over the randomness due to data D_n that governs the estimates $\hat{\theta}_i$ and thus also the weights $\hat{\alpha}$ and α^* used to combines these estimates.

To bound the MSE of OPERA's estimate $\hat{\bar{\theta}}$ observe that,

$$\begin{aligned}
\text{MSE}(\hat{\bar{\theta}}) &:= \mathbb{E}_{D_n}\left[\left(\hat{\bar{\theta}} - \theta^*\right)^2\right] \\
&= \mathbb{E}_{D_n}\left[\left(\hat{\bar{\theta}} - \bar{\theta} + \bar{\theta} - \theta^*\right)^2\right] \\
&\leq \underbrace{\mathbb{E}_{D_n}\left[(\hat{\bar{\theta}} - \bar{\theta})^2\right]}_{=\Delta_\alpha} + \underbrace{E_{D_n}\left[\left(\bar{\theta} - \theta^*\right)^2\right]}_{=\Delta_c}
\end{aligned} \tag{19}$$

We isolate the error of $\hat{\bar{\theta}}$ into two terms: Δ_α and Δ_c. Δ_c is the gap between the best estimate OPERA can give with $\alpha^\star$ and the true estimate of the policy performance θ^*. If θ^* can be expressed as a linear combination of $\hat{\theta}_i$, then $\Delta_c = 0$. Δ_α is the term we want to further analyze because it depends on the difference between $\hat{\alpha}$ and α^*.

$$\begin{aligned}
\Delta_\alpha &:= \mathbb{E}_{D_n}\left[\left(\hat{\bar{\theta}} - \bar{\theta}\right)^2\right] \\
&= \mathbb{E}_{D_n}\left[\left(\sum_{i=1}^k \hat{\alpha}_i\hat{\theta}_i - \sum_{i=1}^k \alpha^*_i\hat{\theta}_i\right)^2\right] \\
&= \mathbb{E}_{D_n}\left[\left(\sum_{i=1}^k (\hat{\alpha}_i - \alpha^*_i)\hat{\theta}_i\right)^2\right] \\
&\leq \mathbb{E}_{D_n}\left[\left(\sum_{i=1}^k (\hat{\alpha}_i - \alpha^*_i)^2\right)\left(\sum_{i=1}^k \hat{\theta}_i^2\right)\right],
\end{aligned} \tag{20}$$

where the last inequality follows from Cauchy-Schwarz inequality. Now by using the fact that $|\hat{\theta}_i| \leq 1$ and by plugging (18) into (20):

$$\begin{aligned}
\Delta_\alpha &\leq \mathbb{E}_{D_n}\left[k\left(\sum_{i=1}^k (\hat{\alpha}_i - \alpha^*_i)^2\right)\right] \\
&= k\sum_{i=1}^k \mathbb{E}_{D_n}\left[(\hat{\alpha}_i - \alpha^*_i)^2\right] \\
&\leq \frac{k^2}{n^{2\lambda}}.
\end{aligned} \tag{21}$$

Therefore, combining (19) and (21),

$$\text{MSE}(\hat{\bar{\theta}}) \leq \frac{k^2}{n^{2\lambda}} + \Delta_c. \tag{22}$$

This bound factors the MSE using the term Δ_c, which is the best a linear combination of estimators can do. Notice that $\Delta_c \leq \min_i \text{MSE}(\hat{\theta}_i)$, as the best linear combination of the estimators can at least achieve the MSE of the best estimator, by assigning weight of 1 to the best estimator and 0 to the rest. Therefore, the rate of decay of Δ_c is bounded above by the rate of convergence of the best estimator in our ensemble.

The other other term $k^2/n^{2\lambda}$ in (22) results due to the error in estimating α^* because of the bootstrapping process used for estimating $\hat{A}$ of A in (7). This is dependent on the number of estimators k – as we include more estimators in our ensemble, the combination weights $\alpha \in \mathbb{R}^k$ that need to be estimated becomes higher dimensional, thereby introducing more errors. However, the overall term decreases as the dataset size n increases.

