

# Wasserstein gradients for the temporal evolution of probability distributions\*

Yaqing Chen and Hans-Georg Müller

*Department of Statistics, University of California, Davis,  
e-mail: [yaqchen@ucdavis.edu](mailto:yaqchen@ucdavis.edu); [hgmue1ler@ucdavis.edu](mailto:hgmue1ler@ucdavis.edu)*

**Abstract:** Many studies have been conducted on flows of probability measures, often in terms of gradient flows. We utilize a generalized notion of derivatives with respect to time to model the instantaneous evolution of empirically observed one-dimensional distributions that vary over time and develop consistent estimates for these derivatives. Employing local Fréchet regression and working in local tangent spaces with regard to the Wasserstein metric, we derive the rate of convergence of the proposed estimators. The resulting time dynamics are illustrated with time-varying distribution data that include yearly income distributions and the evolution of mortality over calendar years.

**MSC2020 subject classifications:** Primary 62G05; secondary 62G20.

**Keywords and phrases:** Time-varying density functions, Wasserstein metric, dynamics of income distributions, evolution of human mortality.

Received January 2021.

## 1. Introduction

There exists a sizeable literature on flows of probability measures, often described in terms of gradient flows (Ambrosio et al., 2008; Santambrogio, 2017). However, the statistical modeling of the instantaneous evolution of observed distributions that are indexed by time has not yet been explored. Figure 1 shows an example of time-indexed densities, which correspond to demographic age-at-death distributions from 1936 to 2010 in the US, for females and males respectively. Motivated by this and similar data, we study temporal flows for one-dimensional probability distributions. Recently, there has been intensive interest in comparing distributions with the Wasserstein distance, both in theory and applications (e.g. Bolstad et al., 2003; Bigot et al., 2017; Galichon, 2017; Cazelles et al., 2018; Bigot et al., 2019), and in visualization (e.g. Delicado and Vieu, 2017). In the one-dimensional case that we consider here, it is well known that the Wasserstein transport can also be expressed in terms of quantile functions (Hoeffding, 1940; Zhang and Müller, 2011; Chowdhury and Chaudhuri, 2019).

Our goal is to develop statistical models that reflect instantaneous evolution of such temporal flows of distributions. Starting with the Monge–Kantorovich

---

\*Research supported in part by NSF Grants DMS-1712864 and DMS-2014626.

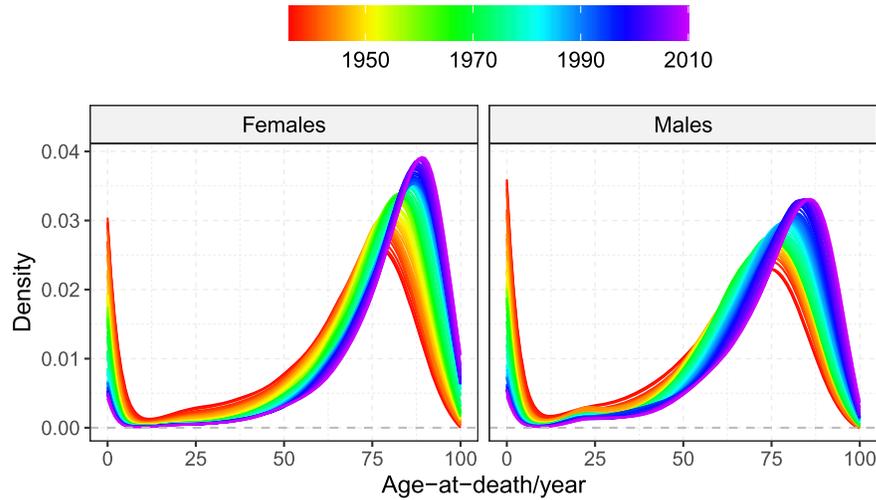


FIG 1. Time-varying densities of age-at-death (in years) for the US from 1936 to 2010.

