PIECEWISE LINEAR APPROXIMATION OF EMPIRICAL DISTRIBUTIONS UNDER A WASSERSTEIN DISTANCE CONSTRAINT

PHILIPP ARBENZ AND WILLIAM GUEVARA-ALARCÓN

August 6, 2018

ABSTRACT. Big data applications and Monte Carlo simulation results can nowadays easily contain data sets in the size of millions of entries. We consider the situation when the information on a large univariate data set or sample needs to be preserved, stored, or transferred. We suggest an algorithm to approximate a univariate empirical distribution through a piecewise linear distribution which requires significantly less memory to store. The approximation is chosen in a computationally efficient manner, such that it preserves the mean, and its Wasserstein distance to the empirical distribution is sufficiently small.

Monte Carlo simulation, empirical distribution, piecewise linear approximation, Wasserstein distance, compression.

1. INTRODUCTION

The continuous improvement in computer processing power and the increase of available digital information, as well as the instruments to register and store it, makes more data available nowadays than ever before and this tendency will grow continuously. On the other hand, Monte Carlo simulation has become an omnipresent tool used in a broad range of disciplines as physics, engineering, biology or finance, to approximate probability distributions of variables of interest that have a random behaviour. The data sets that can be obtained from such big data applications or Monte Carlo simulations often extend to millions of observations. Even though the mechanisms to capture and store this growing influx of data are broadening and ameliorating, the amount of information created increases at a faster rate than the available storage [see 1].

We consider the case of a random variable $X : \Omega \mapsto \mathbb{R}$ on some probability space $(\Omega, \mathfrak{F}, \mathbb{P})$ that has a distribution $X \sim F$. A large univariate sample of size n from distribution F is available and leads to an empirical distribution F_n. We suggest an algorithm to compress F_n through a piecewise linear (PWL) approximating distribution G.

$$F \xrightarrow{sampling} F_n \xrightarrow{compression} G \sim \text{PWL}$$

Note that in this paper, we are not concerned with the sampling algorithm from the distribution F leading to the empirical distribution F_n, but only discuss the approximation step from the empirical distribution F_n to G. Of course, obtaining F_n may also present very challenging technical and mathematical aspects. For instance, testing the hypothesis that the sample underlying F_n comes from a hypothesized distribution F is treated in detail on [2].

PWL distributions are a versatile class of distributions. The class is parsimonious, in the sense that the number of parameters scales with the complexity of the approximated distribution and it can represent distributions with any shape. The proposed algorithm is efficient and reduces significantly the amount of memory required to store the distribution compared to storing the full sample. At the same time, our approach preserves the shape of the distribution and allows to have a controlled error with respect to the sample distribution, which is not possible when storing only particular key statistics, such as the first few moments.

The algorithm preserves the mean and is chosen such that the Wasserstein distance between the empirical distribution and its approximation $W(F_n, G)$ is small. In fact, the error introduced because of the PWL approximation can be controlled as a function of the sampling error and is selected to be noticeably smaller than the latter. This can be achieved thanks to the estimator $\hat{W}(F, F_n)$ in Section 4.5 for which it is not necessary to have information about distribution F other than its first moment to be finite. In this way, a PWL approximation obtained by means of the algorithm presented in this work can be used without significant additional error than the one introduced by the sampling.

This publication is inspired by [3], an article dealing with the same type of piecewise linear approximation in a finance and insurance context and some initial ideas in [4]. The algorithm in [3] uses a constraint on a set of risk measures, namely tail value at risk. In this publication, we consider a constraint on the Wasserstein distance, which is more often used in a statistical environment compared to risk measures. The described methodology can also be seen as the estimation of a histogram on a large dataset with a bound on its Wasserstein distance. A review on different methodologies to construct histograms is done in [5]. In previous works, [6] describe a strategy for selecting a piecewise linear approximation of an empirical distribution, with a predetermined number of segments, using the Wasserstein distance of order 2. Compared with that work, we use the Wasserstein distance of order 1 instead and look for an approximation whose distance $W(F_n, G)$ does not exceed a predefined value, without fixing its number of segments. Optimal quantization deals with a similar problem, to approximate a distribution by a discrete random variable with a predetermined maximum number of atoms, such that the Wasserstein distance is minimized [see 7]. We relate our approach to optimal quantization in Section 4.7. Approximation of density functions is described among others in [8] through piecewise constant functions using a constraint in L_2 distance; in [9] through piecewise polynomial functions using a constraint in total variation and in [10] through histograms that minimize relative error measures.

The paper is structured as follows. Section 2 introduces piecewise linear distributions. Section 3 focuses on the Wasserstein distance and defines admissible approximation distributions. Section 4 specifies the approximation algorithm. In Section 5, we give examples of results and implementation and conclude in Section 6.

2. Piecewise linear distributions

In this section, we define the class of random variables with a piecewise linear distribution that is used as approximation to the empirical distribution function. A PWL distribution has both its cumulative distribution function (cdf) and its quantile function composed by linear segments. For ease of understanding, we start with an illustrative example and define a parametrization subsequently.

Example 2.1. Consider the following cdf and quantile function, defining a PWL distribution:

$$G(x) = \begin{cases} 0, & \text{if } x < 1, \\ \frac{(x-1)0.6}{3}, & \text{if } 1 \le x < 4, \\ 0.8, & \text{if } 4 \le x < 6, \\ \frac{(x-6)0.2}{3} + 0.8, & \text{if } 6 \le x < 9, \\ 1, & \text{if } 9 \le x, \end{cases} \qquad G^{\leftarrow}(t) = \begin{cases} 2.5 + 1.5\left(\frac{2t}{0.6} - 1\right), & \text{if } 0 < t \le 0.6, \\ 4, & \text{if } 0.6 < t \le 0.8, \\ 7.5 + 1.5\left(2\frac{t-0.8}{0.2} - 1\right), & \text{if } 0.8 < t \le 1. \end{cases}$$

The distribution $G(x)$ is illustrated in Figure 1.

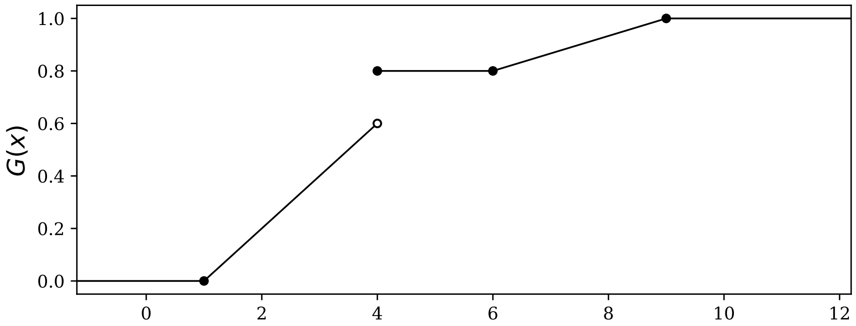

Figure 1. An illustration of the cdf of the PWL distribution defined in Example 2.1. It pertains a positive density on the intervals $(1,4)$ and $(6,9)$, as well as an atom at 4.

It is also possible to parametrize the class of PWL distributions through its interpolation points, i.e., the points where the linear segments connect. However, from an algorithmic point of view, the following equivalent parametrization is easier to use. It is based on the quantile function as a starting point, and it uses average and slope per quantile segment as parameters.

Definition 2.2. A random variable $X : \Omega \to \mathbb{R}$ has a piecewise linear distribution with vector parameters $\mathbf{z} = (z_1, \ldots, z_S) \in [0,1]^S$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{S-1}) \in \mathbb{R}^{S-1}$, and $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{S-1}) \in [0, \infty)^{S-1}$ for $S \in \mathbb{N}$, such that

$$\begin{aligned} &0 = z_1 < z_2 < \cdots < z_S = 1, \quad \text{and} \\ &\mu_s + \delta_s \le \mu_{s+1} - \delta_{s+1} \quad \text{for} \quad s = 1, 2, \ldots, S-2, \end{aligned} \tag{2.1}$$

if its quantile function $G^{\leftarrow}(t)$ on the intervals $(z_{s+1}, z_s]$ is given by:

$$G^{\leftarrow}(t) = \mu_s + \delta_s \left(\frac{2t - (z_{s+1} + z_s)}{z_{s+1} - z_s} \right) \quad \text{for} \quad t \in (z_s, z_{s+1}]. \tag{2.2}$$

The vector $\mathbf{z} = (z_1, z_2, \dots, z_S)$ is called basis of G: $\text{basis}(G) = (z_1, z_2, \dots, z_S)$. The semi-closed intervals $(z_s, z_{s+1}]$ between two consecutive points z_s and z_{s+1} in the basis are called segments of G and $S-1$ denotes the number of segments in the approximation.

Its cdf $G(x) = \mathbb{P}[X \le x]$ $(x \in \mathbb{R})$ is equal to:

$$G(x) = \begin{cases} 0, & \text{if } x < \mu_1 - \delta_1, \\ \frac{(x-\mu_s+\delta_s)(z_{s+1}-z_s)}{2\delta_s} + z_s, & \text{if } \mu_s - \delta_s \le x < \mu_s + \delta_s, \\ z_{s+1}, & \text{if } \mu_s + \delta_s < x < \mu_{s+1} - \delta_{s+1} \text{ or } x = \mu_s - \delta_s = \mu_s + \delta_s, \\ 1, & \text{if } x \ge \mu_{S-1} + \delta_{S-1} \end{cases}$$

The interpolation points of such a PWL distribution are given by $(\mu_1 - \delta_1, z_1), (\mu_1 + \delta_1, z_2), (\mu_2 - \delta_2, z_2), (\mu_2 + \delta_2, z_3), \dots, (\mu_{S-1} + \delta_{S-1}, z_S)$.

Note that if $J : \Omega \to \{1, 2, \dots, S-1\}$ and $U_s : \Omega \to [0,1]$ are independent random variables with $\mathbb{P}[J = s] = z_{s+1} - z_s$ and $U_s \sim U(0,1)$ for $s = 1, \dots, S-1$, then the variable X with distribution $G \sim \text{PWL}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\delta})$ has the following stochastic representation as a discrete mixture of $S-1$ uniform random variables:

$$X = \sum_{s=1}^{S-1} \mathbb{1}\{J = s\} \left[2\delta_s U_s + \mu_s - \delta_s\right] \tag{2.3}$$

where $\mathbb{1}\{\cdot\}$ denotes the indicator function, and the mixing weights are $z_{s+1} - z_s > 0$, $\sum_{s=1}^{S-1}(z_{s+1} - z_s) = z_S - z_0 = 1$. In the case that $\delta_s = 0$, the s-th variable in the mixture would be a degenerate variable given by the constant μ_s.

