Off-Policy Evaluation and Learning for the Future under Non-Stationarity

Tatsuhiro Shimizu
Yale University
Connecticut, USA
tatsuhiro.shimizu@yale.edu

Kazuki Kawamura
Sony Group Corporation
Tokyo, Japan
Kazuki.Kawamura@sony.com

Takanori Muroi
Sony Group Corporation
Tokyo, Japan
Takanori.Muroi@sony.com

Yusuke Narita
Yale University
Connecticut, USA
yusuke.narita@yale.edu

Kei Tateno
Sony Group Corporation
Tokyo, Japan
Kei.Tateno@sony.com

Takuma Udagawa
Sony Group Corporation
Tokyo, Japan
Takuma.Udagawa@sony.com

Yuta Saito
Cornell University
New York, USA
ys552@cornell.edu

arXiv:2506.20417v1 [cs.LG] 25 Jun 2025

Abstract

We study the novel problem of future off-policy evaluation (F-OPE) and learning (F-OPL) for estimating and optimizing the future value of policies in non-stationary environments, where distributions vary over time. In e-commerce recommendations, for instance, our goal is often to estimate and optimize the policy value for the upcoming month using data collected by an old policy in the previous month. A critical challenge is that data related to the future environment is not observed in the historical data. Existing methods assume stationarity or depend on restrictive reward-modeling assumptions, leading to significant bias. To address these limitations, we propose a novel estimator named ***Off-Policy Estimator for the Future Value (OPFV)***, designed for accurately estimating policy values at any future time point. The key feature of OPFV is its ability to leverage the useful structure within time-series data. While future data might not be present in the historical log, we can leverage, for example, seasonal, weekly, or holiday effects that are consistent in both the historical and future data. Our estimator is the first to exploit these time-related structures via a new type of importance weighting, enabling effective F-OPE. Theoretical analysis identifies the conditions under which OPFV becomes low-bias. In addition, we extend our estimator to develop a new policy-gradient method to proactively learn a good future policy using only historical data. Empirical results show that our methods substantially outperform existing methods in estimating and optimizing the future policy value under non-stationarity for various experimental setups.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '25, August 3–7, 2025, Toronto, ON, Canada
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1245-6/25/08
https://doi.org/10.1145/3690624.3709237

CCS Concepts

• **Computing methodologies** → **Batch learning**.

Keywords

Off-Policy Evaluation and Learning, Non-stationarity.

ACM Reference Format:
Tatsuhiro Shimizu, Kazuki Kawamura, Takanori Muroi, Yusuke Narita, Kei Tateno, Takuma Udagawa, and Yuta Saito. 2025. Off-Policy Evaluation and Learning for the Future under Non-Stationarity. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 37 pages. https://doi.org/10.1145/3690624.3709237

1 Introduction

Various decision-making problems, ranging from recommendation systems to precision medicine, increasingly use logged interaction data to implement optimal decisions. Most of these systems operate under a *contextual bandit protocol*, where they observe the context, take actions based on a bandit policy, and observe the resulting rewards. In such an automated decision-making problem, a crucial counterfactual question is how effective a new policy would have been compared to the one actually implemented in the past. This problem is known as *Off-Policy Evaluation (OPE)* [8, 18, 38]. Answering this question using historical logged data is important, as conducting an online experiment may entail ethical issues and significant expense [1, 38].

While there has been considerable progress in OPE within stationary environments, where the distributions and the value of policies remain constant over time, most practical applications of OPE are concerned with the future effectiveness of new policies under varying distributions [4, 14, 48]. For instance, when developing a new recommendation policy, we are interested in its expected reward during the upcoming week or month. However, the historical data available for this estimation was collected by an older (logging) policy in previous weeks or months. A challenge here is the absence of observed data from future distributions. Most

existing methods, such as Inverse Propensity Scoring (IPS) [32] and Doubly Robust (DR) [8], estimate the value of a new policy based on past data (when the logs were collected) and naively apply their estimates to predict the future policy value, assuming a stationary distribution. This assumption of stationarity, however, is often unrealistic and violated in practical situations [4, 14, 27]. For instance, in content recommendations, conversion rates (rewards) tend to vary over time. A typical example can be observed in music streaming services, where users' preferences for songs or podcasts change between weekdays and weekends, as well as across different seasons. This variability renders the standard formulation ineffective [4]. Some previous methods have attempted to address this limitation of the typical formulation. In particular, Chandak et al. [4] tackle abrupt non-stationarity, where expected rewards can incur non-smooth jumps over time. Under such a formulation, they propose a method named Prognosticator to address both OPE and OPL in terms of future policy performance. Prognosticator first estimates the value of a new policy at each past time step using typical OPE estimators, and then predicts the future value by learning its time-series trend via a parametric regression. While this approach often outperforms typical OPE, it still introduces significant bias in its regression stage due to the difficulty of learning a regressor that generalizes well to the future using only past data.

To overcome the limitations of previous methods, we first formally formulate the *Future OPE (F-OPE)* problem, which aims to estimate the future value of new policies using only historical data provided by a logging policy in the past, under non-stationary (context and reward) distributions. In particular, we formulate F-OPE, considering the timestamp as a new random variable. This approach allows us to address both smooth and abrupt non-stationarity, previously studied somewhat independently, under a unified framework. Based on this new formulation, we introduce a new estimator for F-OPE, named ***O**ff-**P**olicy Estimator for the **F**uture **V**alue (**OPFV**)*. A key feature of OPFV is its ability to leverage the useful structure in time-series data such as seasonal, weekly, or holiday effects, that are applicable both in the past and in the future. By doing so, OPFV can more accurately estimate the future value of policies with only historical data. Technically, our estimator applies a new type of importance weighting to unbiasedly estimate the effects characterized by the structure in the time series (which we call *time-series features*). It also deals with the residual effect that cannot be fully explained by time-series features alone, using a regression model to further reduce the bias and variance depending on the model's accuracy. In our theoretical analysis, we identify conditions under which our estimator can provide a low-bias estimation of the future values of new policies. We also demonstrate that the choice of time-series features affects the bias-variance tradeoff of OPFV and introduce a simple data-driven procedure for selecting time-series features to optimize the accuracy of OPFV. Additionally, we propose a novel policy-gradient method for learning good future policies using only historical data. This involves extending the OPFV estimator to estimate policy gradients regarding future policy values. Finally, we conducted experiments with both synthetic and real-world recommendation data under non-stationary conditions, demonstrating that OPFV is more substantially effective in evaluating and learning future policies than existing methods.

2 Typical Formulation

We first formulate the problem of OPE within the contextual bandit protocol under the typical assumption of stationarity. Let $x \in \mathcal{X}$ denote a context vector and $a \in \mathcal{A}$ denote a (discrete) action, such as content recommendation in a movie streaming service. Let $r \in [0, r_{\max}]$ represent a reward variable, which is independently sampled from an unknown conditional distribution $p(r \mid x, a)$. A decision-making policy is modeled as a distribution over the action space, i.e., $\pi : \mathcal{X} \to \Delta(\mathcal{A})$.

In conventional OPE, we are given logged data $\mathcal{D} := \{(x_i, a_i, r_i)\}_{i=1}^n$, consisting of n independent draws from the *logging policy* π_0. OPE aims to estimate the expected reward as the *value* of a given *evaluation policy* π_e (, which is different from π_0):

$$V(\pi_e) := \mathbb{E}_{(x,a,r)\sim p(x)\pi_e(a|x)p(r|x,a)}[r]. \tag{1}$$

The value represents the ground-truth performance of the evaluation policy if deployed in an environment of interest. OPE aims to develop an estimator $\hat{V}(\pi_e; \mathcal{D})$ to approximate $V(\pi_e)$ based only on logged data $\mathcal{D}$. A typical choice for $\hat{V}$ is IPS:

$$\hat{V}_{\mathrm{IPS}}(\pi_e; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_e(a_i \mid x_i)}{\pi_0(a_i \mid x_i)} r_i,$$

Beyond IPS, significant efforts have been made to enable more accurate OPE [8, 17, 20, 29, 40, 43, 52].

These conventional estimators evaluate the value of an evaluation policy if it were deployed in the past and apply the estimated value as a prediction of the future policy values. This procedure is valid only in a stationary environment, where distributions (i.e., $p(x), p(r|x, a)$) and policy value ($V(\pi_e)$) remain constant over time. Under such conditions, IPS provides an unbiased estimate of the policy value. However, in most practical situations where the distributions are non-stationary [4, 12, 14], these conventional methods fail to accommodate varying policy values over different time steps, resulting in substantial bias [4].

2.1 Existing Methods for OPE in Nonstationary Environments

Conventional literature on non-stationary OPE differentiates between (1) smooth non-stationarity [14] and (2) abrupt non-stationarity [4, 27, 48]. The former refers to situations where the distributions shift gradually over time, whereas the latter involves scenarios characterized by sudden shifts in the distribution. This section focuses on the most relevant work done by Chandak et al. [4], and Appendix A discusses other relevant work extensively.

Particularly, in abrupt non-stationary environments, the logged data $\mathcal{D}$ can be divided into K distinct subsets $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^K$ where the distribution remains constant within each period. Each subset $\mathcal{D}_k$ is considered generated from a corresponding distribution as

$$\begin{aligned}\mathcal{D}_k &:= \{(x_{i,k}, a_{i,k}, r_{i,k})\}_{i=1}^{n_k} \\ &\sim \prod_{i=1}^{n_k} p_k(x_{i,k})\pi_0(a_{i,k} \mid x_{i,k})p_k(r_{i,k} \mid x_{i,k}, a_{i,k}),\end{aligned}$$

where n_k represents the number of samples observed in period k, and $n = \sum_{k\in[K]} n_k$. For this setting, Chandak et al. [4] proposed the Prognosticator algorithm for OPL, which includes an OPE estimator we regard as a baseline method to evaluate the

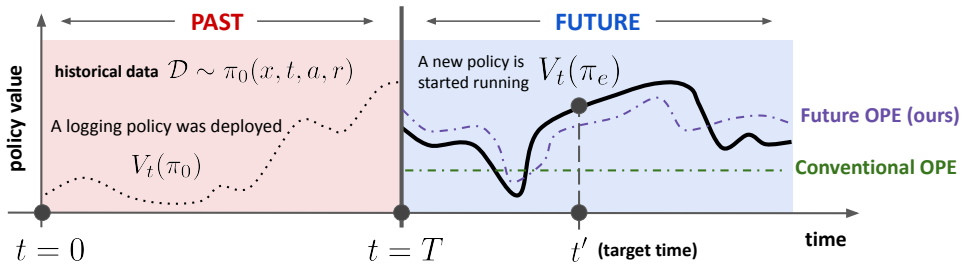


Figure 1: The key concept of Future OPE (F-OPE) and its comparison with conventional OPE, which operates under the assumption of stationarity. In F-OPE, the policy value $V_t(\pi)$ changes over time due to non-stationary distributions, and our goal is to develop an estimator that adapts to this change in the future ($t > T$) yet using only historical logged data $\mathcal{D}$ collected by the logging policy π_0 in the past ($0 \leq t \leq T$).

future performance of an evaluation policy. Prognosticator first performs an estimator $\hat{V}_k(\pi_e;\mathcal{D})$ such as IPS to estimate the past policy value of π_e within each period $k \in [K]$. To estimate the policy value in the future, Prognosticator employs linear regression $\hat{V}_k(\pi_e,\mathcal{D}) \approx \psi(k)^\top w$, where $\psi(k) \in \mathbb{R}^d$ is a basis function encoding the period index, w is a regression coefficient, and $\hat{V}_k(\pi_e,\mathcal{D})$ is an estimated past policy value. The regression is performed using training data (Ψ, Y) where $\Psi := (\psi(1),\cdots,\psi(K))^\top \in \mathbb{R}^{K\times d}$ and $Y := (\hat{V}_1(\pi_e;\mathcal{D}),\cdots,\hat{V}_K(\pi_e;\mathcal{D}))^\top \in \mathbb{R}^K$. The future value estimator for period $K+\delta$ ($\delta = 1,2,\ldots$) is given by:

$$\hat{V}_{K+\delta}(\pi_e;\mathcal{D}) = \psi(K+\delta)^\top(\Psi^\top\Psi)^{-1}\Psi^\top Y,$$

where $w^* := (\Psi^\top\Psi)^{-1}\Psi^\top Y$ is derived from standard ordinary least squares regression. Despite its improved adaptability to non-stationarity compared to conventional OPE, this method faces limitations in both the initial estimation phase and the extrapolation (regression) phase. In the former, Prognosticator experiences significant variance when frequent distribution shifts occur (large K) due to the reduced sample size n_k in each subset $\mathcal{D}_k$. For the extrapolation phase, accurately learning future trends from only the period index k in the past as inputs is extremely challenging with the simple parametric regression, resulting in substantial bias.

3 Our Formulation and Approach

To address the limitations of existing estimators, we first formally formulate the problem of *future OPE*, encompassing both smooth and abrupt changes in distributions. The crux of our new formulation involves considering an unknown probability density $p(t)$ for continuous time $t \in [0,T]$. $[0,T]$ represents the observation window of our historical logged data, where T is the time at which we perform F-OPE of π_e (see Figure 1). In this new formulation, we observe the context x, time t, and action a from the joint distribution $\pi(x,t,a) = p(x,t)\pi(a\,|\,x,t)$. Reward r is then sampled from an unknown conditional distribution $p(r\,|\,x,t,a)$. Note that, in our formulation, the reward distribution can be different for every individual t, and thus **we formulate non-stationarity not just seasonality**. Note also that the conventional formulation of OPE is a special case of our F-OPE formulation with $p(x,t)\pi(a|x,t)p(r|x,t,a) = p(x)\pi(a|x)p(r|x,a),\ \forall t \geq 0$.

The goal in our future OPE problem is to estimate the value of evaluation policy π_e at a given time point t' in the future (which we call the ***target time***), defined as follows.

$$V_{t'}(\pi_e) := \mathbb{E}_{p(x|t')\pi_e(a|x,t')}[q(x,t',a)], \tag{2}$$

where $q(x,t,a) := \mathbb{E}[r\,|\,x,t,a]$ is the expected reward function at time t. Note that we do not impose any specific requirement about the structure of the expected reward $q(x,t,a)$, so this formulation encompasses both gradual and abrupt non-stationarity.

To estimate the value of π_e for a given target time t' ($> T$) , we can use historical logged data $\mathcal{D}$ collected by a logging policy $\pi_0(a\,|\,x,t)$ in the past in the following process:

$$\mathcal{D} := \{(x_i,t_i,a_i,r_i)\}_{i=1}^n \sim \prod_{i=1}^n p(x_i,t_i)\pi_0(a_i\,|\,x_i,t_i)p(r_i\,|\,x_i,t_i,a_i),$$

where $p(x,t) = 0, \forall (x,t)$ for $t > T$ because **we cannot observe any data from future distributions**. This inability to observe any data from the target time $t'(>T)$ makes the F-OPE problem highly non-trivial, and thus the goal of this work is to provide a simple method that is substantially better than existing ideas, rather than to develop a perfect or optimal method in this setup.

It should be noted here that we quantify the accuracy of an estimator by the mean-squared-error (MSE), which can be decomposed into the squared bias and variance as follows.

$$\begin{aligned}\mathrm{MSE}(\hat{V}_{t'}(\pi_e)) &:= \mathbb{E}_{\mathcal{D}}\left[\left(\hat{V}_{t'}(\pi_e;\mathcal{D}) - V_{t'}(\pi_e)\right)^2\right] \\ &= \mathrm{Bias}\left(\hat{V}_{t'}(\pi_e;\mathcal{D})\right)^2 + \mathrm{Var}\left[\hat{V}_{t'}(\pi_e;\mathcal{D})\right].\end{aligned}$$

where $\mathbb{E}_{\mathcal{D}}[\cdot]$ takes the expectation over historical logged data $\mathcal{D}$, and the bias and variance terms are defined as follows.

$$\begin{aligned}\mathrm{Bias}\left[\hat{V}(\pi;\mathcal{D})\right] &:= \mathbb{E}_{\mathcal{D}}\left[\hat{V}(\pi;\mathcal{D})\right] - V(\pi) \\ \mathrm{Var}\left[\hat{V}(\pi;\mathcal{D})\right] &:= \mathbb{E}_{\mathcal{D}}\left[\left(\hat{V}(\pi;\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\left[\hat{V}(\pi;\mathcal{D})\right]\right)^2\right].\end{aligned}$$

3.1 The OPFV Estimator

A core challenge in our estimation problem is that we cannot observe any data (x,t,a,r) from the target time $t'(>T)$ because the historical dataset $\mathcal{D}$ was collected in the past, specifically within the time range $t \in [0,T]$. Despite this seemingly intractable situation, we propose the OPFV estimator, which enables a more accurate estimation of the future value of evaluation policies using only $\mathcal{D}$. The key behind our estimator lies in the utilization

of time series features denoted by the function $\phi(t)$. Essentially, ϕ can be any function that forms a clustering of time t with typical choices including the day of the week, the month, the season, or a combination of these.[1] Even if the target time $t'(>T)$ does not appear in the historical data $\mathcal{D}$, we may still find many timestamps observed in $\mathcal{D}$ that share the same time features as the target time, i.e., $i : \phi(t_i) = \phi(t')$. Leveraging this structure in the time series, we can estimate the future value of new policies at least substantially more accurately than conventional methods such as IPS, DR, and Prognosticator of Chandak et al. [4], if not perfectly.

To develop a new estimator, for a given time feature function ϕ, we consider the following decomposition of the reward function:

$$q(x,t,a) = \underbrace{g(x,\phi(t),a)}_{\textit{time feature effect}} + \underbrace{h(x,t,a)}_{\textit{residual effect}}, \tag{3}$$

where we decompose $q(x,t,a)$ into the *time feature effect* and *residual effect*. Note that we do not impose any restrictions on the functional forms of g and h, and thus Eq. (3) is not an assumption.

Based on Eq. (3), even when considering a target time in the future, we can unbiasedly estimate the time feature effect, by applying importance weighting to the datapoints that share the same feature with the target time ($i : \phi(t_i) = \phi(t')$). This idea leads us to propose the OPFV estimator:

$$\hat{V}^{\mathrm{OPFV}}_{t'}(\pi_e;\mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{I}_\phi(t_i,t')}{p(\phi(t'))}\frac{\pi_e(a_i\,|\,x_i,t')}{\pi_0(a_i\,|\,x_i,t_i)}\Big(r_i-\hat{f}(x_i,t_i,a_i)\Big) + \mathbb{E}_{\pi_e(a|x_i,t')}\left[\hat{f}(x_i,t',a)\right]\right\}, \tag{4}$$

where $\mathbb{I}_\phi(t,t') := \mathbb{I}\{\phi(t)=\phi(t')\}$ denotes the indicator that outputs 1 if t and t' have an identical time feature and 0 otherwise. $p(\phi(t)) := \int_{s\in[0,T]} p(s)\mathbb{I}_\phi(s,t)ds$ is the marginal probability density of ϕ. OPFV consists of two terms where the first term applies the new weight $\mathbb{I}_\phi(t_i,t')/p(\phi(t'))$ in addition to the original importance weight $\pi_e(a\,|\,x,t')/\pi_0(a\,|\,x,t_i)$ to estimate the time feature effect non-parametrically by using only the samples in $\mathcal{D}$ whose time feature is identical to that of the target time t'. The second term aims to deal with the residual effect in Eq. (3) based on a reward regressor $\hat{f}$ obtained via solving $\min_f \sum_{(x,t,a,r)\in\mathcal{D}}(r-f(x,t,a))^2$. We do not expect this regressor to be highly accurate, but it often reduces bias compared to completely ignoring the residual effect.

Intuitively, OPFV can estimate the future policy value at arbitrary target time t', even in the distant future, with much greater precision than existing methods, which do not leverage any structure within time-series. Another advantage of OPFV is its simplicity; it can readily be implemented on top of DR [8] by applying the additional importance weighting factor regarding the time-series feature, yet leading to substantial improvements in the F-OPE problem as we will demonstrate. In addition, our estimator not only enables much more accurate estimations, but also enables a novel and more insightful analysis of the F-OPE problem as done below.

[1] For example, to formulate a weekly effect, we can use the function $\phi : [0,t'] \to \{\text{Monday}, \text{Tuesday}, \ldots, \text{Sunday}\}$. Note that ϕ may not necessarily represent some type of seasonal effects; we can use any function as ϕ, and we will later discuss how to optimize it based only on historical data.

3.2 Theoretical Analysis

To analyze the bias and variance of OPFV, we are based on the following regularity conditions.

Condition 3.1 (Common Support)**.** The logging policy π_0 is said to satisfy the common support for a given π_e and t' if $\pi_e(a\,|\,x,t') > 0 \implies \pi_0(a\,|\,x,t) > 0$ for any $x \in \mathcal{X}$, $t \in [0,T]$, and $a \in \mathcal{A}$.

Condition 3.1 requires that π_0 assigns non-zero probabilities to the actions whose probabilities are positive under π_e, which is a standard condition in most OPE estimators [8, 32, 34].

Condition 3.2 (Common Time Feature Support)**.** The time feature function ϕ is said to satisfy the common time feature support for a given target time t' if $p(\phi(t')) > 0$.

Condition 3.2 requires that we can find a datapoint i that shares the same time feature with the target time with some positive probability. Given that we can define ϕ by ourselves, this condition is within our control.

We can now derive the bias and variance of OPFV as follows.

Theorem 3.3 (Bias Analysis)**.** *Under Conditions 3.1 and 3.2, OPFV has the following bias.*

$$\mathrm{Bias}\left(\hat{V}^{\mathrm{OPFV}}_{t'}(\pi_e;\mathcal{D})\right) = \mathbb{E}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\left(\Delta_q(x,t,t',a)-\Delta_{\hat{f}}(x,t,t',a)\right)\right], \tag{5}$$

where the expectation is taken over $p(x,t)\pi_e(a\,|\,x,t')$. $\Delta_q(x,t,t',a) := q(x,t,a) - q(x,t',a)$ is the relative difference in the expected rewards between timestamps t,t' given x and a, and $\Delta_{\hat{f}}(x,t,t',a) := \hat{f}(x,t,a) - \hat{f}(x,t',a)$ is its estimate based on $\hat{f}(x,t,a)$. See Appendix H.2 for the proof.

Proposition 3.4 (Variance of OPFV)**.** *Under Conditions 3.1 and 3.2, the variance of OPFV is*

$$\begin{aligned} & n\,\mathrm{Var}\left[\hat{V}^{\mathrm{OPFV}}_{t'}(\pi_e;\mathcal{D})\right] \\ &= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\right)^2\sigma^2(x,t,a)\right] \\ &\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)\right]\right] \\ &\quad + \mathbb{V}_{p(t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right]\mathbb{E}_{p(x)}\left[\mathbb{E}_{\pi_e(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)\right]^2\right] \\ &\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_e(a|x,t')}\left[q(x,t',a)\right]\right] \end{aligned} \tag{6}$$

where $\Delta_{q,\hat{f}}(x,t,a) := q(x,t,a) - \hat{f}(x,t,a)$ is the estimation error of $\hat{f}$, and $\sigma^2(x,t,a) := \mathbb{V}[r\,|\,x,t,a]$ is the conditional variance of the reward. See Appendix H.3 for the proof.

