Efficient learning of nonlinear prediction models with time-series privileged information

Bastian Jung  
Chalmers University of Technology  
mail@bastianjung.com

Fredrik D. Johansson  
Chalmers University of Technology  
fredrik.johansson@chalmers.se

Abstract

In domains where sample sizes are limited, efficient learning algorithms are critical. Learning using privileged information (LuPI) offers increased sample efficiency by allowing prediction models access to auxiliary information at training time which is unavailable when the models are used. In recent work, it was shown that for prediction in linear-Gaussian dynamical systems, a LuPI learner with access to intermediate time series data is never worse and often better in expectation than any unbiased classical learner. We provide new insights into this analysis and generalize it to nonlinear prediction tasks in latent dynamical systems, extending theoretical guarantees to the case where the map connecting latent variables and observations is known up to a linear transform. In addition, we propose algorithms based on random features and representation learning for the case when this map is unknown. A suite of empirical results confirm theoretical findings and show the potential of using privileged time-series information in nonlinear prediction.

1 Introduction

In data-poor domains, making the best use of all available information is central to efficient and effective machine learning. Despite this, supervised learning is often applied in such a way that informative data is ignored. A good example of this is learning to predict the condition of a patient at a set follow-up time based on information of the first medical examination. Classical supervised learning makes use only of the initial data to predict the disease status at the follow-up, although in many cases data about medications, laboratory tests or vital signs are routinely collected about patients also at intermediate time points. This information is privileged (Vapnik and Vashist, 2009), as it is unavailable at the time of prediction, but can be used for training a model.

Learning using privileged information (LuPI) (Vapnik and Vashist, 2009), generalized distillation (Lopez-Paz et al., 2016) and multi-view learning (Rosenberg and Bartlett, 2007) have been proposed to increase the sample efficiency by leveraging privileged information in learning. Theoretical results guarantee improved learning rates (Pechyony and Vapnik, 2010; Vapnik et al., 2015) or tighter generalization bounds (Wang, 2019) for large sample sizes under appropriate assumptions. However, privileged information is not always beneficial; it must be related to the task at hand (Jonaschowski et al., 2015). Previous works do not identify such settings. Moreover, existing analyses do not state when learning with privileged information is preferable to classical learning for problems with small sample sizes—which is where efficiency is needed the most.

Karlsson et al. (2022) studied LuPI in the context of predicting an outcome observed at the end of a time series based on variables collected at the first time step. They showed that making use of data from intermediate time steps in particular settings always leads to lower or equal prediction risk—for any sample size—compared to the best unbiased model which does not make use of this privileged information. However, their method called learning using privileged time series (LuPTS) was limited to settings where the outcome function, and estimators of it, are linear functions of baseline features.
We study the regression setting with univariate outcomes, \(L\) hypothesizes \(h\) of all outcomes observed in \(D\) with respect to a loss \(L : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}\) over training sets of fixed size \(m\) drawn from \(p\),

\[
\overline{R}_p(\mathcal{A}) := \mathbb{E}_D[ R_p(\hat{h}) ] \quad \text{where} \quad \hat{h} = \mathcal{A}(D) \quad \text{and} \quad R_p(\hat{h}) := \mathbb{E}_p[ L(\hat{h}(X_1), Y) ] .
\]  

We study the regression setting with \(L\) the squared error, \(L(y, y') = \| y - y' \|^2_2\). In analysis, we focus on univariate outcomes, \(q = 1\), but all results extend to multivariate outcomes, \(q > 1\).
We analyze and compare learners from the privileged and classical paradigms which produce estimates of the form $\hat{h}(x_1) = \hat{\theta}^T \Phi(x_1)$, as motivated by Assumption 1. We let each use representations $\Phi: \mathcal{X} \to \mathcal{Z} \subseteq \mathbb{R}^d$ from a family $\mathcal{F}$ shared by both paradigms, so that the hypothesis class $\mathcal{H} \ni \hat{h}$ is shared as well.

