Optimizing for the Future in Non-Stationary MDPs (Supplementary Material)

A. Proofs for the Properties of the NIS and NWIS Estimators

Here we provide proofs for the properties of the NIS and NWIS estimators. While NIS and NWIS are developed for the non-stationary setting, these properties ensure that these estimators generalize to the stationary setting as well. That is, when used in a stationary setting, the NIS estimator is both unbiased and consistent like the PDIS estimator, and the NWIS estimator is biased and consistent like the WIS estimator.

Our proof technique draws inspiration from the results presented by Mahmood et al. (2014). The key modification that we make to leverage their proof technique is that instead of using the features of the state as the input and the observed return from that corresponding state as the target to the regression function, we use the features of the *time index of an episode* as the input and the observed return for that corresponding episode as the target. In their setup, because states are drawn stochastically from a distribution, their analysis is not directly applicable to our setting where inputs are time indices that form a deterministic sequence. For analysis of our estimators, we leverage techniques discussed by Greene (2003) for analyzing properties of the ordinary least squares estimator.

Before proceeding, we impose the following constraints on the set of policies, and the basis functions $\phi_i : \mathbb{N} \to \mathbb{R}$ used for encoding the time index in both Ψ and $\Psi^\ddagger$, with $\phi(\cdot) = [\phi_1(\cdot), ..., \phi_{d-1}(\cdot), 1]$.

(a) $\phi(\cdot)$ always contains 1 to incorporate a bias coefficient in least-squares regression (for example, $\phi(\cdot) = [\phi_1(\cdot), ..., \phi_{d-1}(\cdot), 1]$, where $\forall i \in [1, d-1]$, $\phi_i(\cdot)$ is a basis function).[4]

(b) There exists a finite constant C_1, such that $\forall i,\ |\phi_i(\cdot)| < C_1$.

(c) Φ has full column rank such that $(\Phi^\top \Phi)^{-1}$ exists.

(d) We only consider a set of policies Π that have non-zero probability of taking any action in any state. That is, $\exists C_2 > 0$, such that $\forall \pi \in \Pi, \forall s \in \mathcal{S}, \forall a \in \mathcal{A}, \pi(a|s) > C_2$.

Satisfying condition (a) is straightforward as it is typically already satisfied by all basis functions. This constraint ensures that the regression based forecasting function can capture a fixed constant that is required to model the absence of any trend. This constraint is useful for our purpose as in the stationary setting there exists no trend in the expected performance across episodes for any given policy.

Conditions (b) and (c) are also readily satisfied by popular basis functions. For example, features from the Fourier basis are bounded by $[-1, 1]$, and features from polynomial/identity bases are also bounded when inputs are adequately normalized. Further, when the basis function does not repeat any feature, and there are more samples than the number of features ($k \geq d$), condition (c) is satisfied. This ensures that the least-squares problem is well-defined and has a unique-solution.

Condition (d) ensures that the denominator in any importance ratio is always bounded below, such that the importance ratios are bounded above. This implies that the importance sampling estimator for any policy has finite variance. Use of entropy regularization with common policy parameterizations (softmax/Gaussian) can prevent violation of this condition.

In the following, we first establish the finite-sample properties and then we establish the large-sample properties for the NIS and NWIS estimators. Before proceeding further, recall from (4) and (7) that the NIS and NWIS estimators are given by:

$$\hat{J}_{k+\delta}(\pi) = \phi(k+\delta)w = \phi(k+\delta)(\Phi^\top \Phi)^{-1}\Phi^\top Y$$
$$\hat{J}^\ddagger_{k+\delta}(\pi) = \phi(k+\delta)w^\ddagger = \phi(k+\delta)(\Phi^\top \Lambda \Phi)^{-1}\Phi^\top \Lambda Y.$$

[4] If additional domain knowledge is available to select an appropriate basis function that can be used to represent the performance trend of all the policies for the given non-stationary environment, then all the following finite-sample and large-sample properties can be extended for that environment as well, using that basis function.

A.1. Finite Sample Properties

In this subsection, finite sample properties of NIS and NWIS are presented. Specifically, it is established that NIS is an unbiased estimator, whereas NWIS is a biased estimator of $J(\pi)$, where $J(\pi)$ is the performance of a policy π in a stationary MDP.

Theorem 1 (Unbiased NIS)**.** *Under Assumptions 1 and 2, for all $\delta \geq 1$, $\hat{J}_{k+\delta}(\pi)$ is an unbiased estimator of $J(\pi)$. That is, $\mathbb{E}[\hat{J}_{k+\delta}(\pi)] = J(\pi)$.*

Proof. Recall from (4) that

$$\hat{J}_{k+\delta}(\pi) = \phi(k+\delta)w = \phi(k+\delta)(\Phi^\top\Phi)^{-1}\Phi^\top Y.$$

Therefore, the expected value of $\hat{J}_{k+\delta}(\pi)$ is

$$\begin{aligned}\mathbb{E}[\hat{J}_{k+\delta}(\pi)] &= \mathbb{E}\left[\phi(k+\delta)(\Phi^\top\Phi)^{-1}\Phi^\top Y\right] \\ &= \phi(k+\delta)\left(\Phi^\top\Phi\right)^{-1}\left(\Phi^\top\mathbb{E}\left[Y\right]\right). \end{aligned} \tag{8}$$

As $Y = [\hat{J}_0(\pi), ..., \hat{J}_k(\pi)]^\top$ and the MDP is stationary, the expected value of each element of Y is $J(\pi)$. Further, since $\phi(\cdot)$ always contains the bias co-efficient, and the performance of any policy in invariant to the episode number in a stationary MDP (Assumption 1), the optimal parameter for the regression model is $w^* = [0, 0, ..., 0, J(\pi)]^\top$, such that for any k,

$$\phi(k)w^* = [\phi_1(k), ..., \phi_{d-1}(k), 1][0, ..., 0, J(\pi)]^\top = J(\pi). \tag{9}$$

Therefore, $\mathbb{E}[Y] = \Phi w^*$. Using this observation in (8),

$$\begin{aligned}\mathbb{E}[\hat{J}_{k+\delta}(\pi)] &= \phi(k+\delta)\left(\Phi^\top\Phi\right)^{-1}\left(\Phi^\top\Phi w^*\right) \\ &= \phi(k+\delta)\left(\Phi^\top\Phi\right)^{-1}\left(\Phi^\top\Phi\right)w^* \\ &= \phi(k+\delta)w^* \\ &= J(\pi).\end{aligned}$$

