Incremental Uncertainty-aware Performance Monitoring with Active Labeling Intervention

Alexander Koebler[*1,2] **Thomas Decker**[*1,3,4] **Ingo Thon**[1] **Volker Tresp**[3,4] **Florian Buettner**[2,5,6]
[1]Siemens AG [2]Goethe University Frankfurt [3]LMU Munich [4]Munich Center for Machine Learning (MCML)
[5]German Cancer Research Center (DKFZ) [6]German Cancer Consortium (DKTK)

Abstract

We study the problem of monitoring machine learning models under gradual distribution shifts, where circumstances change slowly over time, often leading to unnoticed yet significant declines in accuracy. To address this, we propose Incremental Uncertainty-aware Performance Monitoring (IUPM), a novel label-free method that estimates performance changes by modeling gradual shifts using optimal transport. In addition, IUPM quantifies the uncertainty in the performance prediction and introduces an active labeling procedure to restore a reliable estimate under a limited labeling budget. Our experiments show that IUPM outperforms existing performance estimation baselines in various gradual shift scenarios and that its uncertainty awareness guides label acquisition more effectively compared to other strategies.

1 INTRODUCTION

Deployed machine learning models often face the critical challenge of distribution shifts, where the data encountered in production deviates from the data used during training. Many relevant shift scenarios involve changes over time, which are often gradual and continuous (Yao et al., 2022; Xie et al., 2024). These shifts are characterized by the fact that the statistical properties of the data or the environment change progressively rather than abruptly. This property can make gradual shifts more insidious, as they may not be immediately apparent but can still lead to substantial degradation in prediction quality over time (Gama

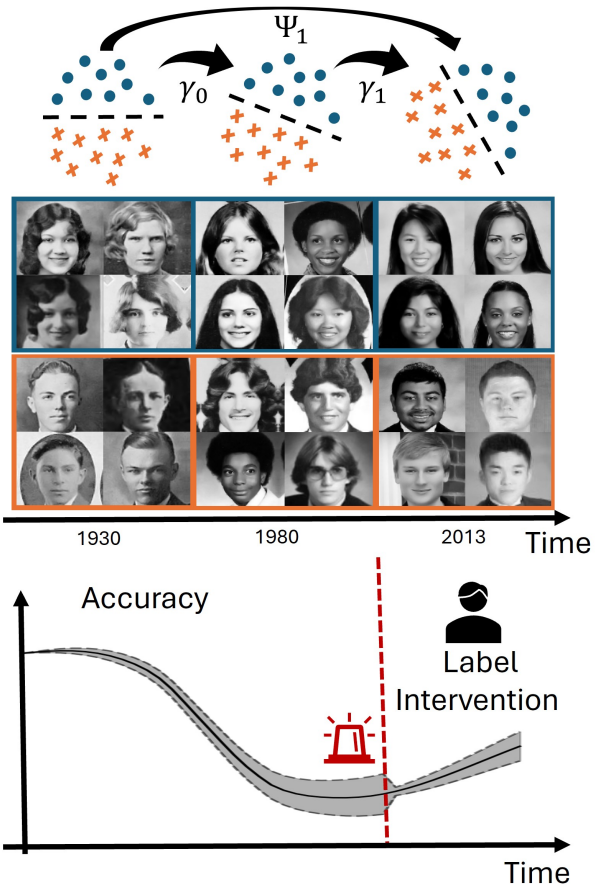

Figure 1: Illustration of Incremental Uncertainty-aware Performance Monitoring (IUPM) to estimate performance changes over time using only labels from the initial training distribution. By iteratively linking unlabeled data points using optimal transport couplings γ and combining them into an overall transition map Ψ, it can anticipate the true model performance under gradual shifts. IUPM also provides an inherent uncertainty measure and an active labeling procedure to efficiently reduce uncertainty and improve estimation reliability under a limited labeling budget.

[*]Equal contribution. Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

et al., 2014). Therefore, anticipating and understanding temporal performance changes is essential for ensuring the reliability and effectiveness of a machine-learning model in dynamic environments. However, directly monitoring the performance during deployment is challenging as labeled data is often unavailable in production. Moreover, obtaining labels can be cumbersome, time-consuming, and costly, leading to delays in the assessment. Therefore, an increasing number of label-free estimation methods have been proposed that aim to anticipate the model performance purely based on unlabeled data available at runtime (Yu et al., 2024). Such methods leverage diverse strategies to approach this task by for instance estimating performance changes based on feature statistics (Deng and Zheng, 2021), the agreement of multiple models (Jiang et al., 2022; Baek et al., 2022) or by analyzing the model's confidence (Guillory et al., 2021; Garg et al., 2022). However, none is explicitly tailored to the particular nature of gradual shifts, which are ubiquitous in practice. On top of that, existing techniques suffer from two fundamental limitations. First, they cannot quantify any uncertainty related to the performance estimate, causing ambiguity about when to trust the estimate and when to rather collect extra labels to obtain a more accurate assessment. Second, current methods do not guide how to best support performance estimation in monitoring situations where a limited number of labels could be acquired. This neglects the fact that actively selecting specific data points of interest to be labeled can greatly improve the effectiveness of model evaluations under a limited labeling budget (Kossen et al., 2021). To address these limitations we make the following contributions:

- We propose **Incremental Uncertainty-aware Performance Monitoring (IUPM)**, as a novel label-free performance estimation method tailored to gradual distribution shifts with an inherent notion of uncertainty.
- We introduce an active intervention step to increase the reliability of the performance estimate by labeling examples contributing the highest uncertainty to the performance estimate.
- We show that IUPM works provably well for a general class of gradual shifts and demonstrate its benefits over existing baselines on different datasets under synthetic and real gradual shifts.

2 BACKGROUND AND RELATED WORK

Label-free performance estimation Label-free performance estimation methods aim to anticipate the predictive quality of a machine learning model in out-of-distribution settings purely based on unlabeled data (Yu et al., 2024). Many existing approaches with theoretical guarantees are typically restricted to certain shifts, such as considering a pure covariate or label shift (Sugiyama et al., 2008; Chen et al., 2021; Garg et al., 2020), or require additional properties of the model's output confidence (Guillory et al., 2021; Lu et al., 2023). On the other hand, a growing number of methods have been introduced that demonstrate promising empirical results for specific model classes and shift types. Such techniques leverage diverse strategies based on feature statistics (Deng and Zheng, 2021), the model's behavior under test-time augmentations (Deng et al., 2021), thresholding the model's confidence (Garg et al., 2022), the agreement between different models (Jiang et al., 2022; Baek et al., 2022), or model differences after retraining (Yu et al., 2022). However, none of them is explicitly tailored to gradual distribution shifts and their inherent structure or comes with an applicable notion of uncertainty, which is the scope of our work.

Gradual Domain adaptation While the special characteristics of gradual shifts have not been considered for performance estimation, they have already been explored in the context of domain adaptation. In general, domain adaptation addresses the problem of increasing the performance of a model trained on a labeled source domain when applied to a novel target domain (Wilson and Cook, 2020). Classical methods typically act in a one-shot fashion, where adaption is performed directly between the two domains. However, this approach can struggle in situations where the encountered distribution shift is particularly strong (He et al., 2024). Alternatively, gradual domain adaptation (Kumar et al., 2020; Wang et al., 2020, 2022) works sequentially by introducing a series of intermediate domains. These domain interpolations serve as stepping stones to reduce the complexity of the overall distribution shift, which can help to increase the adaptation effectiveness (Abnar et al., 2021; Chen and Chao, 2021). When monitoring a model during deployment, however, it is unclear at which point adaptation is truly necessary without having a faithful estimate of the current model performance at runtime. This additionally motivates the need for performance estimation methods that explicitly consider the gradual nature of shifts over time.

Active risk estimation In contrast to label-free performance estimation, the field of active risk estimation (Sawade et al., 2010) is concerned with reducing the number of labels required to yield a reliable estimate of the model's performance under an exist-

ing but limited labeling budget. A variety of methods (Sawade et al., 2010; Kossen et al., 2021, 2022; Lee et al., 2024) have been developed to approach this task analog to active learning by introducing different sampling strategies to query and label test samples. Most prominently, the authors in (Sawade et al., 2010) introduce an importance sampling approach whereas the authors in (Kossen et al., 2021, 2022) utilize a surrogate model using Gaussian process models or Bayesian neural networks to estimate which samples would contribute most to the performance estimation or even use the surrogate to predict the loss of the target model directly (Kossen et al., 2022).

Optimal Transport Optimal Transport (OT) aims at finding the cost-minimizing way to transform one probability measure into another (Peyré et al., 2019). Consider having n_0 samples from a domain $\Omega_0 = \{x_0^i\}_{i=1}^{n_0}$ and n_1 samples from another domain $\Omega_1 = \{x_1^i\}_{i=1}^{n_1}$ with corresponding empirical distributions

$$\hat{p}_0 = \sum_{x_0 \in \Omega_0} \frac{1}{n_0} \delta_{x_0} \qquad \hat{p}_1 = \sum_{x_1 \in \Omega_1} \frac{1}{n_1} \delta_{x_1}$$

where δ_x denotes the Dirac measure. For a cost function $c : \Omega_0 \times \Omega_1 \to \mathbb{R}^+$, the transformation of $\hat{p}_0$ into $\hat{p}_1$ can be formalized by a coupling γ which represents a valid distributions over $(\Omega_0 \times \Omega_1)$ with marginals corresponding to $\hat{p}_0$ and $\hat{p}_1$. Identifying the cost-optimal coupling reads:

$$\hat{\gamma} = \arg\min_{\gamma \in \Gamma} \sum_{x_0 \in \Omega_0} \sum_{x_1 \in \Omega_1} c(x_0, x_1)\gamma(x_0, x_1) \quad \text{with}$$
$$\Gamma = \{\gamma \in \mathbb{R}^{n_0 \times n_1} \mid \gamma \mathbf{1}_{n_1} = \hat{p}_0, \gamma^T \mathbf{1}_{n_0} = \hat{p}_1\}$$

which can be solved using different algorithmic approaches (Peyré et al., 2019). In the discrete sample case the obtained $\gamma \in \mathbb{R}^{n_0 \times n_1}$ simply is a matrix with entries $\gamma(x_0, x_1)$. Moreover, the conditional coupling

$$\gamma(X_0 = x_0 | X_1 = x_1) = \frac{\gamma(x_0, x_1)}{\sum_{x_0 \in \Omega_0} \gamma(x_0, x_1)}$$

is a left-stochastic matrix whose entries can be interpreted as transition probabilities when moving from samples of X_1 to samples of X_0 following the most cost-efficient path. OT couplings can further be used to characterize how similar two distributions are using the corresponding Wasserstein distance:

$$\mathcal{W}_p(P_0, P_1) = \inf_{\gamma \in \Gamma} \mathbb{E}_{(x_0, x_1) \sim \gamma}[c(x_0, x_1)^p]^{1/p}$$

Intuitively, the Wasserstein distance describes the expected transportation cost under the optimal coupling. In the following, we will only consider the case where $p = 1$ such that we denote $\mathcal{W} := \mathcal{W}_1$.