3 Efficient learning with time-series privileged information

In this section, we assume that privileged and classical learners have access to a common, fixed representation function $\Phi$ which is related to the true representation function through a linear transform, meaning there exists a matrix $B$ such that $\hat{\Phi}(\cdot) = B\Phi(\cdot)$.
We now state a result which says that under Assumption 1, for an appropriate fixed representation

\[ \mathcal{A}(D) = \hat{h}_c(\cdot) := \hat{\theta}_c^\top \Phi(\cdot) \quad \text{where} \quad \hat{\theta}_c := (Z_1^\top Z_1)^{-1}Z_1^\top Y. \]  

(3)

When \( m \geq d \) and \( \Phi \) is known up to linear transform, \( \hat{h}_c \) is the minimum-risk unbiased estimator of \( h \) which does not use privileged information. To accommodate the underdetermined case where \( m < d \), the matrix inverse is replaced by the Moore-Penrose pseudo inverse \( (\cdot)^\dagger \) [Penrose 1956].

**Generalized LuPTS.** For privileged learning, we first compute \( \hat{Z}_t = [\hat{\Phi}(x_{1,t}), \ldots, \hat{\Phi}(x_{m,t})]^\top \) for \( t = 1, \ldots, T \). Then, we independently fit parameter estimates \( \hat{A}_1, \ldots, \hat{A}_T, \hat{\beta} \) of the dynamical system shown in Figure 1 by minimizing the squared error in single-step predictions. This equates to a series of OLS estimates in \( \hat{Z} \). At test time, baseline variables \( x_1 \) are embedded with \( \Phi \) and the latent dynamical system is simulated for \( T \) time steps to form a prediction \( \hat{h}_p(x_1) = (\hat{A}_1 \cdots \hat{A}_T \hat{\beta})^\top \hat{\Phi}(x_1) \). Putting this together, we arrive at Algorithm I which we call generalized LuPTS. We may apply a simple matrix identity and replace terms \( \hat{Z}\hat{Z}_1^\top \) in Algorithm I by the Gram matrices \( \hat{K}_t = \kappa(X_t, X_t) \) of a reproducing kernel \( \kappa \) with corresponding (implicit) feature map \( \hat{\Phi} \), \( \kappa(x, x') = \langle \hat{\Phi}(x), \hat{\Phi}(x') \rangle \) [Smola and Schölkopf 1998]. This variant allows for learning with unknown \( \Phi \).

We now state a result which says that under Assumption 1 for an appropriate fixed representation \( \hat{\Phi} \) or kernel \( \kappa \), generalized LuPTS is never worse in expectation than the classical learner in (3).

**Theorem 1.** Let \( D \) be a data set of size \( m \) drawn from \( \mu \), consistent with Assumption 2. Assume that the left inverse \( \hat{\Phi} \) is known up to a linear transform, explicitly or through a kernel \( \kappa(x, x') = \langle \hat{\Phi}(x), \hat{\Phi}(x') \rangle \), i.e., there exists a matrix \( B \) with linearly independent columns such that \( \hat{\Phi}(x) = B \hat{\Phi}(x) \) and independent \( \hat{\Phi} \) on \( \mathcal{X} \rightarrow \hat{\mathcal{Z}} \) for all \( x \in \mathcal{X} \). Then, it holds for the privileged learner \( \mathcal{A}_p(D) = \hat{h}_p \) (Algorithm 1) and the classical learner \( \mathcal{A}_c(D) = \hat{h}_c \).

\[ \overline{\mathcal{R}}(\mathcal{A}_p) = \overline{\mathcal{R}}(\mathcal{A}_c) - \mathbb{E}_{\hat{h}_p, X_1}[\text{Var}_D(\hat{h}_c(X_1) \mid \hat{h}_p)]. \]  

(4)

**Proof sketch.** First, we show that the predictions made by generalized LuPTS are invariant to a linear transform \( B \) applied to \( \hat{\Phi}(\cdot) \) during training, see Appendix A. Then we consider two cases: (i) When \( \hat{d} \leq m \) we may re-purpose the proof of Theorem 1 in [Karlsson et al. 2022], (ii) When \( m < \hat{d} \) Proposition below directly implies \( \mathbb{E}_{\hat{h}_p, X_1}[\text{Var}_D(\hat{h}_c(X_1) \mid \hat{h}_p)] = 0 \) and \( \overline{\mathcal{R}}(\mathcal{A}_p) = \overline{\mathcal{R}}(\mathcal{A}_c) \).

