# Predicting the Future Behavior of a Time-Varying Probability Distribution

Christoph H. Lampert
IST Austria (Institute of Science and Technology Austria), Am Campus 1, 3400 Klosterneuburg, Austria

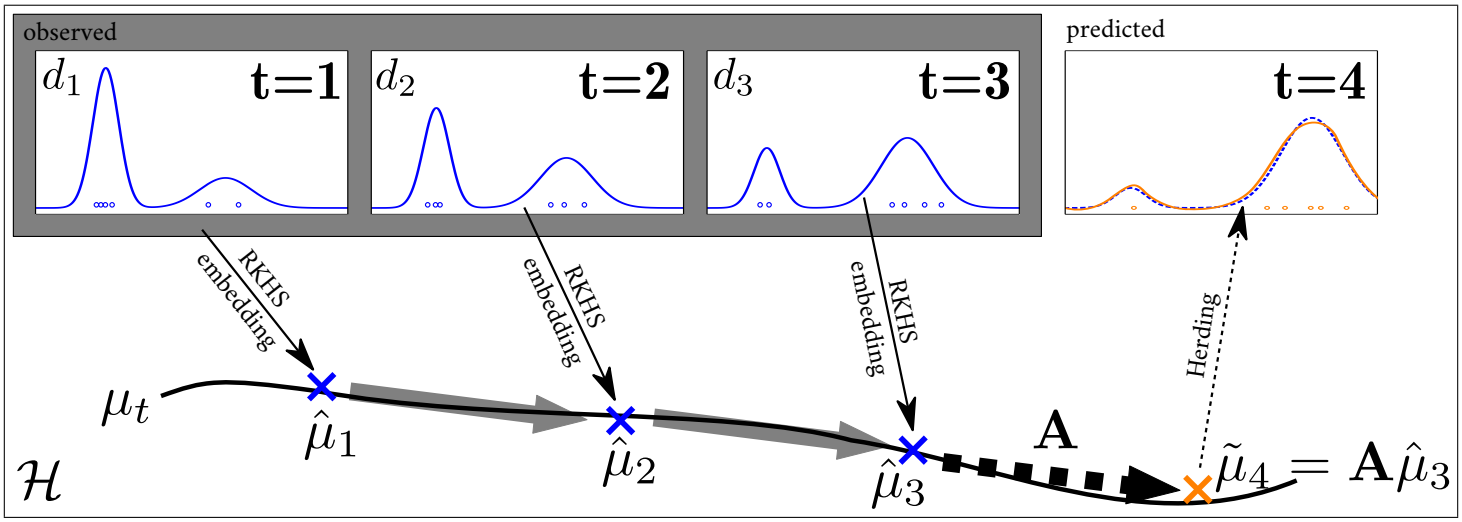


Figure 1: Schematic illustration of EDD. we observe samples sets, $S_t$ (blue dots), from a time varying probability distribution, $d_t$, at different points of time (blue curves). Using the framework of RKHS embeddings, we compute their empirical kernel mean maps, $\hat{\mu}_t = \frac{1}{|S_t|}\sum_{z\in S_t}\phi(z)$ in a Hilbert space $\mathcal{H}$. We learn an operator $\mathbf{A}: \mathcal{H} \to \mathcal{H}$ that approximates the dynamics from any $\hat{\mu}_t$ to $\hat{\mu}_{t+1}$ by vector-valued regression (thick gray arrows). By means of $\mathbf{A}$ we extrapolate the distribution dynamics beyond the last observed distribution (thick dashed arrow), thereby obtaining a prediction, $\tilde{\mu}_4$, for the embedding of the unobserved *target* distribution $d_4$ (dotted blue curve). If desired, we apply *herding* (thin dashed arrow) to produce a new sample set (orange dots) for the predicted distribution (orange curve).

It is a long lasting dream of humanity to build a machine that can predict the future. In this work we aim at making a first step towards giving computers such abilities, at least on very short time scales.

We study the situation of a time-varying probability distribution from which sample sets at different time points are observed. Our main result is a method for learning an operator that captures the dynamics of the time-varying data distribution based on two recent machine learning techniques: the embedding of probability distributions into a reproducing kernel Hilbert space and vector-valued regression. By extrapolating the learned dynamics into the future we obtain an estimate of the future distribution in form of a (potentially weighted) set of samples.