□

Proof. (Alternate) Here we present an alternate proof for Theorem 1 which does not require invoking w^*.

$$\begin{aligned}\mathbb{E}\left[\hat{J}_{k+\delta}(\pi)\right] &= \mathbb{E}\left[\phi(k+\delta)(\Phi^\top\Phi)^{-1}\Phi^\top Y\right] \\ &\overset{(a)}{=} \mathbb{E}\left[\sum_{i=0}^{k}\left[\phi(k+\delta)(\Phi^\top\Phi)^{-1}\Phi^\top\right]_i Y_i\right] \\ &\overset{(b)}{=} \sum_{i=0}^{k}\left[\phi(k+\delta)(\Phi^\top\Phi)^{-1}\Phi^\top\right]_i \mathbb{E}\left[Y_i\right],\end{aligned}$$

where (a) is the dot product written as summation, and (b) holds because the multiplicative constants are fixed values, as given in (6). Since the environment is stationary, $\forall i\ \mathbb{E}\left[Y_i\right] = J(\pi)$, therefore,

$$\mathbb{E}\left[\hat{J}_{k+1}(\pi)\right] = J(\pi)\sum_{i=0}^{k}\left[\phi(k+\delta)(\Phi^\top\Phi)^{-1}\Phi^\top\right]_i. \tag{10}$$

In the following we focus on the terms inside the summation in (10). Without loss of generality, assume that for a given matrix of features Φ, the feature corresponding to value 1 is in the last column of Φ. Let $\mathbf{A} := (\Phi^\top\Phi)^{-1}\Phi^\top \in \mathbb{R}^{d\times k}$, and let $B := \Phi[1:k, 1:d-1] \in \mathbb{R}^{k\times(d-1)}$ be the submatrix of Φ such that $\mathbf{B}$ has all features of Φ except the ones column, $\mathbb{1} \in \mathbb{R}^{k\times 1}$. Let $\mathbf{I}$ be the identity matrix in $\mathbb{R}^{d\times d}$, then it can seen that $(\Phi^\top\Phi)^{-1}(\Phi^\top\Phi)$ can be expressed as,

$$\left[\ \mathbf{A}\ \right]\left[\ \mathbf{B}\ \middle|\ \mathbb{1}\ \right] = \mathbf{I},$$

This implies $[\mathbf{AB} \;\; \mathbf{A}\mathbb{1}] = \mathbf{I}$. Therefore, as the j^{th} row in last column of $\mathbf{I}$ corresponds to the dot product of the j^{th} row of $\mathbf{A}$, $\mathbf{A_j}$, with $\mathbb{1}$,

$$\mathbf{A}_j \mathbb{1} = \begin{cases} 0 & j \neq d, \\ 1 & j = d. \end{cases} \tag{11}$$

Equation (11) ensures that the summation of all rows of $\mathbf{A}$, except the last, sum to 0, and the last one sums to 1. Now, let $\phi(k+\delta) := [\phi_1(k+\delta), \phi_2(k+\delta), ..., \phi_{d-1}(k+\delta), 1] \in \mathbb{R}^{1\times d}$. Therefore,

$$\begin{aligned}
\sum_{i=1}^{k} \left[\phi(k+\delta)(\Phi^\top\Phi)^{-1}\Phi^\top\right]_i &= \sum_{i=1}^{k} \left[\phi(k+\delta)\mathbf{A}\right]_i \\
&= \sum_{i=1}^{k}\sum_{j=1}^{d} \left[\phi(k+\delta)\right]_j \mathbf{A}_{j,i} \\
&= \sum_{j=1}^{d} \left[\phi(k+\delta)\right]_j \sum_{i=1}^{k} \mathbf{A}_{j,i} \\
&= \left(\sum_{j=1}^{d-1} \left[\phi(k+\delta)\right]_j \sum_{i=1}^{k} \mathbf{A}_{j,i}\right) + \left(\left[\phi(k+\delta)\right]_d \sum_{i=1}^{k} \mathbf{A}_{d,i}\right) \\
&= \left(\sum_{j=1}^{d-1} \left[\phi(k+\delta)\right]_j \left(\mathbf{A}_j \mathbb{1}\right)\right) + \left(\left[\phi(k+\delta)\right]_d \left(\mathbf{A}_d \mathbb{1}\right)\right) \\
&= \left(\sum_{j=1}^{d-1} \left[\phi(k+\delta)\right]_j \cdot 0\right) + \left(\left[\phi(k+\delta)\right]_d \cdot 1\right) \\
&= \left[\phi(k+\delta)\right]_d \\
&= 1.
\end{aligned} \tag{12}$$

Therefore, combining (12) with (10), $\mathbb{E}\left[\hat{J}_{k+\delta}(\pi)\right] = J(\pi)$. □

Theorem 2 (Biased NWIS)**.** *Under Assumptions 1 and 2, for all $\delta \geq 1$, $\hat{J}^{\ddagger}_{k+\delta}(\pi)$ may be a biased estimator of $J(\pi)$. That is, it is possible that $\mathbb{E}[\hat{J}^{\ddagger}_{k+\delta}(\pi)] \neq J(\pi)$.*

Proof. We prove this result using a simple counter-example. Consider the following basis function, $\phi(\cdot) = [1]$:

$$\begin{aligned}
J^{\ddagger}_{k+\delta}(\pi) &= \phi(k+\delta) w^{\ddagger} \\
&= \phi(k+\delta) \operatorname*{argmin}_{c\in\mathbb{R}^{1\times 1}} \frac{1}{n}\sum_{i=1}^{n} \rho_i(0,T)(G_i - c\phi(i))^2 \\
&= \operatorname*{argmin}_{c\in\mathbb{R}^{1\times 1}} \frac{1}{n}\sum_{i=1}^{n} \rho_i(0,T)(G_i - c)^2 \\
&= \frac{\sum_{i=1}^{n} \rho_i(0,T) G_i}{\sum_{i=1}^{n} \rho_i(0,T)},
\end{aligned}$$

which is the WIS estimator. Therefore, as WIS is a biased estimator (Precup, 2000), NWIS is also a biased estimator of $J(\pi)$. □

A.2. Large Sample Properties

In this subsection, large sample properties of NIS and NWIS are presented. Specifically, it is established that both NIS and NWIS are consistent estimators of, $J(\pi)$, the performance of a policy π for a stationary MDP.