Theorem I implies that the privileged learner is at least as sample efficient as the classical one since \( \text{Var}(\cdot) \geq 0 \) and thus \( \overline{\mathcal{R}}(\mathcal{A}_p) \leq \overline{\mathcal{R}}(\mathcal{A}_c) \) for the same number of training samples \( m \). The result is a direct generalization of the main result in [Karlsson et al. 2022]. We also observe that generalized LuPTS and the classical learner coincide under certain conditions when \( m \leq \hat{d} \).

**Proposition 1.** Let \( \hat{\Phi} : \mathcal{X} \rightarrow \hat{\mathcal{Z}} \subseteq \mathbb{R}^d \) be any map with corresponding kernel \( \kappa \). Let \( \hat{K}_t \) be the Gram matrix of \( \kappa \) applied to \( X_t \) and let \( \hat{h}_c, \hat{h}_c \) be classical and privileged (Algorithm I) estimates. Then,

\[ \hat{K}_t \text{ is invertible for all } t \implies \hat{h}_p = \hat{h}_c. \]

\( \hat{K}_t \) is noninvertible whenever \( m > \hat{d} \), assuming linearly independent features.

**Proof sketch.** When the Gram matrices \( \hat{K}_t \) are invertible, the pseudo-inverse coincides with the inverse, and factors \( \hat{K}_t \hat{K}_t^\dagger \) in the LuPTS estimators cancel, making LuPTS and CL equal.

**Remarks.** Theorem I extends the applicability of LuPTS to a) nonlinear prediction through a fixed feature map or b) kernel estimation and c) to the underdetermined case of \( m < \hat{d} \). While previous work is restricted to observed linear systems our result considers the case of latent linear dynamics which only need to be identified up to a linear transform (see Figure 1). Proposition I does not claim that no preferable privileged learner exists; it is a statement only about generalized LuPTS. We may relate the result to the double descent characteristic previously observed for other linear estimators for fixed \( m \) and varying \( \hat{d} \) [Loog et al. 2020]. After a phase transition around \( \hat{d} = m \) LuPTS’s variance reduces for a second time when it becomes equivalent to the classical learner (see Figure 4).
3.2 Random feature maps for unknown representations

When the true $\Phi$ is entirely unknown, as is often the case in practice, using a poor representation $\hat{\Phi}$ may yield biased results for both classical and privileged learners. A common solution in non-linear prediction is to use a universal kernel, such as the Gaussian-RBF kernel. These have dense reproducing-kernel Hilbert spaces which allow approximation of any continuous function. However, universal kernels also have positive-definite and thus invertible Gram matrices (Holmann et al., 2008), which according to Proposition 1 eliminates any gain in sample efficiency of generalized LuPTS.

Instead, we combine our algorithm with an approximation of universal kernels—random feature maps. These methods project inputs $x$ onto $\tilde{d}$ features by a random linear map $W \in \mathbb{R}^{k \times \tilde{d}}$, and a nonlinear element-wise activation function. By choosing $\bar{d} < m$, we can benefit from the function approximation properties of universal kernels (see discussion below) and the variance reduction of generalized LuPTS. Popular random features include random Fourier features (RFF) (Rahimi and Recht 2007) and random ReLU features (RRF) (Sun et al., 2018)

\[ \hat{\Phi}_{\text{RFF}}(x) = \sqrt{2/\tilde{d}} \left[ \cos(\sqrt{2\gamma}W_N^T x + b) \right] \text{ and } \hat{\Phi}_{\text{RRF}}^\gamma(x) = f_+(\gamma W_U^T [x; 1]). \]

where $f_+(z) = \max(0, z)$ is the rectifier (ReLU) function and $\gamma > 0$ is a bandwidth hyperparameter.

