Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly

Benjamin Guedj, Le Li

Sequential Learning of Principal Curves: Summarizing Data Streams on the Fly

Benjamin Guedj* and Le Li†

May 8, 2019

Abstract

When confronted with massive data streams, summarizing data with dimension reduction methods such as PCA raises theoretical and algorithmic pitfalls. Principal curves act as a nonlinear generalization of PCA and the present paper proposes a novel algorithm to automatically and sequentially learn principal curves from data streams. We show that our procedure is supported by regret bounds with optimal sublinear remainder terms. A greedy local search implementation (called `slpc`, for Sequential Learning Principal Curves) that incorporates both sleeping experts and multi-armed bandit ingredients is presented, along with its regret computation and performance on synthetic and real-life data.

Keywords sequential learning, principal curves, data streams, regret bounds, greedy algorithm, sleeping experts. MSC 2010: 68T10, 62L10, 62C99.

1 Introduction

Numerous methods have been proposed in the statistics and machine learning literature to sum up information and represent data by condensed and simpler-to-understand quantities. Among those methods, Principal Component Analysis (PCA) aims at identifying the maximal variance axes of data. This serves as a way to represent data in a more compact fashion and hopefully reveal as well as possible their variability. PCA has been introduced by Pearson (1901) and Spearman (1904) and further developed by Hotelling (1933). This is one of the most widely used procedures in multivariate exploratory analysis targeting dimension reduction or features extraction. Nonetheless, PCA is a linear procedure and the need for more sophisticated nonlinear techniques has led to the notion of principal curve. Principal curves may be seen as a nonlinear generalization of the first principal component. The goal is to obtain a curve which passes "in the middle" of data, as illustrated by Figure 1. This notion of skeletonization of data clouds has been at the heart of numerous applications in many different domains, such as physics (Friedsam and Oren, 1989; Brunsdon, 2007), character and speech recognition (Reinhard and Niranjan, 1999; Kégl and Krzyżak, 2002), mapping and geology (Banfield and Raftery, 1992; Stanford and Raftery, 2000; Brunsdon, 2007), to name but a few.

*Inria and University College London – corresponding author, benjamin.guedj@inria.fr
†Université d'Angers and iAdvize

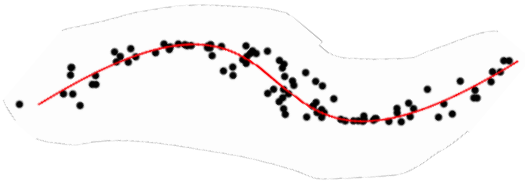


Figure 1: A principal curve.

1.1 Earlier works on principal curves

The original definition of principal curve dates back to Hastie and Stuetzle (1989). A principal curve is a smooth (C^∞) parameterized curve $\mathbf{f}(s) = (f_1(s), \ldots, f_d(s))$ in $\mathbb{R}^d$ which does not intersect itself, has finite length inside any bounded subset of $\mathbb{R}^d$ and is self-consistent. This last requirement means that $\mathbf{f}(s) = \mathbb{E}[X|s_{\mathbf{f}}(X) = s]$, where $X \in \mathbb{R}^d$ is a random vector and the so-called projection index $s_{\mathbf{f}}(x)$ is the largest real number s minimizing the squared Euclidean distance between $\mathbf{f}(s)$ and x, defined by

$$s_{\mathbf{f}}(x) = \sup\left\{s : \|x - \mathbf{f}(s)\|_2^2 = \inf_{\tau} \|x - \mathbf{f}(\tau)\|_2^2\right\}.$$

Self-consistency means that each point of $\mathbf{f}$ is the average (under the distribution of X) of all data points projected on $\mathbf{f}$, as illustrated by Figure 2. However, an unfortunate conse-

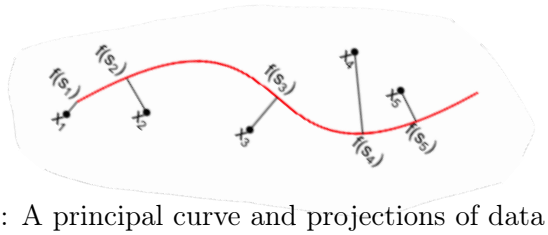


Figure 2: A principal curve and projections of data onto it.

quence of this definition is that the existence is not guaranteed in general for a particular distribution, let alone for an online sequence for which no probabilistic assumption is made. Kégl (1999) proposed a new concept of principal curves which ensures its existence for a large class of distributions. Principal curves $\mathbf{f}^\star$ are defined as the curves minimizing the expected squared distance over a class $\mathcal{F}_L$ of curves whose length is smaller than $L > 0$, namely,

$$\mathbf{f}^\star \in \arg\inf_{\mathbf{f}\in\mathcal{F}_L} \Delta(\mathbf{f}),$$

where

$$\Delta(\mathbf{f}) = \mathbb{E}\left[\Delta\left(\mathbf{f}, X\right)\right] = \mathbb{E}\left[\inf_{s} \|\mathbf{f}(s) - X\|_2^2\right].$$

If $\mathbb{E}\|X\|_2^2 < \infty$, $\mathbf{f}^\star$ always exists but may not be unique. In practical situation where only i.i.d copies $X_1, \ldots, X_n$ of X are observed, Kégl (1999) considers classes $\mathcal{F}_{k,L}$ of all polygonal lines with k segments and length not exceeding L, and chooses an estimator $\hat{\mathbf{f}}_{k,n}$ of $\mathbf{f}^\star$ as the one within $\mathcal{F}_{k,L}$ which minimizes the empirical counterpart

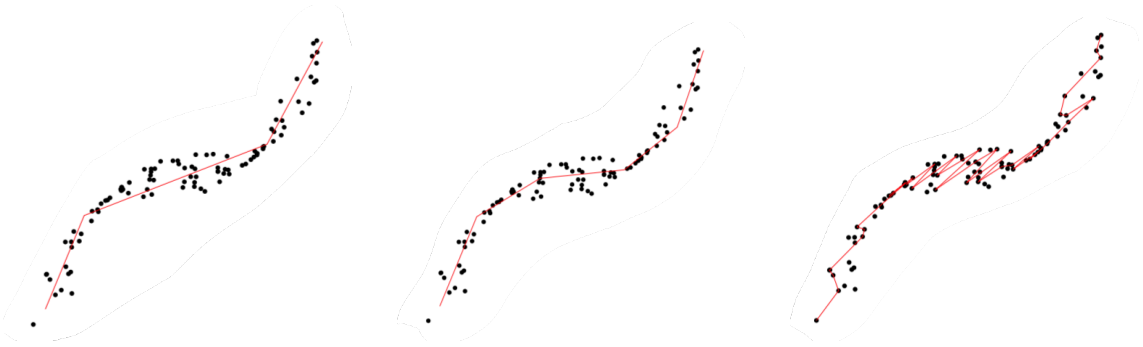
$$\Delta_n(\mathbf{f}) = \frac{1}{n}\sum_{i=1}^{n} \Delta\left(\mathbf{f}, X_i\right)$$

of $\Delta(\mathbf{f})$. It is proved in Kégl et al. (2000) that if X is almost surely bounded and $k \propto n^{1/3}$, then

$$\Delta\left(\hat{\mathbf{f}}_{k,n}\right) - \Delta\left(\mathbf{f}^\star\right) = \mathcal{O}\left(n^{-1/3}\right).$$

As the task of finding a polygonal line with k segments and length at most L that minimizes $\Delta_n(\mathbf{f})$ is computationally costly, Kégl et al. (2000) proposes the Polygonal Line algorithm. This iterative algorithm proceeds by fitting a polygonal line with k segments and considerably speeds up the exploration part by resorting to gradient descent. The two steps (projection and optimization) are similar to what is done by the k-means algorithm. However, the Polygonal Line algorithm is not supported by theoretical bounds and leads to variable performance depending on the distribution of the observations.

As the number k of segments plays a crucial role (a too small k leads to a poor summary of data while a too large k yields overfitting, see Figure 3), Biau and Fischer (2012) aim to fill the gap by selecting an optimal k from both theoretical and practical perspectives. Their approach relies strongly on the theory of model selection by penalization introduced



(a) A too small k. (b) Right k. (c) A too large k.

Figure 3: Principal curves with different number k of segments.

by Barron et al. (1999) and further developed by Birgé and Massart (2007). By considering countable classes $\{\mathcal{F}_{k,\ell}\}_{k,\ell}$ of polygonal lines with k segments, total length $\ell \leq L$ and whose vertices are on a lattice, the optimal $(\hat{k}, \hat{\ell})$ is obtained as the minimizer of the criterion

$$\mathrm{crit}(k, \ell) = \Delta_n \left(\hat{\mathbf{f}}_{k,\ell} \right) + \mathrm{pen}(k, \ell),$$

where

$$\mathrm{pen}(k, \ell) = c_0 \sqrt{\frac{k}{n}} + c_1 \frac{\ell}{n} + c_2 \frac{1}{\sqrt{n}} + \delta^2 \sqrt{\frac{w_{k,\ell}}{2n}}$$

is a penalty function where δ stands for the diameter of observations, $w_{k,\ell}$ denotes the weight attached to class $\mathcal{F}_{k,\ell}$ and with constants c_0, c_1, c_2 depending on δ, the maximum length L and the dimension of observations. Biau and Fischer (2012) then prove that

$$\mathbb{E}\left[\Delta(\hat{\mathbf{f}}_{\hat{k},\hat{\ell}}) - \Delta(\mathbf{f}^\star)\right] \leq \inf_{k,\ell} \left\{ \mathbb{E}\left[\Delta(\hat{\mathbf{f}}_{k,\ell}) - \Delta(\mathbf{f}^\star)\right] + \mathrm{pen}(k, \ell) \right\} + \frac{\delta^2 \Sigma}{2^{3/2}} \sqrt{\frac{\pi}{n}}, \tag{1}$$

where Σ is a numerical constant. The expected loss of the final polygonal line $\hat{\mathbf{f}}_{\hat{k},\hat{\ell}}$ is close to the minimal loss achievable over $\mathcal{F}_{k,\ell}$ up to a remainder term decaying as $1/\sqrt{n}$.

1.2 Motivation

The big data paradigm—where collecting, storing and analyzing massive amounts of large and complex data becomes the new standard—commands to revisit some of the classical statistical and machine learning techniques. The tremendous improvements of data acquisition infrastructures generates new continuous streams of data, rather than batch datasets.

This has drawn a large interest to sequential learning. Extending the notion of principal curves to the sequential settings opens immediate practical application possibilities. As an example, path planning for passengers' location can help taxi companies to better optimize their fleet. Online algorithms that could yield instantaneous path summarization would be adapted to the sequential nature of geolocalized data. Existing theoretical works and practical implementations of principal curves are designed for the batch setting (Kégl, 1999; Kégl et al., 2000; Kégl and Krzyżak, 2002; Sandilya and Kulkarni, 2002; Biau and Fischer, 2012) and their adaptation to the sequential setting is not a smooth process. As an example, consider the algorithm in Biau and Fischer (2012). It is assumed that vertices of principal curves are located on a lattice, and its computational complexity is of order $\mathcal{O}(nN^p)$ where n is the number of observations, N the number of points on the lattice and p the maximum number of vertices. When p is large, running this algorithm at each epoch yields a monumental computational cost. In general, if data is not identically distributed or even adversary, algorithms that originally worked well in the batch setting may not be ideal when cast onto the online setting (see Cesa-Bianchi and Lugosi, 2006, Chapter 4). To the best of our knowledge, very little effort has been put so far into extending principal curves algorithms to the sequential context (to the notable exception of Laparra and Malo, 2016, in a fairly different setting and with no theoretical results). The present paper aims at filling this gap: our goal is to propose an online perspective to principal curves by automatically and sequentially learning the best principal curve summarizing a data stream. Sequential learning takes advantage of the latest collected (set of) observations and therefore suffers a much smaller computational cost.

Sequential learning operates as follows: a blackbox reveals at each time t some deterministic value $x_t, t = 1, 2, \ldots$, and a forecaster attempts to predict sequentially the next value based on past observations (and possibly other available information). The performance of the forecaster is no longer evaluated by its generalization error (as in the batch setting) but rather by a regret bound which quantifies the cumulative loss of a forecaster in the first T rounds with respect to some reference minimal loss. In sequential learning, the velocity of algorithms may be favored over statistical precision. An immediate use of aforecited techniques (Kégl et al., 2000; Sandilya and Kulkarni, 2002; Biau and Fischer, 2012) at each time round t (treating data collected until t as a batch dataset) would result in a monumental algorithmic cost. Rather, we propose a novel algorithm which adapts to the sequential nature of data, *i.e.*, which takes advantage of previous computations.

The contributions of the present paper are twofold. We first propose a sequential principal curves algorithm, for which we derive regret bounds. We then move towards an implementation, illustrated on a toy dataset and a real-life dataset (seismic data). The sketch of our algorithm procedure is as follows. At each time round t, the number of segments of k_t is chosen automatically and the number of segments k_{t+1} in the next round is obtained by only using information about k_t and a small amount of past observations. The core of our procedure relies on computing a quantity which is linked to the mode of the so-called Gibbs quasi-posterior and is inspired by quasi-Bayesian learning. The use of quasi-Bayesian estimators is especially advocated by the PAC-Bayesian theory which originates in the machine learning community in the late 1990s, in the seminal works of Shawe-Taylor and Williamson (1997) and McAllester (1999a,b). The PAC-Bayesian theory has been successfully adapted to sequential learning problems, see for example Li et al. (2018) for online clustering.

The paper is organized as follows. Section 2 presents our notation and our online principal curve algorithm, for which we provide regret bounds with sublinear remainder

terms in Section 3. A practical implementation is proposed in Section 4 and we illustrate its performance on synthetic and real-life data sets in Section 5. Proofs to all original results claimed in the paper are collected in Section 6.