Theorem 3 (Consistent NIS)**.** *Under Assumptions 1 and 2, for all $\delta \geq 1$, $\hat{J}_{k+\delta}(\pi)$ is a consistent estimator of $J(\pi)$. That is, as $N \to \infty$, $\hat{J}_{N+\delta}(\pi) \xrightarrow{a.s.} J(\pi)$.*

Proof. The proof follows from the standard consistency result for ordinary least-squares regression (Greene, 2003). Formally, using (4),

$$\begin{aligned}\lim_{N\to\infty} \quad \hat{J}_{N+\delta}(\pi) &= \lim_{N\to\infty} \quad \phi(N+\delta)w \\ &= \lim_{N\to\infty} \quad \phi(N+\delta)(\Phi^\top\Phi)^{-1}\Phi^\top Y.\end{aligned}$$

Since $Y = [\hat{J}_0(\pi), ..., \hat{J}_N(\pi)]^\top$ and the MDP is stationary, each element of Y is an unbiased estimate of $J(\pi)$. In other words, $\forall i \in [0, N], \hat{J}_i(\pi) = J(\pi) + \epsilon_i$, where ϵ_i is a mean zero error. Let $\epsilon \in \mathbb{R}^{N\times 1}$ be the vector containing all the error terms ϵ_i. Now, using (9),

$$\begin{aligned}\lim_{N\to\infty} \quad \hat{J}_{N+\delta}(\pi) &= \lim_{N\to\infty} \quad \phi(N+\delta)\left(\Phi^\top\Phi\right)^{-1}\left(\Phi^\top(\Phi w^* + \epsilon)\right) \\ &= \lim_{N\to\infty} \quad \phi(N+\delta)\left(\Phi^\top\Phi\right)^{-1}\left(\left(\Phi^\top\Phi\right)w^* + \left(\Phi^\top\epsilon\right)\right) \\ &= \lim_{N\to\infty} \quad \phi(N+\delta)w^* + \phi(N+\delta)\left(\Phi^\top\Phi\right)^{-1}\left(\Phi^\top\epsilon\right) \\ &= \lim_{N\to\infty} \quad J(\pi) + \phi(N+\delta)\left(\Phi^\top\Phi\right)^{-1}\left(\Phi^\top\epsilon\right) \\ &= \lim_{N\to\infty} \quad J(\pi) + \phi(N+\delta)\left(\frac{1}{N}\Phi^\top\Phi\right)^{-1}\left(\frac{1}{N}\Phi^\top\epsilon\right).\end{aligned} \tag{13}$$

If both $Q^{-1} := \left(\lim_{N\to\infty}\frac{1}{N}\Phi^\top\Phi\right)^{-1}$ and $\left(\lim_{N\to\infty}\frac{1}{N}\Phi^\top\epsilon\right)$ exist, then using Slutsky's Theorem,

$$\lim_{N\to\infty} \quad \hat{J}_{N+\delta}(\pi) = J(\pi) + \phi(N+\delta)Q^{-1}\left(\lim_{N\to\infty} \quad \frac{1}{N}\Phi^\top\epsilon\right). \tag{14}$$

To validate conditions for Slutsky's Theorem, notice that it holds from Grenander's conditions that Q^{-1} exists. Informally, Grenander's conditions require that no feature degenerates to a sequence of zeros, no feature of a single observation dominates the sum of squares of its series, and the $\Phi^\top\Phi$ matrix always has full rank. These conditions are easily satisfied for most popular basis functions used to create input features. For formal definitions of these conditions, we refer the reader to Chpt. 5, Greene (2003).

In the following, we restrict our focus to the term inside the brackets in the second term of (14) and show that it exists, so that (14) is valid. Notice that the mean of that term is,

$$\mathbb{E}\left[\frac{1}{N}\Phi^\top\epsilon\right] = \frac{1}{N}\Phi^\top\mathbb{E}\left[\epsilon\right] = 0.$$

Since the mean is 0, the variance is,

$$\mathbb{V}\left[\frac{1}{N}\Phi^\top\epsilon\right] = \frac{1}{N^2}\mathbb{V}\left[\Phi^\top\epsilon\right] = \frac{1}{N^2}\mathbb{E}\left[\left(\Phi^\top\epsilon\right)\left(\Phi^\top\epsilon\right)^\top\right] = \frac{1}{N^2}\left(\Phi^\top\mathbb{E}\left[\epsilon\epsilon^\top|\Phi\right]\Phi\right).$$

As each policy has a non-zero probability of taking any action in any state, the variance of PDIS (or the standard IS) estimator is bounded and thus each element of $\mathbb{E}[\epsilon\epsilon^\top|\Phi]$ is bounded. Further, as $\phi_i(\cdot)$ is bounded, each element of Φ is also bounded. Therefore,

$$\lim_{N\to\infty} \quad \mathbb{V}\left[\frac{1}{N}\Phi^\top\epsilon\right] \to 0.$$

Since the mean is 0 and the variance asymptotes to 0, by Kolmogorov's strong law of large numbers it follows that as $N \to \infty$, $\frac{1}{N}\Phi^\top\epsilon \xrightarrow{\text{a.s.}} 0$. Combining this with (14),

$$\lim_{N\to\infty} \hat{J}_{N+\delta}(\pi) \xrightarrow{a.s.} J(\pi) + \phi(N+\delta)Q^{-1}0 = J(\pi).$$

□

Theorem 4 (Consistent NWIS). *Under Assumptions 1 and 2, for all $\delta \geq 1$, $\hat{J}^{\ddagger}_{k+\delta}(\pi)$ is a consistent estimator of $J(\pi)$. That is, as $N \to \infty$, $\hat{J}^{\ddagger}_{N+\delta}(\pi) \xrightarrow{a.s.} J(\pi)$.*

Proof. Recall from (7) that

$$\hat{J}^{\ddagger}_{N+\delta}(\pi) = \phi(N+\delta)w^{\ddagger} = \phi(N+\delta)(\Phi^\top \Lambda \Phi)^{-1}\Phi^\top \Lambda Y.$$