For large enough numbers of random features and training samples, any continuous function $h$ can be approximated up to arbitrary precision by a linear map $w$ applied to the random features, e.g., $\hat{h}(x) = w^T \hat{\Phi}_{\text{RFF}}(x)$. Applying the same argument to the step-wise estimators of LuPTS we can justify using random features in Algorithm 1 by the following observation: Under appropriate assumptions, we can construct a privileged learner using random feature maps which is a universally consistent estimator of $h(x_1) = E[Y|x_1]$. The precise construction deviates somewhat from Algorithm 1 but follows the same structure. We refer to Appendix B for a precise statement. As universal consistency describes the asymptotic behaviour in the limit of infinite samples and random features, this offers only limited insight into the benefits of privileged information in small sample settings, where performance will be a bias-variance trade-off.

Variance reduction & bias amplification. Generalized LuPTS is only guaranteed lower variance compared to the classical estimator under Assumption 1 although our empirical results (Section 3) suggest this applies more widely. When $\hat{\Phi}$ is a bad approximation of $\Phi$, generalized LuPTS may amplify bias, increasing with the number of privileged time points, compared to classical learning. We show this theoretically in Appendix C and also empirically in Appendix D. Whether generalized LuPTS is still preferable to classical learning in terms of prediction risk appears to depend on the amount of bias that gets amplified. Our experiments imply the variance reduction mostly dominates when using random features, whereas this is not always the case for linear LuPTS. The phenomenon of bias amplification is familiar from e.g., model-based and model-free reinforcement learning (Kober et al., 2013). As the bias with random features may still be high for small sample settings, we next present privileged representation learning algorithms to trade off bias and variance more efficiently.

3.3 Privileged time-series representation learning

Up until now the representation $\hat{\Phi}$ was considered fixed, either because $\Phi$ was known up to a linear transform (explicitly or implicitly) or because of the use of random feature methods. Generalized LuPTS (Algorithm 1) produces minimizers $\{A_t\}$, $\tilde{\beta}$ of the following objective for fixed $\Phi$.

\[ \mathcal{L}_{\text{SRL}}(\hat{\Phi}, \{A_t\}, \tilde{\beta}) := \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \frac{1}{q} \|A_t \hat{\Phi}(x_{i,t}) - \hat{\Phi}(x_{i,t+1})\|_2^2 + \frac{1}{q} \|\tilde{\beta}^T \hat{\Phi}(x_{i,T}) - y_i\|_2^2. \]  

(5)

Objective (5) and the systems described by Assumption 1 lend themselves to methods which also learn the representation $\hat{\Phi}$ in addition to the latent dynamics $\{A_t, \tilde{\beta}\}$. Next, we present three algorithms which combine the ideas of generalized LuPTS with the expressiveness of deep representation learning. All learners use equivalent encoders to represent $\hat{\Phi}(\cdot)$ and linear layers to model the relations between the latent variables $\{Z_t\}$ and the outcome $Y$. The classical learner predicts the outcome linearly from $\bar{Z}_1$. All architectures under consideration are visualized jointly in Figure 3a.
We compare classical learning to variants of generalized LuPTS (Algorithm 1) and the privileged

make test-time predictions using a linear output layer. The results presented were found to be robust to small changes in training

Assumption 1, the expected outcome is linear in the latent state at time step $t$. We verify our theoretical findings by analyzing the sample efficiency and bias-variance characteristics of the given algorithms; (ii) demonstrate that generalized LuPTS with random features succeeds in settings where linear LuPTS suffers from large bias; (iii) point out that privileged representation learners offer even greater sample efficiency in practice and (iv) study how well these algorithms recover the true latent variables $Z_t$ and how this relates to predictive accuracy.