2 Notation

A parameterized curve in $\mathbb{R}^d$ is a continuous function $\mathbf{f} : I \longrightarrow \mathbb{R}^d$ where $I = [a, b]$ is a closed interval of the real line. The length of $\mathbf{f}$ is given by

$$\mathcal{L}(\mathbf{f}) = \lim_{M \to \infty} \left\{ \sup_{a=s_0<s_1<\cdots<s_M=b} \sum_{i=1}^{M} \|\mathbf{f}(s_i) - \mathbf{f}(s_{i-1})\|_2 \right\}.$$

Let $x_1, x_2, \ldots, x_T \in B(0, \sqrt{d}R) \subset \mathbb{R}^d$ be a sequence of data, where $B(\mathbf{c}, R)$ stands for the ℓ_2-ball centered in $\mathbf{c} \in \mathbb{R}^d$ with radius $R > 0$. Let Ω_δ be a grid over $B(0, \sqrt{d}R)$, *i.e.*, $\Omega_\delta = B(0, \sqrt{d}R) \cap \Gamma_\delta$ where Γ_δ is a lattice in $\mathbb{R}^d$ with spacing $\delta > 0$. Let $L > 0$ and define for each $k \in [\![1, p]\!]$ the collection $\mathcal{F}_{k,L}$ of polygonal lines $\mathbf{f}$ with k segments whose vertices are in Ω_δ and such that $\mathcal{L}(\mathbf{f}) \leq L$. Denote by $\mathcal{F}_p = \cup_{k=1}^{p} \mathcal{F}_{k,L}$ all polygonal lines with a number of segments $\leq p$, whose vertices are in Ω_δ and whose length is at most L. Finally, let $\mathcal{K}(\mathbf{f})$ denote the number of segments of $\mathbf{f} \in \mathcal{F}_p$. This strategy is illustrated by Figure 4.

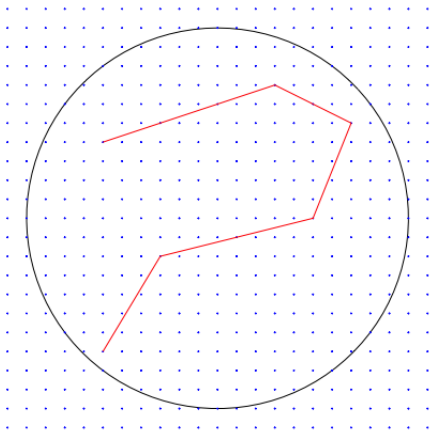


Figure 4: An example of a lattice Γ_δ in $\mathbb{R}^2$ with $\delta = 1$ (spacing between blue points) and $B(0, 10)$ (black circle). The red polygonal line is composed with vertices in $\Omega_\delta = B(0, 10) \cap \Gamma_\delta$.

Our goal is to learn a time-dependent polygonal line which passes through the "middle" of data and gives a summary of all available observations $x_1, \ldots, x_{t-1}$ (denoted by $(x_s)_{1:(t-1)}$ hereafter) before time t. Our output at time t is a polygonal line $\hat{\mathbf{f}}_t \in \mathcal{F}_p$ depending on past information $(x_s)_{1:(t-1)}$ and past predictions $(\hat{\mathbf{f}}_s)_{1:(t-1)}$. When x_t is revealed, the instantaneous loss at time t is computed as

$$\Delta\left(\hat{\mathbf{f}}_t, x_t\right) = \inf_{s \in I} \|\hat{\mathbf{f}}_t(s) - x_t\|_2^2. \tag{2}$$

In what follows, we investigate regret bounds for the cumulative loss based on (2). Given a measurable space Θ (embedded with its Borel σ-algebra), we let $\mathcal{P}(\Theta)$ denote the set of probability distributions on Θ, and for some reference measure π, we let $\mathcal{P}_\pi(\Theta)$ be the set of probability distributions absolutely continuous with respect to π.

For any $k \in [\![1, p]\!]$, let π_k denote a probability distribution on $\mathcal{F}_{k,L}$. We define the *prior* π on $\mathcal{F}_p = \cup_{k=1}^{p} \mathcal{F}_{k,L}$ as

$$\pi(\mathbf{f}) = \sum_{k \in [\![1,p]\!]} w_k \pi_k(\mathbf{f}) \mathbb{1}_{\{\mathbf{f} \in \mathcal{F}_{k,L}\}}, \quad \mathbf{f} \in \mathcal{F}_p,$$

where $w_1, \ldots, w_p \geq 0$ and $\sum_{k \in [\![1,p]\!]} w_k = 1$.

We adopt a quasi-Bayesian-flavored procedure: consider the Gibbs quasi-posterior (note that this is not a proper posterior in all generality, hence the term "quasi")

$$\hat{\rho}_t(\cdot) \propto \exp(-\lambda S_t(\cdot))\pi(\cdot),$$

where

$$S_t(\mathbf{f}) = S_{t-1}(\mathbf{f}) + \Delta(\mathbf{f}, x_t) + \frac{\lambda}{2}\big(\Delta(\mathbf{f}, x_t) - \Delta(\hat{\mathbf{f}}_t, x_t)\big)^2,$$

as advocated by Audibert (2009) and Li et al. (2018) who then consider realisations from this quasi-posterior. In the present paper, we will rather focus on a quantity linked to the mode of this quasi-posterior. Indeed, the mode of the quasi-posterior $\hat{\rho}_{t+1}$ is

$$\arg\min_{\mathbf{f} \in \mathcal{F}_p} \left\{ \underbrace{\sum_{s=1}^{t} \Delta(\mathbf{f}, x_s)}_{(i)} + \underbrace{\frac{\lambda}{2} \sum_{s=1}^{t} \big(\Delta(\mathbf{f}, x_t) - \Delta(\hat{\mathbf{f}}_t, x_t)\big)^2}_{(ii)} + \underbrace{\frac{\ln \pi(\mathbf{f})}{\lambda}}_{(iii)} \right\},$$

where *(i)* is a cumulative loss term, *(ii)* is a term controlling the variance of the prediction $\mathbf{f}$ to past predictions $\hat{\mathbf{f}}_s, s \leq t$, and *(iii)* can be regarded as a penalty function on the complexity of $\mathbf{f}$ if π is well chosen. This mode hence has a similar flavor to follow the best expert or follow the perturbed leader in the setting of prediction with experts (see Hutter and Poland, 2005 and Cesa-Bianchi and Lugosi, 2006, Chapters 3 and 4) if we consider each $\mathbf{f} \in \mathcal{F}_p$ as an expert which always delivers constant advice. These remarks yield Algorithm 1.

Algorithm 1 Sequentially learning principal curves

1: **Input parameters**: $p > 0, \eta > 0, \pi(z) = \mathrm{e}^{-z}\mathbb{1}_{\{z>0\}}$ and penalty function $h : \mathcal{F}_p \to \mathbb{R}^+$
2: **Initialization**: For each $\mathbf{f} \in \mathcal{F}_p$, draw $z_{\mathbf{f}} \sim \pi$ and $\Delta_{\mathbf{f},0} = \frac{1}{\eta}(h(\mathbf{f}) - z_{\mathbf{f}})$
3: **For** $t = 1, \ldots, T$
4: Get the data x_t
5: Obtain

$$\hat{\mathbf{f}}_t = \arg\inf_{\mathbf{f} \in \mathcal{F}_p} \left\{ \sum_{s=0}^{t-1} \Delta_{\mathbf{f},s} \right\},$$

where $\Delta_{\mathbf{f},s} = \Delta(\mathbf{f}, x_s)$, $s \geq 1$.
6: **End for**

3 Regret bounds for sequential learning of principal curves

We now present our main theoretical results.

Theorem 1. *For any sequence $(x_t)_{1:T} \in B(0, \sqrt{d}R)$, $R \geq 0$ and any penalty function $h : \mathcal{F}_p \to \mathbb{R}^+$, let $\pi(z) = \mathrm{e}^{-z}\mathbb{1}_{\{z>0\}}$. Let $0 < \eta \leq \frac{1}{d(2R+\delta)^2}$, then the procedure described in Algorithm 1 satisfies*

$$\sum_{t=1}^{T} \mathbb{E}_\pi \left[\Delta(\hat{\mathbf{f}}_t, x_t)\right] \leq (1 + c_0(\mathrm{e}-1)\eta)\, S_{T,h,\eta} + \frac{1}{\eta}\left(1 + \ln \sum_{\mathbf{f}\in\mathcal{F}_p} \mathrm{e}^{-h(\mathbf{f})}\right),$$

where $c_0 = d(2R+\delta)^2$ and

$$S_{T,h,\eta} = \inf_{k\in[\![1,p]\!]} \left\{ \inf_{\substack{\mathbf{f}\in\mathcal{F}_p \\ \mathcal{K}(\mathbf{f})=k}} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + \frac{h(\mathbf{f})}{\eta} \right\} \right\}.$$

The expectation of the cumulative loss of polygonal lines $\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_T$ is upper-bounded by the smallest penalised cumulative loss over all $k \in \{1, \ldots, p\}$ up to a multiplicative term $(1 + c_0(\mathrm{e}-1)\eta)$ which can be made arbitrarily close to 1 by choosing a small enough η. However, this will lead to both a large $h(\mathbf{f})/\eta$ in $S_{T,h,\eta}$ and a large $\frac{1}{\eta}(1 + \ln \sum_{\mathbf{f}\in\mathcal{F}_p} \mathrm{e}^{-h(\mathbf{f})})$. In addition, another important issue is the choice of the penalty function h. For each $\mathbf{f} \in \mathcal{F}_p$, $h(\mathbf{f})$ should be large enough to ensure a small $\sum_{\mathbf{f}\in\mathcal{F}_p} \mathrm{e}^{-h(\mathbf{f})}$ while not too large to avoid overpenalization and a larger value for $S_{T,h,\eta}$. We therefore set

$$h(\mathbf{f}) \geq \ln(p\mathrm{e}) + \ln \left| \{\mathbf{f} \in \mathcal{F}_p, \mathcal{K}(\mathbf{f}) = k\} \right| \tag{3}$$

for each $\mathbf{f}$ with k segments (where $|M|$ denotes the cardinality of a set M) since it leads to

$$\sum_{\mathbf{f}\in\mathcal{F}_p} \mathrm{e}^{-h(\mathbf{f})}) = \sum_{k\in[\![1,p]\!]} \sum_{\substack{\mathbf{f}\in\mathcal{F}_p \\ \mathcal{K}(\mathbf{f})=k}} \mathrm{e}^{-h(\mathbf{f})} \leq \sum_{k\in[\![1,p]\!]} \frac{1}{p\mathrm{e}} \leq \frac{1}{\mathrm{e}}.$$

The penalty function $h(\mathbf{f}) = c_1\mathcal{K}(\mathbf{f}) + c_2 L + c_3$ satisfies (3), where c_1, c_2, c_3 are constants depending on R, d, δ, p (this is proven in Lemma 3, in Section 6). We therefore obtain the following corollary.

Corollary 1. *Under the assumptions of Theorem 1, let*

$$\eta = \min\left\{ \frac{1}{d(2R+\delta)^2}, \sqrt{\frac{c_1 p + c_2 L + c_3}{c_0(\mathrm{e}-1)\inf_{\mathbf{f}\in\mathcal{F}_p} \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t)}} \right\}.$$

Then

$$\sum_{t=1}^{T} \mathbb{E}\left[\Delta(\hat{\mathbf{f}}_t, x_t)\right] \leq \inf_{k\in[\![1,p]\!]} \left\{ \inf_{\substack{\mathbf{f}\in\mathcal{F}_p \\ \mathcal{K}(\mathbf{f})=k}} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + \sqrt{c_0(\mathrm{e}-1) r_{T,k,L}} \right\} \right\}$$
$$+ \sqrt{c_0(\mathrm{e}-1) r_{T,p,L}} + c_0(\mathrm{e}-1)(c_1 p + c_2 L + c_3),$$

where $r_{T,k,L} = \inf_{\mathbf{f}\in\mathcal{F}_p} \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t)(c_1 k + c_2 L + c_3)$.

Proof. Note that

$$\sum_{t=1}^{T} \mathbb{E}\left[\Delta(\hat{\mathbf{f}}_t, x_t)\right] \leq S_{T,h,\eta} + \eta c_0(\mathrm{e}-1) \inf_{\mathbf{f}\in\mathcal{F}_p} \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + c_0(\mathrm{e}-1)(c_0 p + c_2 L + c_3),$$

and we conclude by setting

$$\eta = \sqrt{\frac{c_1 p + c_2 L + c_3}{c_0(\mathrm{e}-1)\inf_{\mathbf{f}\in\mathcal{F}_p} \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t)}}.$$

□

Sadly, Corollary 1 is not of much practical use since the optimal value for η depends on $\inf_{\mathbf{f}\in\mathcal{F}_p} \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t)$ which is obviously unknown, even more so at time $t = 0$. We therefore provide an adaptive refinement of Algorithm 1 in the following Algorithm 2.