Theorem 3.3 shows that there are two key factors that characterize the bias of OPFV. The first factor is the coarseness of the time feature function. We can see in Eq. (5) that the bias occurs only when time $t \in [0,T]$ shares the same time feature with the target time, i.e., $i : \mathbb{I}_\phi(t,t') = 1$. Therefore, when we use a fine-grained time feature ϕ and $\mathbb{I}_\phi(t,t') = 1$ is less likely, the bias of OPFV becomes closer to zero. The second factor is the accuracy of the

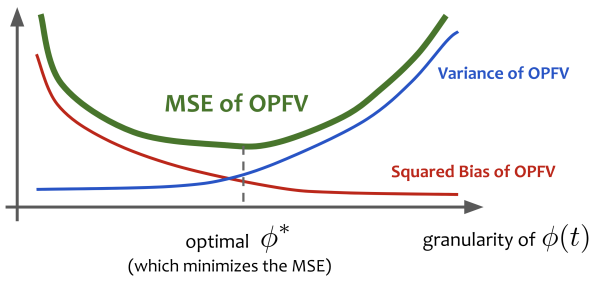


Figure 2: The bias-variance tradeoff of OPFV is controlled by the granularity or the cardinality of the time feature, $|\phi|$. A coarser time feature reduces variance at the cost of introducing some bias. On the other hand, a finer-grained time feature decreases bias but increases variance. An optimal time feature function ϕ^* that minimizes the MSE is often the one that balances the squared bias and variance.

regressor $\hat{f}$ used in OPFV regarding the estimation of $\Delta_q(x,t,t',a)$ for pair of timestamps t and t' that share the same time feature. When t and t' have the same time feature, i.e., $\phi(t) = \phi(t')$, the difference in their expected rewards is attributed to the difference in their residual effects, i.e., $\Delta_q(x,t,t',a) = h(x,t,a) - h(x,t',a)$. Therefore, the second factor $\Delta_q(x,t,t',a) - \Delta_{\hat{f}}(x,t,t',a)$ characterizes how accurately the regressor $\hat{f}$ estimates the variation of the residual effect. Overall, Theorem 3.3 suggests that the bias of OPFV arises only from the estimation error of the residual effect, because the importance weighting term already unbiasedly estimates the time feature effect. As the time feature becomes finer, the residual effect becomes smaller, so does the bias of OPFV. Appendix D takes advantage of this bias analysis and discusses how to optimize the regression model $\hat{f}$ to directly minimize the bias of OPFV.

Next, Proposition 3.4 suggests that the variance of the weight $\mathbb{I}_\phi(t,t')/p(\phi(t'))$ contributes to the variance of OPFV. This observation implies that the granularity of the time feature function ϕ influences the variance of OPFV via the variance of $\mathbb{I}_\phi(t,t')/p(\phi(t'))$. Specifically, using a coarser time feature leads to less variation in the weight $\mathbb{I}_\phi(t,t')/p(\phi(t'))$, resulting in a smaller variance for OPFV. Conversely, employing a finer-grained time feature in OPFV tends to increase its variance.

Coupled with the bias analysis described earlier, we now know that the granularity of the time feature function dictates the bias-variance tradeoff in OPFV. A coarser time feature reduces variance at the cost of introducing some bias. On the other hand, a finer-grained time feature decreases bias but increases variance. Therefore, optimizing the function ϕ, is crucial, so OPFV achieves minimal MSE (as depicted in Figure 2).

3.3 Data-Driven Optimization of Time Feature

The theoretical analysis tells us that the accuracy of the OPFV estimator depends on what time feature function we use. To be able to leverage this fact to further improve OPFV, we provide a data-driven method to optimize the time feature function ϕ using only the historical data. Ideally, we aim to optimize it via $\phi^* \in \arg\min_{\phi\in\Phi} \mathrm{MSE}(\hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D},\phi))$ where Φ is a candidate set of time feature functions that we set. However, the true MSE is unknown as it depends on the true future value ($V_{t'}(\pi_e)$), so we are based on its estimate to perform a data-driven optimization of ϕ:

$$\hat{\phi} \in \arg\min_{\phi\in\Phi} \widehat{\mathrm{Bias}}(\phi)^2 + \widehat{\mathrm{Var}}(\phi), \tag{7}$$

where $\widehat{\mathrm{Bias}}(\phi) := \hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D},\phi) - \hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D},\phi_\infty)$ is an estimate of the bias of OPFV and ϕ_∞ is the finest time feature function in Φ, which leads to the lowest bias. To estimate the bias, we can also use other methods with theoretical guarantees such as the ones proposed by [5, 10, 35, 41, 43, 44, 50]. The variance of OPFV can be estimated by its sample variance [52]. Note that it is straightforward to satisfy Condition 3.2 even after Eq. (7) by including only the candidate functions that satisfy this condition in Φ. In addition, Eq. (7) is akin to hyperparameter optimization in supervised learning, and its computational cost is within our control provided that we can control the cardinality $|\Phi|$.

3.4 Extension to Future Off-Policy Learning

We have thus far focused on evaluating the future value of a new policy in a non-stationary environment. In this section, we extend the problem of F-OPE to F-OPL, aimed at optimizing the expected reward in the future. Specifically, for the F-OPL problem, we aim to optimize the expected reward at some future target time t' ($> T$). In particular, we use a policy-based approach to learn the parameter ζ of a parameterized policy $\pi_\zeta(a\,|\,x)$ to maximize the future value, i.e., $\zeta^* = \arg\max_\zeta V_{t'}(\pi_\zeta)$. The policy-based approach updates the policy parameter via iterative gradient ascent as $\zeta_{\tau+1} \leftarrow \zeta_\tau + \eta\nabla_\zeta V_{t'}(\pi_{\zeta_\tau})$, where η is the learning rate. Since we have no access to the true policy gradient $\nabla_\zeta V_{t'}(\pi_\zeta) = \mathbb{E}_{p(x|t')\pi_\zeta(a|x)}\left[q(x,t',a)\nabla_\zeta \log \pi_\zeta(a\,|\,x)\right]$, we need to estimate it using historical data while accounting for non-stationarity. We can achieve this as

$$\begin{aligned}&\nabla_\zeta \hat{V}_{t'}^{\mathrm{OPFV}}(\pi_\zeta;\mathcal{D})\\ &:= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{I}_\phi(t_i,t')}{p(\phi(t'))}\frac{\pi_\zeta(a_i\,|\,x_i)}{\pi_0(a_i\,|\,x_i,t_i)}\Big(r_i - \hat{f}(x_i,t_i,a_i)\Big)\nabla_\zeta \log \pi_\zeta(a_i\,|\,x_i)\right.\\ &\qquad\left. + \mathbb{E}_{\pi_\zeta(a|x_i)}\left[\hat{f}(x_i,t',a)\nabla_\zeta \log \pi_\zeta(a\,|\,x_i)\right]\right\}.\end{aligned} \tag{8}$$

Appendix F derives the bias and variance of this policy-gradient estimator where we obtain similar theoretical observations with those for the analysis regarding F-OPE. Note that we call the policy gradient (PG) method for F-OPL based on the estimator in Eq. (8) as OPFV-PG. It is also straightforward to extend OPFV-PG to incorporate pessimism by combining it with relevant methods [15, 22, 25, 28].

3.5 Discussion on Relevant Techniques

This section discusses some relevant techniques around OPE that might look similar to the contributions of this work.

First, we differentiate our contributions from those of Uehara et al. [51], which studied OPE and OPL under covariate shift, where the context distributions differ between the logged data ($p^{\mathrm{hist}}(x)$) and in the evaluation environment ($p^{\mathrm{eval}}(x)$), while the reward

distribution $p(r|x,a)$ does not change. Our setup might appear similar to that of Uehara et al. [51] when we formulate the combined feature $\tilde{x} = (x,t)$. However, the most critical difference is that our focus is on evaluating and optimizing the policy value for *future* times. This implies that our target time $t' (> T)$ has zero density in the logged data $\mathcal{D}$, making the density ratio $p^{\mathrm{eval}}(\tilde{x}')/p^{\mathrm{hist}}(\tilde{x}') = p^{\mathrm{eval}}(x,t')/p^{\mathrm{hist}}(x,t')$ nonexistent. In contrast, Uehara et al. [51] assume the existence of the density ratio of $p(\tilde{x})$ (i.e., the assumption of strong overlap, $p^{\mathrm{eval}}(\tilde{x}') > 0 \implies p^{\mathrm{hist}}(\tilde{x}') > 0$ as described in their Assumptions 1 and 2), a condition our setup inherently does not satisfy as $p^{\mathrm{hist}}(x,t') = 0$ for all $x \in \mathcal{X}$ due to the inability to observe any data from future distributions. We have managed to address and analyze this seemingly intractable situation for the first time through the concepts of reward function decomposition and importance weighting regarding the time feature function ϕ. Thus, our problem setup, approach, and associated analysis (particularly the bias analysis given in Theorem 3.3) are conceptually and technically different from Uehara et al. [51]. Another notable difference is that Uehara et al. [51] assumes access to the realizations of the context from the distribution in the evaluation period ($p^{\mathrm{eval}}(\tilde{x}')$) to estimate the density ratio, but this is not possible in our setup where we cannot observe any data from the future distributions. Therefore, the method from Uehara et al. [51] (and its variants) is simply infeasible in our problem of F-OPE.

Second, we discuss critical differences between our contributions and those of Saito et al. [40], which develops the OffCEM estimator to deal with OPE in large action spaces in a stationary setup. Even though our main idea of reward function decomposition is inspired by OffCEM, its application to deal with the non-stationarity is non-trivial and was never mentioned in Saito et al. [40]. Moreover, we discuss how to identify an appropriate time-future function in a data-driven fashion as in Section 3.3 and perform associated empirical evaluations, which are missing in Saito et al. [40]. We also propose an extension of our estimator to an OPL method while Saito et al. [40] focused solely on the OPE problem. Therefore, our work is the first to formulate the F-OPE/L problem and offer multiple unique contributions methodologically and empirically.

4 Empirical Evaluation

This section empirically compares our proposed methods with existing methods on both synthetic and real-world data.[2]

Synthetic Data. In our synthetic experiment, we consider F-OPE and F-OPL problems where the logged data are collected during **Year 1**, and we are interested in estimating the future values of a new policy at some target times in **Year 2**. To generate historical datasets synthetically, we first sample Unix timestamps uniformly within the range from January 1st ($t = 0$) to December 31st ($t = T$) in **Year 1**. We then sample a 10-dimensional context vector x from the standard normal distribution and synthesize the expected reward function as

$$q(x,t,a;\lambda) = \lambda \cdot g(x, \phi_{\mathrm{true}}(t), a) + (1-\lambda) \cdot h(x,t,a), \qquad (9)$$

where λ is an experimental parameter that controls the informativeness of the true time feature ϕ_{true}. The true time feature considers seasonality, evenly dividing a year into eight different seasons (i.e., $\phi_{\mathrm{true}} : t \to \{\text{season } 1, \ldots, \text{season } 8\}$).[3] Note that the h function in Eq. (9) produces the effect of individual timestamp t, so **timestamps that have the same time feature do not have the same expected reward, producing non-stationarity**. The concrete definitions of the g and h functions are provided in Appendix I.1.1.

We then synthesize the logging policy π_0, which produced the historical dataset $\mathcal{D}$ during Year 1, by applying the softmax function to the expected reward $q(x,t,a)$ over the range $t \in [0,T]$ as

$$\pi_0(a \mid x,t;\beta) = \frac{\exp\left(\beta \cdot q(x,t,a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\beta \cdot q(x,t,a')\right)},$$

where β is an experimental parameter to define π_0. We use $\beta = 0.1$ and set the cardinality of the action space to $|\mathcal{A}| = 10$ throughout the experiment. Lastly, we sample the reward r from a normal distribution with mean $q(x,t,a)$ and a standard deviation of $\sigma = 1.0$. The logged data $\mathcal{D} = \{(x_i, t_i, a_i, r_i)\}_{i=1}^n$ is generated by repeating the above sampling procedure n times independently.

We employ an epsilon-greedy policy to define π_e that is to be deployed in the future ($t' > T$):

$$\pi_e(a \mid x,t';\epsilon) = (1-\epsilon) \cdot \left\{a = \arg\max_{a' \in \mathcal{A}} q(x,t',a')\right\} + \epsilon/|\mathcal{A}|,$$

where $\epsilon \in [0,1]$ is a noise parameter that determines the quality of π_e. We default to using $\epsilon = 0.2$.

For F-OPE, we compare OPFV with both true (ϕ_{true}) and tuned ($\hat{\phi}$) time features against OPE estimators for stationary environments (IPS and DR) and Prognosticator [4] as baselines. For F-OPL, we compare OPFV-PG with the regression-based (RegBased) and policy-based (IPS-PG and DR-PG) methods as baselines for stationary environments, and Prognosticator as a baseline that considers non-stationarity. The baseline methods in the F-OPL experiment are defined rigorously in Appendix I.1.4. We use a Random Forest model to estimate $q(x,a)$ for DR and $q(x,t,a)$ for OPFV. Note that we use $n = 1000$ for F-OPE, $n = 8000$ for F-OPL, $\lambda = 0.5$ for both F-OPE and F-OPL experiments as default configurations.

Figure 3 (i) illustrates the true future values of π_e and the respective estimated values for varying target times t', spanning from season 1 to season 8 in Year 2, which is after the end of the collection of the logged data.[4] At each target time t', OPFV with both true (ϕ_{true}) and tuned ($\hat{\phi}$) time features consistently tracks the fluctuations of the true future value $V_{t'}(\pi_e)$ by selectively leveraging the samples whose timestamps share the identical time feature with t'. In contrast, Prognosticator occasionally fails substantially, such as seasons 1 and 2, due to its unreliable regression phase. Figure 3 (ii) compares the estimators' accuracy under varying levels of informativeness of the time features (λ in Eq. (9)). The figure clearly shows that, with λ increasing, OPFV becomes increasingly more accurate than other estimators that do not leverage the time features. In contrast, when the time features are less useful ($\lambda \leq 0.2$), all estimators perform similarly, which could be seen as a limitation of our method; it performs similarly to the conventional methods if any useful structure within time-series does not exist. However,

[2] The experiment code is available at https://github.com/sony/ds-research-code/tree/master/kdd2025-opfv .

[3] Specifically, "season 1" corresponds to Jan 1st to Feb 14th, and "season 8" corresponds to Nov 16th to Dec 31st. These eight seasons are repeated in Year 1 and Year 2.

[4] Note that here we plot the true future policy values and estimated policy values averaged within each season to ease the discussion, but there also exists non-stationarity within each season in our experiments, as we can see in Eq. (9).

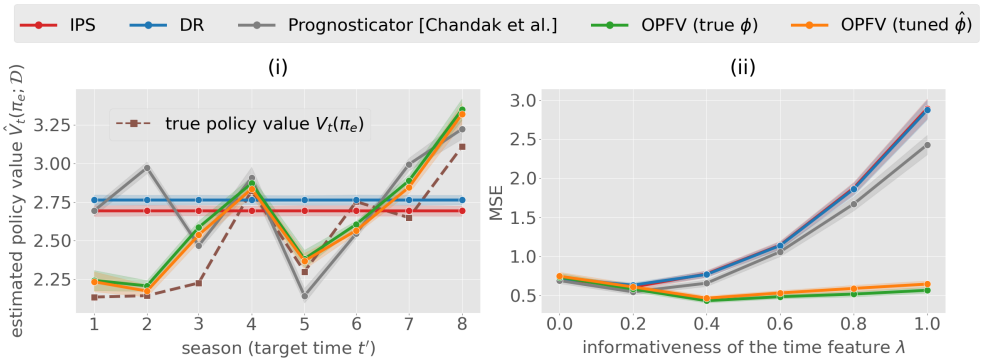


Figure 3: Comparing (i) the true future policy value $V_t(\pi_e)$ and the estimated future policy values $\hat{V}_t(\pi_e; \mathcal{D})$ with varying target times t' in the future and (ii) the MSE of estimators with varying λ values that control the informativeness of the time feature effect as in Eq. (9).



Figure 4: Comparison of the MSE, squared bias, and variance of estimators with varying cardinalities of time features $|\hat{\phi}|$ used in OPFV (which can be different from the true time feature ϕ_{true} that defines the reward function) in Eq. (9) where the cardinality of the true time features is fixed at eight.

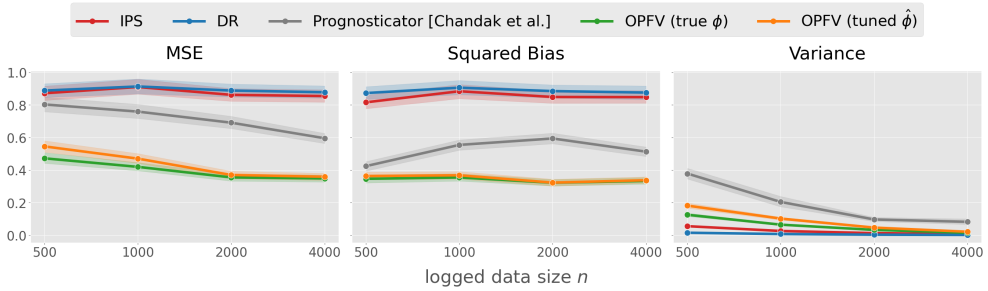


Figure 5: Comparison of the MSE, squared bias, and variance of estimators in F-OPE with varying logged data sizes n.

this also suggests that the worst case performance of OPFV with no useful structure in the time-series reduces to that of existing methods, and therefore there is no reason to rely on existing methods under non-stationarity. We can also see that OPFV with the tuned time feature ($\hat{\phi}$) is almost as effective as OPFV with the true time feature (ϕ_{true}), suggesting our methods' applicability even without the availability of the true function ϕ_{true}.

Figure 4 shows the effect of the cardinality of the time features $|\hat{\phi}|$ used for OPFV on its bias-variance tradeoff. Note that we change only the cardinality of the time features used in OPFV, $|\hat{\phi}|$, while the true time feature ϕ_{true} that defines the reward function in Eq. (9) is fixed in this experiment. We can see from the figure that, as the cardinality of the used time features increases, the bias of OPFV decreases until the time feature becomes finer than the true one (i.e., $|\hat{\phi}| \geq |\phi_{\text{true}}| = 8$), which is consistent with our bias analysis in Theorem 3.3. Conversely, the variance of OPFV slightly increases with the cardinality of the time features, which is also consistent with our theoretical observations. It also implies that the choice of ϕ may greatly affect the MSE of OPFV, which highlights the importance of the data-driven optimization procedure, which we described in Section 3.3.

Figure 5 reports the accuracy of the estimators when we vary the logged data sizes from 500 to 4000. Overall, OPFV performs more accurately than the baseline methods irrespective of the data size, because it can at least unbiasedly estimate the time feature effect, leading to the lowest bias as we can see in the middle figure. Note that OPFV has higher variance than IPS and DR due to the additional weighting factor to estimate the time feature effect, but its bias reduction advantage is much larger, making OPFV the most effective. IPS and DR produce the largest bias due to their unrealistic

Table 1: Comparison of the future policy values learned by F-OPL methods in the test set of the Kuairec dataset. Mean and standard deviation of the future policy values over 10 runs are reported.

	RegBased	IPS-PG	DR-PG	Prognosticator	OPFV-PG (**w/o** tuned ϕ)	OPFV-PG (**w/** tuned ϕ)
Mean Value(± SD)	1.48(± 0.31)	1.02(± 0.70)	1.15(± 0.57)	1.39(± 0.59)	1.60(± 0.56)	**1.67(± 0.63)**

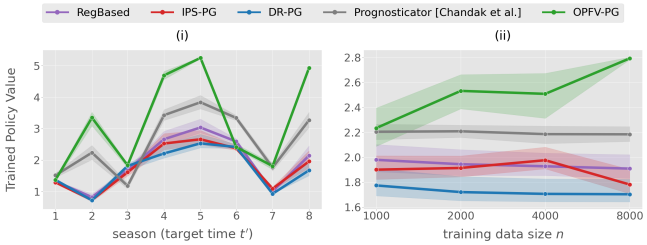


Figure 6: Comparison of the policy values (higher the better) achieved by F-OPL methods with (left) varying target times t' and (right) training data sizes n.

stationarity assumption. Prognosticator fails particularly when the data size is limited due to the significant variance produced when performing OPE in each period in the historical data.

Figure 6 (i) presents the future policy values achieved by five OPL methods with varying target times t'. We observe that OPFV-PG mostly achieves the highest value adaptive to different target times t' due to its effective usage of the time features in estimating the policy gradient for future values. Prognosticator rarely produces a better policy than OPFV-PG and it ends up with a policy even worse than the conventional OPL methods in season 3. In contrast, OPFV-PG always performs better than the conventional methods. Figure 6 (ii) reports the earned future policy values with varying training data sizes. As the training data size n increases, OPFV-PG is able to learn increasingly better policies because of its low-bias estimation of the policy gradient $\nabla_\zeta V_{t'}(\pi_\zeta)$ for any target time t'. In contrast, the baseline methods are based on biased policy gradients due to their inability to deal with non-staitonarity, so they do not necessarily improve the future policy values even with increased training data. Appendix I provides additional results about F-OPL showing that OPFV-PG outperforms existing methods particularly when the g function in Eq. (9) has larger effect (larger λ) and the cardinality of the used time feature $|\phi_{\text{true}}|$ is not too small.

Real-World Data. This section assesses the real-world applicability of our method using the KuaiRec dataset [11], which consists of recommendation logs of the video-sharing app, Kuaishou, with timestamps from July 5th to September 5th, 2020. Each record in the dataset includes a user ID, timestamp t, recommended video ID as action a, and the watch ratio of the recommended video as reward $r \in [0, \infty)$, which represents the play duration divided by the video duration. Each user and video is associated with user and video features, which we consider as context x and features of action a, respectively. We omit variables in the user and video features that are too sparse or have missing values. To construct an action space, we uniformly subsample 100 actions from the original set of 3327 unique videos due to computational efficiency reasons, but non-stationarity of the dataset remains after this pre-processing as we validate in Appendix J.1. The range of the timestamps in the logged data spans from July 5th (corresponding to $t = 0$) to August 4th (corresponding to $t = T$), while the test data covers the period from August 5th to September 5th (considered as the future) to perform an F-OPL experiment under non-stationarity.

We compare the future policy values achieved by OPFV-PG (**w/** tuned ϕ), OPFV-PG (**w/o** tuned ϕ), Reg-based, IPS-PG, DR-PG, and Prognosticator. All methods train a policy parameterized by a neural network with three hidden layers. We used the day of the week as the time feature for OPFV-PG (**w/o** tuned ϕ). We optimized the time feature function used for OPFV-PG (**w/** tuned ϕ) based only on the training data following the procedure described in Section 3.3, and we refer the reader to Appendix J.2 for more details about the tuning procedure employed in the real-world experiment.

