iPDP: On Partial Dependence Plots in Dynamic Modeling Scenarios

Maximilian Muschalik[1,2,*], Fabian Fumagalli[3,*], Rohit Jagtani[1], Barbara Hammer[3], and Eyke Hüllermeier[1,2]

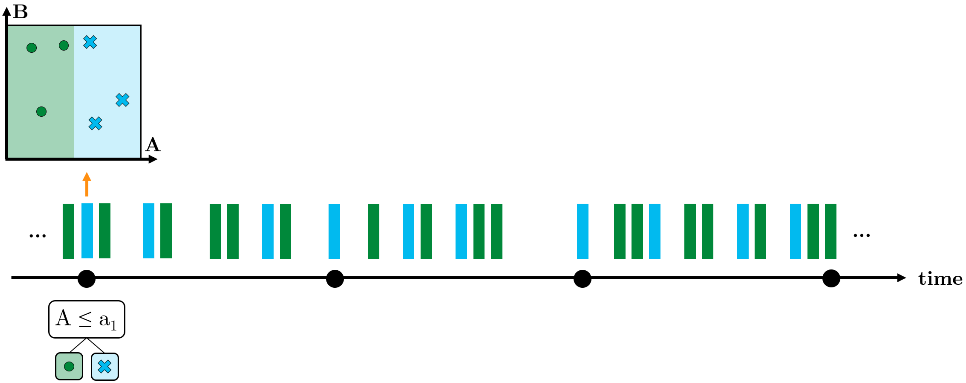
✉ maximilian.muschalik@lmu.de

✉ ffumagalli@techfak.uni-bielefeld.de

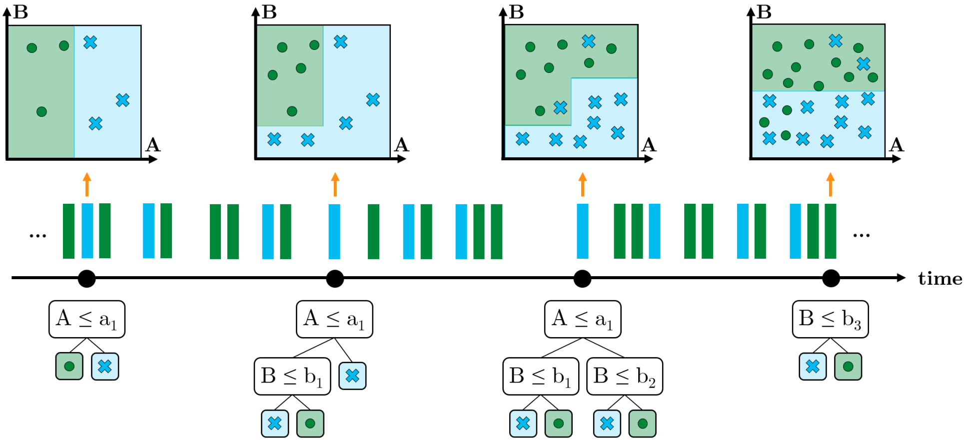
[1] LMU Munich, [2] MCML Munich, [3] Bielefeld University

* denotes equal contribution

Online Models are Learning Incrementally from Data Streams

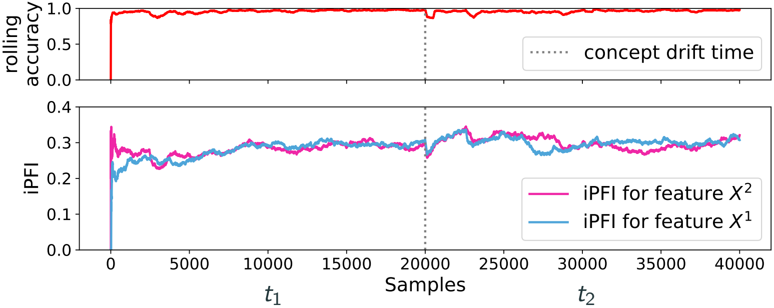


Online Models are Learning Incrementally from Data Streams



Various applications: Bifet and Gavaldà 2007, Gama et al. 2014, Davari et al. 2021, etc.

Still Changes Remain Unnoticed Despite Incremental XAI ...