Algorithm 2 Sequentially and adaptively learning principal curves

1: **Input parameters**: $p > 0$, $L > 0$, π, h and $\eta_0 = \frac{\sqrt{c_1 p + c_2 L + c_3}}{c_0\sqrt{\mathrm{e}-1}}$

2: **Initialization**: For each $\mathbf{f} \in \mathcal{F}_p$, draw $z_{\mathbf{f}} \sim \pi$, $\Delta_{\mathbf{f},0} = \frac{1}{\eta_0}(h(\mathbf{f}) - z_{\mathbf{f}})$ and $\hat{\mathbf{f}}_0 = \operatorname*{arg\,inf}_{\mathbf{f}\in\mathcal{F}_p} \Delta_{\mathbf{f},0}$

3: **For** $t = 1, \ldots, T$

4: Compute $\eta_t = \frac{\sqrt{c_1 p + c_2 L + c_3}}{c_0\sqrt{(\mathrm{e}-1)t}}$

5: Get data x_t and compute $\Delta_{\mathbf{f},t} = \Delta(\mathbf{f}, x_t) + \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)(h(\mathbf{f}) - z_{\mathbf{f}})$

6: Obtain

$$\hat{\mathbf{f}}_t = \operatorname*{arg\,inf}_{\mathbf{f}\in\mathcal{F}_p} \left\{ \sum_{s=0}^{t-1} \Delta_{\mathbf{f},s} \right\}. \tag{4}$$

7: **End for**

Theorem 2. *For any sequence $(x_t)_{1:T} \in B(0, \sqrt{d}R), R \geq 0$, let $h(\mathbf{f}) = c_1\mathcal{K}(\mathbf{f}) + c_2 L + c_3$ where c_1, c_2, c_3 are constants depending on $R, d, \delta, \ln p$. Let $\pi(z) = \mathrm{e}^{-z}\mathbb{1}_{\{z>0\}}$ and*

$$\eta_0 = \frac{\sqrt{c_1 p + c_2 L + c_3}}{c_0\sqrt{\mathrm{e}-1}}, \quad \eta_t = \frac{\sqrt{c_1 p + c_2 L + c_3}}{c_0\sqrt{(\mathrm{e}-1)t}},$$

where $t \geq 1$ and $c_0 = d(2R + \delta)^2$. Then the procedure described in Algorithm 2 satisfies

$$\sum_{t=1}^{T} \mathbb{E}\left[\Delta(\hat{\mathbf{f}}_t, x_t)\right] \leq \inf_{k\in[\![1,p]\!]} \left\{ \inf_{\substack{\mathbf{f}\in\mathcal{F}_p \\ \mathcal{K}(\mathbf{f})=k}} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + c_0\sqrt{(\mathrm{e}-1)T(c_1 k + c_2 L + c_3)} \right\} \right\}$$
$$+ 2c_0\sqrt{(\mathrm{e}-1)T(c_1 p + c_2 L + c_3)}.$$

The message of this regret bound is that the expected cumulative loss of polygonal lines $\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_T$ is upper-bounded by the minimal cumulative loss over all $k \in \{1, \ldots, p\}$, up to an additive term which is sublinear in T. The actual magnitude of this remainder term is $\sqrt{kT}$. When L is fixed, the number k of segments is a measure of complexity of the retained polygonal line. This bound therefore yields the same magnitude than (1) which is the most refined bound in the literature so far (Biau and Fischer, 2012, where the optimal values for k and L are obtained in a model selection fashion).

4 Implementation

The argument of the infimum in Algorithm 2 is taken over $\mathcal{F}_p = \cup_{k=1}^{p} \mathcal{F}_{k,L}$ which has a cardinality of order $|\mathcal{Q}_\delta|^p$, making any greedy search largely time-consuming. We instead turn to the following strategy: given a polygonal line $\hat{\mathbf{f}}_t \in \mathcal{F}_{k_t,L}$ with k_t segments, we consider, with a certain proportion, the availability of $\hat{\mathbf{f}}_{t+1}$ within a neighbourhood $\mathcal{U}(\hat{\mathbf{f}}_t)$ (see the formal definition below) of $\hat{\mathbf{f}}_t$. This consideration is well suited for the principal curves setting since if observation x_t is close to $\hat{\mathbf{f}}_t$, one can expect that the polygonal line which well fits observations $x_s, s = 1, \ldots, t$ lies in a neighbourhood of $\hat{\mathbf{f}}_t$. In addition, if each polygonal line $\mathbf{f}$ is regarded as an action, we no longer assume that all actions are available at all times, and allow the set of available actions to vary at each time. This is a model known as "sleeping experts (or actions)" in prior work (Auer et al., 2003; Kleinberg et al., 2008). In this setting, defining the regret with respect to the best action in the whole set of actions in hindsight remains difficult since that action might sometimes be unavailable. Hence it is natural to define the regret with respect to the best ranking of all actions in the hindsight according to their losses or rewards, and at each round one chooses among the available actions by selecting the one which ranks the highest. Kleinberg et al. (2008) introduced this notion of regret and studied both the full-information (best action) and partial-information (multi-armed bandit) settings with stochastic and adversarial rewards and adversarial action availability. They pointed out that the **EXP4** algorithm (Auer et al., 2003) attains the optimal regret in adversarial rewards case but has a runtime exponential in the number of all actions. Kanade et al. (2009) considered full and partial information with stochastic action availability and proposed an algorithm that runs in polynomial time. In what follows, we materialize our implementation by resorting to "sleeping experts" *i.e.,* a special set of available actions that adapts to the setting of principal curves.

Let σ denote an ordering of $|\mathcal{F}_p|$ actions, and $\mathcal{A}_t$ a subset of the available actions at round t. We let $\sigma(\mathcal{A}_t)$ denote the highest ranked action in $\mathcal{A}_t$. In addition, for any action $\mathbf{f} \in \mathcal{F}_p$ we define the reward $r_{\mathbf{f},t}$ of $\mathbf{f}$ at round $t, t \geq 0$ by

$$r_{\mathbf{f},t} = c_0 - \Delta(\mathbf{f}, x_t).$$

It is clear that $r_{\mathbf{f},t} \in (0, c_0)$. The convention from losses to gains is done in order to facilitate the subsequent performance analysis. The reward of an ordering σ is the cumulative reward of the selected action at each time

$$\sum_{t=1}^{T} r_{\sigma(\mathcal{A}_t),t},$$

and the reward of the best ordering is $\max_\sigma \sum_{t=0}^{T} r_{\sigma(\mathcal{A}_t),t}$ (respectively, $\mathbb{E}\left[\max_\sigma \sum_{t=1}^{T} r_{\sigma(\mathcal{A}_t),t}\right]$ when $\mathcal{A}_t$ is stochastic).

Our procedure starts with a **partition** step which aims at identifying the "relevant" neighbourhood of an observation $x \in \mathbb{R}^d$ with respect to a given polygonal line, and then proceeds with the definition of the **neighbourhood** of an action $\mathbf{f}$. We then provide the full implementation and prove a regret bound.

Partition For any polygonal line $\mathbf{f}$ with k segments, we denote by $\overrightarrow{\mathbf{V}} = (v_1, \ldots, v_{k+1})$ its vertices and by $s_i, i = 1, \ldots, k$ the line segments connecting v_i and v_{i+1}. In the sequel, we use $\mathbf{f}(\overrightarrow{\mathbf{V}})$ to represent the polygonal line formed by connecting consecutive vertices in $\overrightarrow{\mathbf{V}}$ if no confusion arises. Let $V_i, i = 1, \ldots, k+1$ and $S_i, i = 1, \ldots, k$ be the Voronoi partitions of $\mathbb{R}^d$ with respect to $\mathbf{f}$, *i.e.,* regions consisting of all points closer to vertex v_i or segment s_i. Figure 5 shows an example of Voronoi partition with respect to $\mathbf{f}$ with 3 segments.

Neighbourhood For any $x \in \mathbb{R}^d$, we define the neighbourhood $\mathcal{N}(x)$ with respect to $\mathbf{f}$ as the union of all Voronoi partitions whose closure intersects with two vertices connecting the projection $\mathbf{f}(s_{\mathbf{f}}(x))$ of x to $\mathbf{f}$. For example, for the point x in Figure 5, its neighbourhood $\mathcal{N}(x)$ is the union of S_2, V_3, S_3 and V_4. In addition, let $\mathcal{N}_t(x) = \{x_s \in \mathcal{N}(x), s = 1, \ldots, t.\}$ be the set of observations $x_{1:t}$ belonging to $\mathcal{N}(x)$ and $\bar{\mathcal{N}}_t(x)$ be its average. Let $\mathcal{D}(M) = \sup_{x,y \in M} ||x - y||_2$ denote the diameter of set $M \subset \mathbb{R}^d$. We finally define the local grid $\mathcal{Q}_{\delta,t}(x)$ of $x \in \mathbb{R}^d$ at time t as

$$\mathcal{Q}_{\delta,t}(x) = B\left(\bar{\mathcal{N}}_t(x), \mathcal{D}\left(\mathcal{N}_t(x)\right)\right) \cap \mathcal{Q}_{\delta}.$$

We can finally proceed to the definition of the neighbourhood $\mathcal{U}(\hat{\mathbf{f}}_t)$ of $\hat{\mathbf{f}}_t$. Assume $\hat{\mathbf{f}}_t$

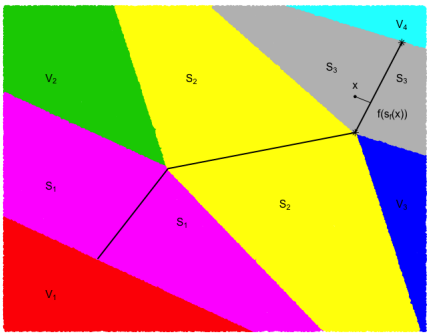


Figure 5: An example of a Voronoi partition.

has $k_t + 1$ vertices $\overrightarrow{\mathbf{V}} = (\underbrace{v_{1:i_t-1}}_{(i)}, \underbrace{v_{i_t:j_t-1}}_{(ii)}, \underbrace{v_{j_t:k_t+1}}_{(iii)})$, where vertices of (ii) belong to $\mathcal{Q}_{\delta,t}(x_t)$ while those of (i) and (iii) do not. The neighbourhood $\mathcal{U}(\hat{\mathbf{f}}_t)$ consists of $\mathbf{f}$ sharing vertices $(i), (iii)$ with $\hat{\mathbf{f}}_t$, but can be equipped with different vertices (ii) in $\mathcal{Q}_{\delta,t}(x_t)$, *i.e.*,

$$\mathcal{U}(\hat{\mathbf{f}}_t) = \left\{\mathbf{f}(\overrightarrow{\mathbf{V}}), \quad \overrightarrow{\mathbf{V}} = (v_{1:i_t-1}, v_{1:m}, v_{j_t:k_t+1})\right\},$$

where $v_{1:m} \in \mathcal{Q}_{\delta,t}(x_t)$ and m is given by

$$m = \begin{cases} j_t - i_t - 1 & \text{reduce segments by 1 unit,} \\ j_t - i_t & \text{same number of segments,} \\ j_t - i_t + 1 & \text{increase segments by 1 unit.} \end{cases}$$

In Algorithm 3, we initiate the principal curve $\hat{\mathbf{f}}_1$ as the first component line segment whose vertices are the two farthest projections of data $x_{1:t_0}$ (t_0 can be set to 2 or 3 in practice) on the first component line. The reward of $\mathbf{f}$ at round t in this setting is therefore $r_{\mathbf{f},t} = c_0 - \Delta(\mathbf{f}, x_{t_0+t})$. Algorithm 3 has an exploration phase (when $I_t = 1$) and an exploitation phase ($I_t = 0$). In the exploration phase, it is allowed to observe rewards of all actions and to choose an optimal perturbed action from the set $\mathcal{F}_p$ of all actions. In the exploitation phase, only rewards of a part of actions can be accessed and rewards of others are estimated by a constant, and we update our action from the neighbourhood $\mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right)$

Algorithm 3 A locally greedy algorithm to sequentially learn principal curves

Input parameters: $p > 0$, $R > 0$, $L > 0$, $\epsilon > 0$, $\alpha > 0$, $1 > \beta > 0$ and any penalty function h

Initialization: Given $(x_t)_{1:t_0}$, obtain $\hat{\mathbf{f}}_1$ as the first principal component

For $t = 2, \ldots, T$

Draw $I_t \sim Bernoulli(\epsilon)$ and $z_{\mathbf{f}} \sim \pi$.

Let

$$\hat{\sigma}_t = \text{sort}\left(\mathbf{f}, \quad \sum_{s=1}^{t-1} \hat{r}_{\mathbf{f},s} - \frac{1}{\eta_{t-1}} h(\mathbf{f}) + \frac{1}{\eta_{t-1}} z_{\mathbf{f}}\right),$$

i.e., sorting all $\mathbf{f} \in \mathcal{F}_p$ in descending order according to their perturbed cumulative reward till $t-1$.

If $I_t = 1$, set $\mathcal{A}_t = \mathcal{F}_p$ and $\hat{\mathbf{f}}_t = \hat{\sigma}^t(\mathcal{A}_t)$ and observe $r_{\hat{\mathbf{f}}_t,t}$

$$\hat{r}_{\mathbf{f},t} = r_{\mathbf{f},t} \quad \text{for} \quad \mathbf{f} \in \mathcal{F}_p.$$

If $I_t = 0$, set $\mathcal{A}_t = \mathcal{U}(\hat{\mathbf{f}}_{t-1})$, $\hat{\mathbf{f}}_t = \hat{\sigma}^t(\mathcal{A}_t)$ and observe $r_{\hat{\mathbf{f}}_t,t}$

$$\hat{r}_{\mathbf{f},t} = \begin{cases} \frac{r_{\mathbf{f},t}}{\mathbb{P}\left(\hat{\mathbf{f}}_t = \mathbf{f} | \mathcal{H}_t\right)} & \text{if} \quad \mathbf{f} \in \mathcal{U}(\hat{\mathbf{f}}_{t-1}) \cap cond(t) \text{ and } \hat{\mathbf{f}}_t = \mathbf{f}, \\ \alpha & \text{otherwise,} \end{cases}$$

where $\mathcal{H}_t$ denotes all the randomness before time t and $cond(t) = \left\{\mathbf{f} \in \mathcal{F}_p : \mathbb{P}\left(\hat{\mathbf{f}}_t = \mathbf{f} | \mathcal{H}_t\right) > \beta\right\}$. In particular, when $t = 1$, we set $\hat{r}_{\mathbf{f},1} = r_{\mathbf{f},1}$ for all $\mathbf{f} \in \mathcal{F}_p$, $\mathcal{U}\left(\hat{\mathbf{f}}_0\right) = \emptyset$ and $\hat{r}_{\hat{\sigma}^1(\mathcal{U}(\hat{\mathbf{f}}_0)),1} \equiv 0$.