Table 1 presents the mean and standard deviation of the values of the learned policies across 10 experimental runs with varying random seeds. It is important to note that the values in Table 1 were estimated by averaging three distinct OPE estimators on the test set to produce reliable results. The results with the three individual estimators (prior to averaging) are detailed in Appendix J.4, and they lead to a similar conclusion. The table illustrates that the two versions of OPFV-PG consistently achieve much higher future policy values than existing methods even under unknown and possibly complicated non-stationarity of the real-world data. In particular, the promising result of OPFV-PG (**w/** tuned ϕ) demonstrates the real-world applicability of the optimization procedure of the time feature function presented in Section 3.3. We also observe that OPFV-PG (**w/o** tuned ϕ) generally outperforms the baseline OPL methods, which implies that our method is at least substantially more effective than existing methods that do not deal with non-stationarity or leverage the structure in the time-series, even without carefully tuning ϕ.

5 Conclusion

This paper formulated the problem of *future* OPE and OPL for estimating and optimizing the future policy values in non-stationary environments. We then proposed the OPFV estimator, which leverages time series features to unbiasedly estimate the time feature effect via importance weighting. We have demonstrated that OPFV is low-bias when the time feature is fine-grained or the regression component accurately estimates the pairwise differences of the residual effect, resulting in much more effective F-OPE and F-OPL compared to existing ideas. We also showed how our estimator can readily be extended to a policy-gradient estimator to solve the future OPL problem. Experiments demonstrated that our methods outperform existing ones in a variety of environments on both synthetic data with simplified settings and real-world data with non-trivial non-stationarity.

References

[1] A Agarwal, N Jiang, and M Kakade, S. 2020. Reinforcement Learning: Theory and algorithms. *MIT* (2020).

[2] Alina Beygelzimer and John Langford. 2009. The Offset Tree for Learning with Partial Labels. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 129–138.

[3] Yash Chandak, Shiv Shankar, Nathaniel Bastian, Bruno da Silva, Emma Brunskill, and Philip S Thomas. 2022. Off-policy evaluation for action-dependent non-stationary environments. *Advances in Neural Information Processing Systems* 35 (2022), 9217–9232.

[4] Yash Chandak, Georgios Theocharous, Shiv Shankar, Martha White, Sridhar Mahadevan, and Philip Thomas. 2020. Optimizing for the future in non-stationary mdps. In *International Conference on Machine Learning*. PMLR, 1414–1425.

[5] Matej Cief, Michal Kompan, and Branislav Kveton. 2024. Cross-Validated Off-Policy Evaluation. *arXiv preprint arXiv:2405.15332* (2024).

[6] Omar Darwiche Domingues, Pierre Ménard, Matteo Pirotta, Emilie Kaufmann, and Michal Valko. 2021. A kernel-based approach to non-stationary reinforcement learning in metric spaces. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3538–3546.

[7] Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. 2014. Doubly Robust Policy Evaluation and Optimization. *Statist. Sci.* 29, 4 (2014), 485–511.

[8] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*. 1097–1104.

[9] Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. 2018. More Robust Doubly Robust Off-Policy Evaluation. In *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80. PMLR, 1447–1456.

[10] Nicolò Felicioni, Michael Benigni, and Maurizio Ferrari Dacrema. 2024. AutoOPE: Automated Off-Policy Estimator Selection. *arXiv preprint arXiv:2406.18022* (2024).

[11] Chongming Gao, Shijun Li, Wenqiang Lei, Jiawei Chen, Biao Li, Peng Jiang, Xiangnan He, Jiaxin Mao, and Tat-Seng Chua. 2022. KuaiRec: A fully-observed dataset and insights for evaluating recommender systems. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 540–550.

[12] Joey Hong, Branislav Kveton, Manzil Zaheer, Yinlam Chow, and Amr Ahmed. 2021. Non-stationary off-policy optimization. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2494–2502.

[13] Daniel G Horvitz and Donovan J Thompson. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association* 47, 260 (1952), 663–685.

[14] Rolf Jagerman, Ilya Markov, and Maarten de Rijke. 2019. When people change their mind: Off-policy evaluation in non-stationary recommendation environments. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 447–455.

[15] Olivier Jeunen and Bart Goethals. 2021. Pessimistic reward models for off-policy learning in recommendation. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 63–74.

[16] Nathan Kallus, Xiaojie Mao, Kaiwen Wang, and Zhengyuan Zhou. 2022. Doubly robust distributionally robust off-policy evaluation and learning. In *International Conference on Machine Learning*. PMLR, 10598–10632.

[17] Nathan Kallus, Yuta Saito, and Masatoshi Uehara. 2021. Optimal off-policy evaluation from multiple logging policies. In *International Conference on Machine Learning*. PMLR, 5247–5256.

[18] Haruka Kiyohara, Ren Kishimoto, Kosuke Kawakami, Ken Kobayashi, Kazuhide Nakata, and Yuta Saito. 2024. Towards Assessing and Benchmarking Risk-Return Tradeoff of Off-Policy Evaluation. In *The Twelfth International Conference on Learning Representations*.

[19] Haruka Kiyohara, Masahiro Nomura, and Yuta Saito. 2024. Off-policy evaluation of slate bandit policies via optimizing abstraction. In *Proceedings of the ACM on Web Conference 2024*. 3150–3161.

[20] Haruka Kiyohara, Yuta Saito, Tatsuya Matsuhiro, Yusuke Narita, Nobuyuki Shimizu, and Yasuo Yamamoto. 2022. Doubly Robust Off-Policy Evaluation for Ranking Policies under the Cascade Behavior Model. In *Proceedings of the 15th International Conference on Web Search and Data Mining*.

[21] Haruka Kiyohara, Masatoshi Uehara, Yusuke Narita, Nobuyuki Shimizu, Yasuo Yamamoto, and Yuta Saito. 2023. Off-Policy Evaluation of Ranking Policies under Diverse User Behavior. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1154–1163.

[22] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 33. 1179–1191.

[23] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv preprint arXiv:2005.01643* (2020).

[24] Mengbing Li, Chengchun Shi, Zhenke Wu, and Piotr Fryzlewicz. 2022. Testing stationarity and change point detection in reinforcement learning. *arXiv preprint arXiv:2203.01707* (2022).

[25] Dawen Liang and Nikos Vlassis. 2022. Local Policy Improvement for Recommender Systems. *arXiv preprint arXiv:2212.11431* (2022).

[26] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. 2018. Breaking the curse of horizon: infinite-horizon off-policy estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 5361–5371.

[27] Vincent Liu, Yash Chandak, Philip Thomas, and Martha White. 2023. Asymptotically Unbiased Off-Policy Policy Evaluation when Reusing Old Data in Nonstationary Environments. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 5474–5492.

[28] Yifei Ma, Yu-Xiang Wang, and Balakrishnan Narayanaswamy. 2019. Imitation-regularized offline learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2956–2965.

[29] Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. 2021. Subgaussian and Differentiable Importance Sampling for Off-Policy Evaluation and Learning. *Advances in Neural Information Processing Systems* 34 (2021).

[30] Tong Mu, Yash Chandak, Tatsunori B Hashimoto, and Emma Brunskill. 2022. Factored DRO: Factored Distributionally Robust Policies for Contextual Bandits. *Advances in Neural Information Processing Systems* 35 (2022), 8318–8331.

[31] Sindhu Padakandla, Prabuchandran KJ, and Shalabh Bhatnagar. 2020. Reinforcement learning algorithm for non-stationary environments. *Applied Intelligence* 50, 11 (2020), 3590–3606.

[32] Doina Precup, Richard S. Sutton, and Satinder P. Singh. 2000. Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the 17th International Conference on Machine Learning*. 759–766.

[33] James Robins, Mariela Sued, Quanhong Lei-Gomez, and Andrea Rotnitzky. 2007. Comment: Performance of double-robust estimators when" inverse probability" weights are highly variable. *Statist. Sci.* 22, 4 (2007), 544–559.

[34] Noveen Sachdeva, Yi Su, and Thorsten Joachims. 2020. Off-Policy Bandits with Deficient Support. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 965–975.

[35] Noveen Sachdeva, Lequn Wang, Dawen Liang, Nathan Kallus, and Julian McAuley. 2024. Off-policy evaluation for large action spaces via policy convolution. In *Proceedings of the ACM on Web Conference 2024*. 3576–3585.

[36] Yuta Saito, Himan Abdollahpouri, Jesse Anderton, Ben Carterette, and Mounia Lalmas. 2024. Long-term Off-Policy Evaluation and Learning. In *Proceedings of the ACM on Web Conference 2024*. 3432–3443.

[37] Yuta Saito, Shunsuke Aihara, Megumi Matsutani, and Yusuke Narita. 2021. Open Bandit Dataset and Pipeline: Towards Realistic and Reproducible Off-Policy Evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[38] Yuta Saito and Thorsten Joachims. 2021. Counterfactual Learning and Evaluation for Recommender Systems: Foundations, Implementations, and Recent Advances. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 828–830.

[39] Yuta Saito and Thorsten Joachims. 2022. Off-Policy Evaluation for Large Action Spaces via Embeddings. In *International Conference on Machine Learning*. PMLR, 19089–19122.

[40] Yuta Saito, Ren Qingyang, and Thorsten Joachims. 2023. Off-Policy Evaluation for Large Action Spaces via Conjunct Effect Modeling. In *International Conference on Machine Learning*. PMLR, 29734–29759.

[41] Yuta Saito, Jihan Yao, and Thorsten Joachims. 2024. POTEC: Off-Policy Learning for Large Action Spaces via Two-Stage Policy Decomposition. *arXiv preprint arXiv:2402.06151* (2024).

[42] Nian Si, Fan Zhang, Zhengyuan Zhou, and Jose Blanchet. 2020. Distributionally robust policy evaluation and learning in offline contextual bandits. In *International Conference on Machine Learning*. PMLR, 8884–8894.

[43] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. 2020. Doubly Robust Off-Policy Evaluation with Shrinkage. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. PMLR, 9167–9176.

[44] Yi Su, Pavithra Srinath, and Akshay Krishnamurthy. 2020. Adaptive estimator selection for off-policy evaluation. In *International Conference on Machine Learning*. PMLR, 9196–9205.

[45] Yi Su, Lequn Wang, Michele Santacatterina, and Thorsten Joachims. 2019. Cab: Continuous adaptive blending for policy evaluation and learning. In *International Conference on Machine Learning*, Vol. 84. 6005–6014.

[46] Adith Swaminathan and Thorsten Joachims. 2015. The Self-Normalized Estimator for Counterfactual Learning. *Advances in Neural Information Processing Systems* 28 (2015).

[47] Philip Thomas and Emma Brunskill. 2016. Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. In *Proceedings of the 33rd International Conference on Machine Learning*, Vol. 48. PMLR, 2139–2148.

[48] Philip Thomas, Georgios Theocharous, Mohammad Ghavamzadeh, Ishan Durugkar, and Emma Brunskill. 2017. Predictive off-policy policy evaluation for nonstationary decision problems, with applications to digital marketing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31, No. 2. 4740–4745.

[49] George Tucker and Jonathan Lee. 2021. Improved Estimator Selection for Off-Policy Evaluation. *Workshop on Reinforcement Learning Theory at the 38th International Conference on Machine Learning* (2021).

[50] Takuma Udagawa, Haruka Kiyohara, Yusuke Narita, Yuta Saito, and Kei Tateno. 2023. Policy-Adaptive Estimator Selection for Off-Policy Evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36.

[51] Masatoshi Uehara, Masahiro Kato, and Shota Yasui. 2020. Off-policy evaluation and learning for external validity under a covariate shift. *Advances in Neural Information Processing Systems* 33 (2020), 49–61.

[52] Yu-Xiang Wang, Alekh Agarwal, and Miroslav Dudık. 2017. Optimal and adaptive off-policy evaluation in contextual bandits. In *International Conference on Machine Learning*. PMLR, 3589–3597.

[53] Da Xu, Yuting Ye, Chuanwei Ruan, and Bo Yang. 2022. Towards robust off-policy learning for runtime uncertainty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 10101–10109.

Table 2: Comparing Our Proposed Method (OPFV) with Key Existing Methods

	Thomas et al. [48], Chandak et al. [4]	Jagerman et al. [14]	OPFV (ours)
Is it applicable to smooth non-stationarity?	✗	✓	✓
Is it applicable to abrupt non-stationarity?	✓	✗	✓
Does it aim to estimate future policy values?	✓	✗	✓
Can it leverage time-series features?	✗	✗	✓
Is its bias-variance tradeoff analyzable?	✗	✓	✓

Note: Chandak et al. [4] and Thomas et al. [48] proposed the Prognosticator method for abrupt non-stationarity, while Jagerman et al. [14] developed sliding-window and exponential IPS for smooth non-stationarity. These existing methods cannot handle abrupt and smooth non-stationarity simultaneously, whereas our method and its associated F-OPE formulation can address arbitrary types of non-stationarity. Both Chandak et al. [4] and our work focus on estimating and optimizing the future value of policies, in contrast to Jagerman et al. [14], which did not aim to handle future policy values. None of the existing works leverage time-series features as our proposed method does, resulting in them suffering from substantial bias, similar to conventional OPE under stationarity. Moreover, our method is the first to present an analyzable bias-variance tradeoff among the methods that aim to estimate future policy values.

A Related work

Off-policy evaluation and learning under stationarity. Backgrounds of our work lie in conventional literature of off-policy evaluation (OPE) in contextual bandits [7, 9, 17, 19, 21, 26, 29, 36, 39, 40, 43, 45, 52]. In this line of work, there exist three primary approaches. The first one, known as the Direct Method (DM) [2], employs an estimator of the expected reward function to estimate the policy value. Although its variance is typically much lower than other approaches, it often incurs significant bias due to challenges in accurately estimating the expected reward of every action in the action space using only partial feedback (logged bandit data). Inverse Propensity Scoring (IPS) [13], in comparison to DM, achieves unbiasedness under the condition of common support by re-weighting the observed rewards. However, its main drawback is its high variance, attributable to the use of importance weights. Doubly Robust (DR) [8] combines DM and IPS, aiming to mitigate the variance issue of IPS while retaining unbiasedness under the same assumptions as IPS.

Note that the naive application of the DR estimator to our F-OPE problem would look like the following, where we consider the combined context $\tilde{x} := (x, t)$:

$$\begin{aligned}\hat{V}_{\mathrm{DR}}(\pi_e; \mathcal{D}, \hat{q}) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\pi_e(a_i \mid \tilde{x}_i)}{\pi_0(a_i \mid \tilde{x}_i)}(r_i - \hat{q}(\tilde{x}_i, a_i)) + \mathbb{E}_{\pi_e(a|\tilde{x}_i)}[\hat{q}(\tilde{x}_i, a)]\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\pi_e(a_i \mid x_i, t_i)}{\pi_0(a_i \mid x_i, t_i)}(r_i - \hat{q}(x_i, t_i, a_i)) + \mathbb{E}_{\pi_e(a|x_i,t_i)}[\hat{q}(x_i, t_i, a)]\right\}, \end{aligned} \tag{10}$$

where $\hat{q}$ is the estimator of the expected reward function. The most significant difference between our OPFV estimator (Eq. (4)) and a version of the DR estimator introduced above is whether we consider additional importance weight $\mathbb{I}_\phi(t, t')/p(\phi(t))$ in terms of the time feature function ϕ. Since DR overlooks this weighting factor regarding the time-series structure, it incurs non-vanishing bias. In contrast, the bias of our OPFV estimator is presented in Theorem 3.3, which can be small when ϕ is fine-grained and the reward model $\hat{f}$ estimates the relative difference in the expected rewards $\Delta_q(x, t, t', a)$ for pairs of timestamps that share the same time feature ($\phi(t) = \phi(t')$). This is because the new weighting factor ($\mathbb{I}_\phi(t, t')/p(\phi(t))$) unbiasedly estimates the time feature effect g, and the bias only arises due to the estimation error of the regression model $\hat{f}$ against the residual effect. This analysis is specific to our proposed method and is novel because our problem formulation is novel.

It is also worth differentiating our contributions from those of Uehara et al. [51], which studied OPE and OPL under covariate shift, where the context distributions differ between the logged data ($p^{\mathrm{hist}}(x)$) and in the evaluation environment ($p^{\mathrm{eval}}(x)$), while the reward distribution $p(r|x, a)$ does not change. Our setup might appear similar to that of Uehara et al. [51] when we formulate the combined feature $\tilde{x} = (x, t)$. However, the most critical difference is that our focus is on evaluating and optimizing the policy value for *future* times. This implies that our target time $t' (> T)$ has zero density in the logged data $\mathcal{D}$, making the density ratio $p^{\mathrm{eval}}(\tilde{x}')/p^{\mathrm{hist}}(\tilde{x}') = p^{\mathrm{eval}}(x, t')/p^{\mathrm{hist}}(x, t')$ nonexistent. In contrast, Uehara et al. [51] assume the existence of the density ratio of $p(\tilde{x})$ (i.e., the assumption of strong overlap, $p^{\mathrm{eval}}(\tilde{x}') > 0 \implies p^{\mathrm{hist}}(\tilde{x}') > 0$ as described in their Assumptions 1 and 2), a condition our setup inherently does not satisfy as $p^{\mathrm{hist}}(x, t') = 0$ for all $x \in \mathcal{X}$ due to the inability to observe any data from future distributions. We have managed to address and analyze this seemingly intractable situation for the first time through the novel concepts of reward function decomposition and importance weighting regarding the time feature function ϕ. Thus, our problem setup, approach, and associated analysis (particularly the bias analysis given in Theorem 3.3) are conceptually and technically different from Uehara et al. [51]. Another notable difference is that Uehara et al. [51] assumes access to the realizations of the context from the distribution in the evaluation period ($p^{\mathrm{eval}}(\tilde{x}')$) to estimate the density ratio, but this is not

possible in our setup where we cannot observe any data from the future distributions. Therefore, the method from Uehara et al. [51] (and its variants) is simply infeasible in our problem of F-OPE.

For off-policy learning (OPL), there exist two typical approaches: regression-based and policy-based. The regression-based approach [15, 34] estimates the expected reward using off-the-shelf supervised learning methods and outputs a policy based on certain rules that take into account the estimated expected reward, such as the softmax function. Similar to OPE, this approach is vulnerable to high bias due to difficulties in accurately estimating the expected reward from only the logged data. On the other hand, the policy-based approach optimizes a parameterized policy via iterative gradient ascent based on an estimated gradient of the value with respect to the policy parameters. While we can estimate the policy gradient in an unbiased fashion under common support by applying OPE estimators such as IPS and DR, it often results in considerable variance, akin to the issue of importance weighting in OPE. Another issue that often arises in OPL is the uncertainty in learning a policy that is substantially different from the logging policy [15, 23]. Some recent work [25, 28] addressed this issue by employing pessimistic approaches, such as imitating the logging policy to some extent during policy learning. Even though our primary focus is on evaluating and learning the future value of policies, it is straightforward to extend OPFV-PG to incorporate pessimism by combining it with methods proposed by [15, 22, 25, 28].

Off-policy evaluation and learning under non-stationarity. There exist previous efforts related to non-stationary environments, and we first describe the work of Chandak et al. [4] in detail, which is the most relevant to ours.

Chandak et al. [4] is particularly interested in abrupt non-stationarity where the logged data $\mathcal{D}$ can be divided into K distinct subsets $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^K$ where the distribution remains constant within each period. Each subset $\mathcal{D}_k$ is considered generated from a corresponding distribution as follows:

$$\mathcal{D}_k := \{(x_{i,k}, a_{i,k}, r_{i,k})\}_{i=1}^{n_k} \sim \prod_{i=1}^{n_k} p_k(x_{i,k})\pi_0(a_{i,k} \mid x_{i,k})p_k(r_{i,k} \mid x_{i,k}, a_{i,k}),$$

where n_k represents the logged data size in period k, and $n = \sum_{k\in[K]} n_k$. For this setting, Chandak et al. [4] proposed the Prognosticator algorithm for OPL, which includes an OPE estimator we regard as a baseline method to evaluate the future performance of an evaluation policy. Prognosticator first performs an estimator $\hat{V}_k(\pi_e; \mathcal{D})$ (e.g., IPS) to estimate the past policy value of π_e within each period $k \in [K]$. For a future performance evaluation, Prognosticator employs linear regression $\hat{V}_k(\pi_e, \mathcal{D}) \approx \psi(k)^\top w$, where $\psi(k) \in \mathbb{R}^d$ is a basis function encoding the period index, w is a regression coefficient, and $\hat{V}_k(\pi_e, \mathcal{D})$ is an estimated past policy value. The regression is performed using training data (Ψ, Y) where $\Psi := (\psi(1), \cdots, \psi(K))^\top \in \mathbb{R}^{K\times d}$ and $Y := (\hat{V}_1(\pi_e; \mathcal{D}), \cdots, \hat{V}_K(\pi_e; \mathcal{D}))^\top \in \mathbb{R}^K$. The future value estimator for period $K + \delta$ ($\delta = 1, 2, \ldots$) is given by:

$$\hat{V}_{K+\delta}(\pi_e; \mathcal{D}) = \psi(K + \delta)^\top (\Psi^\top \Psi)^{-1} \Psi^\top Y,$$

where $w^* := (\Psi^\top \Psi)^{-1} \Psi^\top Y$ is derived from standard ordinary least squares regression. Despite its improved adaptability to non-stationarity compared to conventional OPE, this method faces limitations in both the initial estimation phase and the extrapolation (regression) phase. In the former, Prognosticator experiences significant variance when frequent distribution shifts occur due to the reduced sample size n_k in each subset $\mathcal{D}_k$. For the regression phase, accurately learning future trends from only the period index k in the past as inputs is extremely challenging with the simple parametric regression, resulting in substantial bias as we empirically demonstrated.

Other than Chandak et al. [4], Thomas et al. [48] proposed essentially the same estimator as Prognosticator [4] and empirically demonstrated their OPE procedure on a real digital marketing dataset. Liu et al. [27] recently proposed the regression-assisted DR estimator to more accurately estimate the current value of a new policy by leveraging historical data from varying distributions under abrupt non-stationarity. Their particular focus is to leverage past data to reduce variance without introducing much bias and to construct valid confidence intervals, rather than estimating and optimizing future policy values. Chandak et al. [3] introduced the concept of *active non-stationarity*, where distributional shifts occur due to policy's interactions with the system in abrupt non-stationary environments. However, their method still relies on a two-stage estimation process similar to Prognosticator for non-stationary OPE, encountering limitations in both stages. Hong et al. [12] focused on policy learning under abrupt non-stationarity. Their algorithm learns a new policy for each cluster of distributions, where clustering is achieved through change point detection (also studied in Padakandla et al. [31] for the online reinforcement learning setup) or a Hidden Markov Model applied to the logged data. However, their algorithm requires online updates during deployment and, therefore, does not fall under offline learning. Moreover, it is constrained to only existing distributions in the past and is incapable of handling unknown distribution shifts that may arise in the future. Liu et al. [24] proposed a procedure for hypothesis testing of stationarity given offline data in reinforcement learning settings and a change point detection method based on the proposed testing. However, their focus was not on OPE and OPL for the future and thus is crucially different from our formulation.