3 INCREMENTAL UNCERTAINTY-AWARE PERFORMANCE MONITORING

Problem Setup Consider a machine learning model $f : \mathcal{X} \to \mathcal{Y}$ that has been trained with labeled data (X_0, Y_0) from the distribution $P_0(X_0, Y_0)$ over the space $(\mathcal{X} \times \mathcal{Y})$. Suppose f is deployed in order to make predictions over time $t > 0$ with respect to data $\{(X_t, Y_t)\}_{t=1}^T$ each distributed with $P_t(X_t, Y_t)$. In this work, we are mainly interested in gradual shifts, which are typically characterized as follows (Kumar et al., 2020; He et al., 2024):

Definition 1. *A distribution shift over $\{(X_t, Y_t)\}_{t=0}^T$ is gradual in time $t = 0, \ldots, T$ if the Wasserstein distance between two consecutive steps is bounded:*

$$\mathcal{W}(P_t(X_t, Y_t), P_{t-1}(X_{t-1}, Y_{t-1})) \leq \Delta_t \quad \forall t = 1, \ldots, T$$

In our setup, we consider a slightly more refined property to characterize gradual shifts:

Definition 2. *A distribution shift over $\{(X_t, Y_t)\}_{t=0}^T$ is called gradually Lipschitz smooth in X_t if there is a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ for which we have $\mathcal{W}(P_t(X_t), P_{t-1}(X_{t-1})) \leq \varepsilon_t$ for all $t = 1, \ldots, T$ and there exist constants $L_t > 0$ such that:*

$$\mathcal{W}\big(P_t(Y_t|X_t), P_{t-1}(Y_{t-1}|X_{t-1})\big) \leq L_t \, c(X_t, X_{t-1})$$

Intuitively this property implies that if two data points in consecutive domains are close in terms of a cost function c, their conditional target distributions should also be similar. Note that this assumption is quite reasonable for gradual shifts in general and one can easily show that any shift that is gradually Lipschitz smooth in X_t is also gradual as defined above:

Proposition 1. *If a distribution shift over $\{(X_t, Y_t)\}_{t=0}^T$ is gradually Lipschitz smooth in X_t with constants L_t, then it is also gradual:*

$$\mathcal{W}(P_t(Y_t, X_t), P_{t-1}(Y_{t-1}, X_{t-1})) \leq (1 + L_t)\, \varepsilon_t := \Delta_t$$

Incremental Performance Estimation Given the setup introduced above, remember that during runtime ($t > 0$) one typically only has access to unlabeled data from X_t and our goal is to estimate how well a model f performs over time solely based on this information. To do so effectively in the context of gradual shifts we propose an incremental approach: Let $\gamma_t(X_{t-1}|X_t)$ be the conditional coupling linking data from X_t to data from X_{t-1} in a cost-efficient way. Further, we define $\Psi_t(X_0|X_t) := \prod_{i=1}^t \gamma_i(X_{i-1}|X_i)$ describing the transition matrix obtained from composing all incremental transition matrices $\gamma_i(X_{i-1}|X_i)$ via

matrix multiplication. It expresses the overall transition probabilities of going back to the labeled data available at $t = 0$ by connecting samples of two subsequent time points incrementally using an individual optimal transport coupling. Based on this we propose the following strategy to estimate missing labels for performance evaluation over time:

$$\hat{P}(Y_t|X_t) = \mathbb{E}_{\Psi_t(X_0|X_t)}\left[P(Y_0|X_0)\right]$$

This means that our label estimate $\hat{P}(Y_t|X_t)$ arises as mixture distribution (Everitt, 2013) combining labeled data in X_0 according to the accumulated incremental coupling results. This strategy is explicitly motivated by temporal shifts as we leverage their gradual nature by modeling subsequent distribution shifts incrementally using Optimal Transport.

Given a loss function $\mathcal{L}$ to measure model performance and a set of samples Ω_t from X_t, the resulting performance estimate for IUPM at time t, denoted by $(\hat{\mathcal{L}}_t^{IUPM})$, is given by:

$$\begin{aligned}\hat{\mathcal{L}}_t^{IUPM} &= \mathbb{E}_{P(X_t)}\mathbb{E}_{\hat{P}(Y_t|X_t)}\left[\mathcal{L}(f(X_t), Y_t)\right] \\ &= \frac{1}{n_t}\sum_{x_t \in \Omega_t} \mathbb{E}_{\hat{P}(Y_t|X_t=x_t)}\left[\mathcal{L}(f(X_t), Y_t)\right]\end{aligned}$$

The following theorem shows that this estimate will be close to the true model performance if the encountered shift is gradually Lipschitz smooth:

Theorem 1. *Let $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function that is 1-Lipschitz in its second argument and denote the true model performance at time t with $\mathcal{L}_t$. If a distribution shift over $\{(X_t, Y_t)\}_{t=1}^T$ is gradually Lipschitz smooth in X_t, then:*

$$|\mathcal{L}_t - \hat{\mathcal{L}}_t^{IUPM}| \leq \sum_{i=1}^{t} L_t \varepsilon_t$$

Note that the worst-case performance estimation error grows only linearly in t matching corresponding bounds on the adaptation error established in the gradual domain adaptation literature (He et al., 2024). This theorem also implies that relying explicitly on the incremental OT couplings to relate unlabeled samples to labeled ones is most effective:

Corollary 1. *If each γ_t is the optimal transport coupling between two consecutive domains X_{t-1} and X_t, then this minimizes the estimation error across all possible incremental couplings.*

As a consequence, IUPM provides a theoretically grounded approach to estimating the performance of machine learning models during deployment facing gradual distribution shifts. The underlying proofs and further mathematical details for all theoretical statements made above are provided in Appendix A.

Uncertainty Estimation and Active Labeling Intervention Another implication of the Theorem above is that the estimation error might grow linearly in time. Therefore it would also be desirable to additionally quantify the uncertainty related to the assessment. In cases of low confidence estimates it is preferable to rather collect true labels under the current data distribution to verify the performance indication. Our approach also provides a simple way to achieve this as IUPM exhibits an intrinsic notion of uncertainty due to the incremental matching procedure. More specifically, the estimate $\hat{P}(Y_t|X_t)$ is an actual predictive distribution that also internalizes uncertainty for cases where linked samples have contradicting labels. Therefore, we can use it to quantify the uncertainty of the anticipated performance using the expected standard deviation (SD) of the sample-wise loss estimates:

$$\mathcal{U}(\hat{\mathcal{L}}_t^{IUPM}) = \mathbb{E}_{P(X_t)}\mathrm{SD}_{\hat{P}(Y_t|X_t)}\left[\mathcal{L}(f(X_t), Y_t)\right]$$

In addition to providing a means for users to consolidate their trust in the performance estimate, the quantified uncertainty $\mathcal{U}(\hat{\mathcal{L}}_t^{IUPM})$ can also be used to automatically trigger efficient relabeling when the uncertainty exceeds a significant threshold. We utilize this property by proposing a novel sampling strategy for active label intervention. With our Uncertainty Intervention (UI) strategy we aim to make the most efficient use of a limited labeling budget allowing us to label only m samples. For this, we query ground truth labels for critical samples x_t that contribute the largest to the uncertainty of the anticipated performance for the current step:

$$\arg\text{top-}m_{x_t \in \Omega_t} \mathcal{U}\left[\mathcal{L}(f(x_t), Y_t)\right]$$

where arg top-m denotes the operator selecting the top m elements maximizing the objective. The new labels are used to update $\Psi_t(X_0|X_t)$ such that $\hat{P}(Y_t|X_t = x_t)$ assigns a fixed label removing the accumulated uncertainty for sample x_t.

4 EXPERIMENTS

In this section, we evaluate our IUPM approach on three different gradual shift scenarios. First, to assess the general functionality of IUPM and to yield insights into our proposed uncertainty intervention sampling strategy, we present results based on synthetic examples with continuous shifts in two-dimensional space. Second, we analyze the capabilities of IUPM to monitor performance changes due to different gradual image perturbations on MNIST (Lecun et al., 1998) and a subset of ImageNet (Howard, 2019). Finally, we apply IUPM to a real-world shift scenario based on yearbook portraits across several decades (Ginosar et al., 2015)

and demonstrate its superiority over several baselines. For this dataset, we also provide guidance on how to validate the underlying theoretical assumptions empirically. The code for IUPM is made available[1], the algorithm is detailed in Appendix B and the setup for each experiment is documented in Appendix C.

4.1 Evaluation Setup

Performance Estimation Baselines Throughout the experiments, we compare our approach to several existing performance estimation methods. Those baselines consist of four confidence-based error estimation methods, as described in (Garg et al., 2022). *Average Confidence (AC)* simply estimates the prediction accuracy as the expectation of the confidence for the predicted class across the data set in step t as:

$$AC_{\Omega_t} = \mathbb{E}_{x\sim\Omega_t}[\max_{j\in\mathcal{Y}} f_j(x)]$$

The more sophisticated *Difference Of Confidence (DOC)* (Guillory et al., 2021) uses the discrepancy between the model confidence on the source and target data sets as an estimate of performance degradation. To obtain an approximation of the performance in step t, the degradation is subtracted from the performance on the initialization data set $t = 0$ according to:

$$DOC_{\Omega_t} = \mathbb{E}_{x,y\sim\Omega_0}[\arg\max_{j\in\mathcal{Y}} f_j(x) \neq y] + \mathbb{E}_{x\sim\Omega_t}[\max_{j\in\mathcal{Y}} f_j(x)] - \mathbb{E}_{x\sim\Omega_0}[\max_{j\in\mathcal{Y}} f_j(x)]$$

Average Threshold Confidence (ATC) (Garg et al., 2022) learns a threshold for model confidence on the initialization data set Ω_0 and estimates accuracy on the current set as the fraction of examples where model confidence exceeds this threshold. Lastly, we consider *Importance re-weighting (IM)*, as proposed by (Chen et al., 2021). We also evaluate the direct mapping $\gamma(X_0|X_t)$ for label transport and performance estimation inspired by (Decker et al., 2024) and call it *Non-Incremental Performance Estimation (NIPM)*. Across all experiments, we consider the model accuracy as the loss criterion $\mathcal{L}$ to evaluate performance.