End for

of the previous action $\hat{\mathbf{f}}_{t-1}$. This local update (or search) greatly reduces computation complexity since $|\mathcal{U}(\hat{\mathbf{f}}_{t-1})| \ll |\mathcal{F}_p|$ when p is large. In addition, this local search will be enough to account for the case when x_t locates in $\mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right)$. The parameter β needs to be carefully calibrated since it should not be too large to ensure that the condition $cond(t)$ is non-empty, otherwise all rewards are estimated by the same constant and thus lead to the same descending ordering of tuples for both $\left(\sum_{s=1}^{t-1} \hat{r}_{\mathbf{f},s}, \mathbf{f} \in \mathcal{F}_p\right)$ and $\left(\sum_{s=1}^{t} \hat{r}_{\mathbf{f},s}, \mathbf{f} \in \mathcal{F}_p\right)$. Therefore, we may face the risk of having $\hat{\mathbf{f}}_{t+1}$ in the neighbourhood of $\hat{\mathbf{f}}_t$ even if we are in the exploration phase at time $t+1$. Conversely, very small β could result in large bias for the estimation $\frac{r_{\mathbf{f},t}}{\mathbb{P}\left(\hat{\mathbf{f}}_t = \mathbf{f} | \mathcal{H}_t\right)}$ of $r_{\mathbf{f},t}$. Note that the exploitation phase is close yet different to the label efficient prediction (Cesa-Bianchi et al., 2005, Remark 1.1) since we allow an action at time t to be different from the previous one. Neu and Bartók (2013) have proposed the *Geometric Resampling* method to estimate the conditional probability $\mathbb{P}\left(\hat{\mathbf{f}}_t = \mathbf{f} | \mathcal{H}_t\right)$ since this quantity often does not have an explicit form. However, due to the simple exponential distribution of $z_{\mathbf{f}}$ chosen in our case, an explicit form of $\mathbb{P}\left(\hat{\mathbf{f}}_t = \mathbf{f} | \mathcal{H}_t\right)$ is straightforward.

Theorem 3. *Assume that* $p > 6$, $T \geq 2|\mathcal{F}_p|^2$ *and let* $\beta = |\mathcal{F}_p|^{-\frac{1}{2}} T^{-\frac{1}{4}}$, $\alpha = \frac{c_0}{\beta}$, $\hat{c}_0 = \frac{2c_0}{\beta}$, $\epsilon = 1 - |\mathcal{F}_p|^{\frac{1}{2}-\frac{3}{p}} T^{-\frac{1}{4}}$ *and*

$$\eta_1 = \eta_2 = \cdots = \eta_T = \frac{\sqrt{c_1 p + c_2 L + c_3}}{\sqrt{T(e-1)\hat{c}_0}}.$$

Then the procedure described in Algorithm 3 satisfies the regret bound

$$\sum_{t=1}^{T} \mathbb{E}\left[\Delta\left(\hat{\mathbf{f}}_t, x_t\right)\right] \leq \inf_{\mathbf{f}\in\mathcal{F}_p} \mathbb{E}\left[\sum_{t=1}^{T} \Delta\left(\mathbf{f}, t\right)\right] + \mathcal{O}(T^{\frac{3}{4}}).$$

The proof of Theorem 3 is presented in Section 6. The regret is upper bounded by a term of order $\left(|\mathcal{F}_p|^{\frac{1}{2}} T^{\frac{3}{4}}\right)$, sublinear in T. The term $(1-\epsilon)c_0 T = c_0 |\mathcal{F}_p|^{\frac{1}{2}} T^{\frac{3}{4}}$ is the price to pay for the local search (with a proportion $1-\epsilon$) of polygonal line $\hat{\mathbf{f}}_t$ in the neighbourhood of the previous $\hat{\mathbf{f}}_{t-1}$. If $\epsilon = 1$, we would have that $\hat{c}_0 = c_0$ and the last two terms in the first inequality of Theorem 3 would vanish, hence the upper bound reduces to Theorem 2. In addition, our algorithm achieves an order that is smaller (from the perspective of both the number $|\mathcal{F}_p|$ of all actions and the total rounds T) than Kanade et al. (2009) since at each time, the availability of actions for our algorithm can be either the whole action set or a neighbourhood of the previous action while Kanade et al. (2009) consider at each time only partial and independent stochastic available set of actions generated from a predefined distribution.

5 Numerical experiments

We illustrate the performance of Algorithm 3 on synthetic and real-life data. Our implementation (hereafter denoted by `slpc` – Sequential Learning of Principal Curves) is conducted with the R language and thus our most natural competitor is the R package `princurve` (which is the algorithm from Hastie and Stuetzle, 1989). We let $p = 20$, $R = \max_{t=1,\ldots,T} ||x||_2/\sqrt{d}$, $L = 0.01p\sqrt{d}R$. The spacing δ of the lattice is ajusted with respect to data scale.

Synthetic data We generate a data set $\left\{x_t \in \mathbb{R}^2, t = 1, \ldots, 100\right\}$ uniformly along the curve $y = 0.05 \times (x-5)^3$, $x \in [0, 10]$. Table 1 shows the regret for the ground truth (sum of squared distances of all points to the true curve), `princurve` (sum of squared distances between observation $t+1$ and fitted `princurve` trained on all past t observations) and `slpc` ($\sum_{t=0}^{T-1} \Delta(\hat{\mathbf{f}}_{t+1}, x_{t+1})$). `slpc` greatly outperforms `princurve` on this example, as illustrated by Figure 8 and Figure 9.

ground truth	`princurve`	`slpc`
0.945 (0)	25.387 (0)	9.893 (0.246)

Table 1: Regret (cumulative loss) on synthetic data (average over 10 trials, with standard deviation in brackets). `princurve` is deterministic, hence the zero standard deviation.

Synthetic data in high dimension We also apply our algorithm on a data set $\left\{x_t \in \mathbb{R}^6, t = 1, 2, \ldots, 200\right\}$ in higher dimension. It is generated uniformly along a parametric curve whose coordinates are

$$\begin{pmatrix} 0.5t\cos(t) \\ 0.5t\sin(t) \\ 0.5t \\ -t \\ \sqrt{t} \\ 2\ln(t+1) \end{pmatrix}$$

where t takes 100 equidistant values in $[0, 2\pi]$. To the best of our knowledge, Hastie and Stuetzle (1989), Kégl (1999) and Biau and Fischer (2012) only tested their algorithm on

2-dimensional data. This example aims at illustrating that our algorithm also works on higher dimensional data. Table 2 shows the regret for the ground truth, `princurve` and `slpc`. In addition, Figure 6 shows the behaviour of `slpc` (green) on each dimension.

ground truth	`princurve`	`slpc`
3.290 (0)	14.204 (0)	6.797 (0.409)

Table 2: Regret (cumulative loss) on synthetic data in higher dimension (average over 10 trials, with standard deviation in brackets). `princurve` is deterministic.

(a) `slpc`, $t = 199$, 1st and 2nd coordinates

(b) `slpc`, $t = 199$, 3th and 5th coordinates

(c) `slpc`, $t = 199$, 4th and 6th coordinates

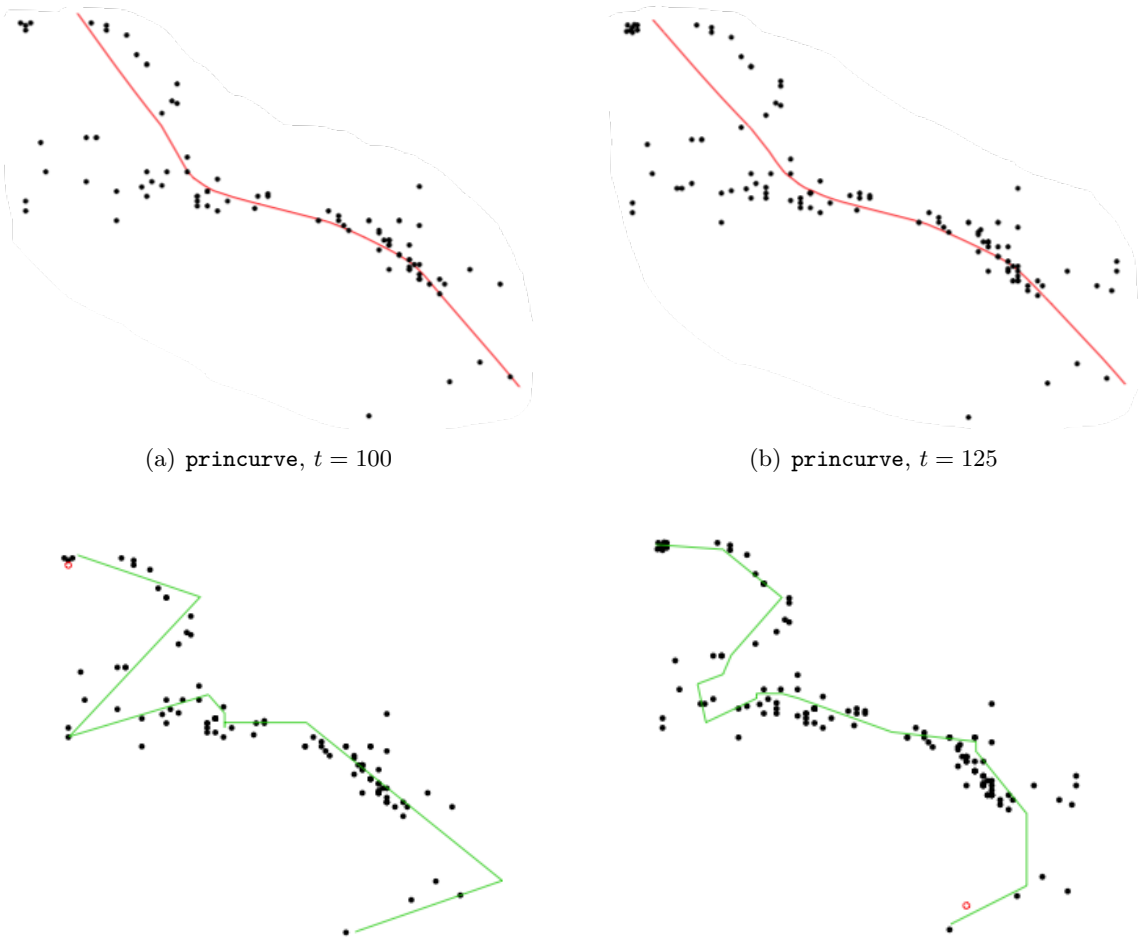
Figure 6: `slpc` (green line) on synthetic data in higher dimension from different perspectives. Black dots represent recordings $x_{1:99}$, the red dot is the new recording x_{200}.

Seismic data Seismic data spanning long periods of time are essential for a thorough understanding of earthquakes. The "Centennial Earthquake Catalog" (Engdahl and Villaseñor, 2002) aims at providing a realistic picture of the seismicity distribution on Earth. It consists in a global catalog of locations and magnitudes of instrumentally recorded earthquakes from 1900 to 2008. We focus on a particularly representative seismic active zone (a lithospheric border close to Australia) whose longitude is between E130° to E180° and latitude between S70° to N30°, with $T = 218$ seismic recordings. As shown in Figure 7, `slpc` recovers nicely the tectonic plate boundary. Lastly, since no ground truth is available, we use the R^2 coefficient to assess the performance (residuals are replaced by the squared distance between data points and their projections onto the principal curve). The average over 10 trials is 0.990.

Back to the synthetic data setting Figure 8 presents the predicted principal curve $\hat{\mathbf{f}}_{t+1}$ for both `princurve` (red) and `slpc` (green). The output of `princurve` yields a curve which does not pass in "the middle of data" but rather bends towards the curvature of the data cloud: `slpc` does not suffer from this behavior. To better illustrate the way `slpc` works between two epochs, Figure 9 focuses on the impact of collecting a new data point on the principal curve. We see that only a local vertex is impacted, whereas the rest of the principal curve remains unaltered. This cutdown in algorithmic complexity is one the key assets of `slpc`.