Figure 2 illustrates a segment of a PWL distribution G parametrized by its basis $\mathbf{z}$, and the vectors $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$ related to the average and slope per segment.

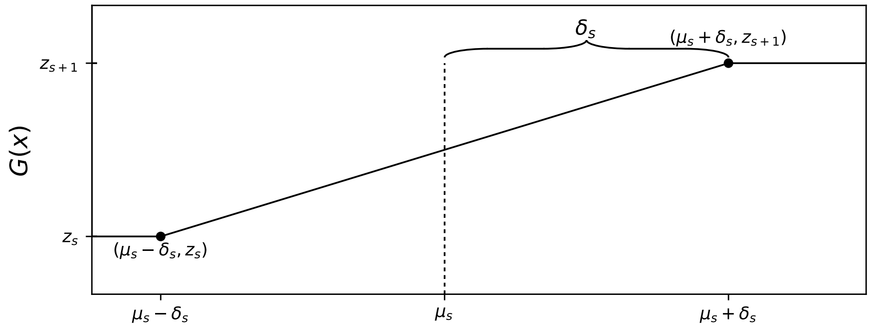

FIGURE 2. Illustration of a PWL distribution segment, parametrized through $\mathbf{z}$, $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$. δ_s equates to the distance between the average of $G^{\leftarrow}(t)$ on $(z_S, z_{s+1}]$ (equal to μ_s) and one of the values $G^{\leftarrow}(z_s)$ (equal to $\mu_s - \delta_s$) or $\lim_{t \uparrow z_{s+1}} G^{\leftarrow}(t)$ (equal to $\mu_s + \delta_s$).

Example 2.3. Consider the PWL distribution G as introduced in Example 2.1. The corresponding parameters are

$$S = 4, \quad \mathbf{z} = (0, 0.6, 0.8, 1), \quad \boldsymbol{\mu} = (2.5, 4, 7.5), \quad \boldsymbol{\delta} = (1.5, 0, 1.5).$$

Note that $\delta_2 = 0$ indicates an atom for the segment. The three segments of G are $(0, 0.6]$, $(0.6, 0.8]$, and $(0.8, 1]$.

A useful property of the class of PWL distributions is that many statistics such as moments can be computed analytically.

Lemma 2.4. *Let $G = \text{PWL}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\delta})$ and $m > 0$. The m-th moment $\mathbb{E}[X^m]$ of a random variable $X \sim G$ is given by*

$$\mathbb{E}[X^m] = \sum_{s=1}^{S-1} \int_{z_s}^{z_{s+1}} G^{\leftarrow}(t)^m \mathrm{d}t,$$

where

$$\int_{z_s}^{z_{s+1}} G^{\leftarrow}(t)^m \mathrm{d}t = \begin{cases} \frac{(\mu_s+\delta_s)^{m+1} - (\mu_s-\delta_s)^{m+1}}{2\delta_s(m+1)}(z_{s+1} - z_s), & \text{if } \delta_s > 0, \\ \mu_s^m (z_{s+1} - z_s), & \text{if } \delta_s = 0. \end{cases}$$

Proof. If U is uniformly distributed on $[0,1]$, then $\mathbb{E}[X^m] = \mathbb{E}[(G^{\leftarrow}(U))^m]$. Therefore,

$$\mathbb{E}[X^m] = \mathbb{E}[G^{\leftarrow}(U)^m] = \int_0^1 G^{\leftarrow}(t)^m \mathrm{d}t = \sum_{s=1}^{S-1} \int_{z_s}^{z_{s+1}} G^{\leftarrow}(t)^m \mathrm{d}t,$$

which leads to the desired result using (2.2). □

Example 2.5. Let $G = \mathrm{PWL}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\delta})$ as in Example 2.1 and $X \sim G$. Then, the mean is given by

$$\mathbb{E}[X] = \frac{4^2 - 1^2}{2 \times 1.5 \times 2}(0.6 - 0) + 4(0.8 - 0.6) + \frac{9^2 - 6^2}{2 \times 1.5 \times 2}(1 - 0.8) = 3.8.$$

The PWL distributions in this paper have a PWL cdf and quantile function, not a PWL density. This type of distribution has been chosen for approximation purposes because of its worthwhile properties:

- PWL distributions allow to easily calculate quantities such as cdf, quantiles, moments or risk measures.
- PWL distributions can be used to approximate any kind of univariate distribution shape; continuous, discrete or mixed (continuous in some parts of the domain with atoms).
- PWL distributions are parsimonious, the size of the parameter vectors $\mathbf{z}$, $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$ varies according to the appearance and complexity of the empirical distribution that is being approximated. For large samples, the number of PWL parameters is massively lower compared to the sample size.

Piecewise functions based on higher degree polynomials can also be used as approximations, but they would make the calculation of quantities more difficult and the algorithm less efficient.

Assumption 2.6. *Throughout the remainder of the paper, we assume that F_n denotes an empirical distribution $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \leq x\}$ of a sample $\{X_1, X_2, \ldots, X_n\}$ with sample size $n \in \mathbb{N}$.*

Definition 2.7. A $\mathrm{PWL}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\delta})$ distribution G is *sample compatible* with respect to an empirical distribution F_n, if $z_s \in \{0/n, 1/n, 2/n, \ldots, n/n\}$ for all $s = 1, 2, \ldots, S$.

The former definition indicates that each linear segment of G corresponds to a specific set of sample points in $\{X_{(1)}, X_{(2)}, \ldots, X_{(n)}\}$, where $X_{(i)}$, $i = 1, \ldots, n$ are the order statistics of the sample.

3. Wasserstein distance and admissibility

In this section, we introduce the Wasserstein distance, a metric that allows to quantify the proximity between two univariate distribution functions. This distance will be the basis to define when a PWL distribution is an admissible approximation of an empirical distribution.

Definition 3.1. The *Wasserstein distance* between F_n and G is given by

$$W(F_n, G) = \int_{-\infty}^{\infty} |F_n(x) - G(x)| \,\mathrm{d}x = \int_0^1 \left|F_n^{\leftarrow}(t) - G^{\leftarrow}(t)\right| \mathrm{d}t. \tag{3.1}$$

The Wasserstein distance above is in fact the minimal L_1 metric between random variables with distributions F_n and G. This distance is known under different names as *Kantorovich metric, Monge-Wasserstein distance, Gini index* or *Earth-Mover's distance* given that several authors from distinct disciplines have worked on it [see 11]. It is also related to optimal transportation problems and it has been used in approximation of probability measures, statistical mechanics and image processing. For historical references on the origin of Wasserstein distance and its use in optimal transportation [see 12]. Definition 3.1 can be extended to the Wasserstein distance of order p as presented for instance in [11]. Under that alternative definition, Equation (3.1) would be the Wasserstein distance of order 1.

Definition 3.2. The *discretized Wasserstein distance $W^*(F_n, G)$* between F_n and a sample compatible G is defined as

$$W^*(F_n, G) = \frac{1}{n}\sum_{i=1}^n \left| X_{(i)} - G^{\leftarrow}\left(\frac{i - 1/2}{n}\right)\right|. \tag{3.2}$$

Note that $\int_{\frac{i-1}{n}}^{\frac{i}{n}} G^{\leftarrow}(t)dt = \frac{1}{n}G^{\leftarrow}\left(\frac{i-1/2}{n}\right)$, thus by defining

$$W_i^* = \left|\int_{\frac{i-1}{n}}^{\frac{i}{n}} \left(F_n^{\leftarrow}(t) - G^{\leftarrow}(t)\right) \mathrm{d}t\right| = \frac{1}{n}\left|X_{(i)} - G^{\leftarrow}\left(\frac{i - 1/2}{n}\right)\right| \tag{3.3}$$

we can write $W^*(F_n, G) = \sum_{i=1}^n W_i^*$.

Note that the empirical quantile function is defined as $F_n^{\leftarrow}(t) = X_{(i)}$ for $\frac{i-1}{n} < t \le \frac{i}{n}$, the inverse of the empirical distribution function. Alternative definitions of the empirical quantile function (discrete or continuous) can be found in [13].

The following theorem calculates the Wasserstein distance $W(F_n, G)$.

Theorem 3.3. *The Wasserstein distance between F_n and a sample compatible $G = \mathrm{PWL}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\delta})$ is given by $W(F_n, G) = \sum_{i=1}^n W_i$, where*

$$W_i = \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left|F_n^{\leftarrow}(t) - G^{\leftarrow}(t)\right| \mathrm{d}t = W_i^* + \frac{1}{2}\left(\frac{1}{n} - \frac{W_i^*}{\delta_s^*}\right) \cdot \max\{\delta_s^* - nW_i^*, 0\}, \quad \textit{and} \tag{3.4}$$

$$\delta_s^* = \frac{\delta_s}{n(z_{s+1} - z_s)} = \left|G^{\leftarrow}\left(\frac{i}{n}\right) - G^{\leftarrow}\left(\frac{i-1/2}{n}\right)\right| = \left|\lim_{t \uparrow \frac{i-1}{n}} G^{\leftarrow}(t) - G^{\leftarrow}\left(\frac{i-1/2}{n}\right)\right|.$$

Note that $W_i = W_i^*$ in case $\delta_s^* \le nW_i^*$.

Proof. We have that $F_n^{\leftarrow}(t)$ is constant for $t \in \left(\frac{i-1}{n}, \frac{i}{n}\right]$: $F_n^{\leftarrow}(t) = X_{(i)}$. $nW_i^* = \left|X_{(i)} - G^{\leftarrow}\left(\frac{i-1/2}{n}\right)\right|$ equates to the distance from $G^{\leftarrow}\left(\frac{i-1/2}{n}\right)$, the middle point of $G^{\leftarrow}$ on $\left(\frac{i-1}{n}, \frac{i}{n}\right]$, to the sample value $X_{(i)}$. Analogously, the distance from that middle point to the end of $G^{\leftarrow}$ on the segment $\left(\frac{i-1}{n}, \frac{i}{n}\right]$ is given by δ_s^*.

We can now distinguish two cases: whether $G^{\leftarrow}(t)$ attains the value $X_{(i)}$ in the interval $\left(\frac{i-1}{n}, \frac{i}{n}\right]$ (case 1) or not (case 2).