Sampling Strategies Throughout all active label intervention experiments, we utilize a fixed threshold for our uncertainty indicator of $\mathcal{U}(\mathcal{L}_t) > 0.1$ to trigger an intervention step. We provide an ablation study detailing the rationale for choosing this threshold value in Appendix D. To evaluate the efficacy of our introduced *Uncertainty Intervention (UI)* sampling strategy introduced in Section 3, we compare two baseline methods. First, we draw from the active testing literature and utilize an active sampling strategy introduced by (Kossen et al., 2021), which selects samples based on their expected loss contribution under the performance estimate measured based on the cross-entropy loss:

$$\arg\text{top-}m_{x_t\in\Omega_t} - \sum_y \hat{P}(y_t|X_t = x_t)\log f(x_t)_y$$

We refer to this sampling strategy as *Cross Entropy Intervention (CEI)*. Note that compared to (Kossen et al., 2021), we only trigger the labeling procedure once our uncertainty threshold is exceeded and not for every performance calculation directly using the m generated labels. Lastly, we introduce *Random Intervention (RI)* as a naïve but effective baseline, where we randomly sample m samples from the available set Ω_t at time step t.

4.2 Experimental Results

Translation and Rotation in Input Space In our first experiment, we use three two-dimensional toy data sets (Figure 2) provided by (Pedregosa et al., 2011). For all three datasets, we train a Random Forest (RF), XGBoost (XGB), and a Multilayer Perceptron (MLP) classifier in the initial source distribution. After training at $t = 0$, all data sets are shifted for

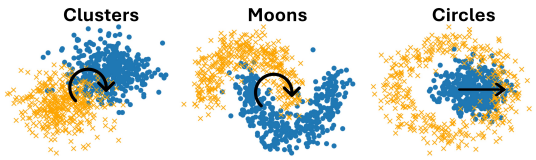

Figure 2: Synthetic two-dimensional toy datasets and corresponding shifts indicated by black arrows.

$t = 1, \ldots, 100$ steps to simulate gradual changes over time. For the "Clusters" and "Moons" data sets, this shift results from rotating both classes by 2° per step. The "Circles" dataset experiences a translation shift by 0.02 in the x-direction only on the inner circle class. By showing the performance estimation over time for monitoring an MLP model on the moons data set facing a gradual rotation, we exemplarily illustrate in Figure 3 that IUPM best estimates the actual performance degradation. While initially, the non-incremental NIPM approach can also approximate the real performance, it deviates strongly as soon as the degree of degradation caused by the shift accelerates. While in early steps, it is still possible to directly match source samples x_0 to target samples x_t sharing the same class, this becomes increasingly difficult as with the rotation, the conditional distributions $P_0(Y_0|X_0)$ and $P_t(Y_t|X_t)$ diverge. The confidence-based approaches fail to sufficiently pick up the gradual performance degradation. These observations are

[1] https://github.com/alexanderkoebler/IUPM

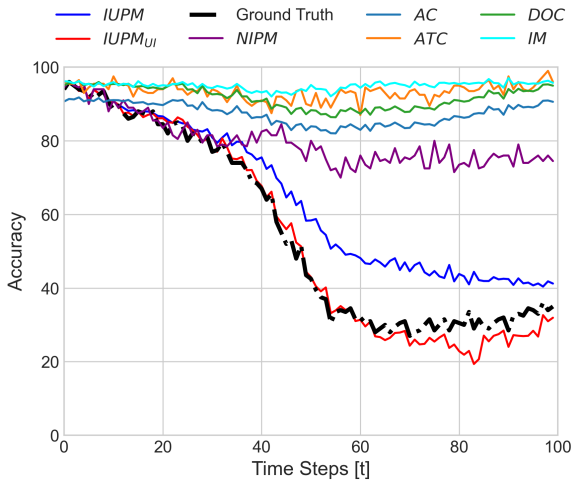

Figure 3: Performance estimation for an MLP model on the synthetic moons data set for a rotational shift over 100 steps resulting in a total rotation of 200°. Our proposed IUPM approach with and without label intervention clearly yields the highest fidelity for the performance estimation.

consistent across three different models and data sets (Table 1). As shown in the previous figure and table, the error of the accuracy estimate by IUPM can further be significantly reduced by introducing active label intervention steps. For this, we compare our proposed UI approach with CEI and RI. In all cases, we relabel 50% of Ω_t when the total uncertainty exceeds a predefined threshold. As described above, this threshold is set to 10% in accuracy deviation. Due to its empirical success, we used the same threshold across all subsequent experiments. As shown in Figure 4, whilst all sampling methods can keep the uncertainty below the predefined threshold, our proposed UI approach requires far fewer intervention steps and, thus, labeling effort. This is underlined by Table 2 showing that UI, in most cases, outperforms the other approaches concerning the error in the performance estimation while consistently requiring far fewer intervention steps. The superiority of UI in terms of intervention efficiency is further confirmed by an ablation study in Appendix D. Figure 5 provides another intuitive illustration of the strength of the proposed sampling strategy. UI selects the most relevant examples that are close to the inherent decision boundary of the dataset.

Monitoring Performance Degradation due to Image Perturbations To make a step towards assessing more complex shifts, we first monitor a model classifying handwritten digits (Lecun et al., 1998) that experience a shift caused by affine transformations, i.e., rotation, used in a related context in (Wang et al., 2020), translation, and scaling of the digits. For this

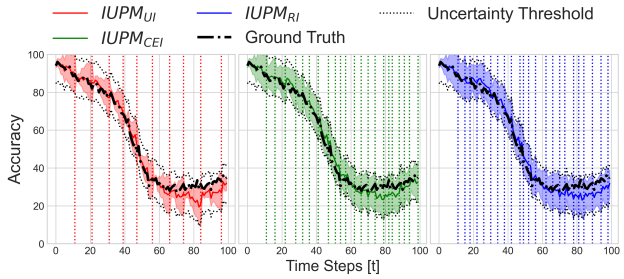

Figure 4: Comparison of sampling strategy using Active Label Intervention on moons data set over 100 steps. All intervention strategies allow keeping the uncertainty below the predefined threshold, however, our proposed Uncertainty Intervention (UI) requires far fewer intervention steps.

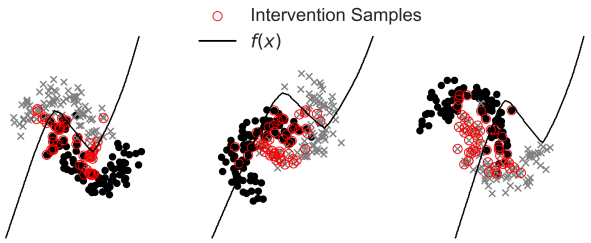

Figure 5: Visualization of 50 samples drawn using our proposed Uncertainty Intervention (UI) sampling strategy before, in the middle, and at the end of 100 time steps. The method consistently samples data points along the decision boundary inherently defined by the moons data set.

task, we trained a LeNet (LeCun et al., 1998) and performed the matching based on representations of the second classification layer of the model, which is a common approach to apply optimal transport to high dimensional data (Courty et al., 2016; He et al., 2024). All shifts are evaluated for 20 steps where we draw a distinct set of 200 samples from the test set in every step. Figure 6 illustrates the performance estimation for a 180° rotation. Even without intervention, IUPM can estimate the performance best. The estimation by NIPM and the other baseline methods can only follow the ground truth sufficiently for a few time steps. The performance estimation error can be further reduced by allowing label intervention with the same settings as in the previous experiment (Figure 6). To keep the uncertainty below the predefined threshold, more interventions are needed in early steps where the shift more strongly impacts the performance. The observations are further supported by the results on two additional shifts in Table 3. To substantiate the findings on MNIST, we have additionally validated our approach on a ResNet-50 (He et al., 2016) classifying 500 samples from the Imagnette validation dataset comprising 10 classes of ImageNet. We analyze several shifts

Table 1: Mean Average Error (MAE) between ground truth and estimated accuracy using baseline methods and IUPM across three synthetic data sets and three different models. The table shows the mean across five random seeds, we refer to Appendix D for confidence intervals.

Method	Clusters			Moons			Circles		
	RF	**XGB**	**MLP**	**RF**	**XGB**	**MLP**	**RF**	**XGB**	**MLP**
ATC	0.4413	0.4664	0.4348	0.3788	0.3594	0.3679	0.3514	0.3442	0.3531
AC	0.4313	0.4875	0.4402	0.3243	0.3741	0.3529	0.2760	0.3325	0.3240
DOC	0.4360	0.4527	0.4488	0.3632	0.3493	0.3701	0.3574	0.3418	0.3473
IM	0.4525	0.4662	0.4705	0.3955	0.3599	0.3800	0.3559	0.3431	0.3488
NIPM	0.4225	0.4186	0.4673	0.2482	0.2286	0.2319	0.0775	0.0791	0.0742
IUPM	0.2914	0.2894	0.3035	0.0781	0.0793	0.1020	0.0352	0.0390	0.0359
IUPM$_{UI}$	**0.0322**	**0.0307**	**0.0331**	**0.0250**	**0.0250**	**0.0230**	**0.0136**	**0.0144**	**0.0138**

Table 2: Mean Average Error (MAE) between ground truth and estimated accuracy and number of triggered label interventions (n_I) for Random Intervention (RI), Cross Entropy Intervention (CEI), and our proposed Uncertainty Intervention (UI) across three synthetic data sets and three different models.

Method		Clusters			Moons			Circles		
		RF	**XGB**	**MLP**	**RF**	**XGB**	**MLP**	**RF**	**XGB**	**MLP**
IUPM$_{RI}$	**MAE**	0.0336	0.0327	0.0296	0.0256	**0.0234**	**0.0198**	**0.0160**	0.0177	0.0176
	n$_I$	32	35	34	23	22	23	31	30	31
IUPM$_{CEI}$	**MAE**	0.0432	0.0397	0.0456	**0.0216**	0.0258	0.0219	0.0187	0.0206	0.0234
	n$_I$	35	34	34	22	21	19	29	27	31
IUPM$_{UI}$	**MAE**	**0.0270**	**0.0272**	**0.0265**	0.0244	0.0242	0.0222	**0.0160**	**0.0157**	**0.0158**
	n$_I$	**15**	**15**	**15**	**10**	**10**	**10**	**13**	**13**	**13**

Method	Rotation	Scaling	Translation
ATC	0.4836	0.1763	0.3111
AC	0.4931	0.2292	0.3842
DOC	0.5181	0.2538	0.4090
IM	0.6282	0.6166	0.5686
NIPM	0.2187	0.0676	0.3110
IUPM	0.0985	0.0442	0.1263
IUPM$_{UI}$	**0.0719**	**0.0438**	**0.0777**

Table 3: Mean Average Error (MAE) between ground truth and estimated accuracy for a LeNet across three different shifts on the MNIST data set. The table shows the mean across five random seeds, we refer to Appendix D for confidence intervals.