Consistency of $\hat{J}^{\ddagger}_{N+\delta}(\pi)$ can be proven similarly to the proof of Theorem 3. Note that here $Y = [G_0, ..., G_k]^\top$ contains the returns for each episode, and ΛY denotes the unbiased estimates for $J(\pi)$. Therefore, similar to (13),

$$\begin{aligned}
\lim_{N\to\infty} \quad \hat{J}^{\ddagger}_{N+\delta}(\pi) &= \lim_{N\to\infty} \quad \phi(N+\delta)(\Phi^\top \Lambda \Phi)^{-1}(\Phi^\top(\Phi w^* + \epsilon)) \\
&= \lim_{N\to\infty} \quad \phi(N+\delta)(\Phi^\top \Lambda \Phi)^{-1}((\Phi^\top \Phi) w^* + \Phi^\top \epsilon) \\
&= \lim_{N\to\infty} \quad \phi(N+\delta)\left(\frac{1}{N}\Phi^\top \Lambda \Phi\right)^{-1}\left(\left(\frac{1}{N}\Phi^\top \Phi\right) w^* + \frac{1}{N}\Phi^\top \epsilon\right). \qquad (15)
\end{aligned}$$

In the following, we will make use of Slutsky's Theorem. To do so, we first restrict our focus to the terms in the first bracket in (15), and show existence of its limit. Let $\tilde{\rho}_k \coloneqq \rho^{\ddagger}_k - \mathbb{E}[\rho^{\ddagger}_k]$ be a mean 0 random variable, then

$$\begin{aligned}
\lim_{N\to\infty} \quad \frac{1}{N}\Phi^\top \Lambda \Phi &= \lim_{N\to\infty} \quad \frac{1}{N}\sum_{k=1}^{N} \rho^{\ddagger}_k \phi(k)^\top \phi(k). \\
&= \lim_{N\to\infty} \quad \left(\frac{1}{N}\sum_{k=1}^{N} \tilde{\rho}_k \phi(k)^\top \phi(k) + \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}\left[\rho^{\ddagger}_k\right] \phi(k)^\top \phi(k)\right). \\
&\overset{(a)}{\to} \lim_{N\to\infty} \quad \frac{1}{N}\sum_{k=1}^{N} \mathbb{E}\left[\rho^{\ddagger}_k\right] \phi(k)^\top \phi(k) \\
&\overset{(b)}{=} \lim_{N\to\infty} \quad \frac{1}{N}\sum_{k=1}^{N} \phi(k)^\top \phi(k) \\
&= \lim_{N\to\infty} \quad \frac{1}{N}\Phi^\top \Phi, \qquad (16)
\end{aligned}$$

where (a) follows from the Kolmogorov's strong law of large numbers. To see this, let $Z_k = \tilde{\rho}_k \phi(k)^\top \phi(k)$. Notice that $\mathbb{E}[Z_k] = \mathbb{E}[\tilde{\rho}_k]\phi(k)^\top \phi(k) = 0$, and as both $\tilde{\rho}$ and $\phi(\cdot)$ are bounded, the variance of Z_k is also bounded. Therefore, $(1/N)\sum \tilde{\rho}_k \phi(k)^\top \phi(k) \to 0$ almost surely as $N \to \infty$. Consequently, (b) is obtained using the fact that the expected value of importance ratios is 1 (Thomas, 2015, Lemma 3). Notice that (16) reduced to Q (which was defined in the proof of Theorem 3) and we know that its limit exists because $\phi(\cdot)$ is bounded. Further, we also know that Q^{-1} and $\left(\lim_{N\to\infty} \frac{1}{N}\Phi^\top \epsilon\right)$ exist (see the proof of Theorem 3). Therefore, using Slutsky's Theorem and substituting (16) in (15),

$$\begin{aligned}
\lim_{N\to\infty} \quad \hat{J}^{\ddagger}_{N+\delta}(\pi) &= \phi(N+\delta)\left(\lim_{N\to\infty} \quad \frac{1}{N}\Phi^\top \Phi\right)^{-1}\left(\left(\lim_{N\to\infty} \quad \frac{1}{N}\Phi^\top \Phi\right) w^* + \lim_{N\to\infty} \quad \frac{1}{N}\Phi^\top \epsilon\right) \\
&= \phi(N+\delta)w^* + \phi(N+\delta)\left(\lim_{N\to\infty} \quad \frac{1}{N}\Phi^\top \Phi\right)^{-1}\left(\lim_{N\to\infty} \quad \frac{1}{N}\Phi^\top \epsilon\right) \\
&= J(\pi) + \phi(N+\delta)\left(\lim_{N\to\infty} \quad \frac{1}{N}\Phi^\top \Phi\right)^{-1}\left(\lim_{N\to\infty} \quad \frac{1}{N}\Phi^\top \epsilon\right) \\
&\xrightarrow{\text{a.s.}} J(\pi), \qquad (17)
\end{aligned}$$

where (17) follows from the simplification used for (14) in the proof of Theorem 3. □

$_t \backslash^l$	0	1	2	...	T
0	$\gamma^0\rho_i(0,0)\Psi_i^0 R_i^0$				
1	$\gamma^1\rho_i(0,1)\Psi_i^0 R_i^1$	$\gamma^1\rho_i(0,1)\Psi_i^1 R_i^1$			
2	$\gamma^2\rho_i(0,2)\Psi_i^0 R_i^2$	$\gamma^2\rho_i(0,2)\Psi_i^1 R_i^2$	$\gamma^2\rho_i(0,2)\Psi_i^2 R_i^2$		
⋮	⋮	⋮	⋮	⋮	⋮
T	$\gamma^T\rho_i(0,T)\Psi_i^0 R_i^T$	$\gamma^T\rho_i(0,T)\Psi_i^1 R_i^T$	$\gamma^T\rho_i(0,T)\Psi_i^2 R_i^T$	...	$\gamma^T\rho_i(0,T)\Psi_i^T R_i^T$

Table 1. let $\Psi_i^t = \partial \log \pi^\theta(A_i^t|S_i^t)/\partial\theta$. This table represents all the terms in (18) required for computing $\nabla \hat{J}_i(\theta)$. Gray color denotes empty cells.