- **Case 1:** $G^{\leftarrow}(t) \neq X_{(i)}$ for all $t \in \left(\frac{i-1}{n}, \frac{i}{n}\right]$. In this case $nW_i^* \ge \delta_s^*$ and $\max\{\delta_s^* - nW_i^*, 0\} = 0$. We know that $F_n^{\leftarrow}(t) - G^{\leftarrow}(t)$ does not change sign for $t \in \left(\frac{i-1}{n}, \frac{i}{n}\right]$. Therefore,
$$W_i = \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left|F_n^{\leftarrow}(t) - G^{\leftarrow}(t)\right| \mathrm{d}t = \left|\int_{\frac{i-1}{n}}^{\frac{i}{n}} (F_n^{\leftarrow}(t) - G^{\leftarrow}(t)) \mathrm{d}t\right| = W_i^*.$$
- **Case 2:** If $G^{\leftarrow}(t)$ attains the value $X_{(i)}$ in the interval $\left(\frac{i-1}{n}, \frac{i}{n}\right]$, then $F_n^{\leftarrow}(t) - G^{\leftarrow}(t)$ changes its sign in that interval and $nW_i^* < \delta_s^*$. As shown in Figure 3, in this case W_i is given by the area of two triangles. The areas of the smaller and larger triangle are given by
$$\frac{1}{2}\left(\delta_s^* - nW_i^*\right)\frac{1}{2n}\left(1 - \frac{nW_i^*}{\delta_s^*}\right) \quad \text{and} \quad \frac{1}{2}\left(\delta_s^* + nW_i^*\right)\frac{1}{2n}\left(1 + \frac{nW_i^*}{\delta_s^*}\right),$$
respectively. Adding up these terms yields
$$\begin{aligned} W_i &= \frac{1}{2}\left(\delta_s^* - nW_i^*\right)\frac{1}{2n}\left(1 - \frac{nW_i^*}{\delta_s^*}\right) + \frac{1}{2}\left(\delta_s^* + nW_i^*\right)\frac{1}{2n}\left(1 + \frac{nW_i^*}{\delta_s^*}\right) \\ &= W_i^* + \frac{1}{2}\left(\frac{1}{n} - \frac{W_i^*}{\delta_s^*}\right) \cdot (\delta_s^* - nW_i^*). \end{aligned}$$
Figure 3 illustrates both cases and the area which equals W_i.

□

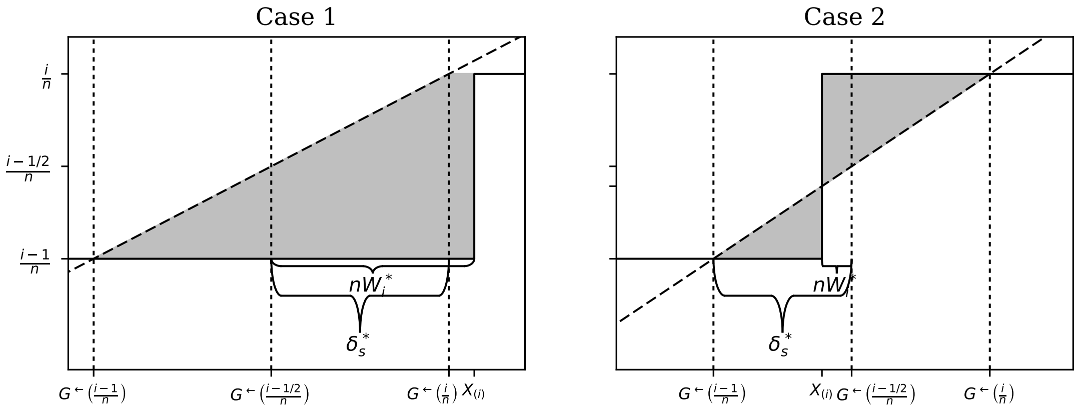
Case 1
Case 2
$\frac{i}{n}$
$\frac{i-1/2}{n}$
$\frac{i-1}{n}$
nW_i^*
δ_s^*
$G^{\leftarrow}(\frac{i-1}{n})$
$G^{\leftarrow}(\frac{i-1/2}{n})$
$G^{\leftarrow}(\frac{i}{n})$
$X_{(i)}$
nW_i^*
δ_s^*
$G^{\leftarrow}(\frac{i-1}{n})$
$X_{(i)}$
$G^{\leftarrow}(\frac{i-1/2}{n})$
$G^{\leftarrow}(\frac{i}{n})$

FIGURE 3. Illustration of W_i (shaded area) for the empirical distribution F_n (solid line) and the PWL approximation G (dashed line) in case that $nW_i^* \ge \delta_s^*$ (left) and $nW_i^* < \delta_s^*$ (right).

The following theorem shows that the discretized Wasserstein distance $W^*(F_n, G)$ is bounded by $W(F_n, G)$.

Theorem 3.4. *For F_n and its approximation $G = \mathrm{PWL}(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\delta})$, we have:*

$$W^*(F_n,G) \le W(F_n,G).$$

Proof. Adding up the components W_i and W_i^* and exchanging absolute value with integral yields

$$\begin{aligned}W^*(F_n,G) &= \sum_{i=1}^n W_i^* = \sum_{i=1}^n \left|\int_{\frac{i-1}{n}}^{\frac{i}{n}} \left(F_n^{\leftarrow}(t) - G^{\leftarrow}(t)\right)\mathrm{d}t\right| \\ &\le \sum_{i=1}^n \int_{\frac{i-1}{n}}^{\frac{i}{n}} \left|F_n^{\leftarrow}(t) - G^{\leftarrow}(t)\right|\mathrm{d}t = \int_0^1 \left|F_n^{\leftarrow}(t) - G^{\leftarrow}(t)\right|\mathrm{d}t = W(F_n,G).\end{aligned}$$

□

Example 3.5. Suppose we have the following sample of size 10:

$$\{X_{(1)}, X_{(2)}, \ldots, X_{(10)}\} = \{1, 1.6, 4.3, 4.6, 6, 7.1, 13, 15.6, 16, 18.8\}.$$

Let F_n be its empirical distribution and $G = \mathrm{PWL}(\mathbf{z},\boldsymbol{\mu},\boldsymbol{\delta})$ with $\mathbf{z} = (0, 0.6, 1)$, $\boldsymbol{\mu} = (4.1, 15.85)$ and $\boldsymbol{\delta} = (3.3, 1.9)$. Figure 4 shows F_n, G and both versions of the Wasserstein distance.

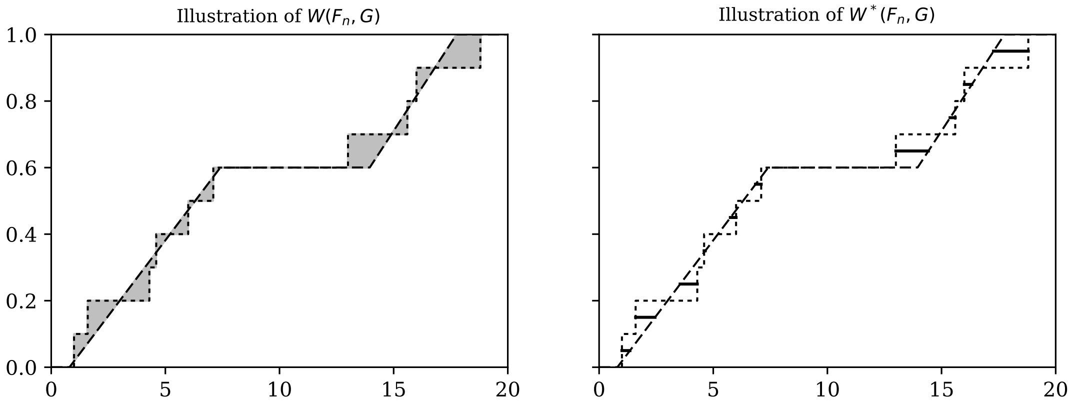

FIGURE 4. Empirical distribution F_n (dotted) and a PWL approximation G (dashed). In this case $W_i = W_i^*$ for $i = 2, 3, 7, 10$. The shaded area in the plot on the left illustrates the Wasserstein distance $W(F_n,G) = 0.6517$. The average length of the solid segments in the plot on the right corresponds to the discretized Wasserstein distance $W^*(F_n,G) = 0.6$.

Definition 3.6. For a sample distribution F_n, a PWL distribution G is called an **admissible approximation** of F_n with accuracy $\epsilon > 0$ if

$$W(F_n,G) \le \epsilon. \tag{3.5}$$

In Section 4.5, we propose a procedure for selecting the ϵ parameter for practical purposes.

Even though we focus on the Wasserstein distance in this paper, an admissible approximation as defined in (3.5) also implies an error bound on the Prokhorov, and Lévy metrics and a risk measure as the tail value at risk.

Definition 3.7. For a distribution F with $\mathbb{E}[F] < \infty$ and $\alpha \in (0,1]$ the tail value at risk (TVaR_α) is equal to $\mathrm{TVaR}_\alpha(F) = \frac{1}{\alpha}\int_{1-\alpha}^1 F^{\leftarrow}(t)\mathrm{d}t$.

Theorem 3.8. *Let the PWL distribution G be an admissible approximation of F_n. Then,*

$$|\mathrm{TVaR}_\alpha(F_n) - \mathrm{TVaR}_\alpha(G)| \le \frac{\epsilon}{\alpha}.$$

Proof.

$$\begin{aligned}\epsilon \ge \int_0^1 \left|F_n^{\leftarrow}(t) - G^{\leftarrow}(t)\right|\mathrm{d}t &\ge \int_{1-\alpha}^1 \left|F_n^{\leftarrow}(t) - G^{\leftarrow}(t)\right|\mathrm{d}t \\ &\ge \left|\int_{1-\alpha}^1 (F_n^{\leftarrow}(t) - G^{\leftarrow}(t))\mathrm{d}t\right| = \alpha|\mathrm{TVaR}_\alpha(F_n) - \mathrm{TVaR}_\alpha(G)|.\end{aligned}$$

□

Note that the converse is not true. For instance, $G = \mathrm{PWL}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\delta})$ with

$$\mathbf{z} = (0, \alpha, 1), \qquad \boldsymbol{\mu} = \left(\frac{1}{1-\alpha}\int_0^{1-\alpha} F_n^{\leftarrow}(t)\mathrm{d}t, \frac{1}{\alpha}\int_{1-\alpha}^{1} F_n^{\leftarrow}(t)\mathrm{d}t\right), \qquad \boldsymbol{\delta} = (0, 0)$$

has the same TVaR_α as F_n, but potentially much larger Wasserstein distance.