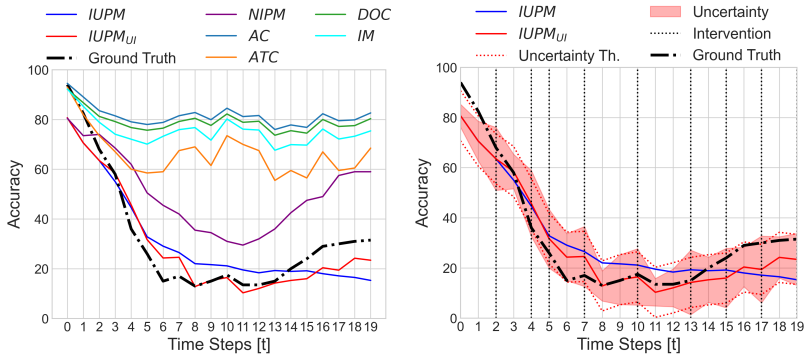

Figure 6: Performance estimation for a rotational shift on the MNIST digits accumulating to a 180° rotation after 20 steps. Comparing IUPM to the different baselines in (Left) illustrates that it offers the highest fidelity for the estimation. (Right) shows the benefit of including Uncertainty Intervention (UI). The intervention steps triggered by exceeding the threshold are indicated as dashed vertical lines.

from `ImageNet-c` (Hendrycks and Dietterich, 2019) that can be considered gradual and further interpolate them over 20 steps. Again, IUPM gives on average the best label-free performance estimate across all considered methods, while only falling short on two shifts. Moreover, relying on our Uncertainty-based labeling interventions (IUPM$_{UI}$) the estimate can further be improved by around 10%.

Method	Blur	Contrast	Brightness	Rotation	Scale	Shear	Translate	Mean
ATC	0.0719	0.0562	0.0260	**0.0564**	0.0871	0.0825	0.0787	0.0655
AC	0.0617	0.0825	0.0551	0.1017	0.1346	0.1695	0.0569	0.0946
DOC	0.0608	0.0827	0.0574	0.1055	0.1349	0.1700	0.0565	0.0954
IM	**0.0456**	0.0936	0.0620	0.1320	0.1166	0.1737	0.0500	0.0962
NIPM	0.1094	0.0805	0.0697	0.1117	0.2127	0.2327	0.1254	0.1346
IUPM	0.0610	**0.0377**	0.0208	0.1024	0.0766	0.0341	0.0469	0.0542
IUPM$_{UI}$	0.0569	0.0417	**0.0173**	0.0712	**0.0752**	**0.0339**	**0.0469**	**0.0490**

Table 4: Mean Average Errors (MAE) between ground truth and estimated accuracy for a ResNet-50 across different gradual shifts from `ImageNet-c`. The results indicate that IUPM gives on average the best label-free performance estimate which can be further improved using our labeling interventions IUPM$_{UI}$.

Monitoring Performance Degradation due to Real-World Temporal Shifts In this experiment, we show the applicability of our approach in a real-world shift scenario. For this, we utilize a gender classification data set consisting of yearbook portraits (Ginosar et al., 2015) across decades, which is commonly used to evaluate methods with respect to gradual shifts (Yao et al., 2022; Kumar et al., 2020; He et al., 2024). We trained a simple convolutional network on the available portraits from 1930 to 1934. By using the samples from 1935 to initialize both our IUPM approach and the baseline methods, we evaluate the performance decline from 1936 to 2013. Unlike previous experiments on synthetic shifts presenting a continuous decline in model performance, Figure 7 shows that the ground truth performance of the model remains almost constant until 1966, when a severe dip in the model performance occurs which partly recovers until 1996. For complex real-world shifts such as the appearance of yearbook photos, the decline in model performance might be linked to various reasons. One such factor fitting very well to the observed performance degradation is quite evident in Figure 1, showing that the hairstyle of male students around 1980 developed to be quite similar to a typical female hairstyle around 1930, which is considered in the models training set. IUPM is the only method that correctly captures this performance drop. Through the introduction of active label intervention steps, our approach is also capable of following a subsequent increase in the model's performance (Figure 7). Further, our method correctly identifies areas with substantial changes and resulting high uncertainties in the performance estimation increasing the frequency of triggering label interventions. In real-world applications, it is highly desirable to identify such time periods to increase the labeling effort rather than relying on a non-faithful estimate.

Empirical Estimation of Gradual Lipschitz Smoothness As a final experiment, we demonstrate how to empirically asses if the underlying assumption of a shift being gradually Lipschitz smooth (Definition 2) holds practice. Again, we consider the portraits classification dataset comprising real college portraits, and we use only a small number of 100 samples per step. To get an estimate $\hat{\varepsilon}$ of the Wasserstein distance $\mathcal{W}$ between two input distributions at consecutive time steps, we solve the underlying transport problem using a linear program solver (Peyré et al., 2019). To scale this approach to high-dimensional real datasets, we quantify this distributional distance based on corresponding network activations of the penultimate layer instead of the raw input, which is a common practice to compare realistic data distribution (Heusel et al., 2017; Zhang et al., 2018). Moreover, since every input image has an unambiguous class label, $\mathcal{W}(P(Y_t|X_t), P(Y_{t-1}|X_{t-1})) = 1$ for samples that have different labels and 0 otherwise. Hence, one can verify the Lipschitz property at time t by identifying the pair (x_t^*, x_{t-1}^*) with different labels but minimal distance within the available samples:

$$(x_t^*, x_{t-1}^*) = \arg\min_{x_t, x_{t-1}} c(x_t, x_{t-1}) \quad \text{st.} \quad y_t \neq y_{t-1}.$$

This results in:

$$\frac{\mathcal{W}(P(Y_t|X_t), P(Y_{t-1}|X_{t-1}))}{c(x_t, x_{t-1})} \leq \frac{1}{c(x_t^*, x_{t-1}^*)} \leq L_t.$$

Therefore, we can use $\hat{L}_t := 1/c(x_t^*, x_{t-1}^*)$ as an empirical lower bound for Lipschitzness at time t. To validate the utility of such estimates, we performed two complementary correlation studies analyzing two theoretical relationships. First, Proposition 1 shows that the constants L_t and ε_t determine the strength of the overall shift through $(1 + L_t)\varepsilon_t$. If our estimates are meaningful, this quantity should correlate with the

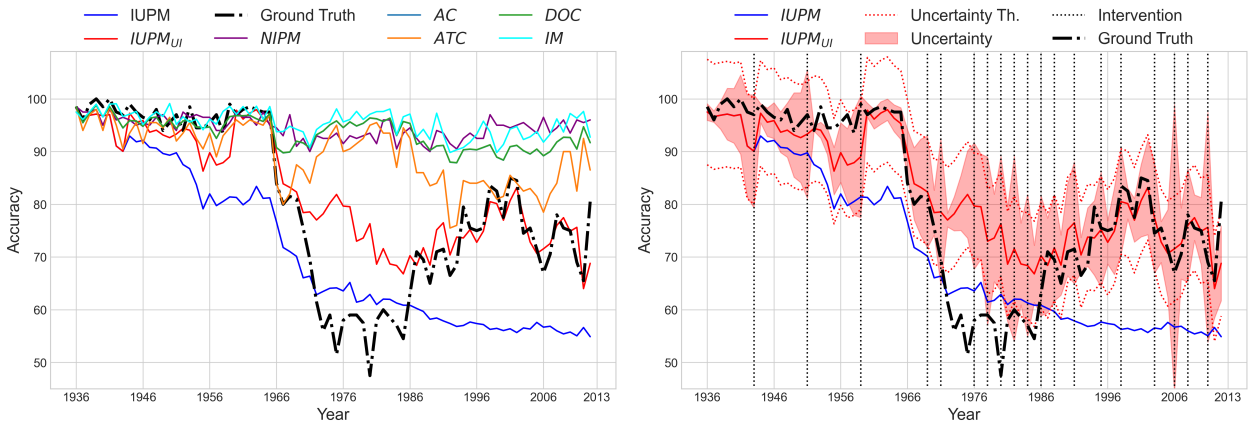

Figure 7: Performance estimation for yearbook model across portraits from 1936 to 2013. The comparison with baselines in (Left) shows that only the estimation by IUPM is consistent with the dip in performance between 1966 and 1996. (Right) illustrates the benefit of including Uncertainty Intervention (UI). The intervention steps triggered by exceeding the threshold are indicated as dashed vertical lines.

actual performance $\mathcal{L}_t$. Second, Theorem 1 provides an upper bound on the performance estimation error $|\mathcal{L}_t - \hat{\mathcal{L}}_t^{IUPM}|$ in terms of these constants. Hence, we would expect a correspondence between this theoretical bound and the actual error if our empirical estimates capture the relevant properties of the observed shifts. Both results indicate a high linear correlation

Table 5: Pearson Correlation (ρ) between performance ($\mathcal{L}_t$) and estimation error ($|\mathcal{L}_t - \hat{\mathcal{L}}_t^{IUPM}|$) at time t with corresponding theoretical upper bounds from Proposition 1 and Theorem 1. The strong linear correlation implies that estimating the underlying quantities can yield a useful indication of the extent to which an observed shift is gradually Lipschitz smooth.

Quantity vs Proxy Estimate	ρ	p-value
$\mathcal{L}_t$ vs. $(1+\hat{L}_t)\hat{\varepsilon}_t$	0.86	< 0.001
$\lvert\mathcal{L}_t - \hat{\mathcal{L}}_t^{IUPM}\rvert$ vs. $\sum_{i=0}^{t} \hat{L}_i\hat{\varepsilon}_t$	0.56	< 0.001

with strong statistical significance between the estimated quantities and two related theoretical expressions. This provides evidence that estimating the underlying quantities can yield a useful indication if an observed shift is gradually Lipschitz smooth.

5 CONCLUSION

In this work, we introduce IUPM, a novel method designed to anticipate the performance of deployed machine learning models under gradual changes over time. We theoretically analyze the underlying assumptions and demonstrate that its estimates closely align with true performance for a broad class of gradual distribution shifts. Additionally, IUPM naturally quantifies uncertainty, enabling more reliable assessments by actively querying for labels to enhance the trustworthiness of its estimates when necessary. Through analysis of simple synthetic datasets, we illustrate the underlying principles. We further validate its effectiveness on both simulated and real gradual and temporal shifts. While IUPM is specifically tailored for gradual changes, it may not be optimal for other types of distribution shifts. When these distribution shifts result in a high uncertainty of the sample-wise loss estimates, our UI approach may require frequent interventions. In many real-world cases, however, providing additional labeled data more frequently is preferable rather than relying on an inaccurate performance estimate without any indication of uncertainty at all.

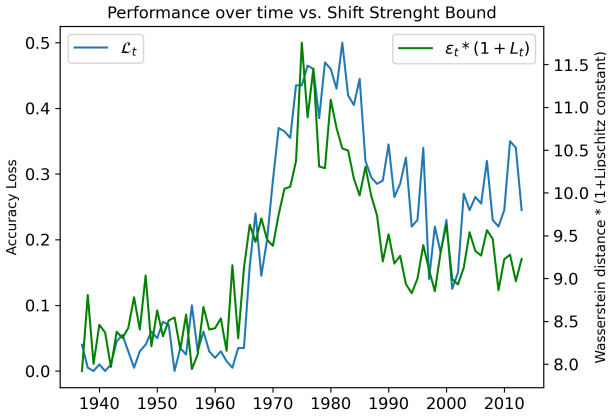

Figure 8: Performance ($\mathcal{L}_t$) and estimated overall shift strength ($(1+\hat{L}_t)\hat{\varepsilon}_t$) over time t.