Back to seismic data Figure 10 is taken from the USGS website[1] and gives the global earthquakes locations on the period 1900–1999. The seismic data (latitude, longitude,

[1] https://earthquake.usgs.gov/data/centennial/



(a) `princurve`, $t = 100$ (b) `princurve`, $t = 125$

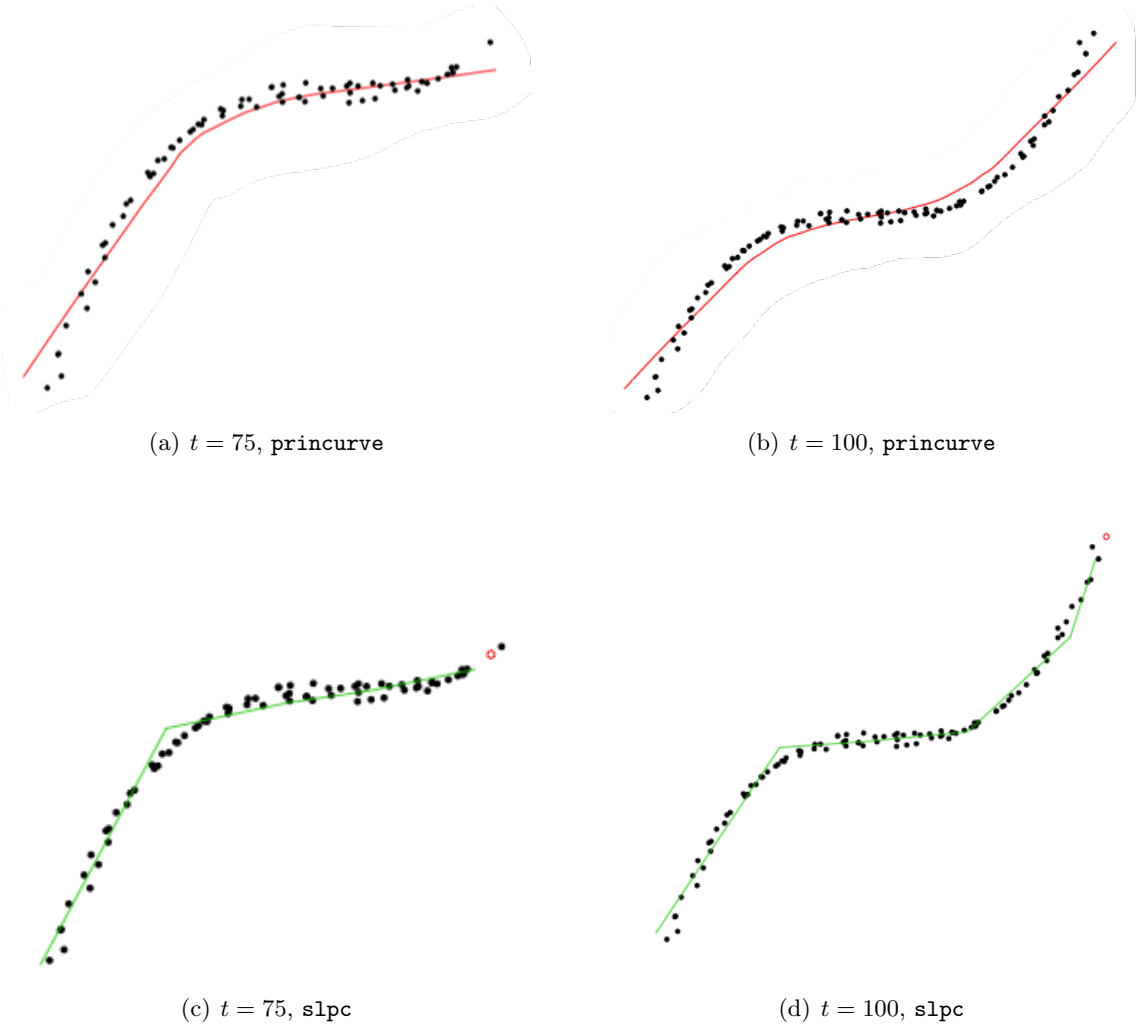
(c) `slpc`, $t = 100$ (d) `slpc`, $t = 125$

Figure 7: Seismic data. Black dots represent seismic recordings $x_{1:t}$, red dot is the new recording x_{t+1}.

magnitude of earthquakes, *etc.*) used in the present paper may be downloaded from this website.

Daily commute data The identification of segments of personal daily commuting trajectories can help taxi or bus companies to optimise their fleets and increase frequencies on segments with high commuting activity. Sequential principal curves appear to be an ideal tool to address this learning problem: we test our algorithm on trajectory data from the University of Illinois at Chicago [2]. The data is obtained from the GPS reading systems carried by two of the lab members during their daily commute for 6 months in the Cook county and the Dupage county of Illinois. Figure 11 presents the learning curves yielded by `princurve` and `slpc` on geolocalization data for the first person, on May 30 in the data set. A particularly remarkable asset of `slpc` is that abrupt curvature in the data sequence is perfectly captured, whereas `princurve` does not enjoy the same flexibility. Again, we use the R^2 coefficient to assess the performance (where residuals are replaced by the squared distance between data points and their projections onto the principal curve). The average

[2] https://www.cs.uic.edu/∼boxu/mp2p/gps_data.html



(a) $t = 75$, `princurve` (b) $t = 100$, `princurve`

(c) $t = 75$, `slpc` (d) $t = 100$, `slpc`

Figure 8: Synthetic data - Black dots represent data $x_{1:t}$, red point is the new observation x_{t+1}. `princurve` (solid red) and `slpc` (solid green).

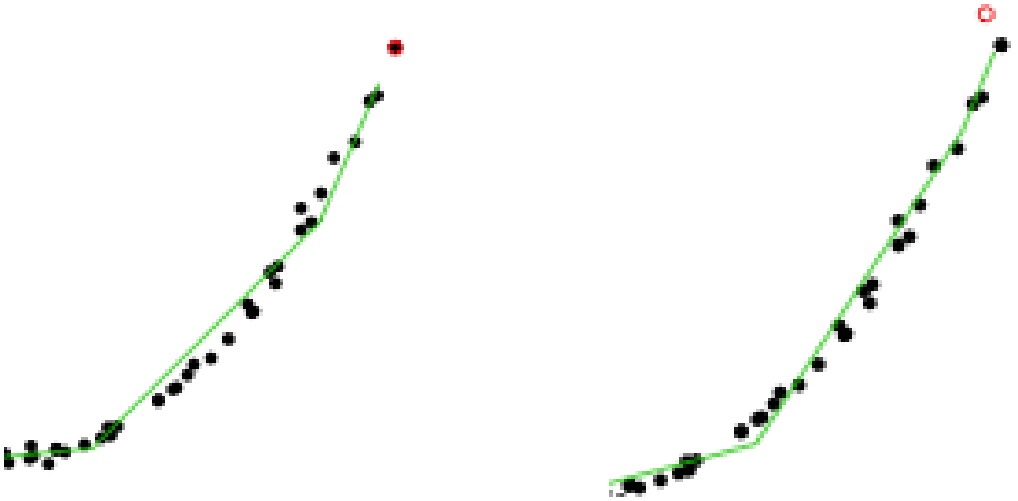
over 10 trials is 0.998.

6 Proofs

This section contains the proof of Theorem 2 (note that Theorem 1 is a straightforward consequence, with $\eta_t = \eta$, $t = 0, \ldots, T$) and the proof of Theorem 3 (which involves intermediary lemmas). Let us first define for each $t = 0, \ldots, T$ the following forecaster sequence $(\hat{\mathbf{f}}_t^\star)_t$

$$\hat{\mathbf{f}}_0^\star = \operatorname*{arg\,inf}_{\mathbf{f} \in \mathcal{F}_p} \{\Delta_{\mathbf{f},0}\} = \operatorname*{arg\,inf}_{\mathbf{f} \in \mathcal{F}_p} \left\{ \frac{1}{\eta_0} h(\mathbf{f}) - \frac{1}{\eta_0} z_{\mathbf{f}} \right\},$$

$$\hat{\mathbf{f}}_t^\star = \operatorname*{arg\,inf}_{\mathbf{f} \in \mathcal{F}_p} \left\{ \sum_{s=0}^{t} \Delta_{\mathbf{f},s} \right\} = \operatorname*{arg\,inf}_{\mathbf{f} \in \mathcal{F}_p} \left\{ \sum_{s=1}^{t} \Delta(\mathbf{f}, x_s) + \frac{1}{\eta_{t-1}} h(\mathbf{f}) - \frac{1}{\eta_{t-1}} z_{\mathbf{f}} \right\}, \quad t \geq 1.$$



(a) At time $t = 97$ (b) And at time $t = 98$

Figure 9: Synthetic data - Zooming in: how a new data point impacts only locally the principal curve.

Note that $\hat{\mathbf{f}}_t^\star$ is an "illegal" forecaster since it peeks into the future. In addition, denote by

$$\mathbf{f}^\star = \operatorname*{arg\,inf}_{\mathbf{f} \in \mathcal{F}_p} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + \frac{1}{\eta_T} h(\mathbf{f}) \right\}$$

the polygonal line in $\mathcal{F}_p$ which minimizes the cumulative loss in the first T rounds plus a penalty term. $\mathbf{f}^\star$ is deterministic while $\hat{\mathbf{f}}_t^\star$ is a random quantity (since it depends on $z_{\mathbf{f}}$, $\mathbf{f} \in \mathcal{F}_p$ drawn from π). If several $\mathbf{f}$ attain the infimum, we choose $\mathbf{f}_T^\star$ as the one having the smallest complexity. We now enunciate the first (out of three) intermediary technical result.

Lemma 1. *For any sequence $x_1, \ldots, x_T$ in $B(0, \sqrt{d}R)$,*

$$\sum_{t=0}^{T} \Delta_{\hat{\mathbf{f}}_t^\star, t} \leq \sum_{t=0}^{T} \Delta_{\hat{\mathbf{f}}_T^\star, t}, \qquad \pi\text{-almost surely.} \tag{5}$$

Proof. Proof by induction on T. Clearly (5) holds for $T = 0$. Assume that (5) holds for $T - 1$:

$$\sum_{t=0}^{T-1} \Delta_{\hat{\mathbf{f}}_t^\star, t} \leq \sum_{t=0}^{T-1} \Delta_{\hat{\mathbf{f}}_{T-1}^\star, t}.$$

Adding $\Delta_{\hat{\mathbf{f}}_T^\star, T}$ to both sides of the above inequality concludes the proof. □

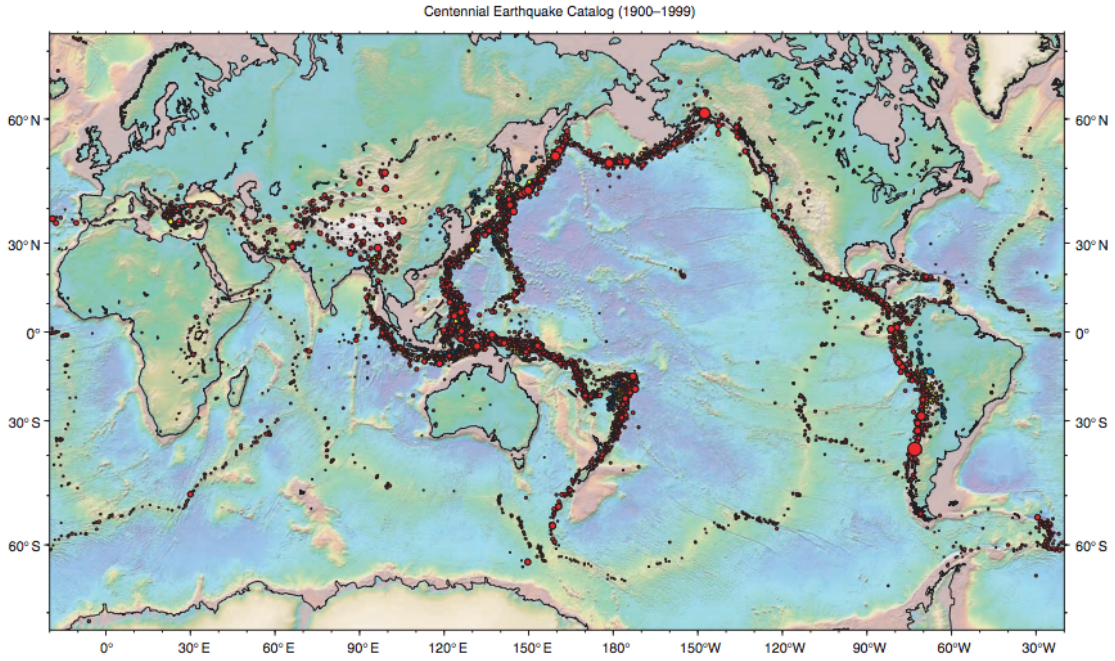


Figure 10: Seismic data from https://earthquake.usgs.gov/data/centennial/

By (5) and the definition of $\hat{\mathbf{f}}_T^\star$, for $k \geq 1$, we have π-almost surely that

$$\begin{aligned}
\sum_{t=1}^{T} \Delta(\hat{\mathbf{f}}_t^\star, x_t) &\leq \sum_{t=1}^{T} \Delta(\hat{\mathbf{f}}_T^\star, x_t) + \frac{1}{\eta_T} h(\hat{\mathbf{f}}_T^\star) - \frac{1}{\eta_T} Z_{\hat{\mathbf{f}}_T^\star} + \sum_{t=0}^{T} \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right) \left(h(\hat{\mathbf{f}}_t^\star) - Z_{\hat{\mathbf{f}}_t^\star}\right) \\
&\leq \sum_{t=1}^{T} \Delta(\mathbf{f}^\star, x_t) + \frac{1}{\eta_T} h(\mathbf{f}^\star) - \frac{1}{\eta_T} Z_{\mathbf{f}^\star} + \sum_{t=0}^{T} \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right) \left(h(\hat{\mathbf{f}}_t^\star) - Z_{\hat{\mathbf{f}}_t^\star}\right) \\
&= \inf_{\mathbf{f} \in \mathcal{F}_p} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + \frac{1}{\eta_T} h(\mathbf{f}) \right\} - \frac{1}{\eta_T} Z_{\mathbf{f}^\star} + \sum_{t=0}^{T} \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right) \left(h(\hat{\mathbf{f}}_t^\star) - Z_{\hat{\mathbf{f}}_t^\star}\right),
\end{aligned}$$