A.3. Proofs on Properties of OPERA

A.3.1. Invariance

In the following, we illustrate an important property of OPERA, that the resulting combined estimate $\hat{\bar{\theta}}$ is invariant to the addition of redundant copies of the base estimators $\{\hat{\theta}_i\}_{i=1}^n$. Without loss of generality, let $\hat{\Theta}_\beta \in \mathbb{R}^{(K+1)\times 1}$ be the stack of unique estimators $\{\hat{\theta}_i\}_{i=1}^k$ with $\hat{\theta}_{k+1}$ being a redundant copy of the $\hat{\theta}_k$,

Theorem 4 (Invariance)**.** *If $\hat{A}$ is positive definite, then $\hat{\bar{\theta}}_\beta = \hat{\bar{\theta}}$, where,*

$$\hat{\bar{\theta}}_\beta := \sum_{i=1}^{k+1} \beta_i^* \hat{\theta}_i \in \mathbb{R}, \quad \textit{where, } \beta^* \in \underset{\beta \in \mathbb{R}^{(k+1)\times 1}}{\arg\min}\ \beta^\top B \beta.$$

Proof. We prove this by contradiction. Recall that $\hat{\alpha} \in \mathbb{R}^k$ are the weights that minimize the bootstrap estimate of MSE of $\hat{\bar{\theta}}$ consisting of k estimators.

$$\widehat{\text{MSE}}(\hat{\alpha}_1\hat{\theta}_1 + ... + \hat{\alpha}_k\hat{\theta}_k) = \hat{\alpha}^\top \hat{A} \hat{\alpha}. \tag{23}$$

As $\hat{\theta}_{k+1}$ is a redundant copy of $\hat{\theta}_k$,

$$\begin{aligned}&\widehat{\text{MSE}}(\beta_1^*\hat{\theta}_1 + ... + \beta_k^*\hat{\theta}_k + \beta_{k+1}^*\hat{\theta}_{k+1}) \\ &= \widehat{\text{MSE}}(\beta_1^*\hat{\theta}_1 + ... + (\beta_k^* + \beta_{k+1}^*)\hat{\theta}_k)\end{aligned} \tag{24}$$

Finally, as $\beta^* \in \mathbb{R}^{k+1}$ is the weight that minimizes the bootstrap estimate of MSE of $\hat{\bar{\theta}}_\beta$. Now, if (23) $<$ (24), then one could assign $\beta_i^* := \hat{\alpha}_i$ for $i \in \{1, ..., k\}$, and $\beta_{k+1}^* = 0$ to make (24) $=$ (23). Further, notice that as both $\hat{\alpha}$ and β^* are within the same feasible set of solutions, the above reassignment is also within the feasible set of solutions. Similarly, if (23) $>$ (24), then one could assign $\hat{\alpha}_i := \beta_i^*$ for $i \in \{1, ..., k-1\}$, and $\hat{\alpha}_k = \beta_k^* + \beta_{K+1}^*$ to make (24) $=$ (23). Hence, if (23) does not equal (24), then either $\hat{\alpha}$ or β^* is not optimal and that would be a contradiction. This ensures that $\widehat{\text{MSE}}(\hat{\bar{\theta}}_\beta) = \widehat{\text{MSE}}(\hat{\bar{\theta}})$.

As $\hat{A}$ is positive definite, it implies that (8) is strictly convex with linear constraints. Thus the minimizer $\hat{\alpha}$ of (8) is unique, and $\hat{\bar{\theta}}_\beta = \hat{\bar{\theta}}$. Note that due to redundancy, B will not be PD despite $\hat{A}$ being PD. This would imply that there can be multiple values of β_k^* and β_{k+1}^*. Nonetheless, since $\beta_k^* + \beta_{k+1}^* = \hat{\alpha}_k$, it implies that $\hat{\bar{\theta}}_\beta = \hat{\bar{\theta}}$.