References

Abnar, S., Berg, R. v. d., Ghiasi, G., Dehghani, M., Kalchbrenner, N., and Sedghi, H. (2021). Gradual domain adaptation in the wild: When intermediate distributions are absent. *arXiv preprint arXiv:2106.06080*.

Baek, C., Jiang, Y., Raghunathan, A., and Kolter, J. Z. (2022). Agreement-on-the-line: Predicting the performance of neural networks under distribution shift. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K., editors, *Advances in Neural Information Processing Systems*.

Chen, H.-Y. and Chao, W.-L. (2021). Gradual domain adaptation without indexed intermediate domains. *Advances in neural information processing systems*, 34:8201–8214.

Chen, M., Goel, K., Sohoni, N. S., Poms, F., Fatahalian, K., and Ré, C. (2021). Mandoline: Model evaluation under distribution shift. In *International conference on machine learning*, pages 1617–1629. PMLR.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Courty, N., Flamary, R., Habrard, A., and Rakotomamonjy, A. (2017). Joint distribution optimal transportation for domain adaptation. *Advances in neural information processing systems*, 30.

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865.

Cuturi, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Decker, T., Koebler, A., Lebacher, M., Thon, I., Tresp, V., and Buettner, F. (2024). Explanatory model monitoring to understand the effects of feature shifts on performance. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 550–561.

Deng, W., Gould, S., and Zheng, L. (2021). What does rotation prediction tell us about classifier accuracy under varying testing environments? In *International Conference on Machine Learning*, pages 2579–2589. PMLR.

Deng, W. and Zheng, L. (2021). Are labels always necessary for classifier accuracy evaluation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15069–15078.

Everitt, B. (2013). *Finite mixture distributions*. Springer Science & Business Media.

Flamary, R., Courty, N., Gramfort, A., Alaya, M. Z., Boisbunon, A., Chambon, S., Chapel, L., Corenflos, A., Fatras, K., Fournier, N., Gautheron, L., Gayraud, N. T., Janati, H., Rakotomamonjy, A., Redko, I., Rolet, A., Schutz, A., Seguy, V., Sutherland, D. J., Tavenard, R., Tong, A., and Vayer, T. (2021). Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM computing surveys (CSUR)*, 46(4):1–37.

Garg, S., Balakrishnan, S., Lipton, Z. C., Neyshabur, B., and Sedghi, H. (2022). Leveraging unlabeled data to predict out-of-distribution performance. In *International Conference on Learning Representations*.

Garg, S., Wu, Y., Balakrishnan, S., and Lipton, Z. (2020). A unified view of label shift estimation. *Advances in Neural Information Processing Systems*, 33:3290–3300.

Genevay, A., Chizat, L., Bach, F., Cuturi, M., and Peyré, G. (2019). Sample complexity of sinkhorn divergences. In *The 22nd international conference on artificial intelligence and statistics*, pages 1574–1583. PMLR.

Ginosar, S., Rakelly, K., Sachs, S., Yin, B., and Efros, A. A. (2015). A century of portraits: A visual historical record of american high school yearbooks. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–7.

Gozlan, N. and Léonard, C. (2010). Transport inequalities. a survey. *arXiv preprint arXiv:1003.3852*.

Guillory, D., Shankar, V., Ebrahimi, S., Darrell, T., and Schmidt, L. (2021). Predicting with confidence on unseen distributions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1134–1144.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

He, Y., Wang, H., Li, B., and Zhao, H. (2024). Gradual domain adaptation: Theory and algorithms. *Journal of Machine Learning Research*, 25(361):1–40.

Hendrycks, D. and Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Howard, J. (2019). Imagenette: A smaller subset of 10 easily classified classes from imagenet.

Jiang, Y., Nagarajan, V., Baek, C., and Kolter, J. Z. (2022). Assessing generalization of SGD via disagreement. In *International Conference on Learning Representations*.

Kossen, J., Farquhar, S., Gal, Y., and Rainforth, T. (2021). Active testing: Sample-efficient model evaluation. In *International Conference on Machine Learning*, pages 5753–5763. PMLR.

Kossen, J., Farquhar, S., Gal, Y., and Rainforth, T. (2022). Active surrogate estimators: An active learning approach to label-efficient model evaluation. *Advances in Neural Information Processing Systems*, 35:24557–24570.

Kumar, A., Ma, T., and Liang, P. (2020). Understanding self-training for gradual domain adaptation. In *International conference on machine learning*, pages 5468–5479. PMLR.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Lee, J., Kolla, L., and Chen, J. (2024). Towards optimal model evaluation: enhancing active testing with actively improved estimators. *Scientific Reports*, 14.

Lu, Y., Qin, Y., Zhai, R., Shen, A., Chen, K., Wang, Z., Kolouri, S., Stepputtis, S., Campbell, J., and Sycara, K. (2023). Characterizing out-of-distribution error via optimal transport. *Advances in Neural Information Processing Systems*, 36:17602–17622.

Mu, N. and Gilmer, J. (2019). Mnist-c: A robustness benchmark for computer vision. *arXiv preprint arXiv:1906.02337*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Sawade, C., Landwehr, N., Bickel, S., and Scheffer, T. (2010). Active risk estimation. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 951–958, Madison, WI, USA. Omnipress.

Sugiyama, M., Suzuki, T., Nakajima, S., Kashima, H., von Bünau, P., and Kawanabe, M. (2008). Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746.

Villani, C. et al. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wang, H., He, H., and Katabi, D. (2020). Continuously indexed domain adaptation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 9898–9907.

Wang, H., Li, B., and Zhao, H. (2022). Understanding gradual domain adaptation: Improved analysis, optimal path and beyond. In *International Conference on Machine Learning*, pages 22784–22801. PMLR.

Wilson, G. and Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 11(5):1–46.

Xie, M., Li, S., Yuan, L., Liu, C., and Dai, Z. (2024). Evolving standardization for continual domain generalization over temporal drift. *Advances in Neural Information Processing Systems*, 36.

Yao, H., Choi, C., Cao, B., Lee, Y., Koh, P. W. W., and Finn, C. (2022). Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems*, 35:10309–10324.

Yu, H., Liu, J., Zhang, X., Wu, J., and Cui, P. (2024). A survey on evaluation of out-of-distribution generalization. *arXiv preprint arXiv:2403.01874*.

Yu, Y., Yang, Z., Wei, A., Ma, Y., and Steinhardt, J. (2022). Predicting out-of-distribution error with the projection norm. In *International Conference on Machine Learning*, pages 25721–25746. PMLR.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes, we include a GitHub link]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A THEORETICAL PROOFS

In this section, we provide additional details on all theoretical results and conduct missing proofs. To prove our main result stated in Theorem 1, we rely on the assumption that the experienced shift is gradually Lipschitz smooth, which we restate below:

Definition 2. *A distribution shift over $\{(X_t, Y_t)\}_{t=0}^T$ is called gradually Lipschitz smooth in X_t if for a cost function $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ we have $\mathcal{W}(P_t(X_t), P_{t-1}(X_{t-1})) \leq \varepsilon_t$ for all $t = 1, \ldots, T$ and there exist constants $L_t > 0$ such that for any realizations x_t, x_{t-1} of X_t, X_{t-1} it holds:*

$$\mathcal{W}\big(P_t(Y_t|X_t = x_t), P_{t-1}(Y_{t-1}|X_{t-1} = x_{t-1})\big) \leq L_t \, c(x_t, x_{t-1})$$

Any shift satisfying this property is also gradual (see Definition 1 in the main paper) in the typically assumed sense:

Proposition 1. *If a distribution shift over $\{(X_t, Y_t)\}_{t=0}^T$ is gradually Lipschitz smooth in X_t with constants L_t, then it is also gradual:*

$$\mathcal{W}(P_t(Y_t, X_t), P_{t-1}(Y_{t-1}, X_{t-1})) \leq (1 + L_t) \, \varepsilon_t := \Delta_t$$

Proof. Measuring the Wasserstein distance between two joint distributions requires a cost metric c_{xy} operating on the product space $(\mathcal{X} \times \mathcal{Y})$. Note that if c_x is a cost metric on $\mathcal{X}$ and c_y on $\mathcal{Y}$, a natural choice is to simply consider c_{xy} to be separable: $c_{xy}((x, y), (x', y')) = c_x(x, x') + c_y(y, y')$. Let $\Pi\big(P_t(Y_t, X_t), P_{t-1}(Y_{t-1}, X_{t-1})\big)$ be the space of all valid couplings over the joint distributions of two consecutive domains. Furthermore let $\Pi_x\big(P_t(X_t), P_{t-1}(X_{t-1})\big)$ bet the set of all couplings over marginals in X and given an tuple $x = (x_t, x_{t-1})$, $\Pi_{y|x}\big(P_t(Y_t|X_t = x_t), P_{t-1}(Y_{t-1}|X_{t-1} = x_{t-1})\big)$ be the set of couplings over all target conditionals. Then we have:

$$\begin{aligned}
\mathcal{W}^{c_{xy}}(P_t(Y_t, X_t), P_{t-1}(Y_{t-1}, X_{t-1})) &= \inf_{\pi \in \Pi} \mathbb{E}_\pi \big[c_x(x_t, x_{t-1}) + c_y(y_t, y_{t-1})\big] \\
&\leq \inf_{\pi_x \in \Pi_x} \mathbb{E}_{(x_t, x_{t-1}) \sim \pi_x(X_t, X_{t-1})} \big[c_x(x_t, x_{t-1})\big] \quad + \\
&\qquad \mathbb{E}_{(x_t, x_{t-1}) \sim \pi_x(X_t, X_{t-1})} \Big[\inf_{\pi_{y|x} \in \Pi_{y|x}} \mathbb{E}_{(y_t, y_{t-1}) \sim \pi_{y|x}(y_t, y_{t-1}|x_t, x_{t-1})} \big[c_x(y_t, y_{t-1})\big] \Big] \\
&= \mathcal{W}^{c_x}(P_t(X_t), P_{t-1}(X_{t-1})) \quad + \\
&\qquad \mathbb{E}_{(x_t, x_{t-1}) \sim \pi(x_t, x_{t-1})} \big[\mathcal{W}^{c_y}(P_t(Y_t|X_t = x_t), P_{t-1}(Y_{t-1}|X_{t-1} = x_{t-1}))\big] \\
&\leq \mathcal{W}^{c_x}(P_t(X_t), P_{t-1}(X_{t-1})) + L_t \, \mathbb{E}_{(x_t, x_{t-1}) \sim \pi_x(X_t, X_{t-1})} \big[c_x(x_t, x_{t-1})\big] \\
&\leq (1 + L_t) \mathcal{W}^{c_x}(P_t(X_t), P_{t-1}(X_{t-1})) \leq (1 + L_t) \varepsilon_t
\end{aligned}$$

□

For more technical details on working with transportation costs in product spaces, we refer to Appendix A of (Gozlan and Léonard, 2010). This implies, that Definition 2 simply describes a gradual shift, where the overall change in the joint $P_t(X_t, Y_t)$ is dominated by the change in $P_t(X_t)$. This assumption is also common in the domain adaptation literature and relates for instance to the property of being *Probabilistic Transfer Lipschitz* with respect to a labeling function analyzed in (Courty et al., 2017).