In smooth non-stationary environments, Jagerman et al. [14] introduced the *sliding window IPS* and *exponential IPS* estimators. Sliding window IPS re-weights observed rewards using data from times close to the target time, at which the value of the evaluation policy is defined. In contrast, exponential IPS uses all data, giving more weight to recent observations. Domingues [6] generalized sliding window and exponential estimators mostly in online settings in reinforcement learning. However, when applied to F-OPE and OPL, the estimators and algorithms based on sliding windows and exponential discounting incur significant bias when estimating the value of an evaluation policy in the distant future. This is because they predominantly weigh recent data and thus fail to capture potential distributional shifts in the future, which might significantly differ from the distributions present when the logged data were collected. Another key formulation is

distributionally robust OPL, where the training and testing data are assumed to be sampled from different distributions [16, 30, 42, 53]. This approach aims to evaluate the value of a policy in the worst-case scenarios among a set of potential distribution shifts and learns the optimal policy that maximizes this worst-case value. Consequently, the resulting policy is often excessively pessimistic in real-world applications. Additionally, this formulation typically involves only two distributions (training and testing), which is a much simpler scenario than our formulation, which considers continual distribution changes.

B Additional theoretical analysis of the OPFV estimator

Among the estimators for OPE under stationary context and reward distributions described in the previous section, one might observe that the form of DR in Eq. (10) is similar to our proposed OPFV estimator in Eq. (4). However, there exists a notable distinction between OPFV and DR since OPFV is particularly tailored for F-OPE. To distinguish the difference between DR and the OPFV estimator, we present the bias reduction of the OPFV estimator against DR in the following theorem.

Theorem B.1 (Bias Reduction of OPFV against DR)**.** *Under Conditions 3.1 and 3.2, the difference in the bias between the DR and OPFV is given by:*

$$\begin{aligned}
&\mathrm{Bias}(\hat{V}_{\mathrm{DR}}(\pi_e;\mathcal{D},\hat{q})) - \mathrm{Bias}(\hat{V}^{\mathrm{OPFV}}_{t'}(\pi_e;\mathcal{D})) \\
&= \underbrace{\mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\left(\Delta_q(x,t,t',a) - \Delta_{\hat{q}}(x,t,t',a)\right) - \frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\left(\Delta_q(x,t,t',a) - \Delta_{\hat{f}}(x,t,t',a)\right)\right]}_{(i)} \\
&\quad + \underbrace{\mathbb{E}_{p(x,t)}\left[\sum_{a\in\mathcal{A}}(\pi_e(a|x,t) - \pi_e(a|x,t'))q(x,t,a)\right]}_{(ii)} + \underbrace{\mathbb{E}_{p(x,t)\pi_e(a|x,t')}[\Delta_{\hat{q}}(x,t,t',a)]}_{(iii)}
\end{aligned}$$

See Appendix H.4 for the proof.

Theorem B.1 reveals that three terms capture the bias reduction of OPFV against DR. Specifically, the first term (i) of the bias reduction implies that the OPFV estimator reduces the bias against DR by the expectation with respect to the evaluation policy of the estimation error of the relative difference $\Delta_q(x,t,t',a)$ in the expected rewards between timestamp t and t' when t does not share the identical time feature to target time t'. We can attribute this bias reduction of the term (i) to the use of the new importance weight $\mathbb{I}_\phi(t_i,t')/p(\phi(t'))$ in the OPFV estimator. The second term (ii) indicates that additional bias is incurred to DR when the evaluation policies at timestamp t and target time t' differ. This bias reduction arises from the difference between the parts of the DR estimator (i.e., $\pi_e(a_i|x_i,\boldsymbol{t_i})$ and $\pi_e(a|x_i,\boldsymbol{t_i})$) and the OPFV estimator (i.e., $\pi_e(a_i|x_i,\boldsymbol{t'})$ and $\pi_e(a|x_i,\boldsymbol{t'})$), which correctly account for the evaluation policy at target time t'. The third term (iii) demonstrates that the OPFV estimator reduces the bias against DR by the expectation with respect to an evaluation policy $\pi_e(a|x,t')$ of the relative difference $\Delta_{\hat{q}}(x,t,t',a)$ in the estimated expected rewards between timestamps t and t'. This bias reduction is due to the fact that the OPFV correctly uses the regression of the reward at the target time $\boldsymbol{t'}$ in the second term of the estimator in Eq. (4), whereas DR naively applies the estimator of the expected reward $\hat{q}(x_i,\boldsymbol{t_i},a)$ in Eq. (10) without aiming to estimate the reward at the target time t'. Thus, the theoretical analysis of the bias reduction of OPFV against DR clearly demonstrates that OPFV evaluates the policy for the future under non-stationarity with considerably less bias than DR, notably by the use of the time feature ϕ and the corresponding new importance weight as well as by the correct use of the evaluation policy and regression of the reward at target time t'. Note that some literature on OPE includes the theoretical analyses of a sample complexity bound, but we do not include such analyses as the problem of F-OPE is distinct from the standard OPE. More specifically, we do not have access to the samples from the distributions $p(x|t')$ and $p(r|x,t',a)$ in the future t'. Thus, it is impossible to construct the estimator $V_{t'}(\pi_e;\mathcal{D})$ that converges to the true policy value $V_{t'}(\pi_e)$ in the future t'. Even in such a challenging situation of the F-OPE problem, our novel theoretical analysis rigorously demonstrates how to optimize the time feature function ϕ to minimize the MSE of the proposed estimator. Hence, our bias-variance analysis is distinct from the theoretical analysis of the standard OPE estimator, providing a unique contribution to the problem of F-OPE.

C OPFV under non-stationary context and reward distributions

C.1 The extended estimator for F-OPE

In this section, we describe the extension of the OPFV estimator to the case where both context and reward are non-stationary. To leverage the logged data, we consider the following decomposition of the non-stationary distribution $p(x|t)$ of the context x:

$$p(x|t) = \alpha \cdot p_1(x|\phi_x(t)) + (1-\alpha)\cdot p_2(x|t),$$

where $p_1(x|\phi_x(t))$ represents the part of the non-stationary distribution of the context that can be explained by the time feature ϕ_x for the context, $p_2(x|t)$ represents the portion of the non-stationary distribution that cannot be captured by the time feature for the context, and α

controls how informative the time feature ϕ_x for the context is. Then, we can decompose the value $V_{t'}(\pi_e)$ of an evaluation policy π_e at the target time t' into the following two terms.

$$V_{t'}(\pi_e) = \underbrace{G_{\phi_x(t')}(\pi_e)}_{\text{time feature effect for context}} + \underbrace{H_{t'}(\pi_e)}_{\text{residual effect for context}}$$
$$G_{\phi_x(t')}(\pi_e) := \alpha \cdot \mathbb{E}_{p_1(x|\phi_x(t'))\pi_e(a|x,t')}[q(x,t',a)]$$
$$H_{t'}(\pi_e) := (1-\alpha) \cdot \mathbb{E}_{p_2(x|t')\pi_e(a|x,t')}[q(x,t',a)],$$

where the time feature effect for context $G_{\phi_x(t')}(\pi_e)$ captures the value for the context whose non-stationarity in the distribution can be explained by the time feature ϕ_x for context x while the residual effect for context $H_{t'}(\pi_e)$ represents the value for context whose non-stationarity in its distribution cannot be explained by the time feature ϕ_x for context. Note that we have decomposed the expected reward function $q(x,t',a)$ into the time feature effect for reward $g(x,\phi_r(t'),a)$ and the residual effect for reward $h(x,t',a)$ in Eq. (3). Thus, under the non-stationarity in the context distribution, we can consider the reward decomposition in Eq. (3) for both the time feature effect and residual effect for context. Then, leveraging some data that have the same time feature $\phi_x(t_i)$ for context with the target time t', we can extend the definition of the OPFV estimator in Eq. (4) to be able to estimate the time feature effect for context $G_{\phi_x(t')}(\pi_e)$ as follows.

$$\hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{I}_{\phi_{x,r}}(t_i,t')}{p(\phi_{x,r}(t'))}\frac{\pi_e(a_i|x_i,t')}{\pi_0(a_i|x_i,t_i)}\left(r_i-\hat{f}(x_i,t_i,a_i)\right)\right.$$
$$\left.+\frac{\mathbb{I}_{\phi_x}(t_i,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a|x_i,t')}[\hat{f}(x_i,t',a)]\right\} \tag{11}$$

where $\phi_{x,r}(t) := (\phi_x(t),\phi_r(t))^\top$ is a vector of time features with respect to the context and reward, $\mathbb{I}_{\phi_{x,r}}(t_i,t) := \mathbb{I}\{\phi_{x,r}(t_i) = \phi_{x,r}(t)\}$ denotes the indicator that outputs 1 if timestamps t_i and t share an identical time feature for context and reward, and $\mathbb{I}_{\phi_x}(t_i,t) := \mathbb{I}\{\phi_x(t_i) = \phi_x(t)\}$ denotes the indicator that outputs 1 if timestamps t_i and t share an identical time feature for context. The marginal probability densities of time features for context and reward $\phi_{x,r}$, and context ϕ_x are defined as $p(\phi_{x,r}(t)) := \int_{s\in[0,T]} p(s)\mathbb{I}\{\phi_{x,r}(s) = \phi_{x,r}(t)\}ds$ and $p(\phi_x(t)) := \int_{s\in[0,T]} p(s)\mathbb{I}\{\phi_x(s) = \phi_x(t)\}ds$, respectively. The extended OPFV estimator consists of two terms as the original OPFV estimator in Eq. (4). For the first term, the extended OPFV employs the new weight $\mathbb{I}_{\phi_{x,r}}(t_i,t')/p(\phi_{x,r}(t'))$ rather than $\mathbb{I}_{\phi_r}(t_i,t')/p(\phi_r(t'))$ used in the original OPFV estimator to be able to estimate the time feature effect for reward $g(x,\phi_r(t'),a)$ in the time feature effect for context $G_{\phi_x(t')}(\pi_e)$. For the second term in Eq. (11), the extended OPFV multiplies the second term of the original OPFV estimator in Eq. (4) by the weight with respect to the context $\mathbb{I}_{\phi_x}(t_i,t')/p(\phi_x(t'))$ to capture the residual effect for reward $h(x,t',a)$ in the time feature effect for context $G_{\phi_x(t')}(\pi_e)$. Thus, the extended OPFV generalizes the original OPFV to accommodate the distributional shifts in both context and reward.

C.2 Theoretical analysis of the extended OPFV estimator

To theoretically analyze the extended OPFV estimator, we introduce the following condition on the distributional shift with respect to the context.

Condition C.1 (Conditional Stationarity for Context)**.** A non-stationary distribution for context $p(x|t) = \alpha \cdot p_1(x|\phi_x(t)) + (1-\alpha)\cdot p_2(x|t)$ is said to satisfy conditional stationarity for context if the following holds for an arbitrary target time t':

$$p_2(x|t) = p_2(x|t')$$

for any $x \in \mathcal{X}$ and $t \in [0,T]$ such that $\mathbb{I}_{\phi_x}(t,t') = 1$

Condition C.1 requires that the part of the non-stationary context that can not be explained by the time feature ϕ_x for context is identical across the time t which shares the same time feature for the context as the target time. Under such a condition, the extended OPFV estimator has the following bias.

Theorem C.2 (Bias of the extended OPFV estimator)**.** *Under Conditions 3.1, 3.2, and C.1, the extended OPFV estimator has the following bias.*

$$\mathrm{Bias}\left(\hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D})\right) = \mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\left(\Delta_q(x,t,t',a)-\Delta_{\hat{f}}(x,t,t',a)\right)\right] \tag{12}$$

See Appendix H.8 for the proof.

The formula for the bias of the extended OPFV is similar to the bias of the original OPFV in Section 3.2. The key difference from the original OPFV is that the bias of the extended OPFV depends not only on the coarseness of the time feature for reward ϕ_r but also on the coarseness of the time feature for reward ϕ_x as the bias in Eq. (5) occurs only when timestamp $t \in [0,T]$ in the logged data shares the identical time feature for reward with target time t' (i.e., $\phi_r(t) = \phi_r(t')$) whereas the bias in Eq. (12) occurs when timestamp t shares the identical time feature for not only reward but also context (i.e., $\phi_{x,r}(t) = \phi_{x,r}(t')$).

Furthermore, we formally introduce the following condition required to prove the unbiasedness and variance of the original OPFV estimator for non-stationary reward and the extended OPFV estimator for non-stationary context and reward distributions.

Condition C.3 (Conditional Pairwise Correctness)**.** A regression model $\hat{f}$ is said to satisfy conditional pairwise correctness if the following holds for a given target time t':

$$\Delta_q(x, t, t', a) = \Delta_{\hat{f}}(x, t, t', a),$$

for all $x \in \mathcal{X}$, $a \in \mathcal{A}$, and $t \in [0, T]$ such that $\mathbb{I}_{\phi_r}(t, t') = 1$.

Condition C.3 requires only the regressor $\hat{f}$ to preserve the pairwise difference of the expected reward that is produced between different timestamps t, t' conditional on x and a. This condition is less restrictive than the correct estimation of the global expected reward given (x, t, a), which is the requirement of the extrapolation (regression) phase of Prognosticator.

Under such an ideal situation, we have the following proposition on the unbiasedness of the extended OPFV estimator.

Proposition C.4 (Unibasedness of the extended OPFV estimator)**.** *Under Conditions 3.1, 3.2, C.1, and C.3, the extended OPFV estimator is unbiased for arbitrary target time t' ($> T$), i.e., $\mathbb{E}_{\mathcal{D}}[\hat{V}^{\mathrm{OPFV}}_{t'}(\pi_e; \mathcal{D})] = V_{t'}(\pi_e)$. See Appendix H.9 for the proof.*

Proposition C.4 suggests that the extended OPFV becomes unbiased as the time feature for the context gets fine-grained.

D Optimization of the regression model

In this section, we provide how to further minimize the MSE via minimizing the error of the estimator of the reward regressor $\hat{f}$ used in our estimator in terms of both bias and variance. The bias analysis of OPFV given in Theorem 3.3 and variance analysis demonstrated in Proposition 3.4 imply that we should optimize the regressor $\hat{f}$ to minimize its error against $\Delta_q(x, t, t', a)$ for bias minimization while we should optimize $\hat{f}$ to minimize its error towards $\Delta_{q,\hat{f}}(x, t, a)$ for variance minimization.

Specifically, we start from decomposing $\hat{f}(x, t, a)$ in the following way:

$$\hat{f}(x, t, a; \theta, \omega) := \underbrace{\hat{g}(x, \phi(t), a; \omega)}_{\text{est. of time feature effect}} + \underbrace{\hat{h}(x, t, a; \theta)}_{\text{est. of residual effect}}$$

where $\omega \in \Omega$ and $\theta \in \Theta$ are the parameters for the estimator of the time feature effect $\hat{g}$ and that of the residual effect $\hat{h}$, respectively. We use the following two-stage regression to learn the optimal parameters ω^* and θ^* that minimize the bias in Eq. (5) and variance in Eq. (6).

(1) Bias Minimization Step

$$\omega^* \in \arg\min_{\omega \in \Omega} \sum_{(x, a, t_j, t_k, r_j, r_k) \in \mathcal{D}_{\mathrm{pair}}} \ell_h\Big(r_j - r_k, \hat{h}(x, t_j, a; \omega) - \hat{h}(x, t_k, a; \omega)\Big)$$

where $\mathcal{D}_{\mathrm{pair}}$ is the augmented logged data for the pairwise regression, defined as

$$\mathcal{D}_{\mathrm{pair}} := \left\{ (x, a, t_j, t_k, r_j, r_k) \,\middle|\, \begin{array}{l} (x_j, t_j, a_j, r_j), (x_k, t_k, a_k, r_k) \in \mathcal{D} \\ x = x_j = x_k, a = a_j = a_k, \mathbb{I}_{\phi}(t_j, t_k) = 1 \end{array} \right\}.$$

(2) Variance Minimization Step

$$\theta^* \in \arg\min_{\theta \in \Theta} \sum_{(x, t, a, r) \in \mathcal{D}} \ell_g\Big(r, \hat{g}_l(x, \phi(t), a; \theta) + \hat{h}(x, t, a; \omega^*)\Big)$$

Note that, for the reward regressor $\hat{f}(x, t', a)$ at target time t' in the future, we cannot accurately estimate the residual effect $\hat{h}(x, t', a; \omega)$ unless we have the data at time t' or assume strong conditions, so we randomly learn it from a neural network. On the other hand, we can learn the estimated time feature effect $\hat{g}(x, \phi(t'), a; \theta)$ from the data at time t whose time feature $\phi(t)$ is identical to the time feature $\phi(t')$ at target time t'.

E Details of the optimization of the time feature function

In Section 3.3, we provided the optimization procedure to tune the time feature function $\hat{\phi}$ to select the most accurate OPFV estimator. This section further covers the details regarding the satisfaction of Condition 3.2 after the optimization of the time feature and introduces some examples. Although the satisfaction of Condition 3.2 is not necessary in practice as the ultimate goal of our problem is the minimization of the MSE if we want to make sure that the tuned time feature function $\hat{\phi}$ obtained by the optimization procedure in Section 3.3 satisfies Condition 3.2, then we can achieve it by considering the set Φ of the time feature functions which meet Condition 3.2. One of the sufficient conditions for the satisfaction of Condition 3.2 is the observation of the time features $\phi(t)$ in the logged data $\mathcal{D}$, which is identical to the time feature $\phi(t')$ of the target time t'. For instance, suppose we want to evaluate or optimize a policy for the upcoming month in 2024,

given logged data collected in 2023, where we observe at least one data point on an arbitrary date in 2023. In such a situation, we can tune the time feature function ϕ by considering the following set

$$\begin{aligned}\Phi := \Big\{ & \phi_{\text{season}}, \phi_{\text{month}}, \phi_{\text{week}}, \phi_{\text{date}}, \phi_{\text{day_of_week}}, \phi_{\text{holiday}}, \\ & \phi_{\text{season}} \bigotimes \phi_{\text{week}}, \phi_{\text{season}} \bigotimes \phi_{\text{date}}, \phi_{\text{season}} \bigotimes \phi_{\text{day_of_week}}, \phi_{\text{season}} \bigotimes \phi_{\text{holiday}}, \\ & \phi_{\text{month}} \bigotimes \phi_{\text{week}}, \phi_{\text{month}} \bigotimes \phi_{\text{date}}, \phi_{\text{month}} \bigotimes \phi_{\text{day_of_week}}, \phi_{\text{month}} \bigotimes \phi_{\text{holiday}}, \\ & \phi_{\text{week}} \bigotimes \phi_{\text{holiday}}, \phi_{\text{day_of_week}} \bigotimes \phi_{\text{holiday}} \Big\}\end{aligned}$$

of the time feature functions where each function $\phi \in \Phi$ is defined as follows.

$$\begin{aligned}\phi_{\text{season}} &: [0, t'] \to \{\text{spring, summer, fall, winter}\} \\ \phi_{\text{month}} &: [0, t'] \to \{\text{Jan, Feb, } \cdots \text{, Dec}\} \\ \phi_{\text{week}} &: [0, t'] \to \{\text{1st week, 2nd week, } \cdots \text{, 6th week}\} \\ \phi_{\text{date}} &: [0, t'] \to \{\text{1st, 2nd, } \cdots \text{, 31st}\} \\ \phi_{\text{day_of_week}} &: [0, t'] \to \{\text{Sun, Mon, Tue, Wed, Thu, Fri, Sat}\} \\ \phi_{\text{holiday}} &: [0, t'] \to \{\text{holiday, not holiday}\}\end{aligned}$$

where, for any $\phi_i, \phi_j \in \{\phi_{\text{season}}, \phi_{\text{month}}, \phi_{\text{week}}, \phi_{\text{date}}, \phi_{\text{day_of_week}}, \phi_{\text{holiday}}\}$, we define the time feature function which combine both ϕ_i and ϕ_j by $\phi_i \bigotimes \phi_j : [0, t'] \to C_i \times C_j$ if the codomains of ϕ_i and ϕ_j are denoted by C_i and C_j, respectively, (i.e., $\phi_i : [0, t'] \to C_i$ and $\phi_j : [0, t'] \to C_j$). This example illustrates that we can consider various candidate time feature functions satisfying Condition 3.2. Thus, it is not challenging to satisfy Condition 3.2 even after the optimization of the time feature functions using a set Φ, which is reasonably large enough to unbiasedly estimate the time feature effect g. Note that the problem of hyperparameter tuning in OPE is a more general topic related not only to the F-OPE problem but also to any OPE problem. Hence the finite sample suboptimality gap between ϕ^* and $\hat{\phi}$ is considered an independent area of interest. If we want to keep track of the gap, then we can utilize SLOPE [44, 49] for the hyperparameter tuning as we discussed in the main text where the suboptimality gap is formally derived in their literature.

F Theoretical analysis of OPFV-PG

In this section, we analyze the statistical properties of the gradient of the OPFV estimator used in F-OPL in Section 3.4. Overall, we have bias and variance similar to the OPFV estimator tailored for F-OPE. Before providing the theoretical properties of the gradient of OPFV, we introduce a new condition for F-OPL.

Condition F.1 (Full Support)**.** The logging policy π_0 is said to satisfy the full support if $\pi_0(a \mid x, t) > 0$ for any $x \in \mathcal{X}$, $t \in [0, T]$, and $a \in \mathcal{A}$.

Then, under the newly introduced condition, we derive the bias of the gradient of the OPFV estimator as follows.

Theorem F.2 (Bias of the Gradient of OPFV)**.** *If Conditions F.1 and 3.2 hold, the gradient of the OPFV estimator has the following bias.*

$$\text{Bias}\left(\nabla_\zeta \hat{V}^{\text{OPFV}}_{t'}(\pi_\zeta; \mathcal{D})\right) = \mathbb{E}_{p(x,t)\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t, t')}{p(\phi(t'))}\left(\Delta_q(x, t, t', a) - \Delta_{\hat{f}}(x, t, t', a)\right) s_\zeta(x, t', a)\right]$$

See Appendix H.6 for the proof.

Theorem F.2 shows that the bias of the gradient of the OPFV can be obtained by multiplying the bias of the OPFV for F-OPE in Eq. (5) by the policy score function $s_\zeta(x, t', a)$. Thus, as the time feature function becomes fine-grained, the bias decreases, as seen in the arguments for the bias of the OPFV estimator in F-OPE in Section 3.2. Furthermore, we observe that the error $\Delta_q(x, t, t', a) - \Delta_{\hat{f}}(x, t, t', a)$ of the relative difference in the expected reward between timestamps t and t' contributes to the bias of the gradient of the OPFV estimator. Hence, in an ideal situation when Condition C.3 is satisfied, the gradient of the OPFV estimator becomes the unbiased estimator as described in the following proposition.