problem (e.g., Ambrosio, 2003; Villani, 2003, 2008), given two probability measures  $p_1$  and  $p_2$ , one aims to transport the pile of mass distributed as in  $p_1$  to that as in  $p_2$  while minimizing the transport cost. The transport map attaining the minimum transport cost defines the optimal transport from  $p_1$  to  $p_2$ . Based on such optimal transport maps, basic concepts such as tangent bundles and exponential and logarithmic maps in Riemannian manifolds can be generalized to the space of univariate probability distributions endowed with the Wasserstein distance, which form a quasi-Riemannian manifold (e.g., Ambrosio et al., 2008; Bigot et al., 2017; Zemel and Panaretos, 2019). The log map, defined as the difference between the optimal transport and identity maps, captures the direction and distance of each small mass along the order-preserving transport from the starting probability measure to the target measure and can be used to quantify the change between the two probability measures. Hence, we utilize temporal derivatives of log maps, the Wasserstein temporal gradients, to model the instantaneous temporal evolution of distributions. For this purpose, we harness local Fréchet regression (Petersen and Müller, 2019a) to first smooth the observed probability measures over time due to the discrepancy between the true conditional Fréchet mean and the observed distributions and then estimate the Wasserstein temporal gradients by difference quotients based on the local Fréchet regression estimates.

The Wasserstein temporal gradients that we target are introduced in Section 2, with estimation and asymptotic theory in Section 3. In Section 4, we discuss implementation details, followed by a simulation study. Applications are demonstrated in Section 5 for longitudinal household income and human mortality data.

## 2. Preliminaries

### 2.1. Optimal transport in the Wasserstein space

Given a bounded interval  $\mathcal{D}$  in  $\mathbb{R}$ , we focus on the Wasserstein space  $\mathcal{W} = \mathcal{W}(\mathcal{D})$  of distributions on  $\mathcal{D}$  (for which second moments are finite), endowed with the  $\mathcal{L}^2$ -Wasserstein distance  $d_W$ . This metric is related to the solution of Monge’s optimal transport problem (Villani, 2003) and has repeatedly been rediscovered, with examples including Mallow’s distance (Mallows, 1972), earth mover’s distance (Rubner et al., 1997) or quantile normalization (Bolstad et al., 2003).

Specifically, the  $\mathcal{L}^2$ -Wasserstein distance between any atomless distributions  $p_1, p_2 \in \mathcal{W}$  is the square root of

$$d_W^2(p_1, p_2) = \inf_{g\#p_1=p_2} \int_{\mathcal{D}} [x - g(x)]^2 dp_1(x), \tag{2.1}$$

where  $g\#p$  is a push-forward measure such that  $g\#p(A) = p(\{x : g(x) \in A\})$ , for any measurable function  $g: \mathbb{R} \rightarrow \mathbb{R}$ , distribution  $p \in \mathcal{W}$  and set  $A \subseteq \mathbb{R}$ . It is well known (e.g., Cambanis et al., 1976) that the minimum in (2.1) is attained at the optimal transport map  $\Upsilon_{p_1,p_2} = F_2^{-1} \circ F_1$  from  $p_1$  to  $p_2$ , and is  $d_W^2(p_1, p_2) = \int_0^1 [F_1^{-1}(u) - F_2^{-1}(u)]^2 du$ . Here,  $F_l$  and  $F_l^{-1}$  are the cumulative distribution function and quantile function of  $p_l$  for  $l = 1, 2$ , where cumulative distribution functions are considered to be right continuous and quantile functions to be left continuous. A distribution  $p$  is atomless if it has a continuous cumulative distribution function.