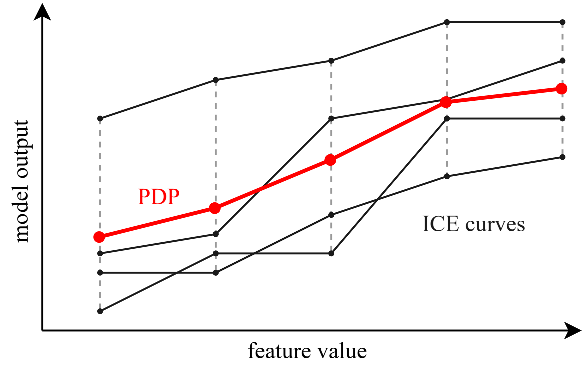
hidden concept drift

$$P_{t_1}(Y|X) \neq P_{t_2}(Y|X)$$

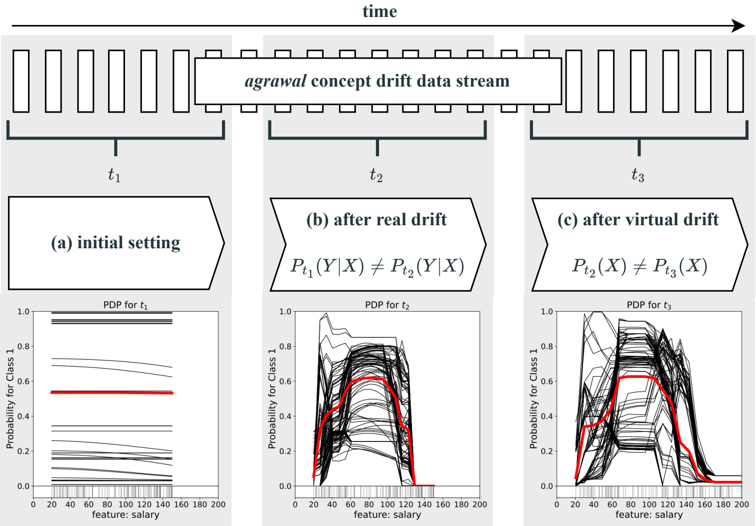
Partial Dependence Plots (PDPs) Explain Feature Effects

Definition of PDP (Friedman 2001)

$$f_S^{\mathrm{PD}}(\mathbf{x}^S) = \mathbb{E}_{X^{\bar{S}}}\left[f(\mathbf{x}^S, X^{\bar{S}})\right] \qquad \text{in practice:} \quad \hat{f}_S^{\mathrm{PD}}(\mathbf{x}^S) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}^S, \mathbf{x}_i^{\bar{S}})$$



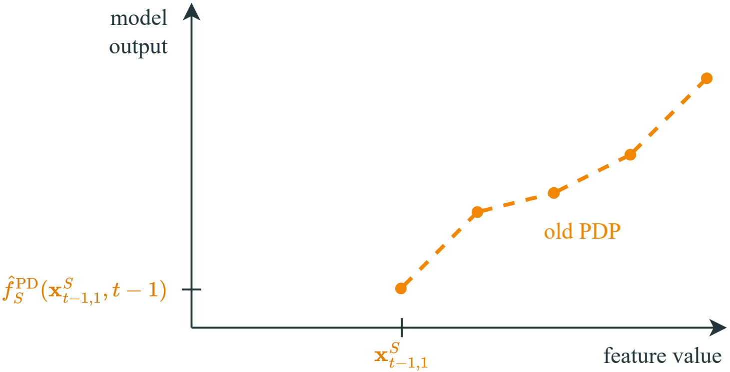
PDP on Virtual and Real Concept Drift



Incremental PDP (iPDP) for Moving Models and Data

Definition of iPDP

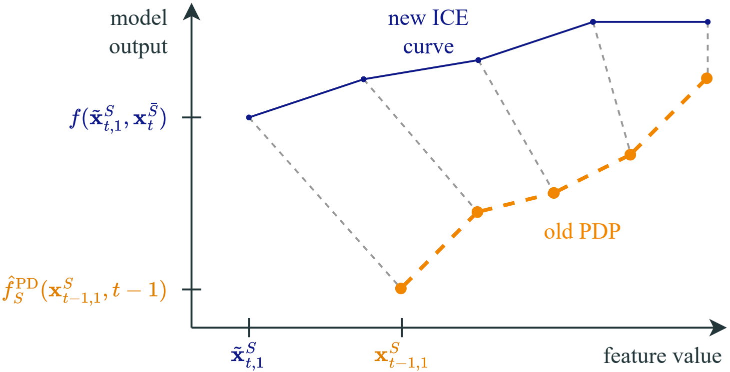
iPDP: $\overbrace{\hat{f}_S^{\text{PD}}(\mathbf{x}_{t-1,k}^S, t-1)}^{\text{old PDP}}$