B. Gradient of PDIS Estimator

Recall that the NIS and NWIS estimators build upon estimators of past performances by using them along with OLS and WLS regression, respectively. Consequently, gradients of the NIS and NWIS estimators with respect to the policy parameters can be decomposed into terms that consist of gradients of the estimates of past performances with respect to the policy parameters. Here we provide complete derivations for obtaining a straightforward equation for computing the gradients of the PDIS estimator with respect to the policy parameters. These might also be of independent interest when dealing with off-policy policy optimization for stationary MDPs.

Property 1 (PDIS Gradient)**.** *Let* $\rho_i(0,l) \coloneqq \prod_{j=0}^{l} \frac{\pi^\theta(A_i^j|S_i^j)}{\beta_i(A_i^j|S_i^j)}$*, then*

$$\nabla \hat{J}_i(\theta) = \sum_{t=0}^{T} \frac{\partial \log \pi^\theta(A_i^t|S_i^t)}{\partial\theta} \left(\sum_{l=t}^{T} \rho_i(0,l)\gamma^l R_i^l \right).$$

Proof. Recall from (2) that,

$$\hat{J}_i(\theta) = \sum_{t=0}^{T} \left(\prod_{l=0}^{t} \frac{\pi^\theta(A_i^l|S_i^l)}{\beta(A_i^l|S_i^l)} \right) \gamma^t R_i^t.$$

Computing the gradient of $\hat{J}_i(\theta)$,

$$\begin{aligned}
\nabla \hat{J}_i(\theta) &= \sum_{t=0}^{T} \frac{\partial}{\partial\theta} \left(\prod_{l=0}^{t} \frac{\pi^\theta(A_i^l|S_i^l)}{\beta(A_i^l|S_i^l)} \right) \gamma^t R_i^t \\
&= \sum_{t=0}^{T} \left(\prod_{l=0}^{t} \frac{\pi^\theta(A_i^l|S_i^l)}{\beta(A_i^l|S_i^l)} \right) \frac{\partial \log \left(\prod_{l=0}^{t} \pi^\theta(A_i^l|S_i^l) \right)}{\partial\theta} \gamma^t R_i^t \\
&= \sum_{t=0}^{T} \left(\prod_{l=0}^{t} \frac{\pi^\theta(A_i^l|S_i^l)}{\beta(A_i^l|S_i^l)} \right) \left(\sum_{l=0}^{t} \frac{\partial \log \pi^\theta(A_i^l|S_i^l)}{\partial\theta} \right) \gamma^t R_i^t \\
&= \sum_{t=0}^{T} \rho_i(0,t) \left(\sum_{l=0}^{t} \frac{\partial \log \pi^\theta(A_i^l|S_i^l)}{\partial\theta} \right) \gamma^t R_i^t. \\
&= \sum_{t=0}^{T} \frac{\partial \log \pi^\theta(A_i^t|S_i^t)}{\partial\theta} \left(\sum_{l=t}^{T} \rho_i(0,l)\gamma^l R_i^l \right),
\end{aligned} \tag{18}$$

where, in the last step, instead of the summation over the partial derivatives of $\log \pi^\theta$ for each weight $\rho(\cdot,\cdot)$, we consider the alternate form where the summation is over the importance weights $\rho(\cdot,\cdot)$ for each partial derivative of $\log \pi^\theta$. To see this step clearly, let $\Psi_i^t = \partial \log \pi^\theta(A_i^t|S_i^t)/\partial\theta$, then Table 1 shows all the terms in (18). The last step above corresponds to taking the column-wise sum instead of the row-wise sum in Table 1.

□

C. Detailed Literature Review

The problem of non-stationarity has a long history. In the operations research community, many dynamic sequential decision-making problems are modeled using infinite horizon *non-homogeneous* MDPs (Hopp et al., 1987). While estimating an optimal policy is infeasible under an infinite horizon setting when the dynamics are changing and a stationary distribution cannot be reached, several researchers have studied the problem of identifying sufficient forecast horizons for performing near-optimal planning (Garcia & Smith, 2000; Cheevaprawatdomrong et al., 2007; Ghate & Smith, 2013) or robust policy iteration (Sinha & Ghate, 2016).

In contrast, non-stationary multi-armed bandits (NMAB) capture the setting where the horizon length is one, but the reward distribution changes over time (Moulines, 2008; Besbes et al., 2014). Many variants of NMAB, like *cascading non-stationary bandits* (Wang et al., 2019b; Li & de Rijke, 2019) and *rotten bandits* (Levine et al., 2017; Seznec et al., 2018) have also been considered. In optimistic online convex optimization, researchers have shown that better performance can be achieved by updating the parameters using predictions (which are based on the past gradients) of the gradient of the future loss (Rakhlin & Sridharan, 2013; Yang & Mohri, 2016; Mohri & Yang, 2016; Wang et al., 2019a).

Non-stationarity also occurs in multiplayer games, like rock-paper-scissors, where each episode is a single one-step interaction (Singh et al., 2000; Bowling, 2005; Conitzer & Sandholm, 2007). Opponent modeling in games has been shown to be useful and regret bounds for multi-player games where players can be replaced with some probability p, i.e., the game changes slowly over time, have also been established (Zhang & Lesser, 2010; Mealing & Shapiro, 2013; Foster et al., 2016; Foerster et al., 2018). However, learning sequential strategies in a non-stationary setting is still an open research problem.

For episodic non-stationary MDPs, researchers have also looked at providing regret bounds for algorithms that exploit oracle access to the current reward and transition functions (Even-Dar et al., 2005; Yu & Mannor, 2009; Abbasi et al., 2013; Lecarpentier & Rachelson, 2019; Li et al., 2019). Alleviating oracle access by performing a count-based estimate of the reward and transition functions based on the recent history of interactions has also been proposed (Gajane et al., 2018; Cheung et al., 2019). For tabular MDPs, past data from a non-stationary MDP can be used to construct a maximum-likelihood estimate model (Ornik & Topcu, 2019) or a full Bayesian model (Jong & Stone, 2005) of the transition dynamics. Our focus is on the setting which is not restricted to tabular representations.

A Hidden-Mode MDP is an alternate setting that assumes that the environment changes are confined to a small number of hidden modes, where each mode represents a unique MDP. This provides a more tractable way to model a limited number of MDPs (Choi et al., 2000; Basso & Engel, 2009; Mahmud & Ramamoorthy, 2013), or perform model-free updates using mode-change detection (Padakandla et al., 2019). In this work, we are interested in the continuously changing setting, where the number of possible MDPs is unbounded.