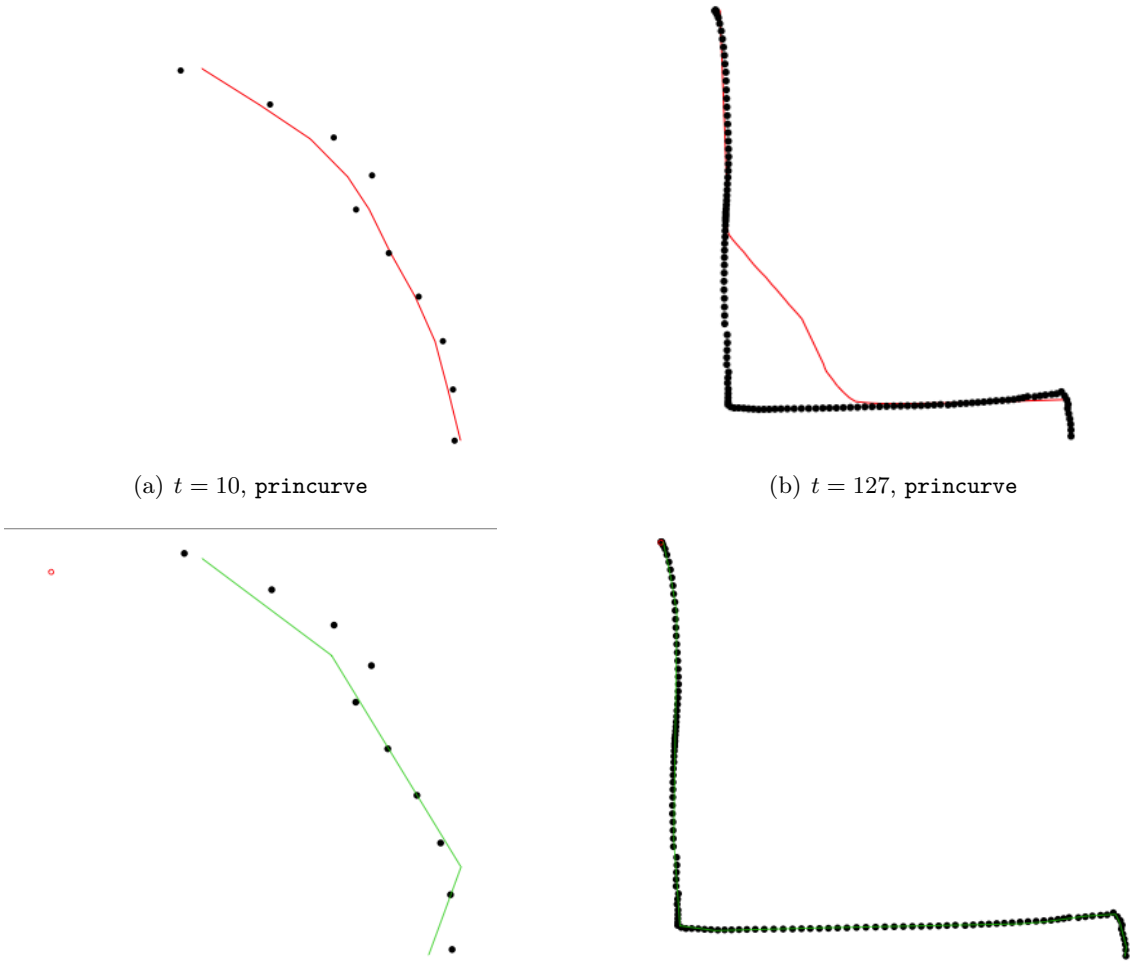
where $1/\eta_{-1} = 0$ by convention. The second and third inequality is due to respectively the definition of $\hat{\mathbf{f}}_T^\star$ and $\mathbf{f}_T^\star$. Hence

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \Delta\left(\hat{\mathbf{f}}_t^\star, x_t\right)\right] &\leq \inf_{\mathbf{f} \in \mathcal{F}_p} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + \frac{1}{\eta_T} h(\mathbf{f}) \right\} - \frac{1}{\eta_T} \mathbb{E}[Z_{\mathbf{f}_T^\star}] + \sum_{t=0}^{T} \mathbb{E}\left[\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \left(-h(\hat{\mathbf{f}}_t^\star) + Z_{\hat{\mathbf{f}}_t^\star}\right)\right] \\
&\leq \inf_{\mathbf{f} \in \mathcal{F}_p} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + \frac{1}{\eta_T} h(\mathbf{f}) \right\} + \sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) \mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_p} \left(-h(\mathbf{f}) + Z_{\mathbf{f}}\right)\right] \\
&= \inf_{\mathbf{f} \in \mathcal{F}_p} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + \frac{1}{\eta_T} h(\mathbf{f}) \right\} + \frac{1}{\eta_T} \mathbb{E}\left[\sup_{\mathbf{f} \in \mathcal{F}_p} \left(-h(\mathbf{f}) + Z_{\mathbf{f}}\right)\right],
\end{aligned}$$

where the second inequality is due to $\mathbb{E}[Z_{\mathbf{f}_T^\star}] = 0$ and $\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) > 0$ for $t = 0, 1, \ldots, T$ since η_t is decreasing in t in Theorem 2. In addition, for $y \geq 0$, one has

$$\mathbb{P}\left(-h(\mathbf{f}) + Z_{\mathbf{f}} > y\right) = \mathrm{e}^{-h(\mathbf{f}) - y}.$$

Hence, for any $y \geq 0$



(a) $t = 10$, `princurve` (b) $t = 127$, `princurve`

(c) $t = 10$, `slpc` (d) $t = 127$, `slpc`

Figure 11: Daily commute data - Black dots represent collected locations $x_{1:t}$, red point is the new observation x_{t+1}. `princurve` (solid red) and `slpc` (solid green).

$$\mathbb{P}\left(\sup_{\mathbf{f}\in\mathcal{F}_p}\left(-h(\mathbf{f})+Z_{\mathbf{f}}\right)>y\right)\leq\sum_{\mathbf{f}\in\mathcal{F}_p}\mathbb{P}\left(Z_{\mathbf{f}}\geq h(\mathbf{f})+y\right)=\sum_{\mathbf{f}\in\mathcal{F}_p}\mathrm{e}^{-h(\mathbf{f})}\mathrm{e}^{-y}=u\mathrm{e}^{-y},$$

where $u = \sum_{\mathbf{f}\in\mathcal{F}_p} \mathrm{e}^{-h(\mathbf{f})}$. Therefore, we have

$$\begin{aligned}
\mathbb{E}\left[\sup_{\mathbf{f}\in\mathcal{F}_p}(-h(\mathbf{f})+Z_{\mathbf{f}}) - \ln u\right] &\leq \mathbb{E}\left[\max\left(0, \sup_{\mathbf{f}\in\mathcal{F}_p}(-h(\mathbf{f})+Z_{\mathbf{f}} - \ln u)\right)\right] \\
&\leq \int_0^\infty \mathbb{P}\left(\max\left(0, \sup_{\mathbf{f}\in\mathcal{F}_p}(-h(\mathbf{f})+Z_{\mathbf{f}} - \ln u)\right) > y\right)\mathrm{d}y \\
&\leq \int_0^\infty \mathbb{P}\left(\sup_{\mathbf{f}\in\mathcal{F}_p}(-h(\mathbf{f})+Z_{\mathbf{f}}) > y + \ln u\right)\mathrm{d}y \\
&\leq \int_0^\infty u\mathrm{e}^{-(y+\ln u)}\mathrm{d}y = 1.
\end{aligned}$$

We thus obtain

$$\mathbb{E}\left[\sum_{t=1}^T \Delta\left(\hat{\mathbf{f}}_t^\star, x_t\right)\right] \leq \inf_{\mathbf{f}\in\mathcal{F}_p}\left\{\sum_{t=1}^T \Delta(\mathbf{f}, x_t) + \frac{1}{\eta_T}h(\mathbf{f})\right\} + \frac{1}{\eta_T}\left(1 + \ln\sum_{\mathbf{f}\in\mathcal{F}_p}\mathrm{e}^{-h(\mathbf{f})}\right). \tag{6}$$

Next, we control the regret of Algorithm 2.

Lemma 2. *Assume that $z_{\mathbf{f}}$ is sampled from the symmetric exponential distribution in $\mathbb{R}$, i.e., $\pi(z) = \mathrm{e}^{-z}\mathbb{1}_{\{z>0\}}$. Assume that $\sup_{t=1,\dots,T}\eta_{t-1} \leq \frac{1}{d(2R+\delta)^2}$, and define $c_0 = d(2R+\delta)^2$. Then for any sequence $(x_t) \in B(0, \sqrt{d}R)$, $t = 1, \dots, T$,*

$$\sum_{t=1}^T \mathbb{E}\left[\Delta\left(\hat{\mathbf{f}}_t, x_t\right)\right] \leq \sum_{t=1}^T (1 + \eta_{t-1}c_0(\mathrm{e}-1))\,\mathbb{E}\left[\Delta\left(\hat{\mathbf{f}}_t^\star, x_t\right)\right]. \tag{7}$$

Proof. Let us denote by

$$F_t(Z_{\mathbf{f}}) = \Delta\left(\hat{\mathbf{f}}_t, x_t\right) = \Delta\left(\operatorname*{arg\,inf}_{\mathbf{f}\in\mathcal{F}}\left(\sum_{s=1}^{t-1}\Delta(\mathbf{f}, x_s) + \frac{1}{\eta_{t-1}}h(\mathbf{f}) - \frac{1}{\eta_{t-1}}Z_{\mathbf{f}}\right), x_t\right)$$

the instantaneous loss suffered by the polygonal line $\hat{\mathbf{f}}_t$ when x_t is obtained. We have

$$\begin{aligned}
\mathbb{E}[\Delta\left(\hat{\mathbf{f}}_t^\star, x_t\right)] &= \int F_t\left(z - \eta_{t-1}\Delta\left(\mathbf{f}, x_t\right)\right)\pi(z)\mathrm{d}z \\
&= \int F_t(z)\pi\left(z + \eta_{t-1}\Delta(\mathbf{f}, x_t)\right)\mathrm{d}z \\
&= \int F_t(z)\mathrm{e}^{-(z+\eta_{t-1}\Delta(\mathbf{f}, x_t))}\mathrm{d}z \\
&\geq \mathrm{e}^{-\eta_{t-1}d(2R+\delta)^2}\int F_t(z)\mathrm{e}^{-z}\mathrm{d}z \\
&= \mathrm{e}^{-\eta_{t-1}d(2R+\delta)^2}\mathbb{E}[\Delta\left(\hat{\mathbf{f}}_t, x_t\right)],
\end{aligned}$$

where the inequality is due to the fact that $\Delta(\mathbf{f}, x) \leq d(2R+\delta)^2$ holds uniformly for any $\mathbf{f} \in \mathcal{F}_p$ and $x \in B(0, \sqrt{d}R)$. Finally, summing on t on both sides and using the elementary inequality $\mathrm{e}^x \leq 1 + (\mathrm{e}-1)x$ if $x \in (0,1)$ concludes the proof. □

Lemma 3. *For $k \in [\![1, p]\!]$, we control the cardinality of set $\left\{\mathbf{f} \in \mathcal{F}_p, \mathcal{K}(\mathbf{f}) = k\right\}$ as*

$$\ln\left|\left\{\mathbf{f} \in \mathcal{F}_p, \mathcal{K}(\mathbf{f}) = k\right\}\right| \leq \left(\ln(8peV_d) + 3d^{\frac{3}{2}} - d\right) k + \left(\frac{\ln 2}{\delta\sqrt{d}} + \frac{d}{\delta}\right) L + d \ln\left(\frac{\sqrt{d}(2R+\delta)}{\delta}\right)$$
$$\triangleq c_1 k + c_2 L + c_3,$$

where V_d denotes the volume of the unit ball in $\mathbb{R}^d$.

Proof. First, let $N_{k,\delta}$ denote the set of polygonal lines with k segments and whose vertices are in $\mathcal{Q}_\delta$. Notice that $N_{k,\delta}$ is different from $\{\mathbf{f} \in \mathcal{F}_p, \mathcal{K}(\mathbf{f}) = k\}$ and that

$$\left|\{\mathbf{f} \in \mathcal{F}_p, \mathcal{K}(\mathbf{f}) = k\}\right| \leq \binom{p}{k} \left|N_{k,\delta}\right|.$$

Hence

$$\begin{aligned}
\ln\left|\{\mathbf{f} \in \mathcal{F}_p, \mathcal{K}(\mathbf{f}) = k\}\right| &\leq \ln\binom{p}{k} + \ln\left|N_{k,\delta}\right| \\
&\leq k \ln\frac{pe}{k} + k\left(\ln 8V_d + 3d^{\frac{3}{2}} - d\right) + \left(\frac{\ln 2}{\sqrt{d}\delta} + \frac{d}{\delta}\right) L + d\ln\left(\frac{\sqrt{d}(2R+\delta)}{\delta}\right) \\
&\leq k\ln(pe) + k\left(\ln 8V_d + 3d^{\frac{3}{2}} - d\right) + \left(\frac{\ln 2}{\sqrt{d}\delta} + \frac{d}{\delta}\right) L + d\ln\left(\frac{\sqrt{d}(2R+\delta)}{\delta}\right),
\end{aligned}$$

where the second inequality is a consequence to the elementary inequality $\binom{p}{k} \leq \left(\frac{pe}{k}\right)^k$ combined with Lemma 2 in Kégl (1999). □

We now have all the ingredients to prove Theorem 1 and Theorem 2.