Definition 3.9. The *Lévy metric* $d_L(F_n, G)$ and the *Prokhorov metric* $d_P(F_n, G)$ between F_n and G are equal to:

$$\begin{aligned} d_L(F_n, G) &= \inf_{\tau>0}\left\{G(x-\tau) - \tau \le F_n(x) \le G(x+\tau) + \tau, \forall x \in \mathbb{R}\right\}, \\ d_P(F_n, G) &= \inf_{\tau>0}\left\{F_n(A) \le G(A^\tau) + \tau, \forall A \in \mathfrak{B}\right\}. \end{aligned}$$

where A^τ is a closed $\tau-$neighbourhood of A and $\mathfrak{B}$ is the Borel sigma algebra.

Theorem 3.10. *Let the PWL distribution G be an admissible approximation of F_n. Then,*

$$d_L(F_n, G) \le d_P(F_n, G) \le \sqrt{\epsilon}.$$

Proof. Immediate consequence of Definition 3.6 and the fact that $d_L(F_n, G) \le d_P(F_n, G)$ and $(d_P(F_n, G))^2 \le W(F_n, G)$ as shown in Section 3 of [14]. □

4. Algorithm

The basic idea on how to implement an algorithm to find an admissible PWL approximation of F_n is very simple: Start with $\mathbf{z} = (0, 1)$ and iteratively insert values into $\mathbf{z}$, until there exists an admissible PWL distribution G with $\mathrm{basis}(G) = \mathbf{z}$. Figure 5 illustrates the iterative decomposition. However, for a concrete implementation, several mathematical and numerical issues related to this basic idea are clarified in this section.

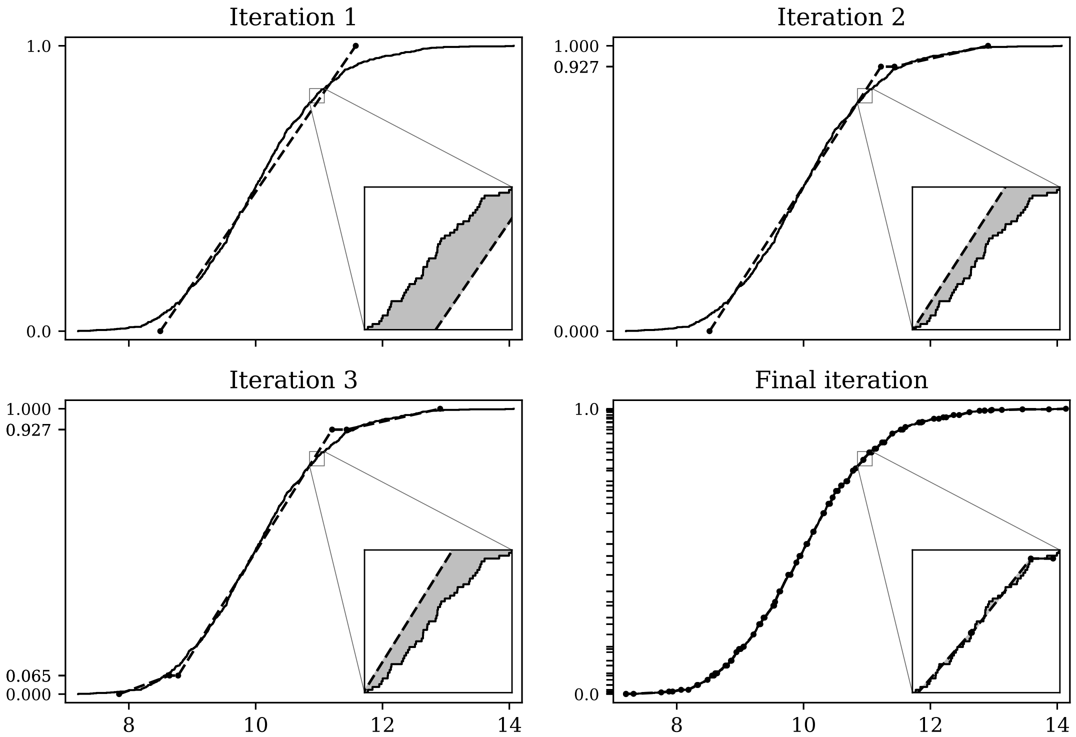

FIGURE 5. Iterative process to find an admissible PWL distribution G. Empirical distribution F_n (solid line) and a PWL approximation G (dashed line). For the first iteration $\mathbf{z} = (0, 1)$, for the second $\mathbf{z} = (0, 0.927, 1)$, for the third $\mathbf{z} = (0, 0.065, 0.927, 1)$ and for the final iteration $\mathbf{z}$ is a vector of size 50 in this case. The Wasserstein distance corresponds to the shaded area in the lower sub-plots.

4.1. **Mean parameter.**

In this subsection, we describe the algorithmic choice of the mean (location) parameter μ_s. The following theorem shows the relationship between the Wasserstein distance and the difference on the expected values of two distributions.

Theorem 4.1. *For two distributions F_n and G with expected values $\mathbb{E}[F_n]$ and $\mathbb{E}[G]$, we have:*

$$|\mathbb{E}[F_n]-\mathbb{E}[G]| \le W(F_n,G) \le |\mathbb{E}[F_n]-\mathbb{E}[G]| + W(F_n-\mathbb{E}[F_n], G-\mathbb{E}[G]). \tag{4.1}$$

Proof.

$$\begin{aligned}
|\mathbb{E}[F_n]-\mathbb{E}[G]| &= \left|\int_0^1 \left(F_n^{\leftarrow}(t)-G^{\leftarrow}(t)\right)dt\right| \le \int_0^1 \left|F_n^{\leftarrow}(t)-G^{\leftarrow}(t)\right|dt = W(F_n,G)\\
&= \int_0^1 \left|F_n^{\leftarrow}(t)-\mathbb{E}[F_n]+\mathbb{E}[F_n]-G^{\leftarrow}(t)-\mathbb{E}[G]+\mathbb{E}[G]\right|dt\\
&\le \int_0^1 |\mathbb{E}[F_n]-\mathbb{E}[G]|\,dt + \int_0^1 \left|F_n^{\leftarrow}(t)-\mathbb{E}[F_n]-G^{\leftarrow}(t)+\mathbb{E}[G]\right|dt\\
&= |\mathbb{E}[F_n]-\mathbb{E}[G]| + W(F_n-\mathbb{E}[F_n], G-\mathbb{E}[G]).
\end{aligned}$$

□

From (4.1), we observe that if we set the PWL approximation to have the same expectation than F_n, $\mathbb{E}[G]=\mathbb{E}[F_n]$, their Wasserstein distance will have a smaller upper bound, so that it will be faster to find an admissible approximation. In order to achieve the same first moment, we define μ_s as the average of all order statistics $X_{(i)}$ belonging to the segment s.

Algorithm 4.2. Given a $\mathbf{z}$ vector parameter of a sample compatible PWL distribution, set $\boldsymbol{\mu}$ such that

$$\mu_s = \frac{1}{n(z_{s+1}-z_s)} \sum_{i=nz_s+1}^{nz_{s+1}} X_{(i)} \quad \text{for} \quad s=1,2,\ldots,S-1. \tag{4.2}$$

Lemma 4.3. *If μ_s is set through Algorithm 4.2, then*

$$\int_{z_s}^{z_{s+1}} G^{\leftarrow}(t)\mathrm{d}t = \int_{z_s}^{z_{s+1}} F_n^{\leftarrow}(t)\mathrm{d}t.$$

Proof. Using Definition 2.2, we get $\int_{z_s}^{z_{s+1}} G^{\leftarrow}(t) = (z_{s+1}-z_s)\mu_s$. So that if μ_s is set as in Algorithm 4.2 $\int_{z_s}^{z_{s+1}} G^{\leftarrow}(t) = \frac{1}{n}\sum_{i=nz_s+1}^{nz_{s+1}} X_{(i)}$ which is equal to $\int_{z_s}^{z_{s+1}} F_n^{\leftarrow}(t)\mathrm{d}t$. □

4.2. **Default slope parameter.**

Given a basis $\mathbf{z}$, since the parameter $\boldsymbol{\mu}$ is given through Algorithm 4.2, the only remaining free parameter is $\boldsymbol{\delta}$. In this section, we provide default values for δ_s to be used. We propose to take the δ_s which minimizes the discretized Wasserstein distance between $F_n^{\leftarrow}$ and $G^{\leftarrow}$ measured on the segment $(z_s, z_{s+1}]$.

Theorem 4.4. *For a segment s of a sample compatible approximation G, let $\omega_s(\delta_s):[0,\infty)\to[0,\infty)$ denote the discretized Wasserstein distance between $G^{\leftarrow}$ and $F_n^{\leftarrow}$ on $(z_s,z_{s+1}]$ as a function of δ_s using μ_s as defined in Algorithm 4.2, $\omega_s(\delta_s) = \sum_{i=nz_s+1}^{nz_{s+1}} W_i^*$. Let δ_s^W the value that minimizes $\omega_s(\delta_s)$:*

$$\delta_s^W = \underset{\delta_s\ge 0}{\operatorname{argmin}}\, \omega_s(\delta_s) = \underset{\delta_s\ge 0}{\operatorname{argmin}} \sum_{i=nz_s+1}^{nz_{s+1}} \frac{1}{n}\left|X_{(i)} - \left(\mu_s + \delta_s\left(\frac{2\left(\frac{i-1/2}{n}\right)-(z_{s+1}+z_s)}{z_{s+1}-z_s}\right)\right)\right|.$$

Then, the minimum is attained for some $\delta_s^W \in \Delta_s = \{\Delta_{s,i} : i = nz_s+1,\ldots,nz_{s+1}\}$, where

$$\Delta_{s,i} = \left(X_{(i)}-\mu_s\right) / \left(\frac{2\left(\frac{i-1/2}{n}\right)-(z_{s+1}+z_s)}{z_{s+1}-z_s}\right). \tag{4.3}$$

Proof. W_i^* is a piecewise linear function in δ_s, being decreasing for $\delta_s < \Delta_{s,i}$ and increasing for $\delta_s > \Delta_{s,i}$; hence it attains its minimum $W_i^* = 0$ when $\delta_s = \Delta_{s,i}$. A special case occurs when there is an odd number of sample points in segment s, and $i = \frac{n(z_s+z_{s+1})+1}{2}$. In that situation $W_i^* = \frac{1}{n}|X_{(i)}-\mu_s|$ is constant for all values of δ_s. Therefore, $\omega_s(\delta_s)$ is a piecewise linear function since it is the sum of piecewise linear functions. This function has left and right derivatives on the full domain but it is not differentiable for the values of $\delta_s \in \Delta_{s,i}$; precisely because its slope changes at those values. Then, ω_s is convex and has a minimum value that is attained at one or some of these points in Δ_s. □

Algorithm 4.5. Given $\mathbf{z}$ and $\boldsymbol{\mu}$, calculate δ_s as follows:

- Determine the set Δ_s and sort its values.