**4 Experiments**

We compare classical learning to variants of generalized LuPTS (Algorithm 1) and the privileged representation learners of Section 3.3 on two synthetic and two real-world time-series data sets. We (i) verify our theoretical findings by analyzing the sample efficiency and bias-variance characteristics of the given algorithms; (ii) demonstrate that generalized LuPTS with random features succeeds in settings where linear LuPTS suffers from large bias; (iii) point out that privileged representation learners offer even greater sample efficiency in practice and (iv) study how well these algorithms recover the true latent variables $Z_t$ and how this relates to predictive accuracy.

**Experimental setup.** We report the mean coefficient of determination ($R^2$), proportional to the squared-error risk $\mathcal{R}$, for varying sample sizes, sequence lengths and prediction horizons. Experiments are repeated and averaged over different random seeds. In each repetition, a given model performs hyperparameter tuning on the training data using random search and five-fold cross-validation before being re-trained on all training data. The test set size is 1000 samples for synthetic data and 20% of all data available for real-world data sets. We consider six privileged learners of two groups. The first group consists of the representation learners SRL, CRL and GRL. For tabular data, their encoder is a multi-layer perceptron with three hidden layers of 25 neurons each. For the image data they use LeNet-5 [LeCun et al. 1989]. The classical learner (Classic Rep.) uses the same encoder with a linear output layer. The results presented were found to be robust to small changes in training

**SRL.** The first privileged representation learner directly optimizes objective (5), just like generalized LuPTS, but now also fitting the representation $\tilde{\Phi}^t$, parameterized by a neural network. We refer to this model as *stepwise representation learner* (SRL). As we will see in experiments, a drawback of this approach is that representations may favor predicting transitions $\tilde{z}_{i,t} \rightarrow \tilde{z}_{i,t+1}$ with small error, while losing information relevant for the target outcome in the process. At test time, for a new input $x_1$, SRL composes the stepwise dynamics to output $\hat{h}_p(x_1) = \beta^T \tilde{A}_T^T \ldots \tilde{A}_1^T \tilde{\Phi}(x_1)$.

**CRL and GRL.** To make sure that the learned representation $\tilde{\Phi}$ retains information about $Y$, we add linear outcome supervision to the representation $\tilde{\Phi}(X_t)$ at each time step $t$. Recall that, by Assumption 1, the expected outcome is linear in the latent state at any time step. We introduce a hyperparameter $\lambda \in [0, 1]$ to trade off the two types of losses and arrive at the combined representation learner (CRL). With $\alpha = (\tilde{\Phi}, \{\tilde{A}_t\}, \{\tilde{\theta}_t\})$ the entire parameter vector, CRL minimizes the objective

$$L_{\text{CRL}}(\alpha) := \frac{\lambda}{NT} \sum_{i,t} \| \tilde{\theta}_i^T \tilde{\Phi}(x_{i,t}) - y_i \|^2 + \frac{1 - \lambda}{N(T - 1)d} \sum_{i,t} \| \tilde{A}_i^T \tilde{\Phi}(x_{i,t}) - \tilde{\Phi}(x_{i,t+1}) \|^2.$$  

We make test-time predictions using $\hat{h}_p(x_1) = \tilde{\theta}_1^T \tilde{\Phi}(x_1)$. In experiments, we highlight the case $\lambda = 1$ where only outcome supervision is used as *greedy representation learner* (GRL). For a precise definition of the GRL objective, see Appendix D. GRL is related to multi-view learning, in which prediction of the same quantity is made from multiple “views” (cf. time points) [Zhao et al. 2017].

![Figure 3: Representation learning architectures (left) and samples from Clocks-LGS (right).](image-url)
We make similar observations on the We also predict the progression of Alzheimer’s disease (AD) as measured by the outcome of the The main goal of our work is to improve learning efficiency by incorporating privileged time-series The Metro Interstate traffic volume data set (Liang et al., 2016), where we predict the air quality in five Chinese cities (see Appendix F). In addition we tested our algorithms on the outcome of interest is the MMSE score 48 months after the first examination. Privileged information obtained through the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (ADNI, 2004). The anonymized data were obtained through the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (ADNI, 2004). The outcome of interest is the MMSE score 48 months after the first examination. Privileged information are the measurements at 12, 24 and 36 months. In addition we tested our algorithms on the PM$_{2.5}$ data set (Liang et al., 2016), where we predict the air quality in five Chinese cities (see Appendix F).