Proposition F.3 (Unbiasedness of the Gradient of OPFV)**.** *Under Conditions F.1, 3.2, and C.3, the gradient of OPFV is unbiased. i.e.,* $\mathbb{E}_{\mathcal{D}}[\nabla_\zeta \hat{V}^{\text{OPFV}}_{t'}(\pi_\zeta; \mathcal{D})] = \nabla_\zeta V_{t'}(\pi_\zeta)$. *See Appendix H.5 for the proof.*

In addition to the bias, we derive the variance of the gradient of OPFV in the following proposition.

Proposition F.4 (Variance of the Gradient of OPFV). *Under Conditions F.1, 3.2, and C.3, the variance of the gradient of the OPFV estimator is as follows.*

$$
\begin{aligned}
n\mathbb{V}_{\mathcal{D}}\left[\nabla_\zeta \hat{V}^{\mathrm{OPFV}}_{t'}(\pi_\zeta;\mathcal{D})\right] = \mathbb{E}_{p(x,t)\pi_0(a|x,t)}&\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}s_\zeta(x,t',a)\right)^2\sigma^2(x,t,a)\right]\\
&+\mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]\right]\\
&+\mathbb{V}_{p(t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right]\mathbb{E}_{p(x)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]^2\right]\\
&+\mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[q(x,t',a)s_\zeta(x,t',a)\right]\right]
\end{aligned}
$$

where $\Delta_{q,\hat{f}}(x,t,a) := q(x,t,a) - \hat{f}(x,t,a)$ *is the estimation error of* $\hat{f}(x,t,a)$ *against the expected reward funciton* $q(x,t,a)$ *and* $\sigma^2(x,t,a) := \mathbb{V}_{p(r|x,t,a)}[r]$ *is the variance of the reward* r *given the context* x*, timestamp* t*, and action* a*. See Appendix H.7 for the proof.*

Proposition F.4 suggests that the term $\mathbb{I}_\phi(t,t')/p(\phi(t'))$ contributes to the variance of the gradient of OPFV. Specifically, using the coarse time feature ϕ leads to a low variance, whereas the fine-grained time feature is susceptible to incurring the variance. Hence, when employed in F-OPL, the gradient of the OPFV needs to consider the bias-variance tradeoff by optimizing the granularity of the time feature to achieve low MSE.

G Extension of OPFV-PG with pessimism

In Appendix A, we discussed that the OPL algorithm based on the importance weight might be susceptible to uncertainty when learning a new policy, particularly if a policy is significantly different from the logging policy. This problem also exists in F-OPL as the optimal policy in the future can be considerably different from a logging policy due to the non-stationarity. Thus, in this section, we further extend the OPFV-PG method by incorporating the pessimistic approach to cope with such uncertainty. Specifically, we add a regularization term denoted by $\mathrm{IML}(\pi_\zeta)$ [28] in the objective of the OPL so that the learned policy does not deviate considerably from the logging policy. Concretely, when we combine OPFV-PG with the pessimistic approach, the objective we want to maximize becomes as follows:

$$
\hat{V}^{\mathrm{OPFV}}_{t'}(\pi_\zeta;\mathcal{D}) + \rho\cdot\mathrm{IML}(\pi_\zeta),
$$

where ρ is a parameter which controls the exploration and exploitation tradeoff in OPL and $\mathrm{IML}(\pi_\zeta) := -\frac{1}{n}\sum_{i=1}^n \log\frac{\pi_\zeta(a_i|x_i)}{\pi_0(a_i|x_i,t_i)}$ is the estimator of the Kullback-Leibler (KL) divergence $\mathrm{KL}(\pi_0||\pi_\zeta) := \mathbb{E}_{p(x,t)\pi_0(a|x,t)}[\log(\frac{\pi_0(a|x,t)}{\pi_\zeta(a|x)})]$ between the logging π_0 and the policy π_ζ. To incorporate this objective in the policy gradient approach, we use the following gradient of the objective with respect to the parameter ζ:

$$
\begin{aligned}
&\nabla_\zeta\left(\hat{V}^{\mathrm{OPFV}}_{t'}(\pi_\zeta;\mathcal{D}) + \rho\cdot\mathrm{IML}(\pi_\zeta)\right)\\
:=&\frac{1}{n}\sum_{i=1}^n\left\{\frac{\mathbb{I}_\phi(t_i,t')}{p(\phi(t'))}\frac{\pi_\zeta(a_i\mid x_i)}{\pi_0(a_i\mid x_i,t_i)}\left(r_i-\hat{f}(x_i,t_i,a_i)\right)\nabla_\zeta\log\pi_\zeta(a_i\mid x_i)\right.\\
&\left.+\mathbb{E}_{\pi_\zeta(a|x_i)}\left[\hat{f}(x_i,t',a)\nabla_\zeta\log\pi_\zeta(a\mid x_i)\right]-\rho\cdot\nabla_\zeta\log\frac{\pi_\zeta(a_i|x_i)}{\pi_0(a_i|x_i,t_i)}\right\}.
\end{aligned}
$$

Then, by the following iterative gradient ascent yields the pessimistic policy: $\zeta_{\tau+1} \leftarrow \zeta_\tau + \eta\cdot\nabla_\zeta\left(\hat{V}^{\mathrm{OPFV}}_{t'}(\pi_{\zeta_\tau};\mathcal{D}) + \rho\cdot\mathrm{IML}(\pi_{\zeta_\tau})\right)$, we can cope with uncertainty even in F-OPL.

H Omitted proofs

In this section, we provide the comprehensive proofs for the theorems and propositions we have introduced on the unbiasedness, bias, and variance of the original OPFV estimator for F-OPE in Section 3.2, the gradient of the OPFV estimator employed in OPFV-PG for F-OPL in Section F, and the extended OPFV estimator for F-OPE under non-stationary context and reward distributions in Section C.2.

H.1 Proof of the unbiasedness of the OPFV estimator

Proof. We prove the unbiasedness of the OPFV estimator in F-OPE under Conditions 3.1, 3.2, and C.3 below.

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[\hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D})] \\
&= \mathbb{E}_{p(x_i,t_i)\pi_0(a_i|x_i,t_i)p(r_i|x_i,t_i,a_i)\forall i\in[n]}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{I}_\phi(t_i,t')}{p(\phi(t'))}\frac{\pi_e(a_i|x_i,t')}{\pi_0(a_i|x_i,t_i)}\left(r_i-\hat{f}(x_i,t_i,a_i)\right)\right.\right. \\
&\quad \left.\left.+\mathbb{E}_{\pi_e(a|x_i,t')}[\hat{f}(x_i,t',a)]\right\}\right] \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)p(r|x,t,a)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \quad \because \text{i.i.d} \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \quad \because \text{def. of } q(x,t,a) \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \quad \because \text{Condition C.3} \\
&= \mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)\right]+\mathbb{E}_{p(x,t)\pi_e(a'|x,t')}[\hat{f}(x,t',a')] \quad \because \text{cancel out } \pi_0(a|x,t) \\
&= \mathbb{E}_{p(x)p(t)\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)\right]+\mathbb{E}_{p(x)\pi_e(a'|x,t')}[\hat{f}(x,t',a')] \quad \because \text{stationary context} \\
&= \mathbb{E}_{p(x)\pi_e(a|x,t')}\left[\int_{t\in[0,T]}p(t)\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)dt\right]+\mathbb{E}_{p(x)\pi_e(a'|x,t')}[\hat{f}(x,t',a')] \\
&= \mathbb{E}_{p(x)\pi_e(a|x,t')}\left[\frac{\Delta_{q,\hat{f}}(x,t',a)}{p(\phi(t'))}\int_{t\in[0,T]}p(t)\mathbb{I}_\phi(t,t')dt\right]+\mathbb{E}_{p(x)\pi_e(a'|x,t')}[\hat{f}(x,t',a')] \\
&= \mathbb{E}_{p(x)\pi_e(a|x,t')}\left[\frac{\Delta_{q,\hat{f}}(x,t',a)}{p(\phi(t'))}p(\phi(t'))\right]+\mathbb{E}_{p(x)\pi_e(a'|x,t')}[\hat{f}(x,t',a')] \quad \because \text{definition of } p(\phi(t')) \\
&= \mathbb{E}_{p(x)\pi_e(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)\right]+\mathbb{E}_{p(x)\pi_e(a'|x,t')}[\hat{f}(x,t',a')] \quad \because \text{cancel out } p(\phi(t')) \\
&= \mathbb{E}_{p(x)\pi_e(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)+\hat{f}(x,t',a)\right] \\
&= \mathbb{E}_{p(x)\pi_e(a|x,t')}\left[q(x,t',a)\right] \quad \because \text{Cancel out } \hat{f}(x,t',a) \\
&= \mathbb{E}_{p(x|t')\pi_e(a|x,t')}\left[q(x,t',a)\right] \quad \because \text{stationary context i.e., } p(x)=p(x|t') \\
&= V_{t'}(\pi_e) \quad \because \text{definition of the value function}
\end{aligned}
$$

□

H.2 Proof of Theorem 3.3

Proof. If the Condition C.3 is violated, the bias of the OPFV estimator in F-OPE under Conditions 3.1 and 3.2 is derived as follows.

$$
\begin{aligned}
&\mathrm{Bias}\left(\hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D})\right)\\
&= \mathbb{E}_{\mathcal{D}}\left[\hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D})\right] - V_{t'}(\pi_e) \quad \because \text{definition of the bias}\\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)p(r|x,t,a)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]\\
&\quad - \mathbb{E}_{p(x|t')\pi_e(a|x,t')}[q(x,t',a)]\\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]\\
&\quad - \mathbb{E}_{p(x|t')\pi_e(a|x,t')}[q(x,t',a)]\\
&= \mathbb{E}_{p(x)p(t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]\\
&\quad - \mathbb{E}_{p(x)\pi_e(a|x,t')}[q(x,t',a)] \quad \because \text{stationary context}\\
&= \mathbb{E}_{p(x)p(t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)\right] - \mathbb{E}_{p(x)\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t',a)]\\
&= \mathbb{E}_{p(x)p(t)\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t,a)\right] - \mathbb{E}_{p(x)\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t',a)] \quad \because \text{cancel out } \pi_0(a|x,t')\\
&= \mathbb{E}_{p(x)\pi_e(a|x,t')}\left[\frac{1}{p(\phi(t'))}\mathbb{E}_{p(t)}[\mathbb{I}_\phi(t,t')\Delta_{q,\hat{f}}(x,t,a)] - \Delta_{q,\hat{f}}(x,t',a)\right]\\
&= \mathbb{E}_{p(x)\pi_e(a|x,t')}\left[\frac{1}{p(\phi(t'))}\left(\mathbb{E}_{p(t)}[\mathbb{I}_\phi(t,t')\Delta_{q,\hat{f}}(x,t,a)] - p(\phi(t'))\Delta_{q,\hat{f}}(x,t',a)\right)\right]\\
&= \mathbb{E}_{p(x)\pi_e(a|x,t')}\left[\frac{1}{p(\phi(t'))}\left(\mathbb{E}_{p(t)}[\mathbb{I}_\phi(t,t')\Delta_{q,\hat{f}}(x,t,a)] - \mathbb{E}_{p(t)}[\mathbb{I}_\phi(t,t')]\Delta_{q,\hat{f}}(x,t',a)\right)\right] \quad \because \text{def. of } p(\phi(t))\\
&= \mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\left(\Delta_{q,\hat{f}}(x,t,a) - \Delta_{q,\hat{f}}(x,t',a)\right)\right]\\
&= \mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\left(\Delta_{q}(x,t,t',a) - \Delta_{\hat{f}}(x,t,t',a)\right)\right]
\end{aligned}
$$

Note that this quantity is bounded by $2r_{\max}$ since $|\Delta_q(x,t,t',a) - \Delta_{\hat{f}}(x,t,t',a)| \leq 2r_{\max}$. □

H.3 Proof of Proposition 3.4

Proof. We can derive the variance of the OPFV estimator for F-OPE under Conditions 3.1, 3.2, and C.3 below.

$$
\begin{aligned}
& n \operatorname{Var}\left[\hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D})\right] \\
&= n\mathbb{V}_{p(x_i,t_i)\pi_0(a_i|x_i,t_i)p(r_i|x_i,t_i,a_i)\forall i\in[n]}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{I}_\phi(t_i,t')}{p(\phi(t'))}\frac{\pi_e(a_i|x_i,t')}{\pi_0(a_i|x_i,t_i)}\left(r_i-\hat{f}(x_i,t_i,a_i)\right)\right.\right. \\
&\quad \left.\left.+\mathbb{E}_{\pi_e(a|x_i,t')}[\hat{f}(x_i,t',a)]\right\}\right] \\
&= \mathbb{V}_{p(x,t)\pi_0(a|x,t)p(r|x,t,a)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \quad \because \text{i.i.d} \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\mathbb{V}_{p(r|x,t,a)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]\right] \\
&\quad +\mathbb{V}_{p(x,t)\pi_0(a|x,t)}\left[\mathbb{E}_{p(r|x,t,a)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]\right] \quad \because \text{total var.} \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\right)^2\sigma^2(x,t,a)\right] \\
&\quad +\mathbb{V}_{p(x,t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \quad \because \sigma^2(x,t,a):=\mathbb{V}_{p(r|x,t,a)}[r] \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\right)^2\sigma^2(x,t,a)\right] \\
&\quad +\mathbb{E}_{p(x,t)}\left[\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]\right] \\
&\quad +\mathbb{V}_{p(x,t)}\left[\mathbb{E}_{\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]\right] \quad \because \text{law of total variance} \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\right)^2\sigma^2(x,t,a)\right] \\
&\quad +\mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)\right]\right] \\
&\quad +\mathbb{V}_{p(x,t)}\left[\mathbb{E}_{\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t,a)\right]+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \quad \because \text{cancel out } \pi_0(a|x,t) \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\right)^2\sigma^2(x,t,a)\right] \\
&\quad +\mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)\right]\right] \\
&\quad +\mathbb{V}_{p(x)p(t)}\left[\mathbb{E}_{\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t,a)\right]+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \quad \because \text{stationary context}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)\right]\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{p(t)}\left[\mathbb{E}_{\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t,a)\right]+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]\right] \\
&\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{p(t)}\left[\mathbb{E}_{\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t,a)\right]+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]\right] \quad \because \text{law of total variance} \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)\right]\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{p(t)}\left[\mathbb{E}_{\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)\right]\right]\right] \\
&\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{p(t)\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)\right]+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \quad \because \text{Condition C.3} \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)\right]\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{E}_{\pi_e(a|x,t)}\left[\Delta_{q,\hat{f}}(x,t',a)\right]^2\right]\cdot\mathbb{V}_{p(t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right] \\
&\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_e(a|x,t')}\left[\int_{t\in[0,T]}p(t)\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)dt\right]+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)\right]\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{E}_{\pi_e(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)\right]^2\right]\cdot\mathbb{V}_{p(t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right] \\
&\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_e(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)\right]+\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \quad \because \text{cancel out } p(\phi(t)) \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)\right]\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{E}_{\pi_e(a|x,t)}\left[\Delta_{q,\hat{f}}(x,t',a)\right]^2\right]\cdot\mathbb{V}_{p(t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right] \\
&\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_e(a|x,t')}\left[q(x,t',a)\right]\right] \quad \because \text{cancel out } \hat{f}(x,t',a)
\end{aligned}
$$

Note that this quantity is well-bounded under the bounded importance weight (*i.e.*, $\pi_e(a|x,t')/\pi_0(a|x,t) < \infty$), the bounded conditional variance of the reward (*i.e.*, $\sigma^2(x,t,a) < \infty$), and the bounded variance of the expected reward in terms of the context (*i.e.*, $\mathbb{V}_{p(x)}[\mathbb{E}_{\pi_e(a|x,t')}[q(x,t',a)]] < \infty$) for all $x \in \mathcal{X}, t \in [0,T], a \in \mathcal{A}$, and $t' > T$. □

H.4 Proof of Theorem B.1

Proof. the bias reduction of the OPFV estimator against DR in F-OPE under Conditions 3.1 and 3.2 is derived as follows. Note that we use the following DR estimator applied to the estimation of a policy at target time t' to distinguish which part of the OPFV estimator contributes to the bias reduction against the naive application of DR:

$$\hat{V}_{t'}^{\mathrm{DR}}(\pi_e;\mathcal{D},\hat{q}) := \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{I}_\phi(t_i,t')}{p(\phi(t'))}\frac{\pi_e(a_i|x_i,t')}{\pi_0(a_i|x_i,t_i)}(r_i-\hat{f}(x_i,t_i,a_i))+\mathbb{E}_{\pi_e(a|x_i,t')}[\hat{f}(x_i,t',a)]\right\}$$

Then, adding and subtracting the bias of the newly introduced DR estimator, we decompose the bias reduction of OPFV compared to DR into three terms in the following way.

$$\begin{aligned}
&\mathrm{Bias}\left(\hat{V}_{\mathrm{DR}}(\pi_e;\mathcal{D},\hat{q})\right)-\mathrm{Bias}\left(\hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D})\right)\\
=\ &\mathrm{Bias}\left(\hat{V}_{\mathrm{DR}}(\pi_e;\mathcal{D},\hat{q})\right)-\mathrm{Bias}\left(\hat{V}_{t'}^{\mathrm{DR}}(\pi_e;\mathcal{D},\hat{q})\right)+\mathrm{Bias}\left(\hat{V}_{t'}^{\mathrm{DR}}(\pi_e;\mathcal{D},\hat{q})\right)-\mathrm{Bias}\left(\hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D})\right)\\
=\ &\mathbb{E}_{\mathcal{D}}\left[\hat{V}_{\mathrm{DR}}(\pi_e;\mathcal{D},\hat{q})\right]-\mathbb{E}_{\mathcal{D}}\left[\hat{V}_{t'}^{\mathrm{DR}}(\pi_e;\mathcal{D},\hat{q})\right]\\
&+\mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\Delta_q(x,t,t',a)-\Delta_{\hat{q}}(x,t,t',a)\right]-\mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\left(\Delta_q(x,t,t',a)-\Delta_{\hat{q}}(x,t,t',a)\right)\right]\\
=\ &\mathbb{E}_{p(x,t)\pi_e(a|x,t)}\left[q(x,t,a)\right]-\mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\Delta_{q,\hat{q}}(x,t,a)+\hat{q}(x,t',a)\right]\\
&+\mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\left(\Delta_q(x,t,t',a)-\Delta_{\hat{q}}(x,t,t',a)\right)-\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\left(\Delta_q(x,t,t',a)-\Delta_{\hat{q}}(x,t,t',a)\right)\right]\\
=\ &\mathbb{E}_{p(x,t)}\left[\sum_{a\in\mathcal{A}}\left\{\pi_e(a|x,t)q(x,t,a)-\pi_e(a|x,t')\left(\Delta_{q,\hat{q}}(x,t,a)+\hat{q}(x,t',a)\right)\right\}\right]\\
&+\mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\left(\Delta_q(x,t,t',a)-\Delta_{\hat{q}}(x,t,t',a)\right)-\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\left(\Delta_q(x,t,t',a)-\Delta_{\hat{q}}(x,t,t',a)\right)\right]\\
=\ &\mathbb{E}_{p(x,t)}\Bigg[\sum_{a\in\mathcal{A}}\Big\{\pi_e(a|x,t)q(x,t,a)-\pi_e(a|x,t')q(x,t,a)\\
&+\pi_e(a|x,t')q(x,t,a)-\pi_e(a|x,t')\left(\Delta_{q,\hat{q}}(x,t,a)+\hat{q}(x,t',a)\right)\Big\}\Bigg]\\
&+\mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\left(\Delta_q(x,t,t',a)-\Delta_{\hat{q}}(x,t,t',a)\right)-\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\left(\Delta_q(x,t,t',a)-\Delta_{\hat{q}}(x,t,t',a)\right)\right]\\
=\ &\mathbb{E}_{p(x,t)}\left[\sum_{a\in\mathcal{A}}\left\{(\pi_e(a|x,t)-\pi_e(a|x,t'))q(x,t,a)+\pi_e(a|x,t')\Delta_{\hat{q}}(x,t,t',a)\right\}\right]\\
&+\mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\left(\Delta_q(x,t,t',a)-\Delta_{\hat{q}}(x,t,t',a)\right)-\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\left(\Delta_q(x,t,t',a)-\Delta_{\hat{q}}(x,t,t',a)\right)\right]
\end{aligned}$$

□

H.5 Proof of Proposition F.3

Proof. We prove the unbiasedness of the gradient of the OPFV estimator for F-OPL under Conditions F.1, 3.2, and C.3 below.