Incremental PDP (iPDP) for Moving Models and Data

Definition of iPDP

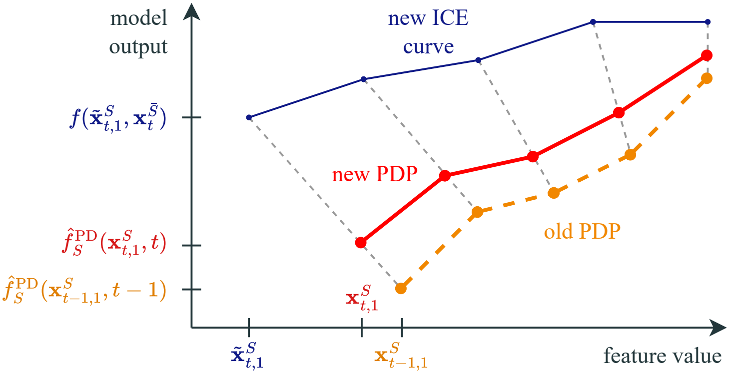
iPDP: $\underbrace{\hat{f}_S^{\text{PD}}(\mathbf{x}_{t-1,k}^S, t-1)}_{\text{old PDP}} \quad \underbrace{f_t(\tilde{\mathbf{x}}_{t,k}^S, \mathbf{x}_t^{\bar{S}})}_{\text{new ICE}}$



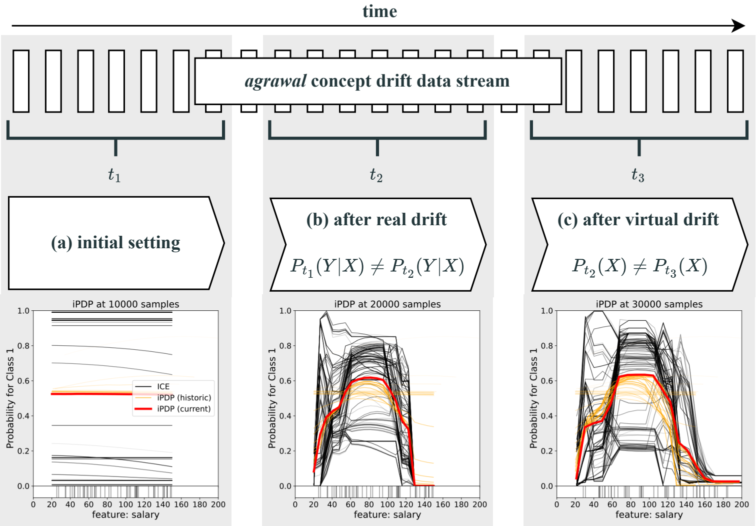
Incremental PDP (iPDP) for Moving Models and Data

Definition of iPDP

$$\textbf{iPDP:}\quad \overbrace{\hat{f}_S^{\text{PD}}(\mathbf{x}_{t,k}^S, t)}^{\text{new PDP}} := (1-\alpha)\cdot \overbrace{\hat{f}_S^{\text{PD}}(\mathbf{x}_{t-1,k}^S, t-1)}^{\text{old PDP}} + \alpha \cdot \overbrace{f_t(\tilde{\mathbf{x}}_{t,k}^S, \mathbf{x}_t^{\bar{S}})}^{\text{new ICE}}$$



iPDP on Virtual and Real Concept Drift



Theoretical Guarantees

Theorem (Reactiveness)

iPDP reacts to real drift and favors recent PD values, as

$$\mathbb{E}[\hat{f}_S^{PD}(\mathbf{x}_{t,k}^S, t)] = \alpha \sum_{i=1}^{t} (1-\alpha)^{t-i} \underbrace{\mathbb{E}_{X_i^{\bar{S}}}\left[f_i(\tilde{\mathbf{x}}_{i,k}^S, X_i^{\bar{S}})\right]}_{\text{PD function at time } i}, \text{ for } k = 1, \ldots, m.$$