4.1 Sample efficiency, bias and variance

The main goal of our work is to improve learning efficiency by incorporating privileged time-series information. Across almost all prediction tasks and sample sizes, nonlinear variants of generalized LuPTS outperform their classical counterpart in terms of sample efficiency as can be seen in Figure 4. On the Traffic prediction task, LuPTS ReLU RF outperforms linear LuPTS as the former appears to exhibit less bias. On the synthetic data of Square-Sign, linear models reach their best accuracy quickly, while they are limited by their lack of expressiveness. Generalized LuPTS amplifies this bias, making linear LuPTS worse than OLS. Random feature methods attain higher accuracy here but are generally less sample efficient. Generalized LuPTS combined with random features manages to decrease this gap significantly. On ADNI, we don’t see a benefit of using nonlinear models in general. We make similar observations on the PM$_{2.5}$ air quality data set, see Appendix E.

Data sets. We briefly describe evaluation data sets and refer to Appendix E for further details. First, we create two synthetic data sets in which latent states and outcomes are generated from linear-Gaussian systems as in Assumption 1. To produce the observations $\{X_t\}$ we use a deterministic nonlinear function $\Psi: Z \rightarrow \mathcal{X}$. In the first synthetic data set, which we call Square-Sign, the nonlinear transformation $\Psi: \mathbb{R}^d \rightarrow \mathbb{R}^{2d}$ maps each latent feature $Z_{(t,k)}$ to a two dimensional vector such that

$$X_t := \Psi(Z_{(t)}) = [Z^2_{(t,1)}, \text{sgn}(Z_{(t,1)}), \ldots, Z^2_{(t,d)}, \text{sgn}(Z_{(t,d)})]^{\top}.$$  

The second synthetic data set uses the same latent linear system with $Z_t \in \mathbb{R}^2$ and produces square images (28×28 pixels) as observations. As the images are reminiscent of clocks we refer to this data set as Clocks-LGS. Example sequences of these observations are presented in Figure 3a. The angle, size and fill of the two clock hands encode the value of the corresponding latent variable. The outcome $Y$ is a linear function of $Z_T$ with $q = 1$. For Square-Sign, $q = 3$.

The Metro Interstate traffic volume data set (Traffic) (Hogue, 2012) contains hourly records of the traffic volume on the interstate 94 between Minneapolis and St. Paul, MN. In addition, the data contains weather features and a holiday indication. We predict the traffic volume for a fixed time horizon given the present observations. Privileged information is observed every four hours.

We also predict the progression of Alzheimer’s disease (AD) as measured by the outcome of the Mini Mental State Examination (MMSE) (Galea and Woodward, 2005). The anonymized data were obtained through the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (ADNI, 2004). The outcome of interest is the MMSE score 48 months after the first examination. Privileged information are the measurements at 12, 24 and 36 months. In addition we tested our algorithms on the PM$_{2.5}$ data set (Liang et al., 2016), where we predict the air quality in five Chinese cities (see Appendix F).

![Figure 4: Predictive accuracy of generalized LuPTS on three data sets, over 60 repetitions. The shaded area represents one standard deviation above and below the mean over repetitions.](https://github.com/Healthy-AI/glupts)
The representation learners proposed in Section 3.3 are evaluated on the same data sets and on the image prediction task Clocks-LGS. Example results are displayed in Figure 5. These demonstrate that directly transferring the LuPTS objective to neural networks in the form of SRL results in subpar performance. SRL does not seem to have a strong enough incentive to learn representations which accurately predict the outcome. Generalized distillation for privileged time series as suggested by Hayashi et al. (2019), does not improve upon classical learning in our tasks. On all experiments displayed in Figure 5, CRL and GRL outperform the classical learner. The predictive accuracy of these models is similar on most tasks, as may be explained by the fact that CRL may reduce to GRL when choosing $\lambda = 1$ in objective 6. Noticeably, the general observation of GRL and CRL being more sample efficient than the classic model, neither appears to depend on the neural architecture used for the encoder, nor does the modality of the data play an important role, as the image prediction task Clocks-LGS (Figure 5c) demonstrates. For additional empirical results, we refer to Appendix F.