- Apply a binary search algorithm to find $\delta_s^W \in \Delta_s$ such that $\omega_s\left(\delta_s^W\right)$ has a non-positive left derivative and a non-negative right derivative.

Note that this algorithm has numerical complexity $O(|\Delta_s|\log|\Delta_s|)$, with $|\Delta_s| = n(z_{s+1} - z_s)$ the number of sample points corresponding to segment s.

Example 4.6. Given the sample distribution F_n and $\mathbf{z} = (0, 0.6, 1)$ and $\boldsymbol{\mu} = (4.1, 15.85)$ as defined in Example 3.5. Then, Algorithm 4.5 yields to:

$$\Delta_1 = \{3.72, 5, -1.2, 3, 3.8, 3.6\}, \qquad \delta_1^W = 3.72,$$

$$\Delta_2 = \{3.8, 1, 0.6, 3.933\}, \qquad \delta_2^W = 3.8.$$

Figure 6 shows F_n and G with this choice for the slope parameters. Figure 7 shows the function $\omega_s(\delta_s)$ for the two segments $(0, 0.6]$ and $(0.6, 1]$.

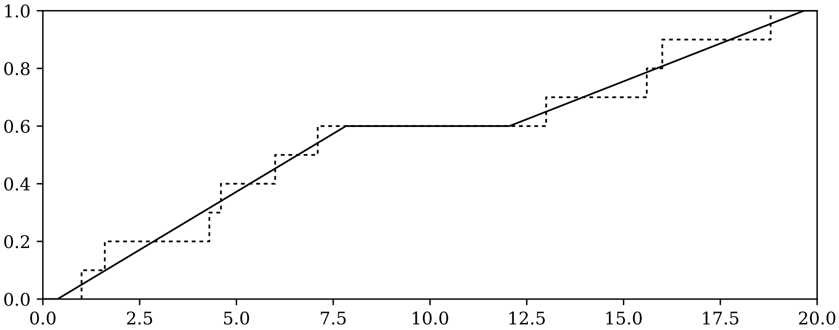

FIGURE 6. F_n (dotted line) and G (solid line) with δ_s^W that minimizes $\omega_s(\delta_s)$. For this parametrization $W^*(F_n, G) =$ 0.332 and $W(F_n, G) = 0.5521$.

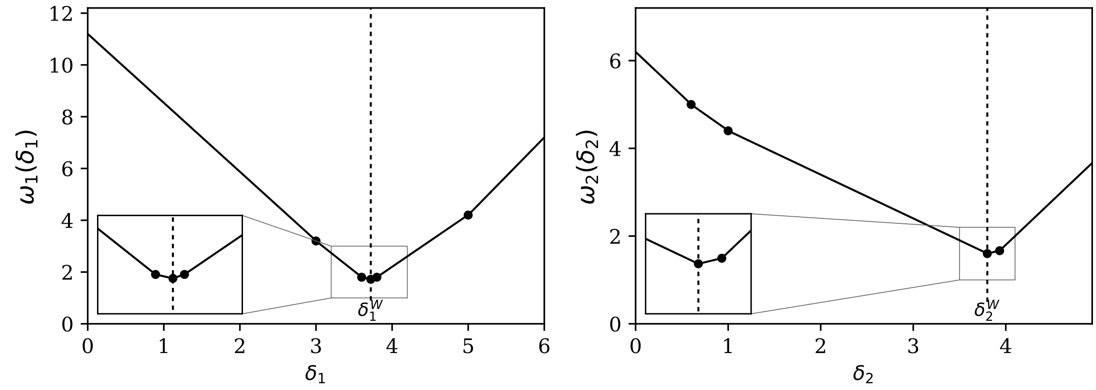

FIGURE 7. On the left $\omega_1(\delta_1)$, for the first segment the minimum $\omega_1\left(\delta_1^W\right) = 1.72$ is obtained for $\delta_1^W = 3.72$. On the right $\omega_2(\delta_2)$, with the minimum $\omega_2\left(\delta_2^W\right) = 1.6$ for $\delta_2^W = 3.8$. The dots in the graphs correspond to the points $(\Delta_{s,i}, \omega(\Delta_{s,i}))$. Note that $\omega_s(\delta_s)$ changes the slope at the values $\delta_s \in \Delta_{s,i}$.

The default parameter δ_s^W proposed in this section corresponds to the parameter from the median (quantile) regression line between $F_n^{\leftarrow}$ and $G^{\leftarrow}$ on $(z_s, z_{s+1}]$ that crosses through $\left(\mu_s, \frac{z_s+z_{s+1}}{2}\right)$. Refer to [15] for a detailed explanation of quantile regression.

Alternative approaches to set the slope parameter δ_s could be taken. For instance, if the method of moments is used, μ_s and δ_s can be obtained by matching $\frac{1}{n}\sum_{i=1}^n X_{(i)}$ with $\mathbb{E}[G]$ and $\frac{1}{n}\sum_{i=1}^n X_{(i)}^2$ with $\mathbb{E}[G^2]$ as given in Lemma 2.4 for $m = 1, 2$. In this way, not only the mean but also the sample variance could be preserved on the approximation G.

4.3. **Segment bisection.**

The basic idea of the approximation algorithm is to start with $\mathbf{z} = (0,1)$ and to insert points into $\mathbf{z}$ until an admissible solution is found. In this section, we describe where to bisect $(z_s, z_{s+1}]$. When choosing among the different segments, the segment s with the biggest value of $\omega_s(\delta_S)$ is selected for bisection.

Algorithm 4.7. If it is required to bisect segment s, insert $\widetilde{z}$ between z_s and z_{s+1}, where

$$\widetilde{z} = \underset{\zeta\in\{z_s+\frac{1}{n},z_s+\frac{2}{n},\dots,z_{s+1}-\frac{1}{n}\}}{\operatorname{argmin}} \left\{\int_{z_s}^{\zeta} \left(L_{(z_s,\zeta]}(t) - F_n^{\leftarrow}(t)\right)^2 \mathrm{d}t + \int_{\zeta}^{z_{s+1}} \left(L_{(\zeta,z_{s+1}]}(t) - F_n^{\leftarrow}(t)\right)^2 \mathrm{d}t\right\} \tag{4.4}$$

with

$$\int_{z_s}^{\zeta} \left(L_{(z_s,\zeta]}(t) - F_n^{\leftarrow}(t)\right)^2 \mathrm{d}t = \frac{1}{n} \sum_{i=nz_s+1}^{n\zeta} X_{(i)}^2 + \delta_I \left[\mu_I(\zeta + z_s) - \frac{2}{n^2(\zeta - z_s)} \sum_{i=nz_s+1}^{n\zeta} X_{(i)}(i - 1/2)\right] - \mu_I^2(\zeta - z_s)$$

Where for the segment $I = [z_s, \zeta)$, μ_I is chosen accordingly to Algorithm 4.2 (but replacing z_{s+1} by ζ) and δ_I corresponds to the δ_s parameter used for the PWL approximation $L_I(t)$ that is the L_2 regression of the sample points on the interval I [see 3, Theorem 5.7]. $\int_{\zeta}^{z_{s+1}} \left(L_{(\zeta,z_{s+1}]}(t) - F_n^{\leftarrow}(t)\right)^2 \mathrm{d}t$ is given by an analogous expression but with the segment $I = [\zeta, z_{s+1})$.

The value $\widetilde{z}$ bisects the segment at the point with the optimal reduction of the L_2 distance. Such a distance is preferred to the Wasserstein in this section of the algorithm, because a closed formula exists for the slope parameter δ_{s_i}, so that Algorithm 4.7 has a $O(|\Delta_s|)$ numerical complexity, whilst using the distance $\omega_{s_i}\left(\delta_{s_i}^W\right)$ would imply a $O\left(|\Delta_s|^2 \log|\Delta_s|\right)$ complexity. The effect of using the Wasserstein distance instead of the L_2 distance in the bisection of segments, in the performance of the algorithm, is examined in Section 5.

Example 4.8. Given the sample distribution F_n as defined in Example 3.5 and $\mathbf{z} = (0,1)$, applying Algorithm 4.7 on the segment $(z_1, z_2] = (0,1]$ yields $\widetilde{z} = 0.6$ since for $z = (0.1, 0.2, \dots, 0.9)$, the value of the right expression in (4.4) evaluates to $(1.868, 1.755, 1.425, 1.354, 1.299, 0.392, 1.521, 2.183, 2.131)$.

4.4. **Ensuring segment compatibility.**

Remember that the idea of the approximation algorithm is to find a PWL$(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\delta})$ distribution G such that $W(F_n, G) \le \epsilon$. Although setting the value of $\delta_s = \delta_s^W$ as described in Section 4.2 will lead in most of the cases to compatible segments, it is possible that the compatibility condition (2.1) is not fulfilled for some segments. This means that

$$\mu_s + \delta_s^W > \mu_{s+1} - \delta_{s+1}^W,$$

for some $s \in \{1, 2, \dots, S-2\}$, i.e. that the PWL function defined by $(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\delta})$ is not a proper distribution function.

The following algorithm has the purpose of correcting this situation.

Algorithm 4.9. Suppose that for some fixed s, we have

$$\mu_s + \delta_s^W > \mu_{s+1} - \delta_{s+1}^W,$$

Then, reset δ_s and δ_{s+1} such that

$$\mu_s + \delta_s = \mu_{s+1} - \delta_{s+1} = \frac{\mu_s + \delta_s^W + \mu_{s+1} - \delta_{s+1}^W}{2}$$

so that the values of δ_s and δ_{s+1} after the adjustment are $\delta_s = \frac{\mu_{s+1}-\mu_s+\delta_s^W-\delta_{s+1}^W}{2}$, $\delta_{s+1} = \frac{\mu_{s+1}-\mu_s+\delta_{s+1}^W-\delta_s^W}{2}$.

The following example illustrates Algorithm 4.9.