Theoretical Guarantees

Theorem (Reactiveness)

iPDP reacts to real drift and favors recent PD values, as

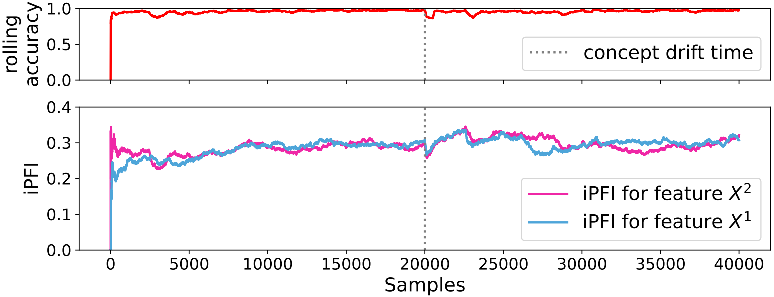
$$\mathbb{E}[\hat{f}_S^{PD}(\mathbf{x}_{t,k}^S, t)] = \alpha \sum_{i=1}^{t} (1-\alpha)^{t-i} \underbrace{\mathbb{E}_{X_i^{\bar{S}}}\left[f_i(\tilde{\mathbf{x}}_{i,k}^S, X_i^{\bar{S}})\right]}_{\text{PD function at time } i}, \text{ for } k = 1, \ldots, m.$$

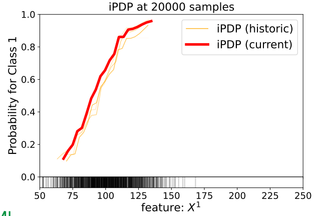
Theorem (Batch PDP Approximation in Static Settings)

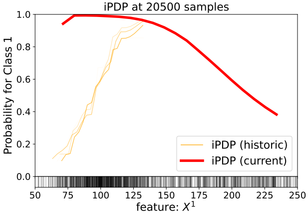
Let observations $(x_0, y_0), \ldots, (x_t, y_t)$ *be iid from* $\mathbb{P}(X, Y)$ *and* $f \equiv f_t$ *be a static model. If* f *is locally linear in the range of temporary model evaluation points* $\{\tilde{\mathbf{x}}_{i,k}^S\}_{i=1}^t$ *for* $k = 1, \ldots, m$, *then*

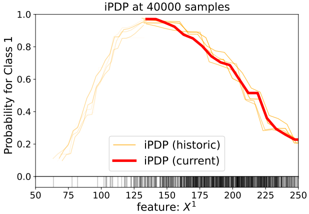
$$\mathbb{E}\left[\hat{f}_S^{PD}(\mathbf{x}_{t,k}^S, t)\right] = f_S^{PD}\left(\mathbf{x}_{t,k}^S\right) \text{ and } \mathbb{E}\left[\frac{\hat{f}_S^{PD}(\mathbf{x}_{t,k}^S, t)}{1-(1-\alpha)^t}\right] = f_S^{PD}\left(\frac{\mathbf{x}_{t,k}^S}{1-(1-\alpha)^t}\right).$$