Basic concepts of Riemannian manifolds can be analogously defined in the Wasserstein space  $\mathcal{W}$  based on optimal transport maps (e.g., Ambrosio et al., 2008; Bigot et al., 2017; Zemel and Panaretos, 2019). Suppose that  $p_0$  is an atomless reference probability measure in  $\mathcal{W}$ . The tangent space at  $p_0$  is defined as (Equation (8.5.1), Ambrosio et al., 2008)

$$\mathcal{T}_{p_0} = \overline{\{\eta(\Upsilon_{p_0,p} - \text{id}) : p \in \mathcal{W}, \eta > 0\}}^{\mathcal{L}^2_{p_0}},$$

where  $\mathcal{L}^2_{p_0} = \mathcal{L}^2_{p_0}(\mathcal{D})$  is the Hilbert space of  $p_0$ -square-integrable functions on  $\mathcal{D} \subset \mathbb{R}$ , with inner product  $\langle \cdot, \cdot \rangle_{p_0}$  and norm  $\| \cdot \|_{p_0}$ ; we reserve the notations without subscripts  $\langle \cdot, \cdot \rangle$  and  $\| \cdot \|$  for the the inner product and norm corresponding to the Lebesgue measure. Due to the atomlessness of  $p_0$ , the tangent space  $\mathcal{T}_{p_0}$  is a subspace of  $\mathcal{L}^2_{p_0}$  equipped with the same inner product and induced norm. The exponential map  $\text{Exp}_{p_0}: \mathcal{T}_{p_0} \rightarrow \mathcal{W}$  is then defined by

$$\text{Exp}_{p_0}g = (g + \text{id})\#p_0, \quad \text{for } g \in \mathcal{T}_{p_0}.$$

Although the exponential map here is not a local homeomorphism as in Riemannian manifolds (Ambrosio et al., 2004), any  $p \in \mathcal{W}$  can be recovered from  $p_0$  by  $\text{Exp}_{p_0}(\Upsilon_{p_0,p} - \text{id})$ , which motivates the definition of the inverse of the exponential map, i.e., the logarithmic map  $\text{Log}_{p_0}: \mathcal{W} \rightarrow \mathcal{T}_{p_0}$ ,

$$\text{Log}_{p_0}p = \Upsilon_{p_0,p} - \text{id}, \quad \text{for } p \in \mathcal{W}.$$

The tangent vector given by log maps quantifies the difference between  $p_0$  and  $p$ . Indeed,  $\|\text{Log}_{p_0} p\|_{p_0} = d_W(p_0, p)$ . Furthermore, the difference between optimal transport maps and the identity map reveals how mass is transported between distributions provided that the order is preserved. Specifically, given  $x \in \mathcal{D}$ , if  $\Upsilon_{p_1, p_2}(x) > x$  (respectively,  $\Upsilon_{p_1, p_2}(x) < x$ ), then  $x$  should be moved to the right (respectively, left) to  $\Upsilon_{p_1, p_2}(x)$  in order to keep its rank, i.e.,  $F_2(\Upsilon_{p_1, p_2}(x)) = F_1(x)$ .

## 2.2. Wasserstein temporal gradients

Let  $(T, P)$  be a pair of random elements in  $\mathcal{T} \times \mathcal{W}$  with joint distribution  $\mathcal{F}$ , where  $\mathcal{T} \subseteq \mathbb{R}$  is the time domain. We assume

(A1)  $P$  is atomless almost surely.

Note that  $\mathbb{E}[d_W^2(P, p) \mid T = t] \leq \text{diam}(\mathcal{D})^2 < \infty$  for all  $p \in \mathcal{W}$  and  $t \in \mathcal{T}$ . Since the Wasserstein space  $\mathcal{W}$  is a Hadamard space (Kloeckner, 2010), there exists a unique minimizer of  $\mathbb{E}[d_W^2(P, \cdot) \mid T = t]$  (Sturm, 2003). Thus, the conditional Fréchet mean  $\mu_{\oplus}(t)$  of  $P$  given  $T = t$  is well-defined; specifically,

$$\mu_{\oplus}(t) = \underset{p \in \mathcal{W}}{\text{argmin}} M(p, t), \quad \text{with } M(p, t) := \mathbb{E}[d_W^2(P, p) \mid T = t],$$

and the quantile function of  $\mu_{\oplus}(t)$  is given by  $F_{\mu_{\oplus}(t)}^{-1}(\cdot) = \mathbb{E}[F_P^{-1}(\cdot) \mid T = t]$ , where  $F_P^{-1}$  is the quantile function of  $P$ .