Proving Theorem 1 and Deriving Corollary 1 Next we conduct the proof of Theorem 1 and discuss the resulting Corollary 1 mentioned in the main paper.

Theorem 1. *Let $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function that is 1-Lipschitz in its second argument and denote the true model performance at time t with $\mathcal{L}_t$. If a distribution shift over $\{(X_t, Y_t)\}_{t=1}^T$ is gradually Lipschitz smooth in X_t, then:*

$$|\mathcal{L}_t - \hat{\mathcal{L}}_t^{IUPM}| \leq \sum_{i=1}^{t} L_t \, \varepsilon_t$$

Proof. Let $\gamma_t(X_{t-1}, X_t)$ be the incremental optimal transport couplings in X and $\Psi_t(X_0|X_t)$ be the composition of all incremental transition probabilities:

$$\Psi_t(X_0|X_t) = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \gamma_1(X_0|x_1)\gamma_2(x_1|x_2) \dots \gamma_t(x_{t-1}|X_t)\, dx_1 \dots dx_{t-1}$$

Note that in the discrete case, all $\gamma_i(X_{i-1}|X_i)$ are matrices and this composition can equivalently be expressed using iterative matrix multiplication: $\Psi_t(X_0|X_t) = \prod_{i=1}^{t} \gamma_i(X_{i-1}|X_i)$. For the estimated target distribution $\hat{P}(Y_t|X_t) = \mathbb{E}_{\Psi_t(X_0|X_t)}\left[P(Y_0|X_0)\right]$ it holds:

$$\begin{aligned}
|\mathcal{L}_t - \hat{\mathcal{L}}_t^{IUPM}| &= |\mathbb{E}_{(x_t,y_t)\sim P_t(X_t,Y_t)}\big[\mathcal{L}(f(x_t), y_t)\big] - \mathbb{E}_{(x_t,\hat{y}_t)\sim P_t(X_t,Y_t)}\big[\mathcal{L}(f(x_t), \hat{y}_t)\big]| \\
&\leq \mathbb{E}_{x_t\sim P_t(X_t)}\big[|\mathbb{E}_{y_t\sim P_t(Y_t|X_t=x_t)}[\mathcal{L}(f(x_t), y_t)] - \mathbb{E}_{\hat{y}_t\sim \hat{P}_t(Y_t|X_t=x_t)}[\mathcal{L}(f(x_t), \hat{y}_t)]|\big]
\end{aligned}$$

Since we assume $\mathcal{L}$ to be 1-Lipschitz in its second argument ($\mathcal{L}(\cdot, y) \in \mathrm{Lip}_1$) we know that there exists a cost function $c_y : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ with $|\mathcal{L}(f(x_t), y) - \mathcal{L}(f(f_x), y')| \leq c_y(y, y')$ for any fixed x_t. Hence, one can apply the Kantorovich-Rubenstein duality (Theorem 5.10 in (Villani et al., 2009)):

$$\begin{aligned}
&\mathbb{E}_{x_t\sim P_t(X_t)}\big[|\mathbb{E}_{y_t\sim P_t(Y_t|X_t=x_t)}[\mathcal{L}(f(x_t), y_t)] - \mathbb{E}_{\hat{y}_t\sim \hat{P}_t(Y_t|X_t=x_t)}[\mathcal{L}(f(x_t), \hat{y}_t)]|\big] \\
&\quad\leq \mathbb{E}_{x_t\sim P_t(X_t)}\big[|\sup_{\phi\in \mathrm{Lip}_1}\Big\{\mathbb{E}_{y_t\sim P_t(Y_t|X_t=x_t)}[\phi(y_t)] - \mathbb{E}_{\hat{y}_t\sim \hat{P}_t(Y_t|X_t=x_t)}[\phi(\hat{y}_t)]\Big\}|\big] \\
&\quad\leq \mathbb{E}_{x_t\sim P_t(X_t)}\Big[\mathbb{E}_{x_0\sim \Psi_t(X_0|X_t)}\big[\sup_{\phi\in \mathrm{Lip}_1}\big\{\mathbb{E}_{y_t\sim P_t(Y_t|X_t=x_t)}[\phi(y_t)] - \mathbb{E}_{y_0\sim P_0(Y_0|X_0=x_0)}[\phi(y_0)]\big\}\big]\Big] \\
&\quad= \mathbb{E}_{x_t\sim P_t(X_t)}\Big[\mathbb{E}_{x_0\sim \Psi_t(X_0|X_t)}\big[\mathcal{W}^{c_y}\big(P_t(Y_t|X_t = x_t), P_0(Y_0|X_0 = x_0)\big)\big]\Big] \\
&\quad\leq \mathbb{E}_{x_t\sim P_t(X_t)}\Big[\mathbb{E}_{x_0\sim \Psi_t(X_0|X_t)}\big[\sum_{i=1}^{t}\mathcal{W}^{c_y}\big(P_i(Y_i|X_i = x_i), P_{i-1}(Y_{i-1}|X_{i-1} = x_{i-1})\big)\big]\Big]
\end{aligned}$$

where the last step follows from the fact that the Wasserstein distance satisfies the triangular inequality. Now one can use the assumption that the shift is gradually Lipschitz smooth in X_t for a cost metric $c_x : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$:

$$\begin{aligned}
&\mathbb{E}_{x_t\sim P_t(X_t)}\Big[\mathbb{E}_{x_0\sim \Psi_t(X_0|X_t)}\big[\sum_{i=1}^{t}\mathcal{W}^{c_y}\big(P_i(Y_i|X_i = x_i), P_{i-1}(Y_{i-1}|X_{i-1} = x_{i-1})\big)\big]\Big] \\
&\qquad\leq \mathbb{E}_{x_t\sim P_t(X_t)}\Big[\mathbb{E}_{x_0\sim \Psi_t(X_0|X_t)}\big[\sum_{i=1}^{t} L_t\; c_x(x_i, x_{i-1})\big]\Big]
\end{aligned}$$

Lastly, by decomposing the overall transition probabilities again into the incremental cost-minimizing ones and rearranging all expectations we have:

$$\begin{aligned}
\mathbb{E}_{x_t\sim P_t(X_t)}\Big[\mathbb{E}_{x_0\sim \Psi_t(X_0|X_t)}\big[\sum_{i=1}^{t} L_t\; c_x(x_i x_{i-1})\big]\Big] &= \big(\sum_{i=1}^{t} L_t\; \mathbb{E}_{(x_i,x_{i-1})\sim\gamma_i(X_i,X_{i-1})}\big[c(x_i, x_{i-1})\big]\big) \\
&= \sum_{i=1}^{t} L_t\; \mathcal{W}^{c_x}\big(P_i(X_i), P_{i-1}(X_{i-1})\big) \leq \sum_{i=1}^{t} L_t\; \varepsilon_t
\end{aligned}$$

□

Notice that if an alternative estimator $\tilde{P}(Y_t|X_t)$ is constructed by composing arbitrary incremental couplings π_i instead of the cost-optimal ones, the entire derivation is the same until the last part. Thus, bounding the error of the resulting performance estimate $\tilde{\mathcal{L}}_t^{\pi}$ yields:

$$|\mathcal{L}_t - \tilde{\mathcal{L}}_t^{\pi}| \leq \big(\sum_{i=1}^{t} L_t\; \mathbb{E}_{(x_i,x_{i-1})\sim\pi_i(X_i,X_{i-1})}\big[c(x_i, x_{i-1})\big]\big)$$

Taking the infimum over all possible couplings π_i on the right-hand side to minimize the upper bound results $\pi_i = \gamma_i$ by the definition of the optimal transport couplings. Hence, choosing incremental couplings based on optimal transport does effectively optimizes an upper bound on the performance estimation error, which implies Corollary 1:

Corollary 1. *If each γ_t is the optimal transport coupling between two consecutive domains X_{t-1} and X_t, then this minimizes the estimation error across all possible incremental couplings.*

B FURTHER DETAILS ON THE IUPM ALGORITHM

In the following, we provide an overview of the IUPM algorithm and subsequently discuss algorithmic complexity.

B.1 Overview of the IUPM algorithm

Algorithm 1 describes the proposed procedure to estimate the performance of the model f over time steps t.

Algorithm 1 IUPM algorithm

procedure IUPM(f, x_{init}, th_U)
  Input: f: ML model, x_{init}: Initialization dataset, th_U: Uncertainty threshold
  Output: $\hat{\mathcal{L}}_t^{IUPM}$: Performance estimation, $\mathcal{U}(\hat{\mathcal{L}}_t^{IUPM})$: Inherent uncertainty measure
  for timesteps t **do**
    Sample x_t from Ω_t
    Calculate $\hat{\gamma}_t = \arg\min_{\gamma \in \Gamma} \sum_{x_{t-1} \in \Omega_{t-1}} \sum_{x_t \in \Omega_t} c(x_{t-1}, x_t)\gamma(x_{t-1}, x_t)$ ▷ Optimal transport matching
    Update $\Psi_t(X_0|X_t) = \prod_{i=1}^{t} \gamma_i(X_{i-1}|X_i)$ ▷ Update transition matrix
    Calculate $\hat{\mathcal{L}}_t^{IUPM} = \frac{1}{n_t} \sum_{x_t \in \Omega_t} \mathbb{E}_{\hat{P}(Y_t|X_t=x_t)} [\mathcal{L}(f(X_t), Y_t)]$
    Get $\mathcal{U}(\hat{\mathcal{L}}_t^{IUPM}) = \mathbb{E}_{P(X_t)} \mathrm{SD}_{\hat{P}(Y_t|X_t)} [\mathcal{L}(f(X_t), Y_t)]$
    if $\mathcal{U}(\hat{\mathcal{L}}_t^{IUPM}) > th_U$ **then**
      Get m samples x_t based on $\arg\text{top-}m_{x_t \in \Omega_t} \mathcal{U} [\mathcal{L}(f(x_t), Y_t)]$
      Query labels y_t for samples x_t from user
      Correct $\Psi_t(X_0|X_t)$ such that $\hat{P}(Y_t|X_t = x_t)$ assigns y_t
    end if
  end for
end procedure

B.2 Discussion of Coupling Estimation and Algorithmic Complexity

The crucial algorithmic components of IUPM are the computations of the incremental transport couplings γ_i. They result from solving an optimal transport coupling for which a variety of different computational approaches exist (Peyré et al., 2019). We utilized a popular approach to increase the estimation efficiency using entropic regularization (Cuturi, 2013; Flamary et al., 2021). Let $KL(\cdot|\cdot)$ denote the KL-divergence between two distributions and let λ be a hyperparameter capturing the regularization strength, then the objective reads:

$$\min_{\gamma \in \Gamma} \sum_{x_0 \in \Omega_0} \sum_{x_1 \in \Omega_1} c(x_0, x_1)\gamma(x_0, x_1) + \lambda KL(\gamma|\hat{p}_0 \otimes \hat{p}_1) \quad \text{with} \quad \Gamma = \{\gamma \in \mathbb{R}^{n_0 \times n_1} \mid \gamma \mathbf{1}_{n_1} = \hat{p}_0, \gamma^T \mathbf{1}_{n_0} = \hat{p}_1\}$$

This can efficiently be solved using the Sinkhorn algorithm (Cuturi, 2013), which has sample complexity of $\mathcal{O}(1/\sqrt{n})$ and time complexity of $\mathcal{O}(n^2)$, where n is the number of samples to be matched from each domain (Genevay et al., 2019). Note that IUPM requires solving one optimal transport problem for every time point of assessment during model deployment.