Example 4.10. Suppose F_n is defined as in Example 3.5. Furthermore, let $\mathbf{z} = (0, 0.3, 1)$. We have

$$\mu_1 = 2.3, \qquad \delta_1^W = 2.475, \qquad \mu_2 = 11.586, \qquad \delta_2^W = 8.417,$$

which leads to

$$4.775 = \mu_1 + \delta_1^W > \mu_2 - \delta_2^W = 3.169.$$

By using Algorithm 4.9, we set $\delta_1 = 1.672$ and $\delta_2 = 7.614$, such that $\mu_1 + \delta_1 = \mu_2 - \delta_2 = 3.972$. After this adjustment, the two segments connect at the midpoint. Figure 8 provides an illustration.

It is important to note that the stochastic representation in (2.3) does not require that the compatibility condition (2.1) is fulfilled. The mixture of uniform variables would add a new segment between $\mu_{s+1} - \delta_{s+1}$ and $\mu_s + \delta_s$ without introducing additional elements to the vector parameters $\mathbf{z}$, $\boldsymbol{\mu}$ and $\boldsymbol{\delta}$. However, the new segment does not guarantee the sample compatibility condition on Definition 2.7. Therefore, the obtained formulae (3.2) and (3.3) for $W^*(F_n, G)$ and (3.4) for $W(F_n, G)$ would not be exact. Because of the former, the

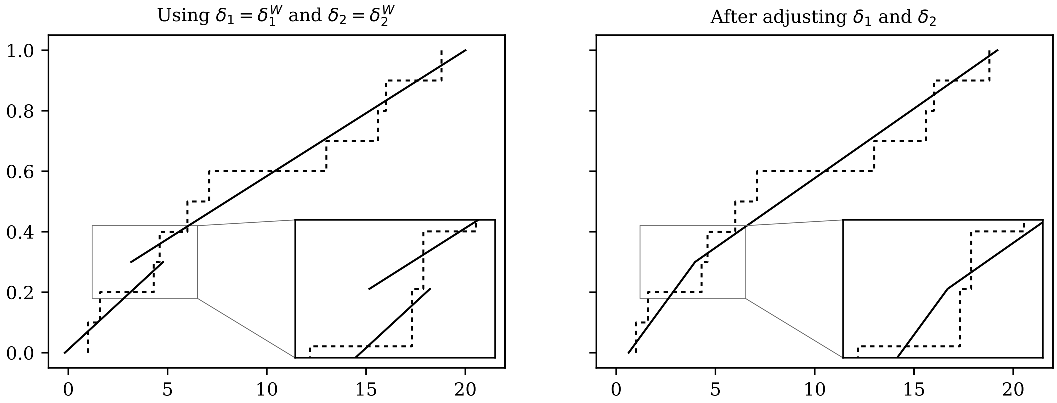

FIGURE 8. Illustration of Algorithm 4.9. Left: With $\delta_1 = \delta_1^W$ and $\delta_2 = \delta_2^W$, the two segments of the PWL approximation (solid) are incompatible. Right: Setting $\delta_1 = 1.672$ and $\delta_2 = 7.614$ yields compatible and connected segments.

use of the stochastic representation to solve the problem with the compatibility condition would imply more involved formulae for the Wasserstein distance, and the full algorithm would lose its simplicity, hence the solution with Algorithm 4.9 is preferred.

4.5. **Selection of accuracy parameter.**

In this section, we describe a statistical approach for the estimation of the accuracy parameter ϵ. To that end, we consider the convergence of F_n to F in terms of the Wasserstein distance, and then set ϵ such that the approximation error is significantly smaller than the sampling error. Most of the convergence theory in this section is based on [16].

Definition 4.11. An estimator of the expected Wasserstein distance between F_n and F is defined as

$$\widehat{W}(F, F_n) = \frac{1}{\sqrt{n}}\sqrt{\frac{2}{\pi}}\int_{-\infty}^{\infty}\sqrt{F_n(t)(1-F_n(t))}\mathrm{d}t.$$

Note that $\widehat{W}(F, F_n)$ depends only on F_n, i.e., it is independent of F and can therefore be calculated from the sample.

The following theorem provides the conditions and mathematical formulation of $\widehat{W}(F, F_n)$ as a consistent estimator of $\mathbb{E}[W(F, F_n)]$.

Theorem 4.12. *Suppose* $\mathbb{E}[|X|^\xi] < \infty$ *for* $X \sim F$ *and some* $\xi > 2$. *Then,*

$$\lim_{n\to\infty}\left|\sqrt{n}\widehat{W}(F, F_n) - \sqrt{n}\mathbb{E}[W(F, F_n)]\right| = 0.$$

Proof. [See 16, Theorem 2.4]. We have $L_\xi \subset L_{2,1}$ [see 16, p.1014 after (2.1')]. For the expectation $\mathbb{E}\int|B(F)| = \sqrt{2/\pi}\int\sqrt{F(1-F)}$ [see 16, p.1038]. del Barrio E. et al. [16, Equation (1.6)] also provide the limiting distribution and variance of $W(F, F_n)$ around the mean in terms of a weighted Brownian motion. □

The careful reader may have observed that Theorem 4.12 requires $\mathbb{E}[|X|^\xi] < \infty$ for some $\xi > 2$. del Barrio E. et al. [16] also provide the convergence behaviour in the other cases:

- If $\mathbb{E}[X^2] = \infty$ and $\mathbb{E}[|X|^\xi] < \infty$ for some $\xi > 1$, then $a_n\mathbb{E}[W(F, F_n)]$ converges to a constant for some appropriately chosen sequence of scaling factors a_n. Under the conditions of Theorem 4.12, $a_n = \sqrt{n}$ was used, but the infinite variance case implies a slower convergence rate. For instance, for a power tailed distribution with a tail decay parameter α (such as Pareto(α)), del Barrio E. et al. [16, Theorem 2.2 and (2.13)] provide $a_n = n^{1-1/\alpha}$. For $\alpha = 1.5$, this gives $a_n = n^{1/3}$. The constant to which $a_n\mathbb{E}[W(F, F_n)]$ converges can be determined as shown by $K(t)$ and [16, (2.24)], but cannot practically be estimated, since it depends on constants which are not known when only having a sample distribution F_n at hand [see 16, (2.22)]. del Barrio E. et al. [16, Theorem 1.1b)] provide the limiting distribution of $a_nW(F, F_n)$ and [16, Theorem 2.4b)] the corresponding moments.
- If $\mathbb{E}[|X|] = \infty$ then $W(F, F_n) = \infty$ for all n. However, F_n is finite, and G is finite as well. Therefore, a "good" approximation can be found, in some sense to be defined by the user. However, ϵ cannot be based on the convergence behaviour of $W(F, F_n)$. A different criterion will be required.
- There are some boundary cases not captured in the three cases mentioned above and in Theorem 4.12. These are covered in [16, Sections 4-6], but are not of practical relevance in this paper.

We propose to generally use the result of Theorem 4.12 in order to practically determine the accuracy parameter ϵ.

Algorithm 4.13. For a given sample distribution F_n, set

$$\widehat{\epsilon} = 0.1 \cdot \widehat{W}(F,F_n) = 0.1 \frac{1}{\sqrt{n}}\sqrt{\frac{2}{\pi}}\sum_{i=1}^{n-1}\sqrt{\frac{i}{n}\cdot\frac{n-i}{n}}\left(X_{(i+1)} - X_{(i)}\right). \tag{4.5}$$

Using Algorithm 4.13 implies that asymptotically the error introduced through the piecewise linear approximation (error between F_n and G) is at least one order of magnitude smaller than the sampling error (error between F and F_n). In fact, if the distribution F satisfies certain conditions, then we can establish an upper bound on the probability that the difference between these two errors is bigger than a fixed value δ.

Lemma 4.14. *Assume that F satisfies a Poincaré inequality, i.e. it is a continuous distribution for which exists a constant $\lambda > 0$, such that for any absolutely continuous function $u : \mathbb{R} \to \mathbb{R}$*

$$\lambda \operatorname{Var}[u(F)] \le \mathbb{E}\left[|u'(F)|^2\right]$$

For the empirical distribution F_n, let G be an admissible PWL approximation with accuracy $\hat{\epsilon}$ according to Algorithm 4.13. Then

$$\mathbb{P}\left[|W(F_n,G) - 0.1W(F,F_n)| \ge \delta\right] \le Ce^{-20\,\delta\sqrt{\lambda n}} \qquad \textit{for any} \qquad \delta > 0, n \in \mathbb{N}$$

where $C > 0$ is an absolute constant.

Proof. This result is a consequence of Definition 3.6, Definition 4.11, and Theorem 7.1 in [17]. □

The upper bound in Lemma 4.14 applies to continuous random variables with finite exponential moments such as Gaussian or gamma, but it excludes sub-exponential distributions with heavy tails as the ones described in [18]. Since the approximation is intended to be used in the case of large sample sizes, the factor 0.1 in Algorithm 4.13 defines for most of the cases in practice the expected proportion of the approximation error relative to the sampling error.

In case Algorithm 4.13 is deemed unsuitable for the situation at hand, one may either adjust the factor 0.1 to some other value, or select ϵ as a value adapted to the specific situation or application. This should notably be considered if the underlying distribution is known to have infinite variance or if F_n is not derived from an i.i.d. sample of F.

4.6. **Full algorithm.**

Using the results of the previous sections, we now have all components to describe the full algorithm.

Algorithm 4.15.

(1) Initialize $\mathbf{z} = (0,1)$. Set $\epsilon > 0$ as described in Section 4.5.
(2) Determine $G \sim \mathrm{PWL}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\delta})$ with $\boldsymbol{\mu}$ set through Algorithm 4.2 and $\boldsymbol{\delta}$ set through $\delta_s = \delta_s^W$ as in Algorithm 4.5.
(3) Calculate $W^*(F_n,G)$. If $W^*(F_n,G) > \epsilon$, bisect the segment s with the biggest value $\omega_s(\delta_s)$ using Algorithm 4.7. Incorporate $\tilde{z}$ into $\mathbf{z}$ and return to point (2).
(4) If there are any incompatible segments (i.e., segments breaching condition (2.1)), then adjust $\boldsymbol{\delta}$ by applying Algorithm 4.9.
(5) Calculate $W(F_n,G)$ through Theorem 3.3. If $W(F_n,G) < \epsilon$, then a sample compatible and admissible solution $G \sim \mathrm{PWL}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\delta})$ has been found. Otherwise, bisect the segment s with the largest value of $\omega_s(\delta_s)$ and go back to point (2) but using $W(F_n,G)$ instead of $W^*(F_n,G)$.