Experiment A - Synthetic Data

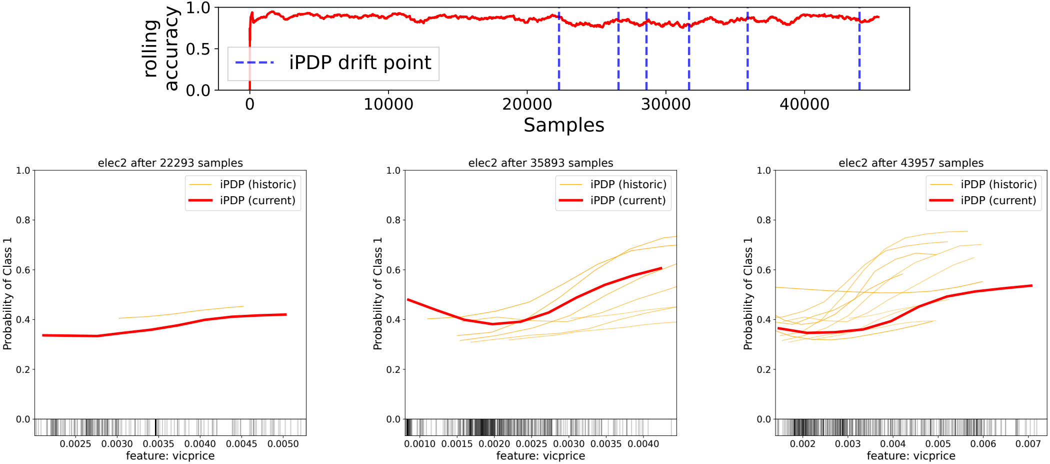








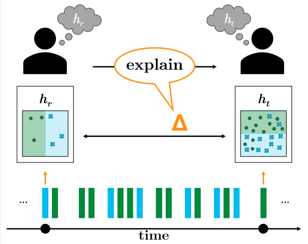
Experiment B - Concept Drift Detection



The Road Ahead and Open Source Implementation

Towards Explaining Change.

- iPDP is a **model-agnostic** XAI method to capture feature effects of models **in flux**.
- iSAGE and iPFI can be used to compute global feature importance incrementally.



iXAI

docs passing | pypi v0.1.3 | status alpha | License MIT

Installation

```
pip install ixai
```

Quickstart

```
>>> for (n, (x, y)) in enumerate(stream, start=1)
...     accuracy.update(y, model.predict_one(x))   # inference
...     incremental_pfi.explain_one(x, y)          # explaining
...     model.learn_one(x, y)                      # learning
```

References

- Bifet, Albert and Ricard Gavaldà (2007). "Learning from Time-Changing Data with Adaptive Windowing". In: *Proceedings of the Seventh SIAM International Conference on Data Mining (SIAM 2007)*, pp. 443–448. DOI: 10.1137/1.9781611972771.42.
- Davari, Narjes et al. (2021). "Predictive Maintenance Based on Anomaly Detection Using Deep Learning for Air Production Unit in the Railway Industry". In: *8th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2021)*. IEEE, pp. 1–10. DOI: 10.1109/DSAA53316.2021.9564181.
- Friedman, Jerome H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine". In: *The Annals of Statistics* 29.5, pp. 1189–1232. ISSN: 00905364. URL: http://www.jstor.org/stable/2699986.

References

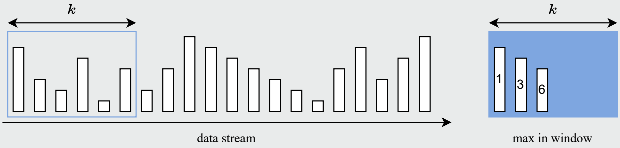
Fumagalli, Fabian et al. (2022). "Incremental Permutation Feature Importance (iPFI): Towards Online Explanations on Data Streams". In: *CoRR* abs/2209.01939. DOI: 10.48550/arXiv.2209.01939. arXiv: 2209.01939. URL: https://doi.org/10.48550/arXiv.2209.01939.

Gama, João et al. (2014). "A Survey on Concept Drift Adaptation". In: *ACM Comput. Surv.* 46.4, 44:1–44:37. DOI: 10.1145/2523813.

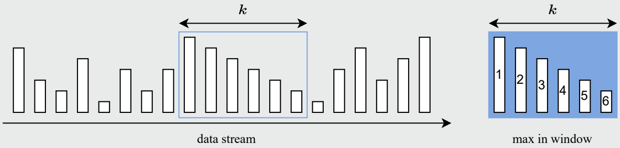
Muschalik, Maximilian et al. (2023). "iSAGE: An Incremental Version of SAGE for Online Explanation on Data Streams". In: *CoRR* abs/2303.01181. arXiv: 2303.01181. URL: https://doi.org/10.48550/arXiv.2303.01181.