Tracking has also been shown to play an important role in non-stationary domains. Thomas et al. (2017) and Jagerman et al. (2019) have proposed policy evaluation techniques in a non-stationary setting by tracking a policy's past performances. However, they do not provide any procedure for searching for a good future policy. To adapt quickly in non-stationary tasks, TIDBD (Kearney et al., 2018) and AdaGain (Jacobsen et al., 2019) perform TD-learning while also automatically (de-)emphasizing updates to (ir)relevant features by modulating the learning rate of the parameters associated with the respective features. Similarly, Abdallah & Kaisers (2016) propose repeating a Q-value update inversely proportional to the probability with which an action was chosen to obtain a transition tuple. In this work, we go beyond tracking and proactively optimize for the future.

D. Implementation details

D.1. Environments

We provide empirical results on three non-stationary environments: diabetes treatment, recommender system, and a goal-reacher task. Details for each of these environments are provided in this section.

Non-stationary Diabetes Treatment: This MDP models the problem of Type-1 diabetes management. A person suffering from Type-1 diabetes does not produce enough *insulin*, a hormone that promotes absorption of glucose from the blood. Consumption of a meal increases the blood-glucose level in the body, and if the blood-glucose level becomes too high, then the patient can suffer from *hyperglycemia*. Insulin injections can reduce the blood-glucose level, but if the level becomes too low, then the patient suffers from *hypoglycemia*. While either of the extremes is undesirable, hypoglycemia is more dangerous and can triple the five-year mortality rate for a person with diabetes (Man et al., 2014).

Autonomous medical support systems have been proposed to decide how much insulin should be injected to keep a person's blood glucose levels near ideal levels (Bastani, 2014). Currently, the parameters of such a medical support system are set by a doctor specifically for each patient. However, due to non-stationarities induced over time as a consequence of changes in the body mass index, the insulin sensitivity of the pancreas, diet, etc., the parameters of the controller need to be readjusted regularly. Currently, this requires revisiting the doctor. A viable reinforcement learning solution to this non-stationary problem could enable the automatic tuning of these parameters for patients who lack regular access to a physician.

To model this MDP, we use an open-source implementation (Xie, 2019) of the U.S. Food and Drug Administration (FDA) approved Type-1 Diabetes Mellitus simulator (T1DMS) (Man et al., 2014) for treatment of Type-1 diabetes, where we induce non-stationarity by oscillating the body parameters between two known configurations. Each episode consists of a day (1440 timesteps, where each timestep corresponds to a minute) in an *in-silico* patient's life and the transition dynamics of a patient's body for each second is governed by a continuous time ordinary differential equation (ODE) (Man et al., 2014). After each minute the insulin controller is used to inject the desired amount of insulin for controlling the blood glucose.

For controlling the insulin injection, we use a parameterized policy based on the amount of insulin that a person with diabetes is instructed to inject prior to eating a meal (Bastani, 2014):

$$\text{injection} = \frac{\text{current blood glucose} - \text{target blood glucose}}{CF} + \frac{\text{meal size}}{CR},$$

where 'current blood glucose' is the estimate of the person's current blood-glucose level, 'target blood glucose' is the desired blood glucose, 'meal size' is the estimate of the size of the meal the patient is about to eat, and CR and CF are two real-valued parameters, that must be tuned based on the body parameters to make the treatment effective.

Non-stationary Recommender System: During online recommendation of movies, tutorials, advertisements and other products, a recommender system needs to interact and personalize for each user. However, interests of a user for different items, among the products that can be recommended, fluctuate over time. For example, interests during online shopping can vary based on seasonality or other unknown factors.

This environment models the desired recommender system setting where reward (interest of the user) associated with each item changes over time. Figure 7 (left) shows how the reward associated with each item changes over time, for each of the considered 'speeds' of non-stationarity. The goal for the reinforcement learning agent is to maximize revenue by recommending the item which the user is most interested in at any time.

Non-stationary Goal Reacher: For an autonomous robot dealing with tasks in the open-world, it is natural for the problem specification to change over time. An ideal system should quickly adapt to the changes and still complete the task.

To model the above setting, this environment considers a task of reaching a non-stationary goal position. That is, the location of the goal position keeps slowly moving around with time. The goal of the reinforcement learning agent is to control the four (left, right, up, and down) actions to move the agent towards the goal as quickly as possible given the real valued Cartesian coordinates of the agent's current location. The maximum time given to the agent to reach the goal is 15 steps.

D.2. Hyper-parameters

For both the variants of the proposed *Prognosticator* algorithms, we use the Fourier basis to encode the time index while performing (ordinary/weighted) least squares estimation. Since the Fourier basis requires inputs to be normalized with $|x| \leq 1$, we normalize each time index by dividing it by $K + \delta$, where K is the current time and δ is the maximum time into the future that we will forecast for. Further, as we are regressing only on time (which are all positive values), is does not matter whether the function for the policy performance over time is odd ($\Psi(x) = -\Psi(-x)$) or not. Therefore, we drop all the terms in the basis corresponding to $\sin(\cdot)$, which are useful for modeling odd functions. This reduces the number of parameters to be estimated by half. Finally, instead of letting $n \in \mathbb{N}$, we restrict it to a finite set $\{1, ..., d-1\}$, where d is a fixed constant that determines the size of the feature vector for each input. In all our experiments, d was a hyper-parameter chosen from $\{3, 5, 7\}$.