First, combining (6) and (7) yields that

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\Delta(\hat{\mathbf{f}}_t, x_t)\right] &\leq \inf_{\mathbf{f}\in\mathcal{F}_p}\left\{\sum_{t=1}^{T}\Delta(\mathbf{f}, x_t) + \frac{1}{\eta_T}h(\mathbf{f})\right\} + \frac{1}{\eta_T}\left(\frac{1}{2} + \ln\sum_{\mathbf{f}\in\mathcal{F}_p}\mathrm{e}^{-h(\mathbf{f})}\right) \\
&\quad + c_0(\mathrm{e}-1)\sum_{t=1}^{T}\eta_{t-1}\mathbb{E}\left[\Delta(\hat{\mathbf{f}}_t^\star, x_t)\right] \\
&\leq \inf_{k\in[\![1,p]\!]}\left\{\inf_{\substack{\mathbf{f}\in\mathcal{F}_p \\ \mathcal{K}(\mathbf{f})=k}}\left\{\sum_{t=1}^{T}\Delta(\mathbf{f}, x_t) + \frac{h(\mathbf{f})}{\eta_T}\right\}\right\} + \frac{1}{\eta_T}\left(\frac{1}{2} + \ln\sum_{\mathbf{f}\in\mathcal{F}_p}\mathrm{e}^{-h(\mathbf{f})}\right) \\
&\quad + c_0(\mathrm{e}-1)\sum_{t=1}^{T}\eta_{t-1}\mathbb{E}\left[\Delta(\hat{\mathbf{f}}_t^\star, x_t)\right].
\end{aligned}$$

Assume that $\eta_t = \eta$, $t = 0, \ldots, T$ and $h(\mathbf{f}) = c_1\mathcal{K}(\mathbf{f}) + c_2 L + c_3$ for $\mathbf{f} \in \mathcal{F}_p$, then $\left(\frac{1}{2} + \sum_{\mathbf{f}\in\mathcal{F}_p}\mathrm{e}^{-h(\mathbf{f})}\right) \leq 0$ and moreover

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\Delta(\hat{\mathbf{f}}_t, x_t)\right] &\leq S_{T,h,\eta} + \frac{1}{\eta}\left(\frac{1}{2} + \ln\sum_{\mathbf{f}\in\mathcal{F}_p}\mathrm{e}^{-h(\mathbf{f})}\right) + c_0(\mathrm{e}-1)\eta\sum_{t=1}^{T}\mathbb{E}\left[\Delta(\hat{\mathbf{f}}_t^\star, x_t)\right] \\
&\leq S_{T,h,\eta} + c_0(\mathrm{e}-1)\eta S_{T,h,\eta} \\
&\leq S_{T,h,\eta} + \eta c_0(\mathrm{e}-1)\inf_{\mathbf{f}\in\mathcal{F}_p}\sum_{t=1}^{T}\Delta(\mathbf{f}, x_t) + c_0(\mathrm{e}-1)(c_1 p + c_2 L + c_3),
\end{aligned}$$

where

$$S_{T,h,\eta} = \inf_{k\in[\![1,p]\!]} \left\{ \inf_{\substack{\mathbf{f}\in\mathcal{F}_p \\ \mathcal{K}(\mathbf{f})=k}} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + \frac{h(\mathbf{f})}{\eta} \right\} \right\}$$

and the second inequality is obtained with Lemma 1. By setting

$$\eta = \sqrt{\frac{c_1 p + c_2 L + c_3}{c_0(\mathrm{e}-1)\inf_{\mathbf{f}\in\mathcal{F}_p} \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t)}}$$

we obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}\left[\Delta(\hat{\mathbf{f}}_t, x_t)\right] \\
&\leq \inf_{k\in[\![1,p]\!]} \left\{ \inf_{\substack{\mathbf{f}\in\mathcal{F}_p \\ \mathcal{K}(\mathbf{f})=k}} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + \sqrt{c_0(\mathrm{e}-1) r_{T,k,L}} \right\} \right\} + \sqrt{c_0(\mathrm{e}-1) L_{T,p,L}} + c_0(\mathrm{e}-1)c_1 p + c_2 L + c_3,
\end{aligned}$$

where $r_{T,k,L} = \inf_{\mathbf{f}\in\mathcal{F}_p} \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t)(c_1 k + c_2 L + c_3)$. This proves Theorem 1.

Finally, assume that

$$\eta_0 = \frac{\sqrt{c_1 p + c_2 L + c_3}}{c_0\sqrt{(\mathrm{e}-1)}} \quad \text{and} \quad \eta_t = \frac{\sqrt{c_1 p + c_2 L + c_3}}{c_0\sqrt{(\mathrm{e}-1)t}}, \qquad t = 1, \ldots, T.$$

Since $\mathbb{E}\left[\Delta(\hat{\mathbf{f}}_t^\star, x_t)\right] \leq c_0$ for any $t = 1, \ldots, T$, we have

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[\Delta(\hat{\mathbf{f}}_t, x_t)\right] &\leq \inf_{k\in[\![1,p]\!]} \left\{ \inf_{\substack{\mathbf{f}\in\mathcal{F}_p \\ \mathcal{K}(\mathbf{f})=k}} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + \frac{h(\mathbf{f})}{\eta_T} \right\} \right\} + \frac{1}{\eta_T} \left(1 + \ln \sum_{\mathbf{f}\in\mathcal{F}_p} \mathrm{e}^{-h(\mathbf{f})} \right) \\
&\quad + c_0^2(\mathrm{e}-1) \sum_{t=1}^{T} \eta_{t-1} \\
&\leq \inf_{k\in[\![1,p]\!]} \left\{ \inf_{\substack{\mathbf{f}\in\mathcal{F}_p \\ \mathcal{K}(\mathbf{f})=k}} \left\{ \sum_{t=1}^{T} \Delta(\mathbf{f}, x_t) + c_0\sqrt{(\mathrm{e}-1)T(c_0 k + c_2 L + c_3)} \right\} \right\} \\
&\quad + 2c_0\sqrt{(\mathrm{e}-1)T(c_0 p + c_2 L + c_3)},
\end{aligned}$$

which concludes the proof of Theorem 2.

Lemma 4. *Using Algorithm 3, if* $0 < \epsilon \leq 1$, $0 < \beta < 1$, $\alpha \geq \frac{(1-\beta)c_0}{\beta}$ *and* $\left|\mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right)\right| \geq 2$ *for all* $t \geq 2$, *where* $\left|\mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right)\right|$ *is the cardinality of* $\mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right)$, *then we have*

$$\sum_{t=1}^{T} \mathbb{E}\left[r_{\hat{\mathbf{f}}_t, t}\right] \geq \sum_{t=1}^{T} \mathbb{E}\left[\hat{r}_{\hat{\sigma}^t(\mathcal{A}_t), t}\right] - 2(1-\epsilon)\alpha\beta \sum_{t=1}^{T} \left|\mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right)\right|.$$

Proof. First notice that $\mathcal{A}_t = \mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right)$ if $I_t = 0$, and that for $t \geq 2$

$$\begin{aligned}
\mathbb{E}\left[r_{\hat{\mathbf{f}}_t,t}\middle|\mathcal{H}_t, I_t = 0\right] =& \mathbb{E}\left[r_{\hat{\sigma}^t(\mathcal{A}_t),t}\middle|\mathcal{H}_t, I_t = 0\right] \\
=& \sum_{\mathbf{f}\in\mathcal{A}_t\cap cond(t)} r_{\mathbf{f},t}\mathbb{P}\left(\hat{\sigma}^t\left(\mathcal{A}_t\right) = \mathbf{f}\middle|\mathcal{H}_t\right) + \sum_{\mathbf{f}\in\mathcal{A}_t\cap cond(t)^c} r_{\mathbf{f},t}\mathbb{P}\left(\hat{\sigma}^t\left(\mathcal{A}_t\right) = \mathbf{f}\middle|\mathcal{H}_t\right) \\
\geq& \sum_{\mathbf{f}\in\mathcal{A}_t\cap cond(t)} r_{\mathbf{f},t} + \sum_{\mathbf{f}\in\mathcal{A}_t\cap cond(t)^c} \alpha\mathbb{P}\left(\hat{\sigma}^t\left(\mathcal{A}_t\right) = \mathbf{f}\middle|\mathcal{H}_t\right) \\
& - (1-\beta)\sum_{\mathbf{f}\in\mathcal{A}_t\cap cond(t)} r_{\mathbf{f},t} - \sum_{\mathbf{f}\in\mathcal{A}_t\cap cond(t)^c} (\alpha - r_{\mathbf{f},t})\,\mathbb{P}\left(\hat{\sigma}^t\left(\mathcal{A}_t\right) = \mathbf{f}\middle|\mathcal{H}_t\right) \\
=& \mathbb{E}\left[\hat{r}_{\hat{\sigma}^t(\mathcal{A}_t),t}\middle|\mathcal{H}_t, I_t = 0\right] - (1-\beta)\sum_{\mathbf{f}\in\mathcal{A}_t\cap cond(t)} r_{\mathbf{f},t} \\
& - \sum_{\mathbf{f}\in\mathcal{A}_t\cap cond(t)^c} (\alpha - r_{\mathbf{f},t})\,\mathbb{P}\left(\hat{\sigma}^t\left(\mathcal{A}_t\right) = \mathbf{f}\middle|\mathcal{H}_t\right) \\
\geq& \mathbb{E}\left[\hat{r}_{\hat{\sigma}^t(\mathcal{A}_t),t}\middle|\mathcal{H}_t, I_t = 0\right] - (1-\beta)c_0\left|\mathcal{A}_t\right| - \alpha\beta\left|\mathcal{A}_t\right| \\
\geq& \mathbb{E}\left[\hat{r}_{\hat{\sigma}^t(\mathcal{A}_t),t}\middle|\mathcal{H}_t, I_t = 0\right] - 2\alpha\beta\left|\mathcal{A}_t\right|,
\end{aligned}$$

where $cond(t)^c$ denotes the complement of set $cond(t)$. The first inequality above is due to the assumption that for all $\mathbf{f} \in \mathcal{A}_t \cap cond(t)$, we have $\mathbb{P}\left(\hat{\sigma}^t\left(\mathcal{A}_t\right) = \mathbf{f}\middle|\mathcal{H}_t\right) \geq \beta$. For $t = 1$, the above inequality is trivial since $\hat{r}_{\hat{\sigma}^1(\mathcal{U}(\hat{\mathbf{f}}_0)),1} \equiv 0$ by its definition. Hence, for $t \geq 1$, one has

$$\begin{aligned}
\mathbb{E}\left[r_{\hat{\mathbf{f}}_t,t}\middle|\mathcal{H}_t\right] &= \epsilon\mathbb{E}\left[r_{\hat{\sigma}^t(\mathcal{F}_p),t}\middle|\mathcal{H}_t, I_t = 1\right] + (1-\epsilon)\mathbb{E}\left[r_{\hat{\sigma}^t(\mathcal{A}_t),t}\middle|\mathcal{H}_t, I_t = 0\right] \\
&\geq \mathbb{E}\left[\hat{r}_{\hat{\mathbf{f}}_t,t}\middle|\mathcal{H}_t\right] - 2\alpha\beta\left|\mathcal{A}_t\right|. \qquad (8)
\end{aligned}$$

Summing on both sides of inequality (8) over t terminates the proof of Lemma 4. □

Lemma 5. *Let* $\hat{c}_0 = \frac{c_0}{\beta} + \alpha$. *If* $0 < \eta_1 = \eta_2 = \cdots = \eta_T = \eta < \frac{1}{\hat{c}_0}$, *then we have*

$$\mathbb{E}\left[\max_{\hat{\sigma}}\left\{\sum_{t=1}^{T}\hat{r}_{\hat{\sigma}(\mathcal{A}_t),t} - \frac{1}{\eta}h\left(\hat{\sigma}\left(\mathcal{A}_t\right)\right)\right\}\right] - \sum_{t=1}^{T}\mathbb{E}\left[\hat{r}_{\hat{\sigma}^t(\mathcal{A}_t),t}\right] \leq \hat{c}_0^2(e-1)\eta T + \hat{c}_0(e-1)\left(c_1 p + c_2 L + c_3\right).$$

Proof. By the definition of $\hat{r}_{\mathbf{f},t}$ in Algorithm 3, for any $\mathbf{f} \in \mathcal{F}_p$ and $t \geq 1$, we have

$$\hat{r}_{\mathbf{f},t} \leq \max\left\{\frac{r_{\mathbf{f},t}}{\mathbb{P}\left(\hat{\mathbf{f}}_t = \mathbf{f}\middle|\mathcal{H}_t\right)}, \alpha, r_{\mathbf{f},t}\right\} \leq \max\left\{\frac{c_0}{\beta}, \alpha\right\} \leq \hat{c}_0,$$

where in the second inequality we use that $r_{\mathbf{f},t} \leq c_0$ for all $\mathbf{f}$ and t, and that $\mathbb{P}\left(\hat{\mathbf{f}}_t = \mathbf{f}\middle|\mathcal{H}_t\right) \geq \beta$ when $\mathbf{f} \in \mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right) \cap cond(t)$. The rest of the proof is similar to those of Lemma 1 and

Lemma 2. In fact, if we define by $\hat{\Delta}(\mathbf{f}, x_t) = \hat{c}_0 - \hat{r}_{\mathbf{f},t}$, then one can easily observe the following relation when $I_t = 1$ (similar relation in the case that $I_t = 0$)

$$\begin{aligned}\hat{\mathbf{f}}_t = \hat{\sigma}^t(\mathcal{F}_p) &= \arg\max_{\mathbf{f}\in\mathcal{F}_p}\left\{\sum_{s=1}^{t-1}\hat{r}_{\mathbf{f},s} + \frac{1}{\eta}(z_{\mathbf{f}} - h(\mathbf{f}))\right\} \\ &= \arg\min_{\mathbf{f}\in\mathcal{F}_p}\left\{\sum_{s=1}^{t-1}\hat{\Delta}(\mathbf{f}, x_s) + \frac{1}{\eta}(h(\mathbf{f}) - z_{\mathbf{f}})\right\}.\end{aligned}$$

Then applying Lemma 1 and Lemma 2 on this newly defined sequence $\hat{\Delta}\left(\hat{\mathbf{f}}_t, x_t\right), t = 1, \dots T$ leads to the result of Lemma 5.