Note that the algorithm is shift invariant since G is chosen to have equal mean as F_n. Furthermore, it is scale invariant since (4.2), (4.3) and (4.5) scale linearly with the sample. Hence, given a sample X_i, we determine an admissible PWL approximation $G_X \sim \mathrm{PWL}(\mathbf{z}_X, \boldsymbol{\mu}_X, \boldsymbol{\delta}_X)$ using Algorithm 4.15, with accuracy $\hat{\epsilon}_X$ as in Algorithm 4.13. If we transform linearly the sample to $Y_i = a + bX_i$ with $a \in \mathbb{R}$ and $b > 0$, which can for instance be seen as a change in currency or an adjustment by inflation; then it is not required to apply again Algorithm 4.15 to the sample Y_i. The admissible PWL approximation G_Y for sample Y_i with accuracy $\hat{\epsilon}_Y = b\hat{\epsilon}_X$ can be deduced adjusting the parameters of G_X accordingly, so that $G_Y \sim \mathrm{PWL}(\mathbf{z}_X, a\mathbf{1} + b\boldsymbol{\mu}_X, b\boldsymbol{\delta}_X)$.

The full algorithm checks first the admissibility condition (3.5) for the discretized Wasserstein distance $W^*(F_n,G)$; once the condition is fulfilled with this distance, it is verified for the Wasserstein distance $W(F_n,G)$. In doing so, the algorithm approximates W_i, the area between the quantile functions $F_n^{\leftarrow}(t)$ and $G^{\leftarrow}(t)$ for $t \in \left(\frac{i-1}{n}, \frac{i}{n}\right]$, through W_i^*, the distance between the empirical quantile function $F_n^{\leftarrow}(t)$ and $G^{\leftarrow}\left(\frac{i-1/2}{n}\right)$ the middle point of $G^{\leftarrow}$ in the interval $\left(\frac{i-1}{n}, \frac{i}{n}\right]$ (see Figure 4). The approximation is exact in the case that the quantile function $G^{\leftarrow}$ does not equal a sample value $X_{(i)}$ in the interval considered, as shown in Theorem 3.3. This allows the full algorithm to be more efficient, given that W_i is only calculated in the last iteration.

When comparing Algorithm 4.15 with the one described in [3], we can appreciate that the admissibility condition on the latter is based on the relative error of the TVaR_α for all levels $\alpha \in (0,1)$, i.e. the entire domain of the quantile function. This admissibility condition can be verified independently for every segment s, analysing only the TVaR_α for $\alpha \in (z_s, z_{s+1}]$. On the contrary, condition (3.5) depends on the values of the distribution in all its domain and has to be verified globally for the entire distribution at each iteration. Therefore, a PWL approximation obtained with the algorithm described in this work is expected to have its points more uniformly spread, with a smaller proportion of segments in the tail, than the approximation based on the relative error of the TVaR.

4.7. **Connection with optimal quantization.**

There is a link between the proposed approximation algorithm, and the optimal quantization problem in one dimension with the L_1 distance as norm. The optimal quantization problem as described in [7] fixes the number $S \in \mathbb{N}$ of approximation points, and then finds an approximation G with a maximum number of points S that minimizes $W(F_n, G)$; whereas in our approach, given an accuracy ϵ the aim is to find an admissible approximation G such that $W(F_n, G) \le \epsilon$. However, optimal quantization is restricted to discrete approximations (or mixtures of Dirac components) while our approach utilizes PWL approximations (or mixtures of uniform random variables). The use of discrete approximations in our approach would correspond to $G \sim \text{PWL}(\mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\delta})$ with $\boldsymbol{\delta} = \mathbf{0}$.

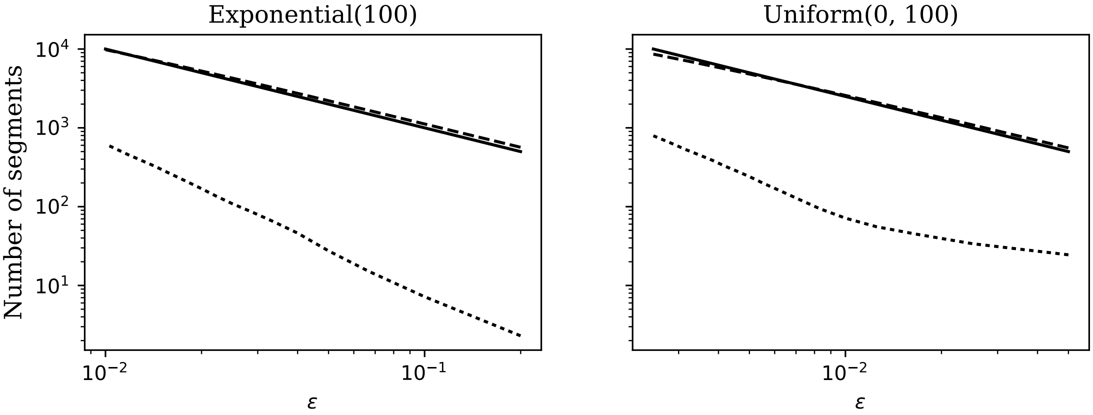

FIGURE 9. Accuracy ϵ as the S-th quantization error against S (solid line) and the number of segments for Algorithm 4.15 with $\delta_s = 0$ (dashed line) and $\delta_s = \delta_s^W$ (dotted line), for $F \sim Exp(100)$ and $F \sim U(0,100)$ over 10 repetitions with $n = 100'000$ (in log log scale).

For a given sample distribution F_n, Algorithm 4.15 provides an admissible approximation G as required in (3.5); but there is no guarantee that the distribution G is the PWL admissible approximation with the minimal number of segments. However, we can fix the accuracy ϵ in Algorithm 4.15, as the minimum Wasserstein distance for a discrete approximation with S points (S-th quantization error) and compare S with the number of segments obtained for the PWL approximation. Figure 9 shows the accuracy parameter ϵ equal to the S-th quantization error (for distributions with explicit formulas of its value), against the number of points S in the quantization problem and the number of segments in the PWL approximation from Algorithm 4.15. We can appreciate that the approximation obtained when Algorithm 4.15 is executed under the same conditions than the optimal quantization problem, i.e. forcing $\delta_s = 0$, has a number of segments on the same order of magnitude than the optimal, in this case the accuracy ϵ decreases as $O\left(S^{-1}\right)$ [see 7, Theorem 6.2]. For a fixed ϵ, when Algorithm 4.15 is executed with $\delta_s = \delta_s^W$, the number of segments in the PWL approximation is of at least one order of magnitude smaller, than in the case $\delta_s = 0$.

5. IMPLEMENTATION AND ILLUSTRATIONS

We provide an implementation of the algorithm in Python. The code is provided under the permissive and free *MIT license*; it can be obtained through the authors or at the following internet address:

`https://sites.google.com/site/philipparbenz/home/pwl-wasserstein`

Example 5.1. The Python implementation of the PWL approximation algorithm takes three arguments: the sample (as a list or `numpy` *array*), the accuracy mode which can be "Relative" (default option) if ϵ is chosen through Algorithm 4.13 or "Absolute" if the value of ϵ is directly entered by the user; and the accuracy parameter that is equal to the percentage that multiplies $\widehat{W}(F, F_n)$ in (4.5) (by default 0.1) in the case of a relative accuracy mode, or equal to the value of ϵ to be used if the absolute accuracy mode is selected.

LISTING 1. Minimal code example in Python

```
from compressor import WassersteinPWLcompressor
Sample = [1, 1.6, 4.3, 4.6, 6, 7.1, 13, 13.4, 16, 18.8]
PWLapprox = WassersteinPWLcompressor(Sample, AccuracyMode = "Relative",
                                     AccuracyParameter = 0.1)
```

To illustrate the result from the algorithm when applied to a large sample, the data set 'brvehins1' for Brazilian vehicle insurance from the CASdatasets package [see 19] of the statistical software R, which contains a collection of insurance related data sets, is used. This data set contains information on risk features, number of claims and claim amounts of 1,965,355 auto mobile insurance policies in Brazil during the year 2011. Two variables from the data set are utilized to illustrate the performance of the algorithm. The first sample corresponds to the total claim amount per payment X_{Tot}, i.e. the total value of the losses paid per policy for the policies that had to pay one or more auto mobile claims. The second sample corresponds to the claim portion paid for partial collision for the policies that had at least one auto mobile claim $X_{PartColl}$. Both samples contain 363,076 observations corresponding to the number of policies with non-zero losses during 2011. Values in the sample X_{Tot} are positive without ties, while $X_{PartColl}$ is a sample with an atom at zero for the cases when there was a claim, but it corresponds to a different coverage than partial collision (robbery, total collision, fire or other guarantees).

Figure 10 shows the result of the algorithm when applied to the empirical distributions F_n^{Tot} and $F_n^{PartColl}$ from the samples X_{Tot} and $X_{PartColl}$ when the accuracy ϵ is selected according to (4.5). In both cases $n = 363,076$ and in the first case $\widehat{W}\left(F, F_n^{Tot}\right) = 104.62$ while for $X_{PartColl}$, we have $\widehat{W}\left(F, F_n^{PartColl}\right) = 36.74$.

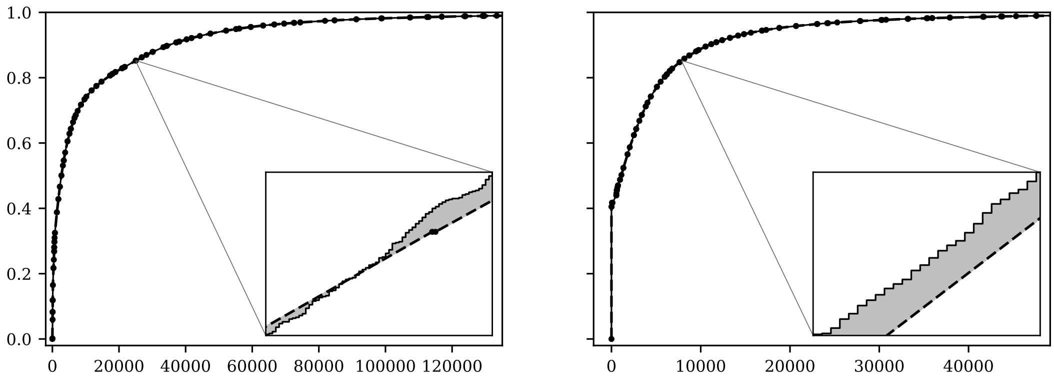

FIGURE 10. Empirical distributions (solid line) and admissible PWL approximations (dashed) with an accuracy $\hat{\epsilon}$ given by Algorithm 4.13. For F_n^{Tot} (left plot) the accuracy used is $\hat{\epsilon} = 10.46$. For $F_n^{PartColl}$ (right plot) the accuracy used is $\hat{\epsilon} = 3.67$. The lower right sub-plot provides a zoom into the quantiles [0.852-0.8525].