To model the instantaneous temporal evolution of probability distributions, we are aiming to generalize the notion of derivatives, which are used to quantify the instantaneous change of differentiable real-valued functions, to the scenario of temporal distribution flows. As discussed in Section 2.1, log maps quantify the discrepancy between two probability distributions. We note that the atomlessness of  $\mu_{\oplus}(t)$  is guaranteed by (A1). Hence, a measure of instantaneous temporal evolution of distributions, the Wasserstein temporal gradient at time  $t \in \mathcal{T}$  can be defined by

$$\begin{aligned} V_t &= \lim_{\Delta \rightarrow 0} \frac{\text{Log}_{\mu_{\oplus}(t)} \mu_{\oplus}(t + \Delta)}{\Delta} \\ &= \lim_{\Delta \rightarrow 0} \frac{F_{\mu_{\oplus}(t+\Delta)}^{-1} \circ F_{\mu_{\oplus}(t)} - \text{id}}{\Delta} \\ &= \frac{\partial F_{\mu_{\oplus}(t)}^{-1}}{\partial t} \circ F_{\mu_{\oplus}(t)}, \quad \mu_{\oplus}(t)\text{-a.e.}, \end{aligned} \tag{2.2}$$

provided that the bivariate function  $(t, u) \mapsto F_{\mu_{\oplus}(t)}^{-1}(u)$  is differentiable with respect to  $t$ . Here,  $F_{\mu_{\oplus}(s)}$  and  $F_{\mu_{\oplus}(s)}^{-1}$  are the cumulative distribution function and quantile function of  $\mu_{\oplus}(s)$  for  $s \in \mathcal{T}$ . If there exists  $g \in \mathcal{L}^1(\mathcal{T})$  such that  $d_W(\mu_{\oplus}(s), \mu_{\oplus}(t)) \leq \int_s^t g(x) dx$ , then  $\mu_{\oplus}$  is an absolutely continuous curve in the Wasserstein space, and  $V_t$  is also referred to as the velocity vector of  $\mu_{\oplus}$  (Ambrosio et al., 2004).

*Example 1.* For  $t \in \mathcal{T}$ , let  $\mu_{\oplus}(t) = \mathcal{N}_{[0,1]}(\zeta_t, v_t^2)$  be a truncated Gaussian distribution on the interval  $[0, 1]$ . Then the Wasserstein temporal gradient at  $t$  is

$$V_t(x) = \zeta'_t + (x - \zeta_t) \frac{v'_t}{v_t} - v_t \frac{F_{\mu_{\oplus}(t)}(x) \left[ \frac{v'_t}{v_t} + \left( \frac{\zeta_t}{v_t} \right)' \right] \varphi \left( \frac{1 - \zeta_t}{v_t} \right) + (1 - F_{\mu_{\oplus}(t)}(x)) \left( \frac{\zeta_t}{v_t} \right)' \varphi \left( \frac{-\zeta_t}{v_t} \right)}{\varphi \circ \Phi^{-1} \left( F_{\mu_{\oplus}(t)}(x) \Phi \left( \frac{1 - \zeta_t}{v_t} \right) + (1 - F_{\mu_{\oplus}(t)}(x)) \Phi \left( \frac{-\zeta_t}{v_t} \right) \right)},$$

where  $F_{\mu_{\oplus}(t)}(x) = [\Phi((x - \zeta_t)/v_t) - \Phi(-\zeta_t/v_t)] / [\Phi((1 - \zeta_t)/v_t) - \Phi(-\zeta_t/v_t)]$ ,  $\Phi$  and  $\varphi$  are the cumulative distribution function and density of standard normal distributions, respectively, and we use the notation  $g'_t = (d/dt)g_t = (d/dt)g(t)$  for a function  $g$ . Densities and Wasserstein temporal gradients of  $\mathcal{N}_{[0,1]}(\zeta_t, v_t^2)$  with different values of  $\zeta_t$  and  $v_t$  are shown in Figure 2.