Let $\mathcal{Z}$ be a data domain and let $d_t(z)$ for $t \in \mathbb{N}$ be a time-varying data distribution over $z \in \mathcal{Z}$. At a fixed point of time, $T$, we assume that we have access to sequences of sets, $S_t = \{z_1^t, \dots, z_{n_t}^t\}$, for $t = 1, \dots, T$, that were sampled i.i.d. from the respective distributions, $d_1, \dots, d_T$. Our goal is to construct a distribution, $\tilde{d}_{T+1}$, that is as close as possible to the so far unobserved $d_{T+1}$. Optionally, we might also want to construct a set, $\tilde{S}$, of samples that are distributed approximately according to $d_{T+1}$. This in work we propose a method for *extrapolating the distribution dynamics (EDD)* that consists of four steps, see Figure 1 for an illustration.

a) We represent each sample set, $S_t$, as a vector in a Hilbert space through its *empirical (kernel mean) embedding* [4]

$$S \mapsto \hat{\mu}(S), \quad \text{for } \hat{\mu}(S) = \frac{1}{|S|}\sum_{z\in S}\phi(z), \tag{1}$$

where $\phi: \mathcal{Z} \to \mathcal{H}$ is the feature map induced by a positive definite kernel function $k: \mathcal{Z}\times\mathcal{Z} \to \mathbb{R}$ with associated reproducing kernel Hilbert space (RKHS) $\mathcal{H}$.

b) We learn an operator that reflects the dynamics between the vectors by solving the regularized vector-valued regression problem [3]

$$\min_{A\in\mathcal{F}} \sum_{t=1}^{T-1} \|\hat{\mu}_{t+1} - A\hat{\mu}_t\|_{\mathcal{H}}^2 + \lambda\|A\|_{\mathcal{F}}^2 \tag{2}$$

for $\lambda \geq 0$. The solution can be written in closed form (see [2])

$$\tilde{\mathbf{A}} = \sum_{t=1}^{T-1}\hat{\mu}_{t+1}\sum_{s=1}^{T-1} W_{ts}\langle\hat{\mu}_s, \cdot\rangle_{\mathcal{H}}, \tag{3}$$

with coefficient matrix $W = (K+\lambda I)^{-1}$, where $K \in \mathbb{R}^{(T-1)\times(T-1)}$ is the kernel matrix with entries $K_{st} = \langle\hat{\mu}_s, \hat{\mu}_t\rangle_{\mathcal{H}}$, and $I$ is the identity matrix of the same size.

c) We extrapolating the distribution dynamics by one step by applying the learned operator, $\tilde{\mathbf{A}}$, to the last vector in the sequence, $\hat{\mu}_T$. The resulting prediction, $\tilde{\mu}_{T+1} = \tilde{\mathbf{A}}\hat{\mu}_T$, can be written as a weighted linear combination of the observed distributions,

$$\tilde{\mu}_{T+1} = \sum_{t=2}^{T}\beta_t\hat{\mu}_t \quad \text{with} \quad \beta_{t+1} = \sum_{s=1}^{T-1} W_{ts}\langle\hat{\mu}_s, \hat{\mu}_T\rangle_{\mathcal{H}}, \tag{4}$$

for $t = 1, \dots, T-1$. The coefficients, $\beta_t$, can be computed from the original sample sets by means of only kernel evaluations, because $\langle\hat{\mu}_s, \hat{\mu}_t\rangle_{\mathcal{H}} = \frac{1}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t} k(z_i^s, z_j^t)$.

d) Optionally, we use kernel herding [1] to create a new sample set, $\bar{S}_{T+1}$, for the extrapolated distribution, i.e. $\hat{\mu}(\bar{S}_{T+1}) \approx \tilde{\mu}_{T+1}$.

Experiments on synthetic and real data show that EDD is in fact able to extrapolate the distribution dynamics and that this can be used for practical tasks, such as domain adaptation in situations when no training examples from the target distribution are available, not even unlabeled ones.

We acknowledge funding from the ERC under grant agreement no 308036.

[1] Yutian Chen, Max Welling, and Alex J. Smola. Super-samples from kernel herding. In *UAI*, 2010.

[2] Heng Lian. Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces. *Canadian Journal of Statistics*, 2007.

[3] Charles A. Micchelli and Massimiliano Pontil. On learning vector-valued functions. *Neural Computation*, 2005.

[4] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *ALT*, 2007.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage: http://www.cv-foundation.org/openaccess/CVPR2015.py