□

The proof of the upcoming Lemma 6 requires the following submartingale inequality: let $Y_0, \dots Y_T$ be a sequence of random variable adapted to random events $\mathcal{H}_0, \dots, \mathcal{H}_T$ such that for $1 \le t \le T$, the following three conditions hold

$$\mathbb{E}[Y_t|H_t] \le 0, \quad \mathrm{Var}(Y_t|H_t) \le a^2, \quad Y_t - \mathbb{E}[Y_t|H_t] \le b.$$

Then for any $\lambda > 0$,

$$\mathbb{P}\left(\sum_{t=1}^{T} Y_t > Y_0 + \lambda\right) \le \exp\left(-\frac{\lambda^2}{2T(a^2+b^2)}\right).$$

The proof can be found in Chung and Lu (2006, Theorem 7.3).

Lemma 6. *Assume that* $0 < \beta < \frac{1}{|\mathcal{F}_p|}, \alpha \ge \frac{c_0}{\beta}$ *and* $\eta > 0$*, then we have*

$$\begin{aligned}&\mathbb{E}\left[\max_{\sigma}\left\{\sum_{t=1}^{T} r_{\sigma(\mathcal{A}_t),t} - \frac{1}{\eta}h(\sigma(\mathcal{A}_t))\right\}\right] - \mathbb{E}\left[\max_{\hat{\sigma}}\left\{\sum_{t=1}^{T}\hat{r}_{\hat{\sigma}(\mathcal{A}_t),t} - \frac{1}{\eta}h(\hat{\sigma}(\mathcal{A}_t))\right\}\right] \\ &\qquad \le (1 - |\mathcal{F}_p|\beta)\sqrt{2T\left[\frac{c_0^2}{\beta} + \alpha^2(1-\beta) + (c_0+2\alpha)^2\right]\ln\left(\frac{1}{\beta}\right)} + |\mathcal{F}_p|\beta c_0 T.\end{aligned}$$

Proof. First, we have almost surely that

$$\max_{\sigma}\left\{\sum_{t=1}^{T} r_{\sigma(\mathcal{A}_t),t} - \frac{1}{\eta}h(\sigma(\mathcal{A}_t))\right\} - \max_{\hat{\sigma}}\left\{\sum_{t=1}^{T}\hat{r}_{\hat{\sigma}(\mathcal{A}_t),t} - \frac{1}{\eta}h(\hat{\sigma}(\mathcal{A}_t))\right\} \le \max_{\mathbf{f}\in\mathcal{F}_p}\sum_{t=1}^{T}(r_{\mathbf{f},t} - \hat{r}_{\mathbf{f},t}).$$

Denote by $Y_{\mathbf{f},t} = r_{\mathbf{f},t} - \hat{r}_{\mathbf{f},t}$. Since

$$\mathbb{E}\left[\hat{r}_{\mathbf{f},t}\middle|\mathcal{H}_t\right] = \begin{cases} r_{\mathbf{f},t} + (1-\epsilon)\alpha\left(1 - \mathbb{P}\left(\hat{\mathbf{f}}_t = \mathbf{f}|\mathcal{H}_t\right)\right) & \text{if} \quad \mathbf{f} \in \mathcal{U}(\hat{\mathbf{f}}_{t-1}) \cap cond(t), \\ \epsilon r_{\mathbf{f},t} + (1-\epsilon)\alpha & \text{otherwise,}\end{cases}$$

and $\alpha > c_0 \ge r_{\mathbf{f},t}$ uniformly for any $\mathbf{f}$ and t, then we have uniformly that $\mathbb{E}[Y_t|\mathcal{H}_t] \le 0$, hence satisfying the first condition.

For the second condition, if $\mathbf{f} \in \mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right) \cap cond(t)$, then

$$\begin{aligned}
\mathrm{Var}(Y_t|\mathcal{H}_t) =& \mathbb{E}\left[\hat{r}^2_{\mathbf{f},t}|\mathcal{H}_t\right] - \left(\mathbb{E}\left[\hat{r}_{\mathbf{f},t}|\mathcal{H}_t\right]\right)^2 \\
\leq& \epsilon r^2_{\mathbf{f},t} + (1-\epsilon)\left[\frac{r^2_{\mathbf{f},t}}{\mathbb{P}\left(\hat{\mathbf{f}}_t = \mathbf{f}|\mathcal{H}_t\right)} + \alpha\left(1 - \mathbb{P}\left(\hat{\mathbf{f}}_t = \mathbf{f}|\mathcal{H}_t\right)\right)\right] \\
& - \left[r_{\mathbf{f},t} + (1-\epsilon)\alpha\left(1 - \mathbb{P}\left(\hat{\mathbf{f}}_t = \mathbf{f}|\mathcal{H}_t\right)\right)\right]^2 \\
\leq& \frac{r^2_{\mathbf{f},t}}{\beta} + \alpha^2(1-\beta) \leq \frac{c_0^2}{\beta} + \alpha^2(1-\beta).
\end{aligned}$$

Similarly, for $\mathbf{f} \notin \mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right) \cap cond(t)$, one can have $\mathrm{Var}(Y_t|\mathcal{H}_t) \leq \alpha^2$. Moreover, for the third condition, since

$$\mathbb{E}\left[Y_{\mathbf{f},t}|\mathcal{H}_t\right] \geq -2\alpha,$$

then

$$Y_{\mathbf{f},t} - \mathbb{E}\left[Y_{\mathbf{f},t}|\mathcal{H}_t\right] \leq r_{\mathbf{f},t} + 2\alpha \leq c_0 + 2\alpha.$$

Setting $\lambda = \sqrt{2T\left[\frac{c_0^2}{\beta} + \alpha^2(1-\beta) + (c_0+2\alpha)^2\right]\ln\left(\frac{1}{\beta}\right)}$ leads to

$$\mathbb{P}\left(\sum_{t=1}^{T} Y_{\mathbf{f},t} \geq \lambda\right) \leq \beta.$$

Hence the following inequality holds with probability $1 - \left|\mathcal{F}_p\right|\beta$

$$\max_{\mathbf{f}\in\mathcal{F}_p} \sum_{t=1}^{T} \left(r_{\mathbf{f},t} - \hat{r}_{\mathbf{f},t}\right) \leq \sqrt{2T\left[\frac{c_0^2}{\beta} + \alpha^2(1-\beta) + (c_0+2\alpha)^2\right]\ln\left(\frac{1}{\beta}\right)}.$$

Finally, noticing that $\max_{\mathbf{f}\in\mathcal{F}_p}\sum_{t=1}^{T}\left(r_{\mathbf{f},t} - \hat{r}_{\mathbf{f},t}\right) \leq c_0 T$ almost surely, we terminate the proof of Lemma 6.

□

Proof of Theorem 3. Assume that $p > 6$, $T \geq 2|\mathcal{F}_p|^2$ and let

$$\beta = |\mathcal{F}_p|^{-\frac{1}{2}} T^{-\frac{1}{4}}, \qquad \alpha = \frac{c_0}{\beta}, \qquad \hat{c}_0 = \frac{2c_0}{\beta},$$
$$\eta_1 = \eta_2 = \cdots = \eta_T = \frac{\sqrt{c_1 p + c_2 L + c_3}}{\sqrt{T(e-1)}\hat{c}_0}, \qquad \epsilon = 1 - |\mathcal{F}_p|^{\frac{1}{2}-\frac{3}{p}} T^{-\frac{1}{4}}.$$

With those values, the assumptions of Lemma 4, Lemma 5 and Lemma 6 are satisfied.

Combining their results lead to the following

$$\begin{aligned}
\sum_{t=1}^{T} \mathbb{E}\left[r_{\hat{\mathbf{f}}_t,t}\right] \geq & \mathbb{E}\left[\max_{\sigma}\left\{\sum_{t=1}^{T} r_{\sigma(\mathcal{A}_t),t} - \frac{1}{\eta} h\left(\sigma\left(\mathcal{A}_t\right)\right)\right\}\right] - 2\alpha\beta(1-\epsilon)\sum_{t=1}^{T}\left|\mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right)\right| \\
& - \hat{c}_0^2(e-1)\eta T - \hat{c}_0(e-1)\left(c_1 p + c_2 L + c_3\right) \\
& - \left(1 - |\mathcal{F}_p|\,\beta\right)\sqrt{2T\left[\frac{c_0^2}{\beta} + \alpha^2(1-\beta) + (c_0+2\alpha)^2\right]\ln\left(\frac{1}{\beta}\right)} - |\mathcal{F}_p|\,\beta c_0 T \\
\geq & \mathbb{E}\left[\max_{\sigma}\left\{\sum_{t=1}^{T} r_{\sigma(\mathcal{A}_t),t} - \frac{1}{\eta} h\left(\sigma\left(\mathcal{A}_t\right)\right)\right\}\right] - (1-\epsilon)\,|\mathcal{F}_p|^{\frac{3}{p}}\, c_0 T \\
& - \hat{c}_0^2(e-1)\eta T - \hat{c}_0(e-1)\left(c_1 p + c_2 L + c_3\right) \\
& - \left(1 - |\mathcal{F}_p|\,\beta\right)\sqrt{2T\left[\frac{c_0^2}{\beta} + \alpha^2(1-\beta) + (c_0+2\alpha)^2\right]\ln\left(\frac{1}{\beta}\right)} - |\mathcal{F}_p|\,\beta c_0 T \\
\geq & \mathbb{E}\left[\max_{\sigma}\left\{\sum_{t=1}^{T} r_{\sigma(\mathcal{A}_t),t} - \frac{1}{\eta} h\left(\sigma\left(\mathcal{A}_t\right)\right)\right\}\right] - \mathcal{O}\left(|\mathcal{F}_p|^{\frac{1}{2}}\, T^{\frac{3}{4}}\right),
\end{aligned}$$

where the second inequality is due to the fact that the cardinality $\left|\mathcal{U}\left(\hat{\mathbf{f}}_{t-1}\right)\right|$ is upper bounded by $|\mathcal{F}_p|^{\frac{3}{p}}$ for $t \geq 1$. In addition, using the definition of $r_{\mathbf{f},t}$ that $r_{\mathbf{f},t} = c_0 - \Delta(\mathbf{f}, x_t)$ terminates the proof of Theorem 3. □

References

J.-Y. Audibert. Fast learning rates in statistical inference through aggregation. *The Annals of Statistics*, 37(4):1591–1646, 2009. 6

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1):48–77, 2003. 9

J. D. Banfield and A. E. Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *Journal of the American Statistical Association*, 87(417):7–16, 1992. 1

A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999. 3

G. Biau and A. Fischer. Parameter selection for principal curves. *IEEE Transactions on Information Theory*, 58(3):1924–1939, 2012. 3, 4, 8, 12

L. Birgé and P. Massart. Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*, 183:33–73, 2007. 3

C. Brunsdon. Path estimation from GPS tracks. In *Proceedings of the 9th International Conference on GeoComputation*, National Centre for Geocomputation, National University of Ireland, Maynooth, Eire, 2007. 1

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning and Games*. Cambridge University Press, New York, 2006. 4, 6

N. Cesa-Bianchi, G. Lugosi, and G. Stoltz. Minimizing regret with label-efficient prediction. *IEEE Transactions on Information Theory*, 51:2152–2162, 2005. 11

F. Chung and L. Lu. Concentration inequalities and martingale inequalities: A survey. *Internet Mathematics*, 3:79–127, 2006. 23

E. R. Engdahl and A. Villaseñor. 41 global seismicity: 1900–1999. *International Geophysics*, 81:665–690, 2002. 13

H. Friedsam and W. A. Oren. The application of the principal curve analysis technique to smooth beamlines. In *Proceedings of the 1st International Workshop on Accelerator Alignment*, 1989. 1

T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84:502–516, 1989. 2, 12

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933. 1

M. Hutter and J. Poland. Adaptive online prediction by following the perturbed leader. *Journal of Machine Learning Research*, 6:639–660, 2005. 6

V. Kanade, B. McMahan, and B. Bryan. Sleeping experts and bandits with stochastic action availability and adversarial rewards. *AISTATS*, 3:1137–1155, 2009. 9, 12

B. Kégl. *Principal curves: learning, design, and applications.* PhD thesis, Concordia University Montreal, Quebec, 1999. 2, 4, 12, 20

B. Kégl and A. Krzyżak. Piecewise linear skeletonization using principal curves. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(1):59–74, 2002. 1, 4

B. Kégl, A. Krzyżak, T. Linder, and K. Zeger. Learning and design of principal curves. *IEEE transactions on pattern analysis and machine intelligence*, 22(3):281–297, 2000. 2, 3, 4

R. D. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. *Regret Bounds for Sleeping Experts and Bandits.* In COLT. Springer, 2008. 9

V. Laparra and J. Malo. Sequential principal curves analysis. arXiv preprint, 2016. URL https://arxiv.org/abs/1606.00856. 4

L. Li, B. Guedj, and S. Loustau. A quasi-Bayesian perspective to online clustering. *Electronic Journal of Statistics*, 12(2):3071–3113, 2018. doi: 10.1214/18-EJS1479. 4, 6

D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999a. 4

D. A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the 12th annual conference on Computational Learning Theory*, pages 164–170. ACM, 1999b. 4

G. Neu and G. Bartók. An efficient algorithm for learning with semi-bandit feedback. In *Lecture Notes in Computer Science*, volume 8139, pages 234–248. Springer, Berlin, Heidelberg, 2013. 11

K. Pearson. On lines and planes of closest fit to systems of point in space. *Philosophical Magazine*, 2(11):559–572, 1901. 1

K. Reinhard and M. Niranjan. Parametric subspace modeling of speech transitions. *Speech Communication*, 27:19–42, 1999. 1

S. Sandilya and S. R. Kulkarni. Principal curves with bounded turn. *IEEE Transactions on Information Theory*, 48:2789–2793, 2002. 4

J. Shawe-Taylor and R. C. Williamson. A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*, pages 2–9. ACM, 1997. doi: 10.1145/267460.267466. 4

C. Spearman. "General Intelligence", Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292, 1904. 1

D. C. Stanford and A. E. Raftery. Finding curvilinear features in spatial point patterns: principal curve clustering with noise. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):601–609, 2000. 1