Other hyper-parameter ranges were common for all the algorithms. The discounting factor γ was kept fixed to 0.99 and learning rate η was chosen from the range $[5 \times 10^{-5}, 5 \times 10^{-2}]$. The entropy regularizer λ was chosen from the range $[0, 1 \times 10^{-2}]$. The batch size δ was chosen from the set $\{1, 3, 5\}$. Inner optimization over past data for the proposed methods and FTRL-PG was run for $\{10, 20, 30\} \times \delta$ iterations. Inner optimization for ONPG corresponds to one iteration over all the trajectories collected in the current batch. Past algorithms have shown that clipping the importance weights can

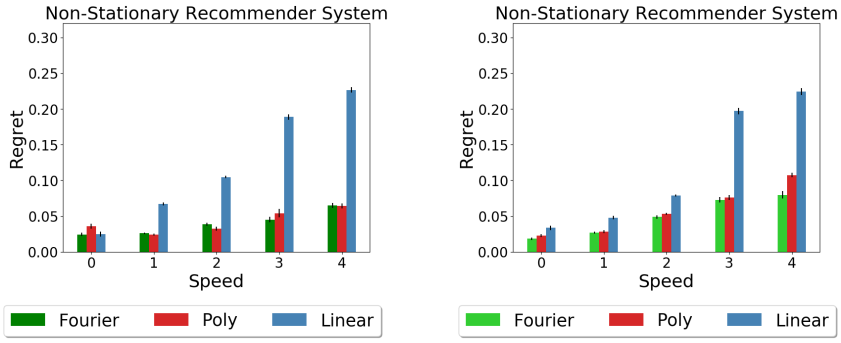

Figure 6. Best performances of all the algorithms for the non-stationary recommender system environment, obtained by conducting a hyper-parameter sweep over 1000 hyper-parameter combinations per algorithm. For each hyper-parameter setting, 30 trials were executed. Error bars correspond to the standard error. (Left) Performance of Pro-OLS with Fourier, polynomial, and linear basis functions. (Right) Performance of Pro-WLS with Fourier, polynomial, and linear basis functions.

improve stability of reinforcement learning algorithms (Schulman et al., 2017). Similarly, we clip the maximum value of the importance ratio to a value chosen from $\{5, 10, 15\}$. As the non-stationary diabetes treatment problem has a continuous action space, the policy was parameterized with a Gaussian distribution having a variance chosen from $[0.5, 2.5]$. For the non-stationary goal-reacher environment, the policy was parameterized using a two-layer neural network with number of hidden nodes chosen from $\{16, 32, 64\}$.

In total, 2000 settings for each algorithm, for each domain, were uniformly sampled (loguniformly for learning rates and λ) from the mentioned hyper-parameter ranges/sets. Results from the best performing settings are reported in all plots. Each hyper-parameter setting was run using 10 seeds for the non-stationary diabetes treatment (as it was time intensive to run a continuous time ODE for each step) and 30 seeds for the other two environments to get the standard error of the mean performances. The authors had shared access to a computing cluster, consisting of 50 compute nodes with 28 cores each, which was used to run all the experiments.[5]

E. Detailed Empirical Results

Complexity analysis (space, time, and sample size) The space requirement for our algorithms and FTRL-PG is linear in the number of episodes seen in the past, whereas it is constant for ONPG as it discards all the past data. The computational cost of our algorithm is also similar to FTRL-PG as the only additional cost is that of differentiating through least-squares estimators which involves computing $(\Phi^\top \Phi)^{-1}$ or $(\Phi^\top \Lambda \Phi)^{-1}$. This additional overhead is negligible as these matrices are of the size $d \times d$, where d is the size of the feature vector for time index and $d << N$, where N is the number of past episodes. Figures 5, 7, and 8 present an empirical estimate for the sample efficiency.

Ablation study: In Figure 6, we show the impact of basis function, $\phi(\cdot)$, on the performance of both of our proposed algorithms: Pro-OLS and Pro-WLS. Dimension d for both the Fourier basis and the polynomial basis was chosen from $\{3, 5, 7\}$. All other hyper-parameters were searched as described in Section D.2. It can be seen that both the Fourier and polynomial basis functions provide sufficient flexibility for modeling the trend, whereas linear basis offers limited flexibility and results in poor performance.

Performance over time: In Figure 5, summary statistics of the results were presented. In this section we present all the results in detail. Figure 7 shows the performances of all the algorithms for individual episodes as the user interests changes over time in the recommender system environment. In this environment, as the true reward for each of the items is directly available, we provide a visual plot for it as well in Figure 7 (left). Notice that the shape of the performance curve for the proposed methods closely resembles the trend of the maximum reward attainable across time.

Figure 8 shows the performances of all the algorithms for the non-stationary goal-reacher and the diabetes treatment environments. In these environments, the maximum achievable performance for each episode is not readily available.

[5] Code for our algorithm can be accessed using the following link: https://github.com/yashchandak/OptFuture_NSMDP.