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\nabla_\zeta \hat{V}^{\mathrm{OPFV}}_{t'}(\pi_\zeta;\mathcal{D})\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{I}_\phi(t_i,t')}{p(\phi(t'))}\frac{\pi_\zeta(a_i|x_i,t')}{\pi_0(a_i|x_i,t_i)}\left(r_i-\hat{f}(x_i,t_i,a_i)\right)s_\zeta(x_i,t',a_i)\right.\right. \\
&\quad \left.\left.+\mathbb{E}_{\pi_\zeta(a|x_i,t')}[\hat{f}(x_i,t',a)s_\zeta(x_i,t',a)]\right\}\right] \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)p(r|x,t,a)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)s_\zeta(x,t',a)\right. \\
&\quad \left.+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right] \quad \because \text{i.i.d} \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right] \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a))s_\zeta(x,t',a)+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right] \quad \because \text{Cond. C.3} \\
&= \mathbb{E}_{p(x,t)\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]+\mathbb{E}_{p(x,t)\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')] \quad \because \text{cancel out } \pi_0 \\
&= \mathbb{E}_{p(x)p(t)\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]+\mathbb{E}_{p(x)\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')] \quad \because \text{stationary } x \\
&= \mathbb{E}_{p(x)\pi_\zeta(a|x,t')}\left[\int_{t\in[0,T]}p(t)\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)dt\right]+\mathbb{E}_{p(x)\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')] \\
&= \mathbb{E}_{p(x)\pi_\zeta(a|x,t')}\left[\frac{\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)}{p(\phi(t'))}\int_{t\in[0,T]}p(t)\mathbb{I}_\phi(t,t')dt\right]+\mathbb{E}_{p(x)\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')] \\
&= \mathbb{E}_{p(x)\pi_\zeta(a|x,t')}\left[\frac{\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)}{p(\phi(t'))}p(\phi(t'))\right]+\mathbb{E}_{p(x)\pi_\zeta(a'|x,t')}[\hat{f}(x,t,a's_\zeta(x,t',a'))] \quad \because \text{def. of } p(\phi(t')) \\
&= \mathbb{E}_{p(x)\pi_\zeta(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]+\mathbb{E}_{p(x)\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')] \quad \because \text{cancel out } p(\phi(t')) \\
&= \mathbb{E}_{p(x)\pi_\zeta(a|x,t')}\left[\left(\Delta_{q,\hat{f}}(x,t',a)+\hat{f}(x,t',a)\right)s_\zeta(x,t',a)\right] \\
&= \mathbb{E}_{p(x)\pi_\zeta(a|x,t')}\left[q(x,t',a)s_\zeta(x,t',a)\right] \quad \because \text{Condition C.3} \\
&= \mathbb{E}_{p(x|t')\pi_\zeta(a|x,t')}\left[q(x,t',a)s_\zeta(x,t',a)\right] \quad \because \text{stationary context i.e., } p(x)=p(x|t') \\
&= \nabla_\zeta V_{t'}(\pi_\zeta) \quad \because \text{definition of the value function}
\end{aligned}
$$

□

H.6 Proof of Theorem F.2

Proof. If the Condition C.3 is violated, the bias of the gradient of the OPFV estimator under Conditions F.1 and 3.2 is derived as follows.

$$
\begin{aligned}
&\mathrm{Bias}\left(\nabla_\zeta \hat{V}^{\mathrm{OPFV}}_{t'}(\pi_\zeta;\mathcal{D})\right)\\
&= \mathbb{E}_{\mathcal{D}}\left[\nabla_\zeta \hat{V}^{\mathrm{OPFV}}_{t'}(\pi_\zeta;\mathcal{D})\right] - \nabla_\zeta V_{t'}(\pi_\zeta) \quad \because \text{definition of the bias}\\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)p(r|x,t,a)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)s_\zeta(x,t',a)+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right]\\
&\quad - \mathbb{E}_{p(x|t')\pi_\zeta(a|x,t')}[q(x,t',a)s_\zeta(x,t',a)] \quad \because \text{definistion of the gradient of the OPFV and i.i.d.}\\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right]\\
&\quad - \mathbb{E}_{p(x|t')\pi_\zeta(a|x,t')}[q(x,t',a)s_\zeta(x,t',a)]\\
&= \mathbb{E}_{p(x)p(t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right]\\
&\quad - \mathbb{E}_{p(x)\pi_\zeta(a|x,t')}[q(x,t',a)s_\zeta(x,t',a)] \quad \because \text{stationary context i.e., } p(x|t)=p(x)\\
&= \mathbb{E}_{p(x)p(t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)\right]-\mathbb{E}_{p(x)\pi_\zeta(a|x,t')}[\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)]\\
&= \mathbb{E}_{p(x)p(t)\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)\right]-\mathbb{E}_{p(x)\pi_\zeta(a|x,t')}[\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)] \quad \because \text{cancel out } \pi_0\\
&= \mathbb{E}_{p(x)\pi_\zeta(a|x,t')}\left[\frac{1}{p(\phi(t'))}\mathbb{E}_{p(t)}[\mathbb{I}_\phi(t,t')\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)]-\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]\\
&= \mathbb{E}_{p(x)\pi_\zeta(a|x,t')}\left[\frac{1}{p(\phi(t'))}\left(\mathbb{E}_{p(t)}[\mathbb{I}_\phi(t,t')\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)]-p(\phi(t'))\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right)\right]\\
&= \mathbb{E}_{p(x)\pi_\zeta(a|x,t')}\left[\frac{1}{p(\phi(t'))}\left(\mathbb{E}_{p(t)}[\mathbb{I}_\phi(t,t')\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)]-\mathbb{E}_{p(t)}[\mathbb{I}_\phi(t,t')]\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right)\right]\\
&= \mathbb{E}_{p(x,t)\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\left(\Delta_{q,\hat{f}}(x,t,a)]-\Delta_{q,\hat{f}}(x,t',a)\right)s_\zeta(x,t',a)\right]\\
&= \mathbb{E}_{p(x,t)\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\left(\Delta_{q}(x,t,t',a)-\Delta_{\hat{f}}(x,t,t',a)\right)s_\zeta(x,t',a)\right]
\end{aligned}
$$

□

H.7 Proof of Proposition F.4

Proof. We derive the variance of the gradient of the OPFV estimator under Conditions F.1, 3.2, and C.3 below.

$$
\begin{aligned}
& n \operatorname{Var}\left[\nabla_\zeta \hat{V}_{t'}^{\mathrm{OPFV}}(\pi_\zeta; \mathcal{D})\right] \\
&= n \mathbb{V}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{I}_\phi(t_i,t')}{p(\phi(t'))}\frac{\pi_\zeta(a_i|x_i,t')}{\pi_0(a_i|x_i,t_i)}\left(r_i-\hat{f}(x_i,t_i,a_i)\right)s_\zeta(x_i,t',a_i)+\mathbb{E}_{\pi_\zeta(a|x_i,t')}[\hat{f}(x_i,t',a)s_\zeta(x_i,t',a)]\right\}\right] \\
&= \mathbb{V}_{p(x,t)\pi_0(a|x,t)p(r|x,t,a)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)s_\zeta(x,t',a)+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right] \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\mathbb{V}_{p(r|x,t,a)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)s_\zeta(x,t',a)+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right]\right] \\
&\quad + \mathbb{V}_{p(x,t)\pi_0(a|x,t)}\left[\mathbb{E}_{p(r|x,t,a)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)s_\zeta(x,t',a)+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right]\right] \\
&\qquad \because \text{law of total variance} \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}s_\zeta(x,t',a)\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{V}_{p(x,t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right] \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}s_\zeta(x,t',a)\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right]\right] \\
&\quad + \mathbb{V}_{p(x,t)}\left[\mathbb{E}_{\pi_0(a|x,t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right]\right] \\
&\qquad \because \text{total var.} \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}s_\zeta(x,t',a)\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)\right]\right] \\
&\quad + \mathbb{V}_{p(x,t)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)\right]+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right] \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}s_\zeta(x,t',a)\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)\right]\right] \\
&\quad + \mathbb{V}_{p(x)p(t)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)\right]+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right] \\
&\qquad \because \text{stationary context}
\end{aligned}
$$

$$
\begin{aligned}
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}s_\zeta(x,t',a)\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)\right]\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{p(t)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)\right]+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right]\right] \\
&\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{p(t)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t,a)s_\zeta(x,t',a)\right]+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right]\right] \quad \because \text{total var.} \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}s_\zeta(x,t',a)\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{V}_{p(t)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]\right]\right] \\
&\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{p(t)\pi_\zeta(a|x,t')}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right] \quad \because \text{Cond. C.3} \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}s_\zeta(x,t',a)\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]^2\right]\cdot\mathbb{V}_{p(t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right] \\
&\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[\int_{t\in[0,T]}p(t)\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)dt\right]+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right] \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}s_\zeta(x,t',a)\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]^2\right]\cdot\mathbb{V}_{p(t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right] \\
&\quad + \mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]+\mathbb{E}_{\pi_\zeta(a'|x,t')}[\hat{f}(x,t',a')s_\zeta(x,t',a')]\right] \quad \because \text{cancel out } p(\phi(t')) \\
&= \mathbb{E}_{p(x,t)\pi_0(a|x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}s_\zeta(x,t',a)\right)^2\sigma^2(x,t,a)\right] \\
&\quad + \mathbb{E}_{p(x,t)}\left[\left(\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right)^2\mathbb{V}_{\pi_0(a|x,t)}\left[\frac{\pi_\zeta(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]\right] \\
&\quad + \mathbb{E}_{p(x)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)s_\zeta(x,t',a)\right]^2\right]\cdot\mathbb{V}_{p(t)}\left[\frac{\mathbb{I}_\phi(t,t')}{p(\phi(t'))}\right]+\mathbb{V}_{p(x)}\left[\mathbb{E}_{\pi_\zeta(a|x,t')}\left[q(x,t',a)s_\zeta(x,t',a)\right]\right] \\
&\quad \because \text{cancel out } \hat{f}(x,t',a)
\end{aligned}
$$

□

H.8 Proof of Theorem C.2

Proof. We derive the bias of the extended OPFV estimator for F-OPE under non-stationary context and reward distributions under Conditions 3.1, 3.2, and C.1 below.

$$
\begin{aligned}
&\mathrm{Bias}\left(\hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D})\right)\\
&=\mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{I}_{\phi_{x,r}}(t_i,t')}{p(\phi_{x,r}(t'))}\frac{\pi_e(a_i|x_i,t')}{\pi_0(a_i|x_i,t_i)}\left(r_i-\hat{f}(x_i,t_i,a_i)\right)+\frac{\mathbb{I}_{\phi_x}(t_i,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a|x_i,t')}[\hat{f}(x_i,t',a)]\right\}\right]-V_{t'}(\pi_e)\\
&=\mathbb{E}_{p(t)p(x|t)\pi_0(a|x,t)p(r|x,t,a)}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]-V_{t'}(\pi_e)\\
&=\mathbb{E}_{p(t)p(x|t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]\\
&\quad-\mathbb{E}_{p(x|t')\pi_e(a|x,t')}[q(x,t',a)]\\
&=\mathbb{E}_{p(t)p(x|t)\pi_e(a|x,t')}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\Delta_{q,\hat{f}}(x,t,a)+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]-\mathbb{E}_{p(x|t')\pi_e(a|x,t')}[q(x,t',a)]\\
&=\mathbb{E}_{p(t)p(x|t)}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t,a)]\right]+\mathbb{E}_{p(t)p(x|t)}\left[\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right]\\
&\quad-\mathbb{E}_{p(x|t')\pi_e(a|x,t')}[q(x,t',a)]\\
&=\mathbb{E}_{p(t)p(x|t)}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t,a)]\right]+\int_{t\in[0,T]}p(t)\int_{x\in\mathcal{X}}p(x|t)\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]dxdt\\
&\quad-\mathbb{E}_{p(x|t')\pi_e(a|x,t')}[q(x,t',a)]\\
&=\mathbb{E}_{p(t)p(x|t)}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t,a)]\right]\\
&\quad+\alpha\int_{t\in[0,T]}p(t)\int_{x\in\mathcal{X}}p_1(x|\phi_x(t))\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]dxdt\\
&\quad+(1-\alpha)\int_{t\in[0,T]}p(t)\int_{x\in\mathcal{X}}p_2(x|t)\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]dxdt-\mathbb{E}_{p(x|t')\pi_e(a|x,t')}[q(x,t',a)]\quad\because\text{Cond. C.1}\\
&=\mathbb{E}_{p(t)p(x|t)}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t,a)]\right]+\alpha\int_{x\in\mathcal{X}}p_1(x|\phi_x(t))\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]dx\\
&\quad+(1-\alpha)\int_{x\in\mathcal{X}}p_2(x|t)\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]dx-\mathbb{E}_{p(x|t')\pi_e(a|x,t')}[q(x,t',a)]\quad\because\text{def. of } p(\phi_x(t')\\
&=\mathbb{E}_{p(t)p(x|t)}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t,a)]\right]+\mathbb{E}_{p(x|t')\pi_e(a|x,t')}[\hat{f}(x,t',a)]-\mathbb{E}_{p(x|t')\pi_e(a|x,t')}[q(x,t',a)]\\
&=\mathbb{E}_{p(t)p(x|t')}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t,a)]\right]-\mathbb{E}_{p(x|t')\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t',a)]\quad\because\text{Cond. C.1}\\
&=\mathbb{E}_{p(x|t')\pi_e(a|x,t')}\left[\frac{1}{p(\phi_{x,r}(t'))}\left(\mathbb{E}_{p(t)}[\mathbb{I}_{\phi_{x,r}}(t,t')\Delta_{q,\hat{f}}(x,t,a)]-p(\phi_{x,r}(t'))\Delta_{q,\hat{f}}(x,t,a)]\right)\right]\\
&=\mathbb{E}_{p(x|t')\pi_e(a|x,t')}\left[\frac{1}{p(\phi_{x,r}(t'))}\left(\mathbb{E}_{p(t)}[\mathbb{I}_{\phi_{x,r}}(t,t')\Delta_{q,\hat{f}}(x,t,a)]-\mathbb{E}_{p(t)}[\mathbb{I}_{\phi_{x,r}}(t,t')]\Delta_{q,\hat{f}}(x,t,a)]\right)\right]\quad\because\text{def. of } p(\phi_{x,r}(t')\\
&=\mathbb{E}_{p(t)p(x|t')\pi_e(a|x,t')}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\left(\Delta_{q,\hat{f}}(x,t,a)-\Delta_{q,\hat{f}}(x,t',a)\right)\right]\\
&=\mathbb{E}_{p(x,t)\pi_e(a|x,t')}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\left(\Delta_{q}(x,t,t',a)-\Delta_{\hat{f}}(x,t,t',a)\right)\right]\quad\because\text{Condition C.1}
\end{aligned}
$$

□

H.9 Proof of Proposition C.4

Proof. We prove that the extended OPFV estimator for F-OPE under non-stationary context and reward distributions is unbiased under Conditions 3.1, 3.2, C.1, and C.3 below.

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[\hat{V}_{t'}^{\mathrm{OPFV}}(\pi_e;\mathcal{D})\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\mathbb{I}_{\phi_{x,r}}(t_i,t')}{p(\phi_{x,r}(t'))}\frac{\pi_e(a_i|x_i,t')}{\pi_0(a_i|x_i,t_i)}\left(r_i-\hat{f}(x_i,t_i,a_i)\right)+\frac{\mathbb{I}_{\phi_x}(t_i,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a|x_i,t')}[\hat{f}(x_i,t',a)]\right\}\right] \\
&= \mathbb{E}_{p(t)p(x|t)\pi_0(a|x,t)p(r|x,t,a)}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\left(r-\hat{f}(x,t,a)\right)+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \quad \because \text{i.i.d.} \\
&= \mathbb{E}_{p(t)p(x|t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t,a)+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \\
&= \mathbb{E}_{p(t)p(x|t)\pi_0(a|x,t)}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\frac{\pi_e(a|x,t')}{\pi_0(a|x,t)}\Delta_{q,\hat{f}}(x,t',a)+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a'|x,t')}[\hat{f}(x,t',a')]\right] \quad \because \text{Condition C.3} \\
&= \mathbb{E}_{p(t)p(x|t)}\left[\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t',a)]+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\hat{f}(x,t',a)]\right] \quad \because \text{cancel out } \pi_0(a|x,t') \\
&= \int_{t\in[0,T]}p(t)\int_{x\in\mathcal{X}}p(x|t)\left\{\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t',a)]+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\hat{f}(x,t',a)]\right\}dxdt \\
&= \int_{t\in[0,T]}p(t)\int_{x\in\mathcal{X}}\{\alpha p_1(x|\phi_x(t))+(1-\alpha)p_2(x|t)\}\left\{\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t',a)]\right. \\
&\quad \left.+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\hat{f}(x,t',a)]\right\}dxdt \\
&= \alpha\int_{t\in[0,T]}p(t)\int_{x\in\mathcal{X}}p_1(x|\phi_x(t'))\left\{\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t',a)]+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\hat{f}(x,t',a)]\right\}dxdt \\
&\quad +(1-\alpha)\int_{t\in[0,T]}p(t)\int_{x\in\mathcal{X}}p_2(x|t)\left\{\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t',a)]+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\hat{f}(x,t',a)]\right\}dxdt \\
&= \alpha\mathbb{E}_{p_1(x|\phi_x(t))\pi_e(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)+\hat{f}(x,t',a)\right] \\
&\quad +(1-\alpha)\int_{t\in[0,T]}p(t)\int_{x\in\mathcal{X}}p_2(x|t)\left\{\frac{\mathbb{I}_{\phi_{x,r}}(t,t')}{p(\phi_{x,r}(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\Delta_{q,\hat{f}}(x,t',a)]+\frac{\mathbb{I}_{\phi_x}(t,t')}{p(\phi_x(t'))}\mathbb{E}_{\pi_e(a|x,t')}[\hat{f}(x,t',a)]\right\}dxdt \\
&\qquad \because \text{cancel out } p(\phi_{x,r}(t')) \text{ and } p(\phi_r(t')) \\
&= \alpha\mathbb{E}_{p_1(x|\phi_x(t'))\pi_e(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)+\hat{f}(x,t',a)\right]+(1-\alpha)\mathbb{E}_{p_2(x|t)\pi_e(a|x,t')}\left[\Delta_{q,\hat{f}}(x,t',a)+\hat{f}(x,t',a)\right] \\
&\qquad \because \text{Condition C.1} \\
&= \alpha\mathbb{E}_{p_1(x|\phi_x(t'))\pi_e(a|x,t')}\left[q(x,t',a)\right]+(1-\alpha)\mathbb{E}_{p_2(x|t)\pi_e(a|x,t')}\left[q(x,t',a)\right] \\
&= G_{\phi_x(t')}(\pi_e)+H_{t'}(\pi_e) \\
&= V_{t'}(\pi_e)
\end{aligned}
$$

□

I Additional setup and results for non-stationary synthetic experiment

I.1 Additional experimental setup for F-OPE and F-OPL under non-stationary reward distribution

I.1.1 Synthetic expected reward function. In this section, we describe how we synthesize the simplified and controllable reward function $q(x,t,a)$ utilized in Section 4 in detail. First, we synthesize the expected reward $q(x,t,a)$ by the weighted sum of the time feature effect $g(x,\phi(t),a)$ and the residual effect $h(x,t,a)$ as follows.

$$q(x,t,a) = \lambda \cdot g(a,\phi(t),a) + (1-\lambda)\cdot h(x,t,a)$$

Specifically, we construct each effect by the following functions.

$$\begin{aligned} g(x,\phi(t),a) &= v_x^\top s_{g_1}(x) + v_{\phi(t)}^\top \text{one_hot}_{\phi(\mathrm{t})} + \text{one_hot}_{\phi(\mathrm{t})}^\top M_{\phi(t),a}\,\text{one_hot}_{\mathrm{a}} \\ &\quad + s_{g_2}(x)^\top M_{x,\phi(t),a}\,\text{one_hot}_{\phi(\mathrm{t}),\mathrm{a}} \\ h(x,t,a) &= \xi_x^\top s_{h_1}(x) + \xi_{\phi_f(t)}^\top \text{one_hot}_{\phi_{\mathrm{f}}(\mathrm{t})} + \xi_a^\top \text{one_hot}_{\mathrm{a}} + \text{one_hot}_{\phi_{\mathrm{f}}(\mathrm{t})}^\top M_{\phi_f(t),a}\,\text{one_hot}_{\mathrm{a}} \\ &\quad + s_{h_2}(x)^\top M_{x,a}\,\text{one_hot}_{\mathrm{a}} + s_{h_3}(x)^\top M_{x,\phi_f(t),a}\,\text{one_hot}_{\phi_{\mathrm{f}}(\mathrm{t}),\mathrm{a}}, \end{aligned}$$

where $\text{one_hot}_{\mathrm{a}} \in \{0,1\}^{|\mathcal{A}|}$, $\text{one_hot}_{\phi(\mathrm{t})} \in \{0,1\}^{|C|}$, $\text{one_hot}_{\phi_{\mathrm{f}}(\mathrm{t})} \in \{0,1\}^{|C_f|}$ are the one-hot encoding of action a, time feature ϕ, finer time feature ϕ_f, respectively and C and C_f are the sets of unique time features used in the time feature effect and residual effect. We use the time feature function ϕ, which divides each year into eight seasons with $|C| = 8$, and the finer time feature ϕ_f that maps the timestamp t to the day of the week with $|C_f| = 7$. $\text{one_hot}_{\phi(\mathrm{t}),\mathrm{a}} \in \{0,1\}^{|C\times\mathcal{A}|}$, and $\text{one_hot}_{\phi_{\mathrm{f}}(\mathrm{t}),\mathrm{a}} \in \{0,1\}^{|C_f\times\mathcal{A}|}$ are the one-hot encoding of time feature $\phi(t)$ and action a and finer time feature $\phi_f(t)$ and action a. Furthermore, we sample each element of the parameters used in the time feature effect, $v_x, v_{\phi(t)}, M_{\phi(t),a}$, and $M_{x,\phi(t),a}$ from a uniform distribution ranging from -3 to 3 and the parameters used in residual effect, $\xi_x, \xi_{\phi_f(t)}, \xi_a, M_{\phi_f(t),a}, M_{x,a}$, and $M_{x,\phi_f(t),a}$ from a uniform distribution ranging from -1 to 1. $s_{g_1}(x), s_{g_2}(x), s_{h_1}(x), s_{h_2}(x)$, and $s_{h_3}(x)$ are defined as follows:

$$\begin{aligned} s_{g_1}(x) &= \left(\mathbb{I}\left\{\sum_{d=1}^{4} x_d < 1.5\right\}, \mathbb{I}\left\{\sum_{d=6}^{9} x_d < -0.5\right\}, \mathbb{I}\left\{\sum_{d=4}^{5} x_d > 3.0\right\}, \mathbb{I}\left\{\sum_{d=7}^{10} x_d > 3.0\right\}\right)^\top \\ s_{g_2}(x) &= \left(\mathbb{I}\left\{\sum_{d=1}^{4} x_d < 4.0\right\}, \mathbb{I}\left\{\sum_{d=6}^{9} x_d > 3.0\right\}, \mathbb{I}\left\{\sum_{d=3}^{10} x_d < -2.5\right\}\right)^\top \\ s_{h_1}(x) &= \left(\mathbb{I}\left\{\sum_{d=1}^{6} x_d < 2.5\right\}, \mathbb{I}\left\{\sum_{d=8}^{9} x_d < -0.5\right\}, \mathbb{I}\left\{\sum_{d=3}^{5} x_d > 2.0\right\}\right)^\top \\ s_{h_2}(x) &= \left(\mathbb{I}\left\{\sum_{d=1}^{4} x_d < 3.0\right\}, \mathbb{I}\left\{\sum_{d=3}^{9} x_d > 2.5\right\}, \mathbb{I}\left\{\sum_{d=2}^{7} x_d < 1.5\right\}, \mathbb{I}\left\{\sum_{d=7}^{10} x_d > -1.5\right\}\right)^\top \\ s_{h_3}(x) &= \left(\mathbb{I}\left\{\sum_{d=1}^{4} x_d < 4.0\right\}, \mathbb{I}\left\{\sum_{d=3}^{9} x_d > 3.5\right\}, \mathbb{I}\left\{\sum_{d=3}^{5} x_d > 1.5\right\}, \mathbb{I}\left\{\sum_{d=6}^{10} x_d < 2.5\right\}\right)^\top, \end{aligned}$$

where x_d is the d-th dimension of the context x.

I.1.2 Optimization of time feature function for the OPFV estimator. This section describes the true time feature function ϕ_{true} used for the data-generating process as well as the set Φ of candidate time features used in the OPFV estimator in the synthetic experiment in Section 4. As briefly described in Section 4, the true time feature function $\phi_{\text{true}} : [0, t'] \to \{s_1, s_2, \cdots, s_8\}$ evenly divide each year into eight seasons where the i-th season is denoted by s_i. Figure 7 illustrates the time feature functions used for the optimization procedure of the OPFV estimator described in Section 3.3. Specifically, we used the set $\Phi := \{\phi_2, \phi_4, \phi_6, \cdots, \phi_{16}\}$, where ϕ_i is strictly finer than ϕ_j for all $i, j \in \{2, 4, 6, \cdots, 16\}$ if $i > j$. For instance, for any $i = 1, 2, 3, 4$, ϕ_{2^i} evenly divide each year into 2^i seasons. We use this set Φ to tune the time feature $\hat{\phi}$ for OPFV (tuned $\hat{\phi}$), and we use the finest time feature function $\phi_\infty = \phi_{16}$ to estimate the bias described in Section 3.3. This definition of the set of time feature functions is helpful in observing the empirical bias-variance tradeoff of the OPFV estimator in terms of the crudeness of the time features to support the claim in theoretical analyses in Section 3.2.