To analyze the bias and variance characteristics of our algorithms, we can estimate the expected squared prediction bias, $E_{X_1}[(E_D[\sigma(D)] | X_1) - E[Y | X_1])^2]$, by computing $E_{Y | X_1}$ in synthetic DGPs in addition to the variance of the different estimators. Figure 6 depicts bias and variance for all models on the Square-Sign data. On the left panel, all variants of generalized LuPTS exhibit lower variance than classical learning, despite being biased. This holds generally: Across all experiments, we never encountered an example where the use of privileged information has not resulted in lower variance compared to classical learning. On the contrary, the privileged learners suffer higher bias than the comparable classical learners because of the bias compounding over the individual prediction steps as shown in Appendix C. For the random feature variants however, the bias decreases with the number of samples. The representation learners display similar characteristics on the right panel of Figure 6. Learning the transitions between latent variables $Z_t$ appears to be associated with low variance and high bias as demonstrated by the results of SRL. GRL however, which does not model these transitions, exhibits the lowest bias and the largest variance of all privileged learners. As CRL is able to trade off between these two objectives, the estimates for its variance and bias lie in between the corresponding values of the other two privileged learners.
4.2 Latent variable recovery

Under Assumption 1, it is sufficient to identify the representation function \( \Phi \) up to a linear transform \( B \) to have provable gains from privileged information over a classical learner. In synthetic data, we can assess to what extent a representation \( \hat{\Phi} \) with this property has been found. As a proxy for the existence of such a transform, we can compute a measure of correlation between \( \hat{\Phi}(X) \) and \( \Phi(X) \), such as the Canonical Correlation Analysis (CCA) (Hotelling, 1936). Raghu et al. (2017) introduced a modified version called Singular Vector Canonical Correlation Analysis (SVCCA) to compute correlations when dealing with noisy dimensions in neural network representations.

The mean SVCCA coefficients \( \bar{r} \) and predictive accuracy of the representation learners are visualized in Figure 7(a). One notices that GRL and CRL produce higher correlation coefficients than the classical learner while also predicting the outcome on the Clocks-LGS task more accurately. Further comparing the representations learned by privileged and classical models, Figure 7(b) shows a visual example of the improved latent recovery of CRL on the same task. Concluding, the use of privileged time-series information does not only increase the sample efficiency of existing algorithms but can also aid the recovery of latent variables in latent dynamical systems.

5 Related work

Existing analyses for learning with privileged information guarantee improvements in asymptotic sample efficiency under strong assumptions (Vapnik and Vashist, 2009) but are insufficient to establish a clear preference for LuPI learners for a fixed sample size. For example, Pechyony and Vapnik (2010) showed that for a specialized problem construction, empirical risk minimization (ERM) using privileged information can achieve fast learning rates, \( O(1/n) \), while classical (non-privileged) ERM can only achieve slow rates, \( O(1/\sqrt{n}) \). We are not aware of any generalization theory tight enough to establish a lower bound on the risk of a classical learner larger than an upper bound for a PI learner. Our problem is also related to multi-task (representation) learning (Maurer et al., 2016), see especially (5)–(6). However, our goal is different in that only a single task is of interest after learning.