C ADDITIONAL DETAILS ON EXPERIMENTS

In this section, we provide additional details on the experiments performed.

C.1 Computational Environment

All numerical experiments are implemented in Python (version 3.9.13) using PyTorch (version 1.13.0) and have been computed on an Nvidia RTX A5000 GPU with CUDA 11.7 and two physical AMD EPYC 7502P 32-Core CPUs running on Linux Ubuntu.

C.2 Method Implementation

Baselines For the four confidence-based baseline methods, we rely on the implementations provided by (Garg et al., 2022). For the ATC method (Garg et al., 2022), we use the author's proposed maximum confidence score function.

IUPM and NIPM For our IUPM and NIPM implementation, we rely on the entropic regularization optimal transport implementation with logarithmic Sinkhorn by (Flamary et al., 2021). As the cost function for calculating optimal transport, we use the squared Euclidean distance throughout all experiments. Similar to the baseline methods, we also consider the model accuracy as the loss criterion $\mathcal{L}$ to evaluate performance in the conducted experiments. In all experiments the uncertainty intervention threshold $\mathcal{U}(\mathcal{L}_t) > 0.1$ is set to trigger a relabeling of 50% of the samples in step t

C.3 Experiments

Translation and Rotation in Input Space As for the synthetic two-dimensional datasets, we used the data generator functionality provided by (Pedregosa et al., 2011). The "Clusters" data set is generated using the make_blobs function with a distance parameter of 1.0. The "Moons" and "Circles" data sets are generated using the corresponding functions with a noise parameter of 0.2 and a circle factor of 0.3. For the training and initialization step, a training set of 800 samples is generated, from which a validation and initialization set Ω_0 of 200 samples is partitioned. In each consecutive step, a set Ω_k with a different random seed is generated. We then apply a shift to the set corresponding to the step k, i.e., $k \cdot 2^\circ$ for rotation and $k \cdot 0.02$ for translation.
For the synthetic data, we use a Random Forest Classifier (RF) and an XGBoost Classifier (XGB) (Chen and Guestrin, 2016) with 50 estimators and a maximum depth of 5 as well as a Multilayer Perceptron (MLP) with a single hidden layer of size 128. We use a regularization parameter of 10^{-4} for optimal transport matching.

Monitoring Performance Degradation due to Image Perturbations For the experiment based on MNIST, we apply three different affine transformations to the original digits. For this, we adapt the corresponding image perturbation implementation introduced in (Mu and Gilmer, 2019) to the continuous setting. The used LeNet model has been trained for 100 epochs with early stopping based on patience of 10 epochs and PyTorch's Adam optimizer with a batch size of 16 and a learning rate of $1e-3$. For optimal transport matching, we use the representations after the second fully connected layer of the LeNet model and a regularization parameter of 1. The sets Ω_t, including the initialization, set Ω_0, each consists of 200 distinct samples from the test set. For ImageNet we fine-tuned a pre-trained ResNet-50 (He et al., 2016) model on the ten classes included in the Imagenette subset (Howard, 2019) of ImageNet. For this we use PyTorch's Adam optimizer with a batch size of 16 and a learning rate of $1e-5$. We fine-tuned the model for 50 epochs with early stopping and a patience of 10 epochs. We analyzed 7 different shifts from `ImageNet-c` (Hendrycks and Dietterich, 2019) that can be considered gradual, each based on 500 samples. Note that `ImageNet-c` provides each shift in five predefined strengths and we have additionally interpolated all shifts for a total number of 20 steps. All incremental couplings have been computed based on the network activation of the last layer before the classification layer with a Sinkhorn regularization parameter of $1e-4$.

Monitoring Performance Degradation due to Real-World Temporal Shifts For the pre-processing steps as well as the network architecture used for the portrait experiment, we rely on the implementation provided by (Yao et al., 2022). By this, images are of shape 32×32 and the used YearbookNetwork consists of four convolutional blocks with 32 channels and a single linear classification layer. The model has been trained for 300 epochs with early stopping based on patience of 5 epochs and PyTorch's Adam optimizer with a batch size of 32 and a learning rate of $1e-3$. For the optimal transport matching, we use the representations after the last convolutional block of the YearbookNetwork and a regularization parameter of $1e-3$.

D ADDITIONAL EXPERIMENTAL RESULTS

In this section, we present additional results that extend our evaluation.

Confidence Intervals To validate the statistical significance of our results on the synthetic and MNIST datasets in Table 1 and Table 3, we provide the confidence intervals in Tables 6 and 7, respectively. Both tables

Table 6: Mean Average Error (MAE) for five different random seeds between ground truth and estimated accuracy using baseline methods and IUPM for the three synthetic data sets and three different models.

Method	**Clusters**			**Moons**			**Circles**		
	RF	**XGB**	**MLP**	**RF**	**XGB**	**MLP**	**RF**	**XGB**	**MLP**
ATC	$0.4413_{\pm 0.0026}$	$0.4664_{\pm 0.0040}$	$0.4348_{\pm 0.0015}$	$0.3788_{\pm 0.0028}$	$0.3594_{\pm 0.0015}$	$0.3679_{\pm 0.0014}$	$0.3514_{\pm 0.0012}$	$0.3442_{\pm 0.0016}$	$0.3531_{\pm 0.0014}$
AC	$0.4313_{\pm 0.0028}$	$0.4875_{\pm 0.0030}$	$0.4402_{\pm 0.0017}$	$0.3243_{\pm 0.0020}$	$0.3741_{\pm 0.0016}$	$0.3529_{\pm 0.0014}$	$0.2760_{\pm 0.0012}$	$0.3325_{\pm 0.0013}$	$0.3240_{\pm 0.0008}$
DOC	$0.4360_{\pm 0.0027}$	$0.4527_{\pm 0.0031}$	$0.4488_{\pm 0.0017}$	$0.3632_{\pm 0.0021}$	$0.3493_{\pm 0.0017}$	$0.3701_{\pm 0.0014}$	$0.3574_{\pm 0.0012}$	$0.3418_{\pm 0.0013}$	$0.3473_{\pm 0.0009}$
IM	$0.4525_{\pm 0.0024}$	$0.4662_{\pm 0.0030}$	$0.4705_{\pm 0.0015}$	$0.3955_{\pm 0.0022}$	$0.3599_{\pm 0.0017}$	$0.3800_{\pm 0.0015}$	$0.3559_{\pm 0.0012}$	$0.3431_{\pm 0.0013}$	$0.3488_{\pm 0.0011}$
NIPM	$0.4225_{\pm 0.0034}$	$0.4186_{\pm 0.0037}$	$0.4673_{\pm 0.0014}$	$0.2482_{\pm 0.0032}$	$0.2286_{\pm 0.0033}$	$0.2319_{\pm 0.0024}$	$0.0775_{\pm 0.0012}$	$0.0791_{\pm 0.0022}$	$0.0742_{\pm 0.0016}$
IUPM	$0.2914_{\pm 0.0132}$	$0.2894_{\pm 0.0139}$	$0.3035_{\pm 0.0120}$	$0.0781_{\pm 0.0091}$	$0.0793_{\pm 0.0085}$	$0.1020_{\pm 0.0075}$	$0.0352_{\pm 0.0044}$	$0.0390_{\pm 0.0040}$	$0.0359_{\pm 0.0040}$
IUPM_{UI}	$0.0322_{\pm 0.0033}$	$0.0307_{\pm 0.0037}$	$0.0331_{\pm 0.0029}$	$0.0250_{\pm 0.0019}$	$0.0250_{\pm 0.0018}$	$0.0230_{\pm 0.0015}$	$0.0136_{\pm 0.0017}$	$0.0144_{\pm 0.0021}$	$0.0138_{\pm 0.0020}$

Table 7: Mean Average Errors (MAE) for five different random seeds between ground truth and estimated accuracy for a LeNet across three different shifts on the MNIST data set.

Method	Rotation	Scaling	Translation
ATC	$0.4836_{\pm 0.0082}$	$0.1763_{\pm 0.0086}$	$0.3111_{\pm 0.0024}$
AC	$0.4931_{\pm 0.0086}$	$0.2292_{\pm 0.0059}$	$0.3842_{\pm 0.0061}$
DOC	$0.5181_{\pm 0.0086}$	$0.2538_{\pm 0.0060}$	$0.4090_{\pm 0.0061}$
IM	$0.6282_{\pm 0.0083}$	$0.6166_{\pm 0.0061}$	$0.5686_{\pm 0.0097}$
NIPM	$0.2187_{\pm 0.0099}$	$0.0676_{\pm 0.0112}$	$0.3110_{\pm 0.0138}$
IUPM	$0.0985_{\pm 0.0064}$	$0.0442_{\pm 0.0077}$	$0.1263_{\pm 0.0084}$
IUPM_{UI}	$0.0719_{\pm 0.0093}$	$0.0438_{\pm 0.0052}$	$0.0777_{\pm 0.0064}$

show that there is very little change across five different random seeds.

Labeling Interventions Reducing the need for human interactions for our proposed IUPM method with human intervention limits the manual effort for a reliable performance estimate. In the main paper, we have already shown in Table 2 that the number of interventions for our proposed Uncertainty Intervention (UI) method is significantly lower than for the other methods, while the quality of the estimate is better or on par. We performed an ablation study in Table 8 to validate that the reverse conclusion that UI provides a better performance estimation for the same number of interventions also holds true. In this experiment, we trigger the interventions at the same steps, taking as a reference the exceeding of the threshold given by the UI method.