I.1.3 Additional baseline estimator for F-OPE. **The Prognosticator ϕ estimator:** To show which components in the OPFV estimator in Section 3 improve the MSE compared to the Prognosticator estimator, we introduce the extended Prognosticator estimator named the Prognosticator ϕ estimator in F-OPE, which uses the time feature function ϕ, that is, one of our contributions to the F-OPE/F-OPL, rather than the basis function ψ originally proposed in Prognosticator Chandak et al. [4]. Concretely, we introduce the time feature function for Prognosticator $\phi_p : [K+\delta] \to C$ where ϕ_p maps the time period $k \in [K+\delta]$ to a time-series feature $c \in C$. Then, the Prognosticator ϕ estimator for estimating the true value $V_{K+\delta}(\pi_e)$ in the future period $K+\delta$ is defined as follows.

$$\hat{V}_{K+\delta}^{\text{Prognosticator }\phi}(\pi_e;\mathcal{D}) := \text{one_hot}_{\phi_{\mathrm{p}}}(K+\delta)^\top(\Phi_p^T\Phi_p)^{-1}\Phi_p^\top Y$$

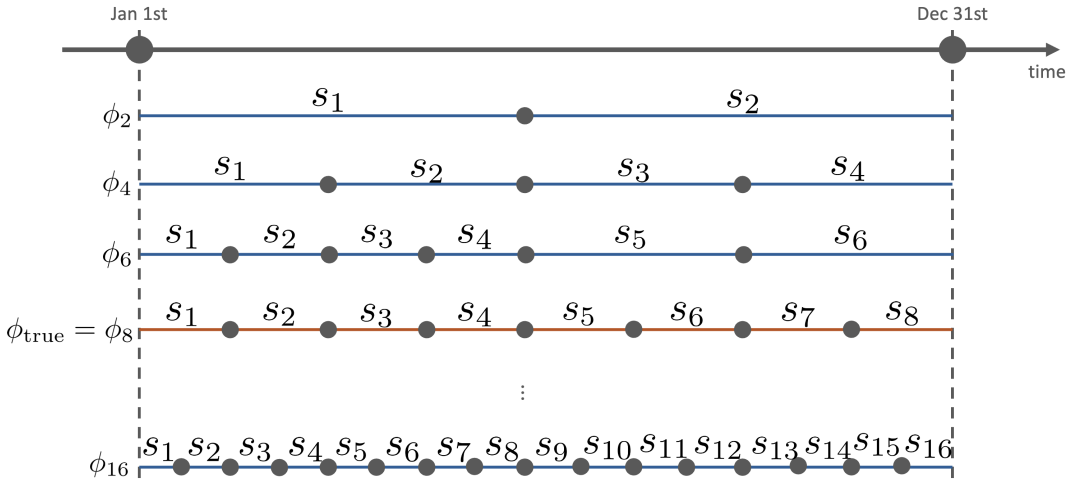


Figure 7: Comparing the true time feature $\phi_{\text{true}}(=\phi_8)$ and the time feature functions $\phi \in \Phi$ used in the optimization procedure in Section 3.3 for the OPFV estimator where $\Phi := \{\phi_{2i} : \forall i = 1, 2, \cdots, 8\}$ for any year

where $\Phi_p := (\text{one_hot}_{\phi_\text{p}}(1), \cdots, \text{one_hot}_{\phi_\text{p}}(K))^\top \in \mathbb{R}^{K \times |C|}$ and $Y := (\hat{V}_1(\pi_e; \mathcal{D}), \cdots, \hat{V}_K(\pi_e; \mathcal{D}))^\top \in \mathbb{R}^K$. Note that $\text{one_hot}_{\phi_\text{p}}(k) \in \mathbb{R}^{|C|}$ is the one-hot encoding of the time features $\phi_p(k)$ of the period $k \in [K + \delta]$. This estimator will show how much improvement in the accuracy of the OPFV estimator compared to the Prognosticator estimator can be attributed to the use of the time feature function ϕ.

As one of the limitations of the Prognosticator estimator $\hat{V}_{K+\delta}(\pi_e; \mathcal{D})$ in Section A was the unreliable extrapolation (regression) phase, which often incurs significant bias because capturing the periodic trends by using time step index k and basis function ψ is challenging, the Prognosticator ϕ is expected to reduce its bias compared to the original Prognosticator estimator in Section A.

I.1.4 Baselines methods for F-OPL. In our synthetic data experiment for F-OPL, we compare OPFV-PG with the Regression-based method and policy gradient (PG) methods, including IPS-PG and DR-PG as the conventional OPL methods for stationary environments and Prognosticator as a baseline method for non-stationary environments. We describe these baseline methods in more detail below.

Regression-based Method: The regression-based method first estimates the expected reward function $q(x, a)$ by some supervised machine learning, where we use a random forest regressor with ten trees. Then, based on the estimated expected reward function $\hat{q}_\zeta(x, a)$ parameterized by ζ, it outputs a stochastic decision $\pi_\zeta(a|x)$. For instance, we can use the softmax function to obtain the learned policy $\pi_\zeta(a|x) = \frac{\exp(\beta \cdot \hat{q}(x,a))}{\sum_{a' \in \mathcal{A}} \exp(\beta \cdot \hat{q}(x,a'))}$ where β is a parameter that controls the stochasticity of the policy π_ζ. In our experiments, we set the parameter $\beta = 10$. One of the statistical advantages of the regression-based method is that it often has a lower bias than methods with importance weights. However, it suffers from significant bias due to the misspecification of the expected reward function $\hat{q}(x, a)$, particularly when the logged data size n is limited compared to the number of pairs of context x and action a in stationary environments. Such a situation is often the case with practical applications [9, 11, 37, 50]. In non-stationary environments, the bias of the estimator of the expected reward $\hat{q}(x, a)$ considerably increases since it does not consider the distributional shift of the context and reward in the $\hat{q}$ function (i.e., $q(x, a) \neq q(x, t', a)$ for target time t'), failing to learn a policy for the future from the logged data effectively.

Inverse Propensity Scoring Policy Gradient (IPS-PG): In contrast to the regression-based method, PG methods use estimators of the gradient of the policy value $\nabla_\zeta V(\pi_\zeta)$ to optimize the policy π_ζ parameterized by ζ via the gradient ascent $\zeta_{\tau+1} \leftarrow \zeta_\tau + \eta \nabla_\zeta V(\pi_{\zeta_\tau})$ where η is a learning rate in stationary environments. One of the typical choices of the estimator is IPS-PG:

$$\nabla_\zeta \hat{V}^{\text{IPS}}(\pi_\zeta; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} \frac{\pi_\zeta(a_i|x_i)}{\pi_0(a_i|x_i)} r_i s_\zeta(x_i, a_i),$$

where $s_\zeta(x, a) := \nabla_\zeta \log \pi_\zeta(a|x)$ is the policy score function. In stationary environments, the IPS-PG estimator is unbiased if the logging policy π_0 assigns a non-zero probability to any action $a \in \mathcal{A}$ for arbitrary context $x \in \mathcal{X}$. In non-stationary environments, however, it is no longer unbiased as the true gradient of the policy value varys over the time (i.e., $\nabla_\zeta V(\pi_\zeta) \neq \nabla_\zeta V_{t'}(\pi_\zeta)$) for target time t'; thus, IPS-PG fails to improve a policy for the future.

Doubly Robust Policy Gradient (DR-PG): DR-PG employs a doubly robust structure to construct an estimator of the gradient of the policy value as follows.

$$\nabla_\zeta \hat{V}^{\text{DR}}(\pi_\zeta; \mathcal{D}, \hat{q}) := \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\pi_\zeta(a_i|x_i)}{\pi_0(a_i|x_i)} (r_i - \hat{q}(x_i, a_i) s_\zeta(x_i, a_i)) + \mathbb{E}_{\pi_\zeta(a|x_i)} [\hat{q}(x_i, a) s_\zeta(x_i, a)] \right\},$$

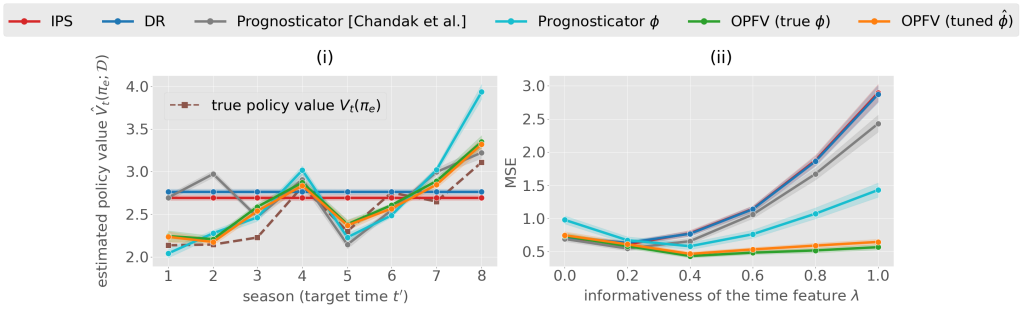


Figure 8: Comparing (i) the true value $V_t(\pi_e)$ and the estimated policy values $\hat{V}_t(\pi_e; \mathcal{D})$ with varying target time range t' and (ii) the MSE of estimators with varying λ values that controls the informativeness of the time feature

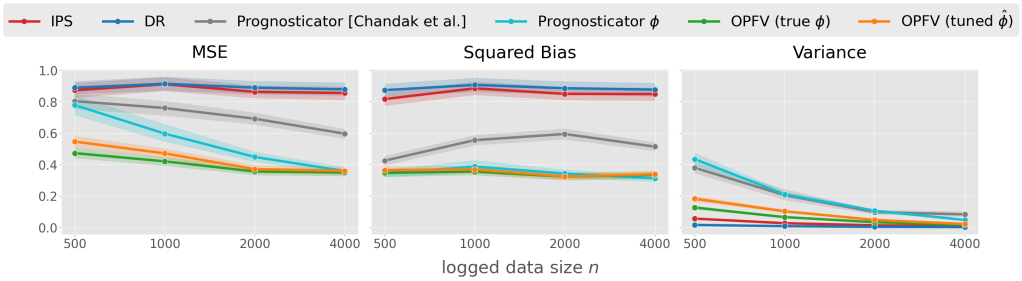


Figure 9: Comparison of the MSE, squared bias, and variance of estimators in an F-OPE problem with varying sample sizes n

The doubly robust structure reduces the estimator's variance by using $q(x, a)$ as a control variate in stationary environments. Nonetheless, the DR-PG estimator incurs significant bias under distributional shifts of the context and reward. The failure to evaluate the gradient of the policy value leads to the unreliable F-OPL.

Prognosticator [4]: One of the baseline methods in abrupt non-stationary environments is Prognosticator. In contextual bandit protocol, Prognosticator first divides the logged data $\mathcal{D}$ into K distinct sub-dataset $\mathcal{D} = \{\mathcal{D}_k\}_{k\in[K]}$, where it assumes the stationarity within each subdata $\mathcal{D}_k$ for any period $k \in [K]$. Then, it uses the following estimator of the gradient of the policy value for future period $K + \delta$:

$$\nabla_\zeta \hat{V}^{\text{Prognosticator}}_{K+\delta}(\pi_\zeta; \mathcal{D}) := \sum_{k=1}^{K} \underbrace{\frac{\partial \hat{V}_{K+\delta}(\pi_\zeta; \mathcal{D})}{\partial \hat{V}^{\text{IPS}}_k(\pi_\zeta; \mathcal{D}_k)}}_{\text{(i) extrapolation phase}} \cdot \underbrace{\nabla_\zeta \hat{V}^{\text{IPS}}_k(\pi_\zeta; \mathcal{D}_k)}_{\text{(ii) estimation phase}}$$

$$= \sum_{k=1}^{K} \left[\psi(K+\delta)(\Psi^\top\Psi)^{-1}\Psi^\top\right]_k \nabla_\zeta \hat{V}^{\text{IPS}}_k(\pi_\zeta; \mathcal{D}_k),$$

where $\left[\psi(K+\delta)(\Psi^\top\Psi)^{-1}\Psi^\top\right]_k$ denotes the k-th dimension of $\psi(K+\delta)(\Psi^\top\Psi)^{-1}\Psi^\top \in \mathbb{R}^K$ and we set $K = 8$ as we consider eight distinct time features in each year in the synthetic data experiments. For a basis function ψ, we use a Fourier basis $\psi(k) := \Big(\sin(2\pi\frac{k}{K+\delta}), \cdots, \sin(2\pi\frac{d'k}{K+\delta}), 1, \cos(2\pi\frac{k}{K+\delta}), \cdot$ $\mathbb{R}^{2d'+1}$. We consider d' as the hyperparameter and let Prognosticator tune this hyperparameter from $d' \in [3, 5, 7]$ (same sets as [4]) by using the true policy value. For (ii) estimation phase of the gradient of the policy value in period k, Prognosticator uses the IPS-PG estimator to estimate the gradient of the policy value for each period $k \in [K]$ as follows.

$$\nabla_\zeta \hat{V}^{\text{IPS}}_k(\pi_\zeta; \mathcal{D}_k) := \frac{1}{n_k}\sum_{i=1}^{n_k} \frac{\pi_\zeta(a_{i,k}|x_{i,k})}{\pi_0(a_{i,k}|x_{i,k})} r_{i,k} s_\zeta(x_{i,k}, a_{i,k}),$$

To estimate the policy gradient for the time step index $K + \delta$, Prognosticator takes the weighted sum of the gradient of the policy value in each period k where the weight $\frac{\partial \hat{V}_{K+\delta}(\pi_\zeta;\mathcal{D})}{\partial \hat{V}^{\text{IPS}}_k(\pi_\zeta;\mathcal{D}_k)}$ represents how much change in the policy value in period k affects the change in the policy value in the future time step index $K + \delta$. Similar to the discussion in Section A, there are two limitations of Prognosticator in both (i) estimation phase and (ii) extrapolation (regression) phase. First, the variance of the estimator $\nabla_\zeta \hat{V}^{\text{IPS}}_k(\pi_\zeta; \mathcal{D}_k)$ of the gradient of the policy value considerably increases in (i) estimation phase when the distributional shifts frequently occur, which is often the case in the real-world

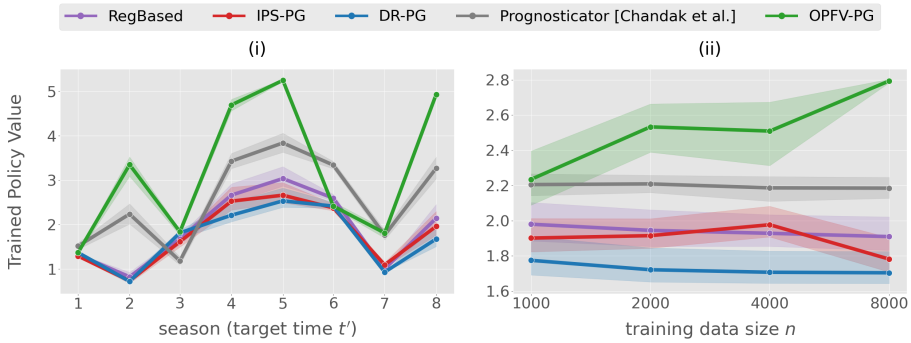


Figure 10: Comparison of the policy values achieved by F-OPL methods with (left) varying target times t' and (right) the logged data sizes (training data) n.

applications. Moreover, it incurs non-negligible bias as capturing the trends by time step index k in (ii) extrapolation (regression) phase with simple linear regression is challenging.

I.2 Additional results for F-OPE under non-stationary reward distribution

Figure 8 (i) compares the accuracy of the Prognosticator ϕ estimator compared to the other estimators in F-OPE with varying target time (t') spanning from target time 1-45 days (season 1) to 320-365 days (season 8) after we collect the logged data T. We observe that using the time feature function ϕ rather than the mere basis function ψ improves the estimation accuracy, particularly during seasons 1 and 2, where the original Prognosticator estimator evaluates the true values as inaccurately as in season 1 or even worse in season 2 than the conventional OPE estimators.

Figure 8 (ii) illustrates the MSE of the Prognosticator ϕ and the other estimators in the main text with varying levels of the informativeness of the time feature (λ) from 0.0 to 1.0 by 0.2. As the time feature ϕ becomes more informative in the expected reward $q(x,t,a)$, the Prognosticator ϕ estimator reduces the MSE of the Prognosticator estimator, demonstrating that the use of the time feature function ϕ contributes to the decrease in the MSE by approximately half the difference in the MSE between OPFV and the Prognosticator estimator.

Figure 9 reports the precision of the estimators when we vary the logged data sizes (n) from 500 to 4000. Overall, OPFV performs more accurately than the baseline methods, irrespective of the data size, because it can at least unbiasedly estimate the time feature effect, leading to the lowest bias, as we can see in the middle figure. Note that OPFV has a higher variance than IPS and DR due to the additional weighting factor to estimate the time feature effect, but its bias reduction advantage is much larger, making OPFV the most effective estimator. IPS and DR produce the largest bias due to their unrealistic stationarity assumption. Prognosticator fails particularly when the data size is limited due to the significant variance produced when performing OPE in each period with a small sub-data size n_k. Moreover, the figure clearly illustrates that using the time feature ϕ reduces the bias of OPFV since the bias of the Prognosticator ϕ is as low as OPFV. In contrast, the Prognosticator ϕ estimator still suffers from as high variance as the original Prognosticator estimator, particularly when the logged data size is limited since they conduct the OPE using the logged data with size n_k in each period $k \in [K]$.

I.3 Additional results for F-OPL under non-stationary reward distribution

We present the additional results on F-OPL under non-stationary reward with (i) varying levels of the informativeness of the time feature (λ) and (ii) varying cardinality of (tuned) time features ($|\hat{\phi}|$) used in the OPFV-PG. We compared three OPL methods (RegBased, IPS-PG, and DR-PG) for stationary environments, Prognosticator as the method for the non-stationary environments, and our proposed OPFV-PG.

Figure 10 (i) presents the policy values achieved by five OPL methods with varying target times t'. We observe that OPFV-PG mostly achieves the highest value adaptive to different target times t' due to its effective usage of the time features in estimating the policy gradient for future values. Prognosticator rarely produces a better policy than OPFV-PG and it ends up with a policy even worse than the conventional OPL methods in season 3. In contrast, OPFV-PG always performs better than the conventional OPL methods.

Figure 10 (ii) reports the earned future policy values with varying logged data sizes. As the logged data size n increases, OPFV-PG is able to learn increasingly better policies because of its low-bias estimation of the policy gradient $\nabla_\zeta V_{t'}(\pi_\zeta)$ for any target time t' in the future. In contrast, the baseline methods are based on biased policy gradients, so they do not necessarily improve even with increased training data.

Figure 11 (i) illustrates the learned policy values when we vary the levels of how informative the time feature is (λ) ranging from 0.0 to 1.0 by 0.2. The results show that OPFV-PG demonstrates a relatively higher learned policy value as the time feature becomes more informative. In instances where the time feature effect is subtle, the policy learned by OPFV-PG is comparable to conventional OPL methods and the Prognosticator.

Figure 11 (ii) demonstrates the impact of varying the cardinalities of (tuned) time features ($|\hat{\phi}|$) on the performance of OPFV-PG, with the true cardinality of time features set at eight ($|\phi_{\text{true}}| = 8$). The results reveal that OPFV-PG enables us to learn a policy with the highest policy value among the baseline OPL methods when the cardinality of the used time feature is from 6 to 16. Even with coarse time features, the

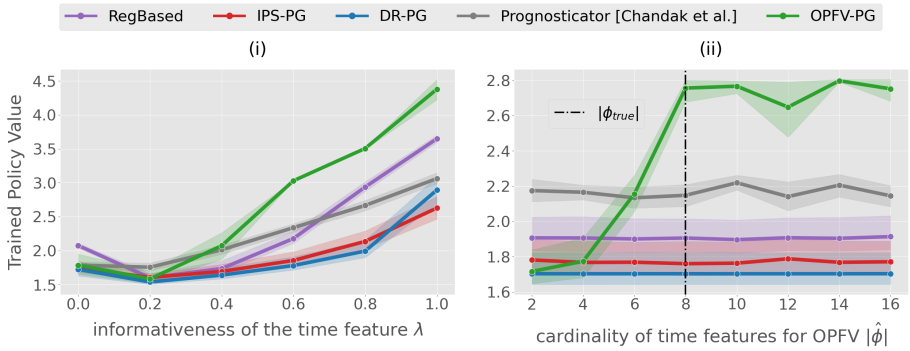


Figure 11: Comparison of the test policy values of the Regression-based approach, IPS and DR Policy Gradient, Prognosticator, and OPFV-PG with (i) varying levels of informativeness of the time feature λ, (ii) varying cardinalities of time features for OPFV-PG, and (iii) varying target time t'

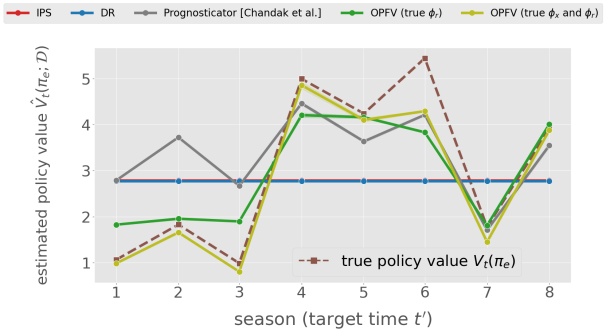


Figure 12: Comparison of the true value and estimated values of IPS, DR, Prognosticator, OPFV (true ϕ_r) for non-stationary reward and OPFV (true ϕ_x and ϕ_r) for non-stationary context and reward distributions with varying target time t'

policy value learned by OPFV-PG is comparable to DR. As the cardinality of time features increases, the value of OPFV-PG grows until it aligns with the true cardinality of time features, which is eight, which is consistent with our bias analysis in Section F. Furthermore, we observe that using finer-grained time features than the true one leads to a decrease in the value of OPFV-PG. This reduction is attributed to an increase in the variance of the estimated policy gradient $\nabla_\zeta \hat{V}^{\mathrm{OPFV}}_{t'}(\pi_\zeta; \mathcal{D})$, while the bias remains unaffected. Consequently, the value of OPFV-PG with overly fine-grained time features slightly decreases compared to the case where the true cardinality of time features is employed. Nevertheless, it maintains the highest value among the five OPL methods, including Prognosticator.