□

A.3.2. Performance Improvement

Theorem 5 (Performance improvement)**.** *If $\hat{\alpha} = \alpha^*$,*

$$\forall i \in \{1, ..., k\}, \quad \textit{MSE}(\hat{\bar{\theta}}) \le \textit{MSE}(\hat{\theta}_i).$$

Algorithm 1: OPERA Score Computation with Bootstrap

Input: offline RL data $\mathcal{D}$; evaluation policy π; a set of OPE estimators $[\text{OPE}_1, \text{OPE}_2, ..., \text{OPE}_k]$; number of bootstrap B; a subsample coefficient $\eta \in [0, 1]$.
Output: estimated π performance s_{OPERA}

for $i \leftarrow 1...K$ **do**
  $s_i^* = \text{OPE}_i(\mathcal{D})$
  $\tilde{s}_i = \emptyset$
  for $j \leftarrow 1...B$ **do**
    $\tilde{n} = |\mathcal{D}|^\eta$
    $\tilde{\mathcal{D}}_j \leftarrow \texttt{Bootstrap}\,(\mathcal{D}, \tilde{n})$
    $\tilde{s}_i = \tilde{s}_i \cup \text{OPE}_i(\tilde{D}_j)$
  end
end
$\tilde{\boldsymbol{M}} \leftarrow [\tilde{s}_1, \tilde{s}_2, ..., \tilde{s}_k] \in \mathbb{R}^{K\times B}$
$\boldsymbol{M} \leftarrow [s_1^*, s_2^*, ..., s_k^*] \in \mathbb{R}^{K\times 1}$
$\boldsymbol{\delta} \leftarrow [(\tilde{s}_1 - s_1^*, \tilde{s}_2 - s_2^*, ..., \tilde{s}_k - s_k^*] \in \mathbb{R}^{K\times B}$
$\boldsymbol{A} \leftarrow \frac{1}{B}\frac{\tilde{n}}{n}\boldsymbol{\delta}\boldsymbol{\delta}^\top \in \mathbb{R}^{K\times K}$
$\boldsymbol{\alpha} = \arg\min_\alpha \boldsymbol{\alpha}\boldsymbol{A}\boldsymbol{\alpha}^\top \quad s.t. \sum \boldsymbol{\alpha} = 1$
$s_{\text{OPERA}} = \boldsymbol{\alpha}^\top \boldsymbol{M}$
return s_{OPERA}

Proof. With a slight overload of notation, we make the dependency of weights α explicit and let $\bar{\theta}(\alpha) = \sum_{i=1}^k \alpha_i \hat{\theta}_i$. Let $\text{MSE}(\bar{\theta}(\alpha)) := \alpha^\top A \alpha$, where A is defined as in (2).

Now from (1) and (2), we know that for $\sum_{i=1}^k \alpha_i = 1$,

$$\alpha^* \in \underset{\alpha \in \mathbb{R}^{k\times 1}}{\arg\min}\, \text{MSE}(\bar{\theta}(\alpha)).$$

Therefore, for any $\lambda \in \mathbb{R}^{k\times 1}$ such that $\sum_{i=1}^k \lambda_i = 1$,

$$\begin{aligned}\text{MSE}(\bar{\theta}(\hat{\alpha})) &= \text{MSE}(\bar{\theta}(\alpha^*)) \qquad \because \hat{\alpha} = \alpha^* \\ &\le \text{MSE}(\bar{\theta}(\lambda)).\end{aligned}$$

Notice that for $e_i := [0, 0, .., 1, .., 0]$, where there is a 1 in the i^{th} position and zero otherwise, $\bar{\theta}(e_i) = \hat{\theta}_i$. Therefore,

$$\begin{aligned}\text{MSE}(\bar{\theta}(\hat{\alpha})) &\le \text{MSE}(\bar{\theta}(e_i)) \qquad \forall i \\ &= \text{MSE}(\hat{\theta}_i) \qquad \forall i.\end{aligned}$$

Therefore, as $\hat{\bar{\theta}} = \bar{\theta}(\hat{\alpha})$, we have the desired result that $\forall i \in \{1, ..., k\}, \quad \text{MSE}(\hat{\bar{\theta}}) \le \text{MSE}(\hat{\theta}_i)$. □

A.4. OPERA Algorithm

We show an illustration of the OPERA algorithm in Figure 1 and we describe the pseudo-code below.