Efficient Access to Feature Distribution over Time

Maximum Value Storage



(a) expected case with iid data

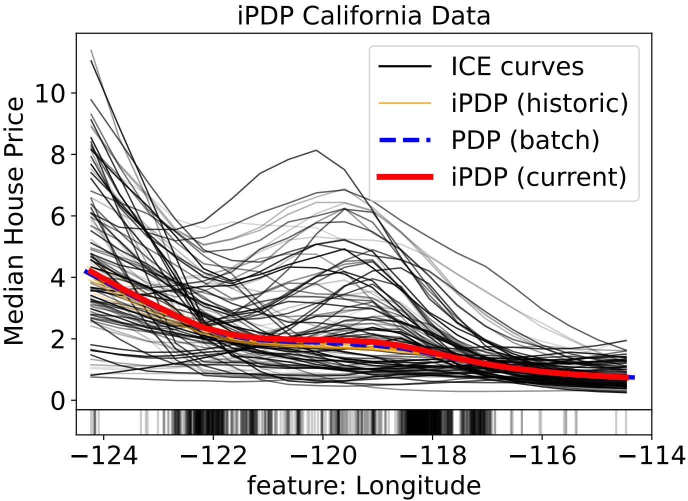


(b) worst case scenario

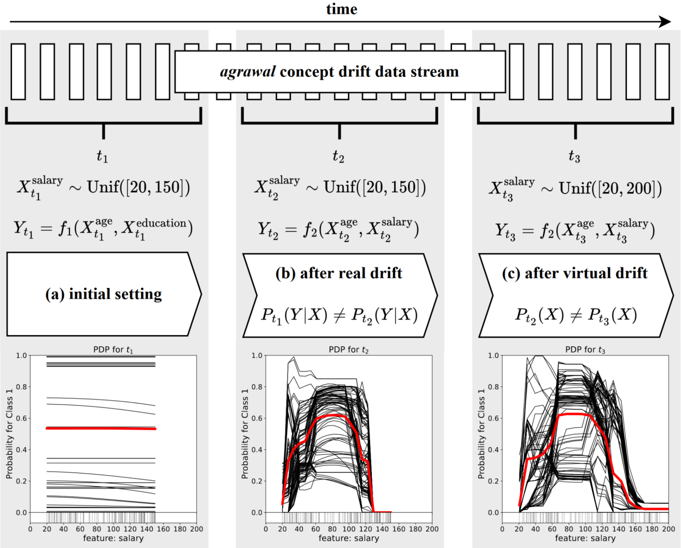
Removal Stratgies

- **Interventional** removal (or categorical features) can be stored in **Geometric Reservoirs** (Fumagalli et al. 2022)
- **Observational** removal can be stored in **Incremental Subgroups** (Muschalik et al. 2023)

Experiment C - Static Model and Data



iPDP on Virtual and Real Concept Drift



iPDP Algorithm

Algorithm 1 iPDP Explanation Procedure

Require: stream $\{\mathbf{x}_t, y_t\}_{t=1}^{\infty}$, model $f_t(.)$, feature set of interest S, smoothing parameter $0 < \alpha \leq 1$, number of grid points m, and storage object R_t

initialize $\hat{f}_S^{\mathrm{PD}}(\mathbf{x}_{0,k}^S, 1) \leftarrow 0$
for all $(\mathbf{x}_t, y_t) \in$ stream **do**
  $\{\tilde{\mathbf{x}}_{t,k}^S\}_{k=1}^m \leftarrow \textsc{GetGridPoints}(R_t, m)$ {e.g., equidistant points, quantiles, etc.}
  for $k = 1, \ldots, m$ **do**
    $\mathbf{x}_{t,k}^S \leftarrow (1-\alpha) \cdot \mathbf{x}_{t-1,k}^S + \alpha \cdot \tilde{\mathbf{x}}_{t,k}^S$ {update grid point}
    $\hat{y}_k \leftarrow f_t\left(\tilde{\mathbf{x}}_{t,k}^S, \mathbf{x}_t^{\bar{S}}\right)$ {evaluate on model evaluation point}
    $\hat{f}_S^{\mathrm{PD}}(\mathbf{x}_{t,k}^S, t) \leftarrow (1-\alpha) \cdot \hat{f}_S^{\mathrm{PD}}(\mathbf{x}_{t-1,k}^S, t-1) + \alpha \cdot \hat{y}_k$ {update point-wise estimates}
  end for
  $R_t \leftarrow \textsc{UpdateStorage}(R_{t-1}, x_t^S)$ {add x_t^S to the storage object}
  Output: $\frac{\hat{f}_S^{\mathrm{PD}}(\mathbf{x}_{t,k}^S, t)}{1-(1-\alpha)^t}, \frac{\mathbf{x}_{t,k}^S}{1-(1-\alpha)^t}$ {debiasing of estimates and grid points}
end for