Table 8: Comparison of the three intervention methods when provided the same annotation budget for the synthetic datasets

Method	Clusters			Moons			Circles		
	RF	XGB	MLP	RF	XGB	MLP	RF	XGB	MLP
IUPM_{RI}	0.0506	0.0485	0.0531	0.0336	0.0319	0.0341	0.0202	0.0165	0.0244
IUPM_{CEI}	0.0600	0.0656	0.0575	0.0406	0.0274	0.0233	0.0230	0.0280	0.0292
IUPM_{UI}	**0.0270**	**0.0272**	**0.0265**	**0.0244**	**0.0242**	**0.0222**	**0.0160**	**0.0157**	**0.0158**

Uncertainty Threshold In the following we discuss the effect of the uncertainty threshold introduced by our method to trigger the intervention steps. Table 9 underlines the intuitive effect that lowering the threshold will lead to a higher number of interventions and thus queried ground truth labels over time. This additional label information helps to correct the estimation, leading to a trade-off between estimation quality and human intervention. In a practical application, this trade-off may be determined by external factors, such as a limited number of possible human interventions, making this hyperparameter useful for tailoring a monitoring system to a specific use case and application.

In our case, we determined the threshold by assessing the number of cumulative interventions combined with the average gain in performance estimation quality per intervention between two steps $\frac{\Delta MAE}{\Delta n_I}$. The value added per intervention begins to saturate between 0.08 and 0.12, so that the benefit of human intervention diminishes. In addition, we believe that an average of about 13 interventions in 100 steps is still reasonable for the observed

Table 9: Evaluation of the relative benefit on performance per intervention step across intervention thresholds from 0.20 to 0.02. This ratio is calculated on the reported average intervention number and average MAE across all three synthetic datasets and evaluated models.

Method	0.20	0.18	0.16	0.14	0.12	0.10	0.08	0.06	0.04	0.02
MAE	0.0356	0.0350	0.0312	0.0264	0.0231	0.0221	0.0213	0.0211	0.0201	0.0222
n_I	5.33	6.67	7.67	8.67	11.00	12.67	16.67	23.00	33.00	57.00
$\frac{\Delta MAE}{\Delta n_I}$		4.23e-04	3.79e-03	4.82e-03	1.43e-03	5.74e-04	2.03e-04	3.62e-05	9.59e-05	-8.71e-05

improvement in the estimate. Based on these results from the synthetic experiments, we chose an uncertainty threshold of 0.1, which proved to be robust in all other experiments.

Further Results on ImageNet-c Figures 9 and 10 provide further insight into the ImageNet-c experiments in Table 4 by presenting the performance estimates and intervention steps over time.

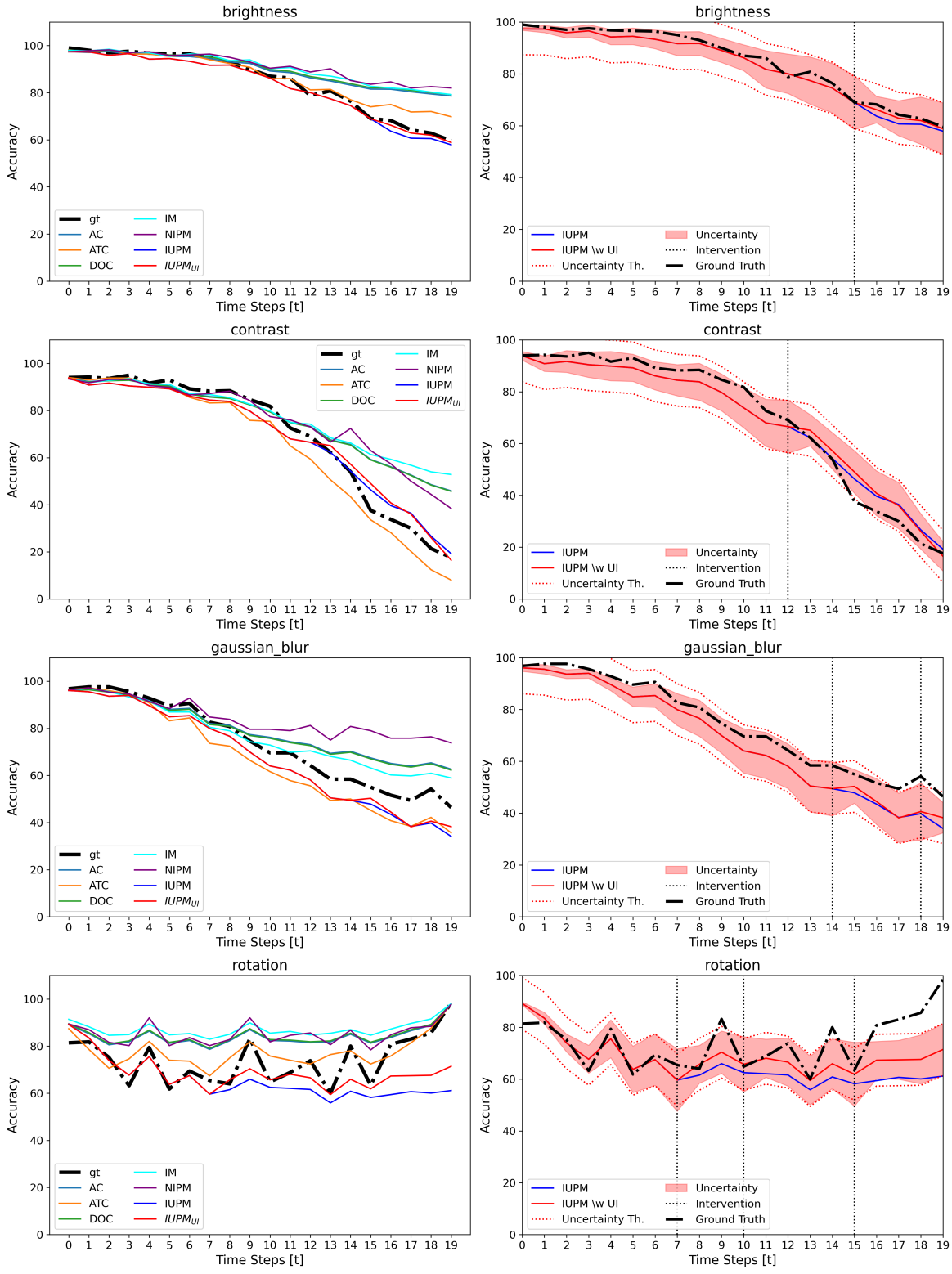

Figure 9: Results on different gradual shifts from ImageNet-c on a ResNet-50 model with 500 samples. Left: Evolution of actual ground truth performance (gt) and different performance estimation methods over time, where at each time point the shift strength increases. Right: Corresponding visualization of IUPM's inherent uncertainty measure and effect of triggered labeling interventions.

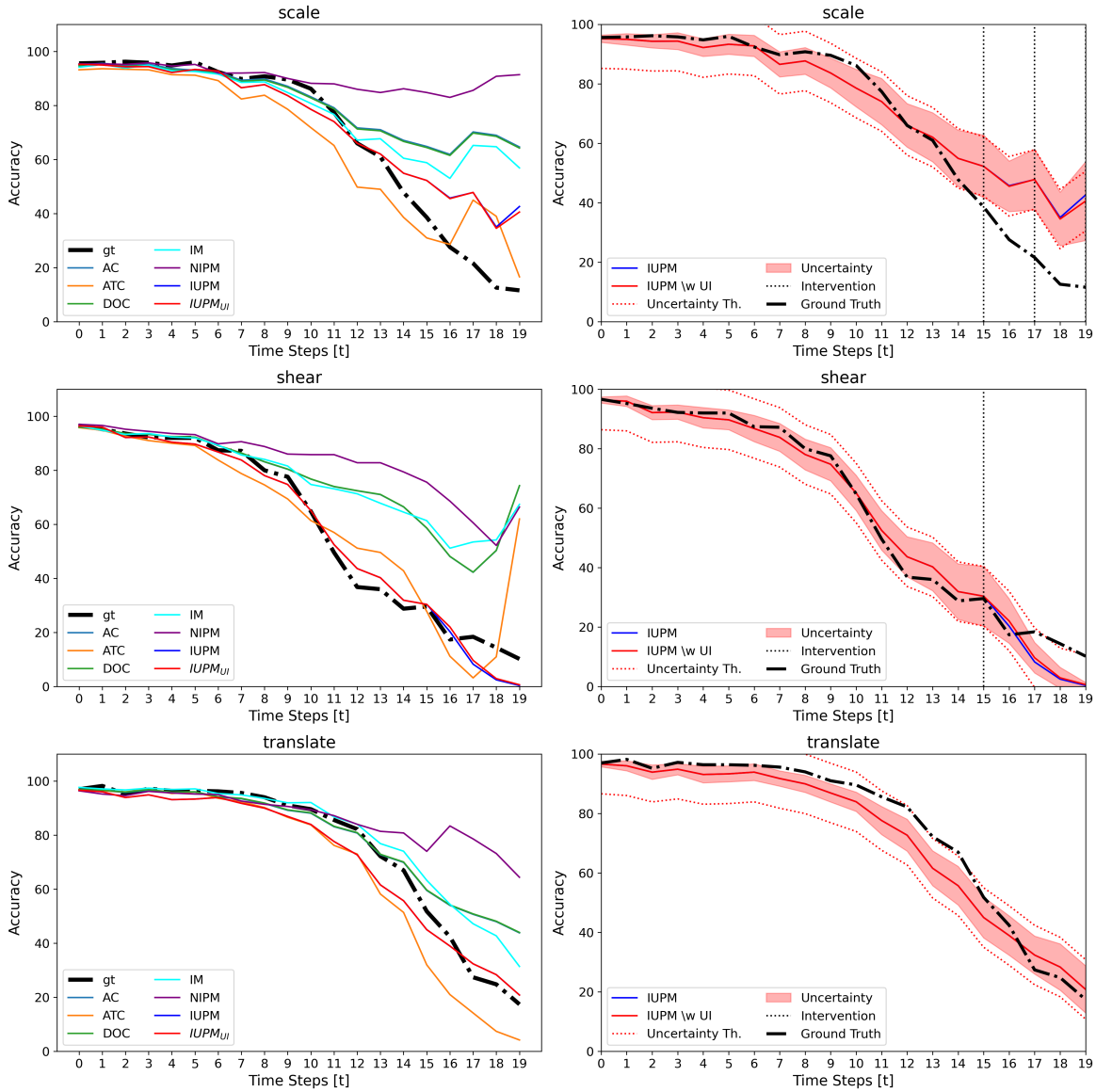

Figure 10: Results on different gradual shifts from ImageNet-c on a ResNet-50 model with 500 samples (continued). Left: Evolution of actual ground truth performance (gt) and different performance estimation methods over time, where at each time point the shift strength increases. Right: Corresponding visualization of IUPM's inherent uncertainty measure and effect of triggered labeling interventions.