In estimation of causal effects, learning from surrogate outcomes (Prentice, 1989) has been proposed as a way to increase sample efficiency. Surrogates are variables related to the outcome which may be available even when the outcome is not (Athey et al., 2019). We can view these as privileged information. While the problem shares structure with ours, the goal is to compensate for missing outcomes and analytical results give no guarantees for improved efficiency when both surrogates and outcomes are always observed (Chen et al., 2008; Kallus and Mao, 2020). Guo and Perkovic (2022) showed that, in the context of a linear-Gaussian system on a directed acyclic graph, a recursive least-squares estimator is the asymptotically most efficient estimator of causal effects using only the sample covariance. In the linear case on a path graph, their estimator coincides with ours. However, no analysis is provided for the nonlinear case or for fixed sample sizes.
6 Discussion

We have presented learning algorithms for predicting nonlinear outcomes by utilizing time-series privileged information during training. We prove that our estimator is preferable to classical learning when data is generated from a latent-variable dynamical system with partially known components. The proof holds for the case when the latent dynamical system is recovered by the representation function up to a linear transform, assuming that a left inverse of the true observation generating function exists. However, this assumption does not appear to be necessary, as our empirical results demonstrate that privileged learning is preferable to classical learning even when these conditions cannot be guaranteed. Consequently, a more general theoretical result where less is known about the latent system seems attainable. For example, one might consider a case in which only a few independent components of the latent variables are recovered by the representation used by the learning algorithm, while other components are treated as noise.

When the latent dynamical system is entirely unknown, we create practical estimators using random feature embeddings which outperform the corresponding classical learner across experiments. We show that a universally consistent learner can be constructed based on this idea, with slightly different form. As a further alternative, we propose representation learning methods of related form using neural networks and demonstrate the empirical benefits also of this estimator over classical learning. In experiments, we analyze how the gap in risk between privileged and classical learning changes for different prediction horizons, as displayed in Figure 16 in Appendix F. The results suggest that the risk advantage of privileged learners grows with the sequence length of the prediction task, despite the fact that the task becomes more difficult at the same time.

Our work focuses on the setting where data is observed as a time series. This setting is chosen for its simple causal structure given by (latent) linear-Gaussian systems and because it can be motivated from many different applications. However, the ideas presented are not specifically tied to time and also apply in the case when all variables are observed simultaneously as long as the causal structure remains sequential. Moreover, we believe that the theory presented here is not limited to sequential settings and generalizes to other causal structures, in particular directed acyclic graphs that connect the baseline covariates to the outcome. In either case, one might only have access to the baseline variable at test time. For example, the timely collection of data for all covariates at test time might be very expensive or even impossible.

As pointed out before, Theorem 1 requires the recovery of the latent variables up to a linear transformation. Nonlinear independent component analysis (ICA) (Hyvarinen and Morioka, 2016) aims to solve precisely this problem and has been applied to time series via time-contrastive learning. This makes for an interesting connection between learning using privileged information and nonlinear ICA, as the experiments of Section 4.2 suggest that privileged time-series information aids the recovery of latent variables. Other remaining challenges include providing risk guarantees for learning with biased representations (including deep neural networks), with regularized estimators, and for more general data generating processes with weaker structural assumptions. We are hopeful that the utility demonstrated in this work will inspire future research to overcome these limitations.

Acknowledgments and Disclosure of Funding

We would like to thank Anton Matsson and Rickard Karlsson for insightful feedback and the Alzheimer’s Neuroimaging Initiative (ADNI) for collecting and providing the data on Alzheimer’s disease used in this project. The present work was funded in part by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

ADNI, the Alzheimer’s Disease Neuroimaging Initiative (ADNI), 2004. URL http://adni.loni.usc.edu


**Checklist**

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes] The introduction refers to the section where each contribution is made.
   (b) Did you describe the limitations of your work? [Yes] See e.g., the discussion in Section 6 and on the applicability of Theorem 1 in Section 3.1 and 3.2.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We have not identified any potential negative societal impact directly related to our work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Assumption 1 and Theorem 1.
   (b) Did you include complete proofs of all theoretical results? [Yes] Yes, see the Appendix.

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] A URL to a code repository is included in the paper.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4 and Appendix E
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Each result figure includes error regions.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Appendix E.2
   (b) Did you mention the license of the assets? [Yes] See Appendix E.2
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Only original code.
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] ADNI data is anonymized.

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]