Table 1 shows the run time, the number of segments and the Wasserstein distance (discretized or not), when the approximation algorithm is applied to F_n^{Tot} and $F_n^{PartColl}$. From this table, we can see that there is only a small difference between the values of $W(G, F_n)$ and $W^*(G, F_n)$ for the final PWL approximation. The number of segments on the approximation is about 3 per 10,000 of the sample size, showing its efficiency in reducing the amount of memory required.

Next, we examine the run time of the algorithm for different sample sizes. To that end, we take sub-samples without replacement from X_{Tot} and $X_{PartColl}$ for sample sizes $n = 10,000$ up to $n = 360,000$ and compute the average run time over 100 repetitions. Figure 11 shows the average run time per sample size n. We observe that the run time is always slightly longer for X_{Tot} than for $X_{PartColl}$, which must be due to the presence of the atom at zero for the second sample whose PWL approximation requires a slightly smaller number of segments. Additionally, we observe a linear behaviour on the increase on time with respect to the sample size. It is important to note that the accuracy parameter $\hat{\epsilon}$ selected through (4.5) decreases when the sample size n increases; being on average 49.21 and 18.98 when $n = 10,000$ for X_{Tot} and $X_{PartColl}$, respectively, and decreasing to the values shown in Table 1 when $n = 363,076$.

TABLE 1. Average run time (in ms), number of segments and Wasserstein distance (discretized or not) of the resulting PWL approximation with $\hat{\epsilon} = 0.1 \cdot \hat{W}\left(F, F_n^X\right)$, when the algorithm is applied to the sample $X = \{X_{Tot}, X_{PartColl}\}$

Sample (X)	Run time (milliseconds)	Number of segments $(S-1)$	$\hat{\epsilon}$	$\hat{W}\left(F, F_n^X\right)$	$W^*\left(G, F_n^X\right)$	$W\left(G, F_n^X\right)$
X_{Tot}	971	108	10.46	104.6	9.66	10.37
$X_{PartColl}$	914	101	3.67	36.7	3.52	3.67

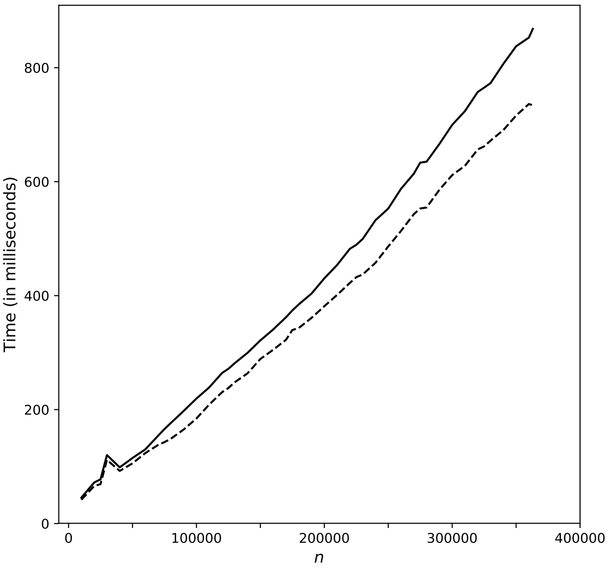

FIGURE 11. Average run time of the algorithm over 100 repetitions when applied to subsamples of size n of the samples X_{Tot} (solid line) and $X_{PartColl}$ (dashed).

Table 2 shows the first five iterations of the algorithm when applied to $F_n^{PartColl}$. The Wasserstein distance decreases rapidly during these initial iterations until an admissible approximation is obtained after 101 iterations. For these initial iterations, the value of the Wasserstein distance and its discretized version is the same up to three decimal digits but their difference increases in the following iterations until reaching the values shown in Table 1 in the final iteration.

TABLE 2. Initial iterations of the Algorithm 4.15 for the sample $X_{PartColl}$

Iteration	$\mathbf{z}$	$W\left(G, F_n^{PartColl}\right)$
1	(0, 1)	3763.40
2	(0, 0.9908, 1)	2816.21
3	(0, 0.9028, 0.9908, 1)	1467.51
4	(0, 0.5872, 0.9028, 0.9908, 1)	832.88
5	(0, 0.5872, 0.9028, 0.9692, 0.9908, 1)	668.54

As discussed in Section 4.3, the distance $\omega_{s_i}\left(\delta_{s_i}^W\right)$ could be used instead of the L_2 distance in Algorithm 4.7. The resulting PWL approximations will be different under the two methods. However, our tests applying these two bisection methods on sub-samples of size $n = 50{,}000$ and $n = 100{,}000$ of X_{Tot} and $X_{PartColl}$ show that the resulting PWL approximations have the same magnitude in the number of segments, differing in a few segments (less than 10% of the number of segments) and without being consistently better for any of the two distances. On the contrary, the run time of Algorithm 4.15 blows up from less than a second when using L_2, to several minutes when using $\omega_{s_i}\left(\delta_{s_i}^W\right)$ in Algorithm 4.7. Therefore, such a change would render the full algorithm unusable for practical purposes.

6. Discussion

This paper introduces an algorithm that computes a piecewise linear approximation G for a large univariate sample distribution F_n, with hundred of thousands or millions data points. The approximation has the same mean as the empirical distribution and a bounded error in terms of the Wasserstein distance. For distributions with a finite second moment, it is proposed to select the error of the approximation as one order of magnitude smaller than the sampling error. The algorithm is efficient, and the resulting distribution has only a couple of hundred points for typical applications with sample sizes on the hundreds of thousands and requires significantly less memory than the original sample. The approximation algorithm can be applied to discrete, continuous or mixed distributions. An efficient and open source software implementation is provided. The approximation is particularly useful in industrial or applied environments where numerous distributions of big samples are repeatedly used and transferred among different systems. In that situation, diverse statistics from the distributions (as moments, quantiles or risk measures) are calculated in each of the stages and the use of the approximations accelerates the operation of the process and maintains the reliability of the quantities calculated.

TABLE 3. Qualitative properties of the different approaches which can be used to retain information on a large univariate sample distribution

	Shape preserving	Memory and bandwidth efficient
Storing key statistics	No	Yes
Storing full sample	Yes	No
PWL approximation	Yes	Yes

Table 3 provides a comparative overview on the qualitative properties of the PWL approximation under a Wasserstein distance constrain compared to two other approaches commonly used when information on large univariate sample distributions should be preserved or transferred between systems: storing some key statistics and storing the full sample. Using a PWL approximation is an approach which preserves shape and statistics, and is memory and bandwidth efficient. The other two approaches do not satisfy both properties.

In future work, we intend to study the approximation of a multivariate sample through a piecewise density under a Wasserstein distance. The Wasserstein distance naturally extends to higher dimensions, but its computation is more involved and the components of the algorithm would need to be adapted.

Disclosure statement

We have no conflicts of interest to disclose.

Acknowledgements

As SCOR Fellow, William Guevara-Alarcón thanks SCOR for financial support. The authors thank Hansjörg Albrecher, Carlos Arocha, Eustasio del Barrio, Sebastian Engelke, Edoardo Luini, Anna Seyfert, Vaclav Varvarovsky, and Johanna Ziegel for their helpful comments. The authors would like to thank the reviewers and the associate editor for their valuable and constructive comments which improved the paper and highlighted connections with other approaches.

References

[1] The Economist. Data, data everywhere, February 25, 2010. [cited 24 May 2018]. Available from : `http://www.economist.com/node/15557443`.

[2] O. Thas. *Comparing distributions*. Springer-Verlag, New York, 2010.

[3] P. Arbenz and W. Guevara-Alarcón. Risk measure preserving piecewise linear approximation of empirical distributions. *European Actuarial Journal*, 6(1):113–148, 2016.

[4] A. Seyfert. Piecewise linear approximation of empirical distributions with bounds based on Wasserstein metric. Master's thesis, Université de Lausanne, 2016.

[5] Y. Ioannidis. The history of histograms (abridged). *Proceedings of the 29th Annual International Conference on Very Large Data Bases*, pages 19–30, 2003.

[6] A. Irpino and E. Romano. Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation. *Revue des Nouvelles Technologies de l'Information RNTI-E-9*, 1:99–110, 2007.

[7] S. Graf and H. Luschgy. *Foundations of quantization for probability distributions.* Springer-Verlag, Berlin, 2000.

[8] J. Acharya, I. Diakonikolas, C. Hegde, J. Li, and L. Schmidt. Fast and near-optimal algorithms for approximating distributions by histograms. *Proceedings of the 34th ACM Symposium on Principles of Database Systems. PODS '15*, pages 249–263, 2015.

[9] S.O. Chan, I. Diakonikolas, R.A. Servedio, and X. Sun. Efficient density estimation via piecewise polynomial approximation. *Proceedings of the 46th Annual ACM Symposium on Theory of Computing. STOC '14*, pages 604–613, 2014.

[10] S. Guha, K. Shim, and J. Woo. REHIST: relative error histogram construction algorithms. *Proceedings of the 30th International Conference on Very Large Data Bases VLDB*, pages 300–311, 2004.

[11] C. Villani. *Optimal transport, old and new.* Springer-Verlag, Berlin, 2009.

[12] R. M. Dudley. *Real analysis and probability.* Cambridge University Press, Cambridge, 2004.

[13] R. J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4): 361–365, 1996.

[14] A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.

[15] R. Koenker. *Quantile regression.* Cambridge University Press, Cambridge, 2005.

[16] del Barrio E., Giné E., and Matrán C. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Annals of Probability*, 27(2):1009–1071, 1999.

[17] S. G. Bobkov and M. Ledoux. One-dimensional empirical measures, order statistics and Kantorovich transport distances. *To appear in Memoirs of the AMS*, Preprint 2016.

[18] C. M. Goldie and C. Klüppelberg. *Subexponential distributions. In: Adler R, Feldman R, Taqqu MS, editors. A practical guide to heavy tails.* Birkhäuser, Boston, 1998.

[19] C. Dutang. *Insurance datasets*, 2016. R package version 1.0-5. Available from: `http://cas.uqam.ca`.

WILLIAM GUEVARA-ALARCÓN (corresponding author)
Université de Lausanne, Department of Actuarial Science, Switzerland
E-Mail: William.GuevaraAlarcon@unil.ch; wmguevaraa@unal.edu.co

PHILIPP ARBENZ
SCOR Switzerland. Ltd and ETH Zürich, Switzerland
E-Mail: philipp.arbenz@gmail.com