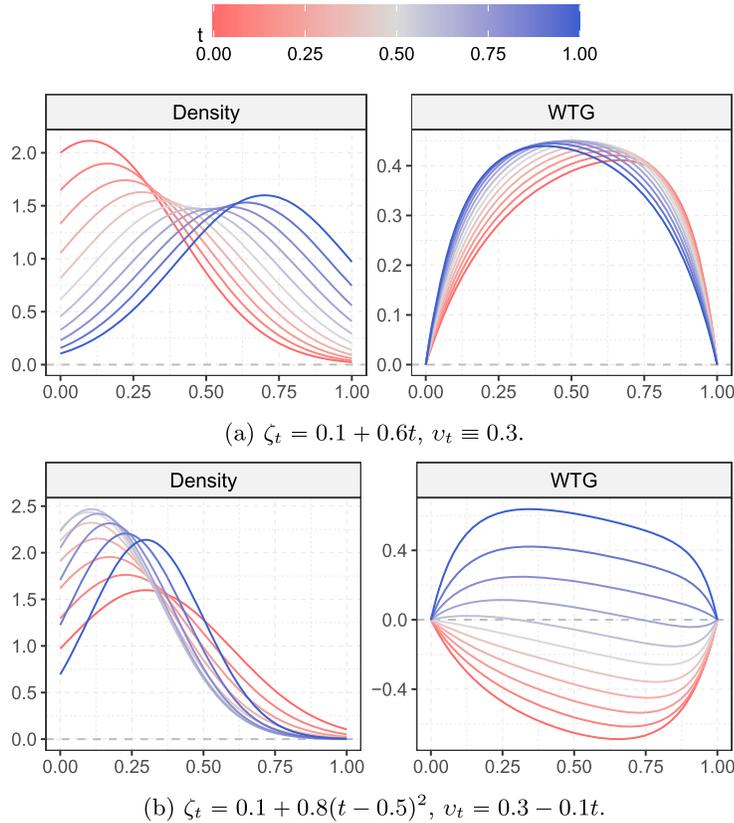


FIG 2. Densities and Wasserstein temporal gradients of  $\mu_{\oplus}(t) = \mathcal{N}_{[0,1]}(\zeta_t, v_t^2)$ , for  $t \in [0, 1]$ .

*Example 2.* For  $t \in \mathcal{T}$ , let  $\mu_{\oplus}(t)$  be atomless distributions in a location-scale family with location and scale parameters being  $\xi_t$  and  $\varsigma_t$ , respectively. Specifically, the cumulative distribution function of  $\mu_{\oplus}(t)$  is given by  $x \mapsto G((x - \xi_t)/\varsigma_t)$ , where  $G: \mathbb{R} \rightarrow [0, 1]$  is a template cumulative distribution function. Then the Wasserstein temporal gradient at  $t \in \mathcal{T}$  is

$$V_t(x) = \xi'_t + (x - \xi_t) \frac{\varsigma'_t}{\varsigma_t}.$$

For a real-valued differentiable function  $g: \mathcal{T} \rightarrow \mathbb{R}$ ,

$$\frac{dF_{\mu_{\oplus}(t)}(g(t))}{dt} = 0 \quad \text{if and only if} \quad g'(t) = V_t(g(t)).$$

Thus, comparing the actual flow  $g'(t)$  for a given longitudinal trajectory  $g(t)$  with the optimal flow  $V_t(g(t))$  provides insights into how the rank of  $g(t)$  changes at each time  $t$ . If  $g'(t) > V_t(g(t))$  (respectively,  $g'(t) < V_t(g(t))$ ), then  $\frac{d}{dt}F_{\mu_{\oplus}(t)}(g(t))$  is positive (respectively, negative), i.e., the rank of  $g(t)$  increases (respectively, decreases) instantaneously at time  $t$ .