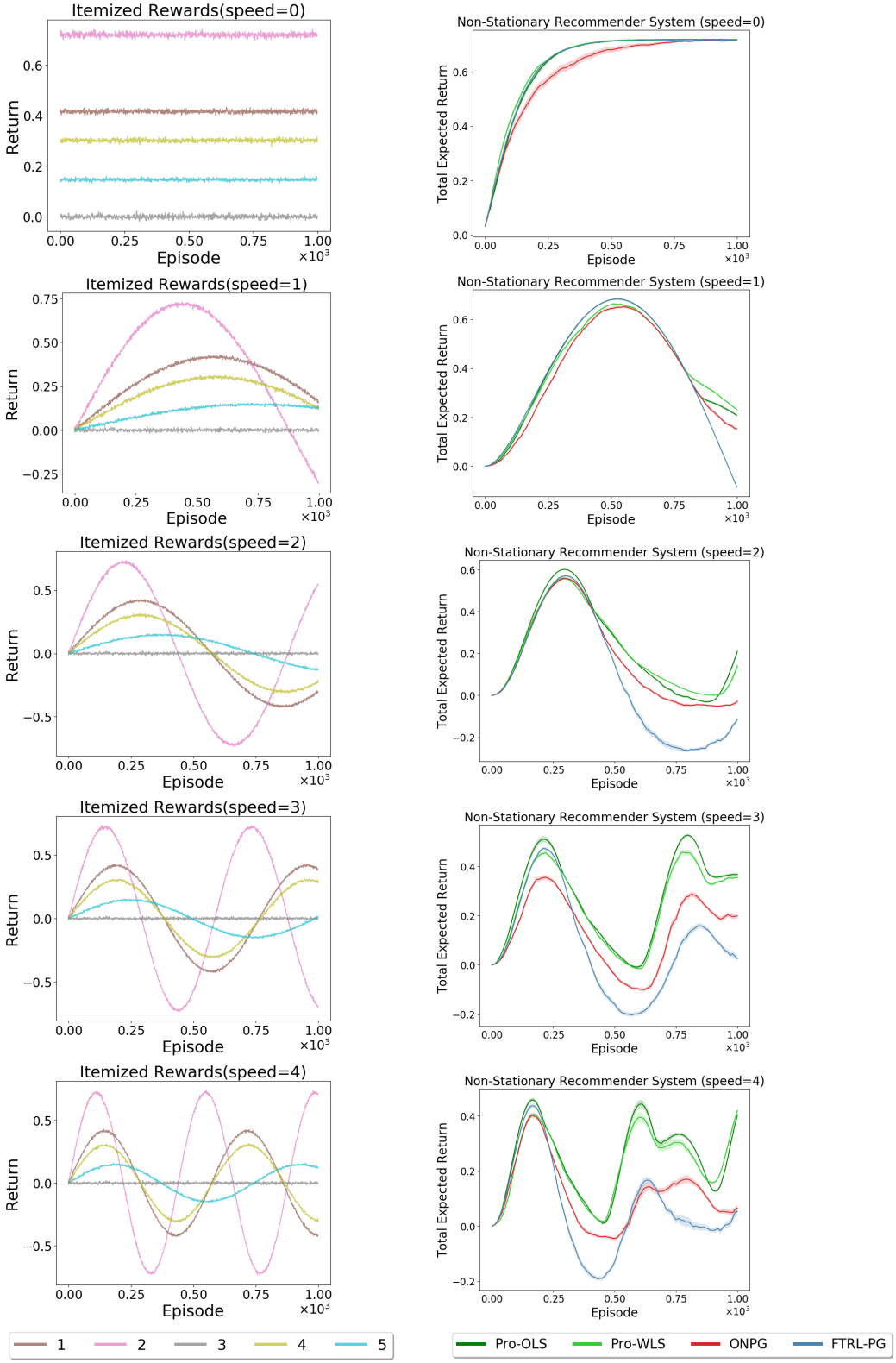

Figure 7. (Left) Fluctuations in the reward associated with each of the 5 items that can be recommended, for different speeds. (Right) Running mean of the best performance of all the algorithms for different speeds; higher total expected return is better. Shaded regions correspond to the standard error of the mean obtained using 30 trials. Notice the shape of the performance curve for the proposed methods, which closely captures the trend of maximum reward attainable over time.

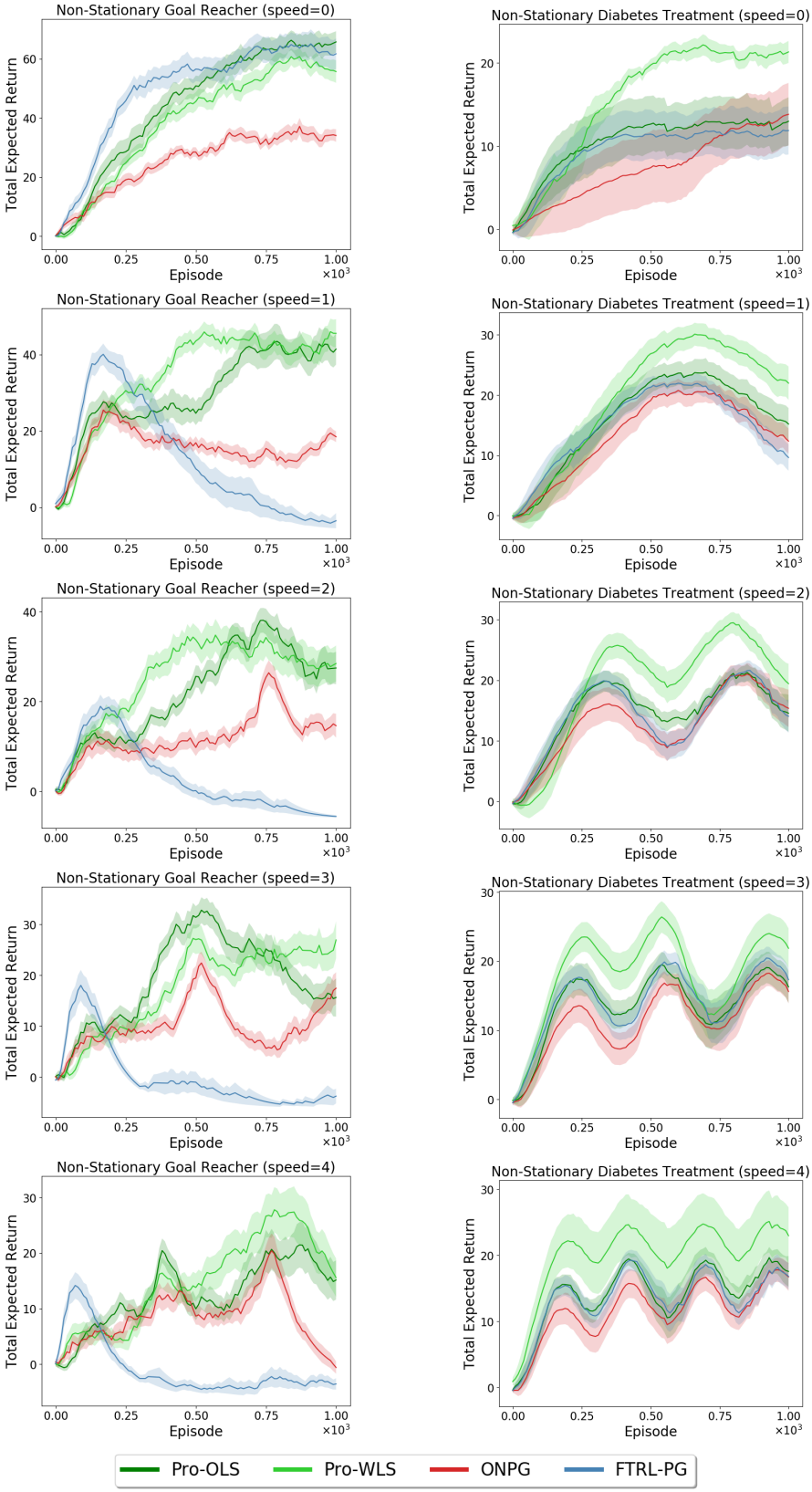

Figure 8. Running mean of the best performance of all the algorithms for different speeds; higher total expected return is better. Shaded regions correspond to the standard error of the mean obtained using 30 trials for NS Goal Reacher and 10 trials for NS Diabetes Treatment.