I.4 Additional experimental setup for F-OPE under non-stationary context and reward distributions

To empirically identify the situations when the extended OPFV estimator proposed in Section C.1 enables us to evaluate a policy in the future under non-stationary context and reward distributions, we conducted a synthetic data experiment. In this section, we describe how we synthesize the non-stationary context. We generate context x from the following non-stationary distribution $p(x|t)$:

$$p(x|t) = \alpha \cdot p_1(x|\phi_x(t)) + (1-\alpha) \cdot p_2(x|t), \tag{13}$$

where $p_1(x|\phi_x(t)) = \prod_{d\in[d_x]} \mathcal{N}(\mu_1, \sigma_1^2)$ and $p_2(x|t) = \prod_{d\in[d_x]} \mathcal{N}(\mu_2, \sigma_2^2)$. For the mean value μ_1 of the component of the non-stationary context depending on the time feature ϕ_x for the context, we use $\mu_1 = \gamma^\top \mathrm{one_hot}_{\phi_x(t)}$ where $\mathrm{one_hot}_{\phi_x(t)}$ is the one hot embedding of the time feature for the context ϕ_x, and γ is a parameter sampled from uniform distribution ranging from -3 to 3. In contrast, for the mean value μ_2, we set it to $\mu_2 = \kappa s(t)$ where $s(t) = \frac{t - \mathrm{t_oldest}}{\mathrm{t_future} - \mathrm{t_oldest}}$ is the normalized time where we use the UNIX timestamps of January 1st in Year 1 for t_oldest and January 1st in Year 3 for t_future, and κ is the parameter sampled from the uniform distribution ranging from -1 to 1. We set the variance of the normal distribution to 1 for σ_1^2 and σ_2^2.

I.5 Additional results for F-OPE under non-stationary context and reward distributions

We compare the precision of OPFV adaptive to the non-stationary reward in the main text denoted by OPFV (true ϕ_r) and the extended OPFV for both non-stationary context and reward distributions denoted by OPFV (true ϕ_x and ϕ_r)) with conventional OPE estimators (IPS and DR), and the Prognosticator estimator. We use the same experimental setting as the non-stationary reward-only environments except

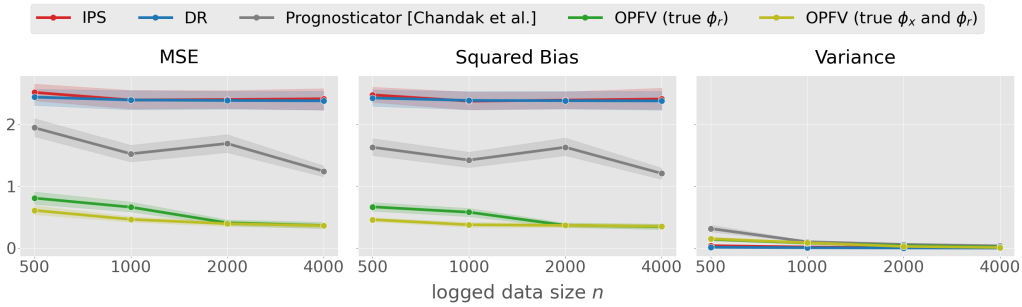


Figure 13: Comparison of the MSE, bias, and variance of IPS, DR, Prognosticator, OPFV (true ϕ_r) for non-stationary reward distributions and OPFV (true ϕ_x and ϕ_r) for non-stationary context and reward distributions with varying the logged data sizes n

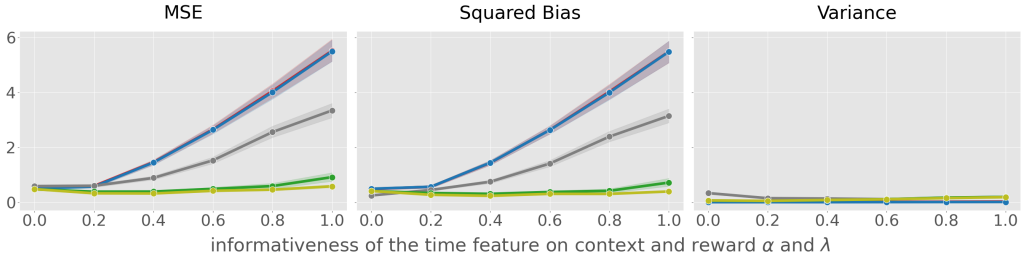


Figure 14: Comparison of the MSE, bias, and variance of IPS, DR, Prognosticator, OPFV (true ϕ_r) for non-stationary reward distributions and OPFV (true ϕ_x and ϕ_r) for non-stationary context and reward distributions with varying levels of the informativeness of the time features for context and reward α and λ

for generating the non-stationary context described in Eq. (13). We set the true cardinality of time feature $|\phi_x|$ for context to eight, the same as that for reward.

Figure 12 presents the true value of an evaluation policy and the estimated values by OPE and F-OPE estimators under the distributional shifts of context and reward with varying target time range (t'). We observe that OPFV (true ϕ_x and ϕ_r) evaluates an evaluation policy at an arbitrary target time t' more correctly than OPFV (true ϕ_r) by adapting to not only the non-stationary reward but also non-stationary context. We also observe that the original OPFV estimator enables us to evaluate the policy value better than the Prognosticator and conventional OPE estimators, even under non-stationary context and reward distributions if the time feature for context ϕ_x and for reward ϕ_r are identical. Prognosticator tries to adapt to the non-stationary environment. Still, it uses the time step index k, failing to represent the effect of the time feature correctly, and hence, incurring as significant bias as IPS and DR for target time ranges 1-45 (season 1) and 92-136 (season 3) and more bias than conventional OPE estimators for the target time 46-91 (season 2). IPS and DR evaluate the policy uniformly regardless of the target time because they do not consider the target time t'.

Figure 13 illustrates the accuracy of the estimators under non-stationary context and reward distributions when we vary the logged data sizes (n) from 500 to 4000. Overall, the extended OPFV for non-stationary context and reward distributions can most precisely estimate the value of an evaluation policy. We also observe that the OPFV adaptive to the non-stationary reward is more precise than the Prognosticator even under the non-stationarity in both the context and reward, provided that the time features for context and reward are identical. This is because the first term of OPFV for non-stationary reward in Eq. (4) becomes the same as the extended OPFV estimator for nonstationary context and reward in Eq. (11).

Figure 14 compares the accuracy of the estimators when we vary the levels of informativeness (α) and (λ) of time features for both contexts and reward. The results illustrate that as the time feature for both context and reward becomes informative, IPS, DR, and the Prognosticator estimator incur significant bias, resulting in an inaccurate future value evaluation, whereas the extended OPFV estimator mitigates this bias by using the time feature. As OPFV (true ϕ_r) is adaptive only to the non-stationarity in reward, it gradually incurs a slight bias due to the non-stationary context. OPFV (true ϕ_x and ϕ_r), however, overcomes this limitation by using the time feature not only for reward but also for context, constantly yielding the lowest MSE with varying levels of the informativeness of the time features for context and reward. When the time feature is less informative ($\alpha \leq 0.2$ and $\lambda \leq 0.2$), OPFV (true ϕ_r) and OPFV (true ϕ_x and ϕ_r) evaluate an evaluation policy as precisely as the conventional OPE estimators and the Prognosticator.

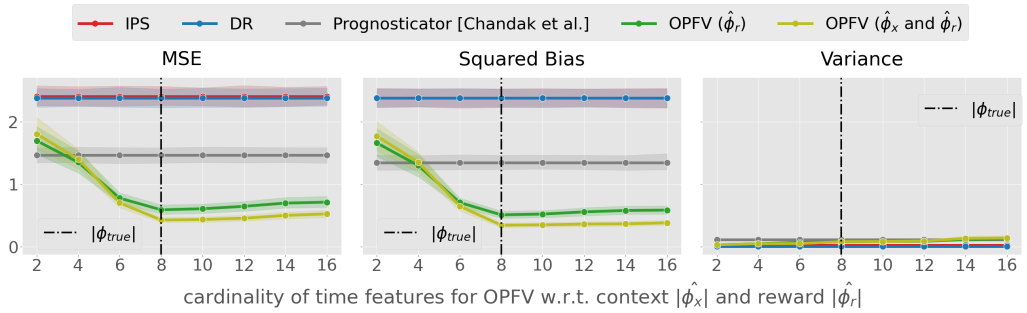


Figure 15: Comparison of the MSE, bias, and variance of IPS, DR, Prognosticator, OPFV ($\hat{\phi}_r$) for non-stationary reward distributions and OPFV ($\hat{\phi}_x$ and $\hat{\phi}_r$) for non-stationary context and reward distributions with varying cardinality of time feature $\hat{\phi}_x$ and $\hat{\phi}_r$ for OPFV

Table 3: Comparison of the average computation times (sec) for DR, OPFV (true ϕ), and OPFV (tuned $\hat{\phi}$) under varying logged data sizes (n)

logged data size (n)	500	1000	2000	4000
DR	0.03599	0.06461	0.12937	0.27898
OPFV (true ϕ)	0.05026	0.08649	0.17060	0.34300
OPFV (tuned $\hat{\phi}$)	0.41265	0.70762	1.39140	2.82739

Figure 15 depicts how the number of (tuned) time features for context and reward ($|\hat{\phi}_x|$ and $|\hat{\phi}_r|$) used in OPFV affects the precision of the estimators in F-OPE, where the true numbers of time features for both context and reward are eight. As the cardinality of time features increases, we observe that the precision of the OPFV improves due to the reduction in the bias with a slight increase in the variance. Moreover, the result demonstrates that OPFV with time features for context and reward is more precise than OPFV with a time feature for reward only, after the cardinality of time features for context and reward becomes large enough to capture the true time feature.

I.6 Investigation on the computational efficiency

This section empirically investigates the computational efficiency of our proposed method in comparison to the doubly robust estimator in F-OPE. Our aim is to demonstrate that the computational cost associated with optimizing the time feature functions used in OPFV is manageable, as described in Section 3.3. Specifically, Table 3 demonstrates the average computation times of three F-OPE algorithms with varying logged data sizes from 500 to 4000. The experimental setting is consistent with that of Figure 5 in Section 4. The table shows that the OPFV estimator completes executions in a few seconds even with the largest data size $n = 4000$ although OPFV with both true ϕ and tuned $\hat{\phi}$ are slower than DR due to the consideration of the additional time feature. Furthermore, in practical applications, it is not always required to complete the executions within minutes. Thus, we can control the computational costs of OPFV by adjusting the cardinality $|\Phi|$ of the set of the candidate time features.

J Additional setup for real-world data experiment

J.1 Non-stationarity in KuaiRec dataset

This section empirically validates that the KuaiRec dataset we used as the real-world data in Section 4 contains non-stationarity. Figure 17 illustrates the value $V_t(\pi)$ of a policy π where $\pi(a|x,t) = \frac{1}{|\mathcal{A}|}$ assigns equal probabilities to all actions throughout the time t in KuaiRec test dataset. Non-stationarity is evident in the dataset, alongside weekly trends. For instance, we observe that the policy's values are relatively high on Sundays, such as on August 16th and 23rd. In contrast, the values on Mondays are relatively low compared to the other weekdays, such as August 17th and 24th. Furthermore, we also observe that the training data also have similar weekly trends, such as high values on Sundays, July 5th, 12th, and August 2nd, and relatively low values on Mondays, July 6th, 13th, and August 3rd. These observations suggest using time features such as the day of the week will enable the capture of periodic fluctuations.

J.2 Optimization of the time feature function for OPFV

The previous section addressed the non-stationarity in the KuaiRec dataset, and we observe that there is a day-of-the-week trend in the dataset. To cope with such non-stationarity, this section provides how the OPFV-PG selected the time feature functions in real-world data experiments. For OPFV-PG (w/o tuned ϕ), we employ the day of the week as the time feature ($i.e.$, $\phi_{\text{day_of_week}} : [0, t'] \rightarrow$

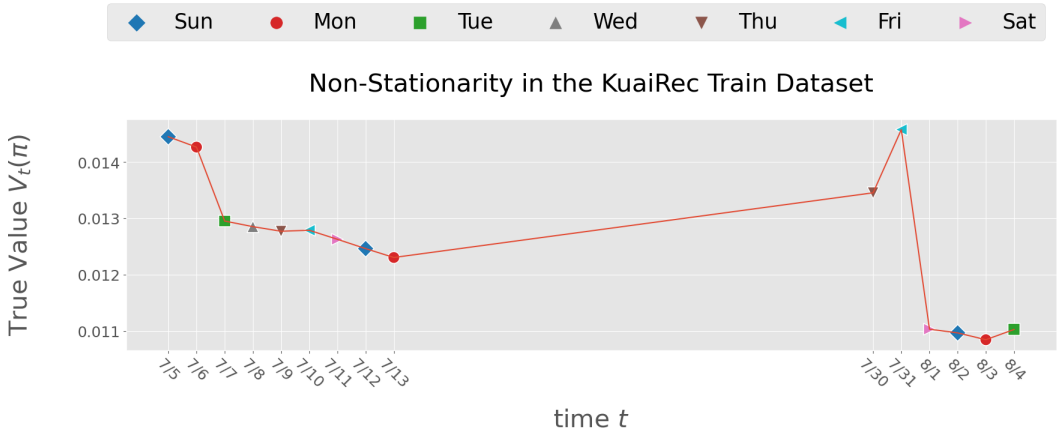


Figure 16: Comparing the true value $V_t(\pi)$ of completely random policy π with varying dates in Kuai Rec training dataset

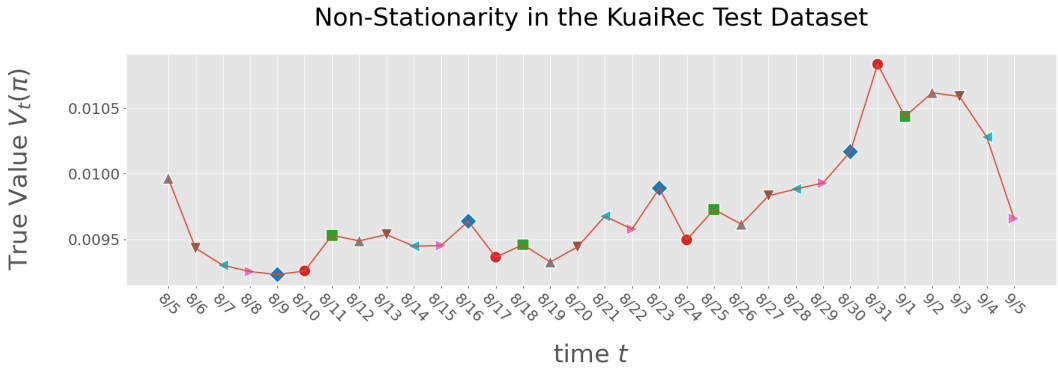


Figure 17: Comparing the true value $V_t(\pi)$ of completely random policy π with varying dates in Kuai Rec test dataset

{Sun, Mon, Tue, Wed, Thu, Fri, Sat}) as we briefly described in Section 4. For OPFV-PG (w/ tuned ϕ), we employ the following set

$$\begin{aligned}\Phi := \Big\{&\phi_{\text{day_of_week}}, \phi_{\text{weekday_weekend}}, \phi_{\text{hour}}, \phi_{\text{four_per_day}}, \phi_{\text{am_pm}},\\ &\phi_{\text{day_of_week}} \bigotimes \phi_{\text{hour}}, \phi_{\text{day_of_week}} \bigotimes \phi_{\text{four_per_day}}, \phi_{\text{day_of_week}} \bigotimes \phi_{\text{am_pm}},\\ &\phi_{\text{weekday_weekend}} \bigotimes \phi_{\text{hour}}, \phi_{\text{weekday_weekend}} \bigotimes \phi_{\text{four_per_day}}, \phi_{\text{weekday_weekend}} \bigotimes \phi_{\text{am_pm}}\Big\},\end{aligned}$$

where the first five functions are defined as follows.

$$\begin{aligned}\phi_{\text{day_of_week}} &: [0, t'] \to \{\text{Sun, Mon, Tue, Wed, Thu, Fri, Sat}\}\\ \phi_{\text{weekday_weekend}} &: [0, t'] \to \{\text{weekday, weekend}\}\\ \phi_{\text{hour}} &: [0, t'] \to \{\text{12:00am–0:59am, 1:00am–1:59am, } \cdots \text{, 11:00pm–11:59pm}\}\\ \phi_{\text{four_per_day}} &: [0, t'] \to \{\text{12:00am–5:59am, 6:00am–11:59am, 12:00pm–5:59pm, 6:00pm–11:59pm}\}\\ \phi_{\text{am_pm}} &: [0, t'] \to \{\text{12:00am–11:59am, 12:00pm–11:59pm}\}\end{aligned}$$

Moreover, for any $\phi_i, \phi_j \in \{\phi_{\text{day_of_week}}, \phi_{\text{weekday_weekend}}, \phi_{\text{hour}}, \phi_{\text{four_per_day}}, \phi_{\text{am_pm}}\}$, we define the time feature function which combine both ϕ_i and ϕ_j by $\phi_i \bigotimes \phi_j : [0, t'] \to C_i \times C_j$ if the codomains of ϕ_i and ϕ_j are denoted by C_i and C_j, respectively, (i.e., $\phi_i : [0, t'] \to C_i$ and $\phi_j : [0, t'] \to C_j$). Then, to execute the optimization, we use $\phi_\infty := \phi_{\text{day_of_week}} \bigotimes \phi_{\text{hour}}$ as the finest time feature since it is strictly finer than any time feature $\phi \in \Phi$.

Table 4: Comparison of estimated values and ranks of policies learned by F-OPL methods in the test set by three different evaluation methods.

Method	Metric	RegBased	IPS-PG	DR-PG	Prognosticator	OPFV-PG (w/o tuned ϕ)	OPFV-PG (w/ tuned ϕ)
DM	Mean Value (± SD)	1.52(± 0.32)	1.12(± 0.22)	1.33(± 0.23)	1.36(± 0.49)	1.63(± 0.56)	**1.68(± 0.61)**
SNIPS	Mean Value (± SD)	1.43(± 0.29)	1.08(± 0.74)	1.16(± 0.56)	1.38(± 0.61)	1.56(± 0.55)	**1.64(± 0.63)**
SNDR	Mean Value (± SD)	1.48(± 0.33)	0.85(± 1.19)	0.95(± 0.99)	1.42(± 0.71)	1.6(± 0.58)	**1.67(± 0.65)**

J.3 Evaluation of the methods in F-OPL

In this section, we describe the evaluation of policies learned by the methods in F-OPL. For each learned policy π_{ζ^*}, we are interested in its value $V(\pi_{\zeta^*}) := \mathbb{E}_{p(t')}[V_{t'}(\pi_{\zeta^*})]$ in the future where $V_{t'}(\pi_{\zeta^*}) := \mathbb{E}_{p(x|t')\pi_{\zeta^*}(a|x,t')}[q(x,t',a)]$ is the value at arbitrary target time t' and $q(x,t',a) := \mathbb{E}_{p(r|x,t',a)}[r]$ is the expected reward function at given target time t'. We estimate the value $V(\pi_{\zeta^*})$ by using the logged data $\mathcal{D}_{\mathrm{te}} := \{(x_i,t_i,a_i,r_i)\}_{i\in[n_{\mathrm{te}}]} \sim \prod_{i\in[n_{\mathrm{te}}]} p(t_i)p(x_i|t_i)\pi_0(a_i|x_i,t_i)p(r_i|x_i,t_i,a_i)$. Specifically, we use Direct Method (DM), Self-Normalized Inverse Propensity Scoring (SNIPS) [46], and Self-Normalized Doubly Robust (SNDR) [33, 47].

Direct Method (DM): DM uses the estimator of the expected reward and is defined as follows.

$$\hat{V}_{\mathrm{DM}}(\pi_{\zeta^*};\mathcal{D}_{\mathrm{te}},\hat{q}) := \frac{1}{n_{\mathrm{te}}}\sum_{i\in[n_{\mathrm{te}}]}\mathbb{E}_{\pi_{\zeta^*}(a|x_i,t_i)}[\hat{q}(x_i,t_i,a)] = \frac{1}{n_{\mathrm{te}}}\sum_{i\in[n_{\mathrm{te}}]}\sum_{a\in\mathcal{A}}\pi_{\zeta^*}(a|x_i,t_i)\hat{q}(x_i,t_i,a),$$

where $\hat{q}(x,t,a)$ is the estimator of the expected reward $q(x,t,a)$. DM often has a lower variance than the estimators, which use importance weights. However, it often incurs bias due to the misspecification of the regression model.

Self-Normalized Inverse Propensity Scoring (SNIPS) [46]: In contrast to DM, SNIPS is one of the model-free approaches defined as follows.

$$\hat{V}_{\mathrm{SNIPS}}(\pi_{\zeta^*};\mathcal{D}_{\mathrm{te}}) := \frac{1}{n_{\mathrm{te}}}\sum_{i\in[n_{\mathrm{te}}]}\frac{w(x_i,t_i,a_i)r_i}{\frac{1}{n_{\mathrm{te}}}\sum_{j\in[n_{\mathrm{te}}]}w(x_j,t_j,a_j)} = \frac{\sum_{i\in[n_{\mathrm{te}}]}w(x_i,t_i,a_i)r_i}{\sum_{j\in[n_{\mathrm{te}}]}w(x_j,t_j,a_j)},$$

where $w(x,t,a) := \pi_{\zeta^*}(a|x,t)/\pi_0(a|x,t)$ is the importance weight and $\frac{1}{n_{\mathrm{te}}}\sum_{j\in[n_{\mathrm{te}}]}w(x_j,t_j,a_j)$ is the empirical mean of the importance weights as the normalizing factor. SNIPS mitigates the high variance incurred in the IPS estimator due to the use of the normalizing factor. Moreover, SNIPS is a consistent estimator and is known to work empirically well [38].

Self-Normalized Doubly Robust (SNDR) [33, 47]: SNDR combines DM and SNIPS as follows.

$$\hat{V}_{\mathrm{SNDR}}(\pi_{\zeta^*};\mathcal{D}_{\mathrm{te}},\hat{q}) := \frac{1}{n_{\mathrm{te}}}\sum_{i\in[n_{\mathrm{te}}]}\left\{\mathbb{E}_{\pi_{\zeta^*}(a|x_i,t_i)}[\hat{q}(x_i,t_i,a)] + \frac{w(x_i,t_i,a_i)(r_i-\hat{q}(x_i,t_i,a_i))}{\frac{1}{n_{\mathrm{te}}}\sum_{j\in[n_{\mathrm{te}}]}w(x_j,t_j,a_j)}\right\},$$

where the first term is the same as DM, and the second term uses $\hat{q}(x,t,a)$ as the control variate to have a lower variance than SNIPS.

J.4 Additional results by DM, SNIPS, and SNDR evaluations

Table 4 compares the mean and standard deviation of the values of the learned policies evaluated by three different OPE estimators across ten experimental runs with different seeds before we take the average of the results of three evaluations shown in Section 4. No matter what OPE estimator we use to evaluate the policies, two versions of OPFV-PG invariably enable us to achieve the highest policy values under unknown and potentially intricate real-world non-stationary environments. More specifically, we observe that the mean values of OPFV-PG (w/ tuned ϕ) are consistently higher than OPFV-PG (w/o tuned ϕ) under all three evaluation methods. These results empirically validate the real-world applicability of the optimization procedure of time feature functions in Section 3.3 for real-world data. Additionally, the table demonstrates that OPFV-PG (w/o tuned ϕ) achieves higher mean values than any baseline OPL methods for any evaluation method. Thus, even without the optimization of the time feature functions, our proposed method for F-OPL enables effective learning under non-stationarity.