### 3. Estimation and theory

#### 3.1. Distribution estimation

In practice, distributions are usually not fully observed. This creates an additional challenge for the implementation of the Wasserstein temporal gradients. This issue can be addressed, for example, by estimating cumulative distribution functions (e.g., Aggarwal, 1955; Read, 1972; Falk, 1983; Leblanc, 2012), or estimating quantile functions (e.g., Parzen, 1979; Falk, 1984; Yang, 1985; Cheng and Parzen, 1997) of the underlying distributions from which the observed data are sampled. Note that with any quantile function estimator  $\widehat{F}^{-1}$  (respectively, cumulative distribution function estimator  $\widehat{F}$ ), the corresponding cumulative distribution function (respectively, quantile function) can be obtained by right (respectively, left) continuous inversion,

$$\begin{aligned} \widehat{F}(x) &= \sup\{u \in [0, 1] : \widehat{F}^{-1}(u) \leq x\}, \quad \text{for } x \in \mathbb{R}, \\ \text{respectively, } \widehat{F}^{-1}(u) &= \inf\{x \in \mathcal{D} : \widehat{F}(x) \geq u\}, \quad \text{for } u \in (0, 1). \end{aligned}$$

Alternatively, one can first estimate densities (Panaretos and Zemel, 2016; Petersen and Müller, 2016) and then obtain the cumulative distribution functions and quantile functions by integration and inversion.

Suppose  $\{(T_i, P_i)\}_{i=1}^n$  are  $n$  independent realizations of  $(T, P)$ . Available observations are samples of independent measurements  $\{X_{ij}\}_{j=1}^{m_i}$  generated from  $P_i$ , respectively, where  $m_i$  are the sample sizes which may vary across distributions  $P_i$ , for  $i = 1, \dots, n$ . Note that the observed data  $X_{ij}$  result from two independent random mechanisms: The first of these generates independently

and identically distributed pairs  $(T_i, P_i)$ ; the second generates samples of observations  $\{X_{ij}\}_{j=1}^{m_i}$  according to each distribution  $P_i$ , i.e.,  $X_{ij} \sim P_i$  independently.

For a given distribution  $p \in \mathcal{W}$ , with a cumulative distribution function estimate  $\hat{F}$  obtained by any estimation method based on a random sample generated from  $p$ , we denote by  $\hat{p} = \pi(\hat{F})$  the distribution associated with  $\hat{F}$ . We make the following assumption on the discrepancy of the estimated and true probability distributions for the theoretical analysis of the Wasserstein temporal gradient estimation.

- (D1) For any distribution  $p \in \mathcal{W}$ , with nonnegative decreasing sequence  $\alpha_m = o(1)$  as  $m \rightarrow \infty$ , the corresponding estimate  $\hat{p}$  based on a sample of size  $m$  generated from  $p$  satisfies

$$\sup_{p \in \mathcal{W}} \mathbb{E}[d_W^2(\hat{p}, p)] = O(\alpha_m).$$

We note that this assumption can be easily satisfied. For example, the density estimator proposed by Panaretos and Zemel (2016) satisfies (D1) with  $\alpha_m = m^{-1/2}$ . If only considering the distributions in  $\mathcal{W}$  with densities satisfying

$$\sup_{x \in \text{supp}(f_p)} \max\{f_p(x), 1/f_p(x), |f'_p(x)|\} \leq C, \text{ uniform across } p,$$

where  $f_p$  is the density function of a distribution  $p \in \mathcal{W}$ ,  $\text{supp}(f_p) = \mathcal{D}$  is the support of  $p$  and  $C > 0$  is a constant, then the empirical measure satisfies (D1) with  $\alpha_m = m^{-1}$ .

In order to deal with the estimation of  $n$  distributions simultaneously, we also require

- (D2) There exists a sequence  $m = m(n)$  such that  $\min_{1 \leq i \leq n} \{m_i\} \geq m$  and  $m \rightarrow \infty$  as  $n \rightarrow \infty$ .

### 3.2. Estimation of Wasserstein temporal gradients

We assume that for each  $i = 1, \dots, n$ , we obtain an estimate  $\hat{F}_{P_i}$  of the cumulative distribution function of  $P_i$  by one of the methods discussed in Section 3.1 from the observed data  $\{X_{ij}\}_{j=1}^{m_i}$ . Denote by  $\hat{P}_i = \pi(\hat{F}_{P_i})$  the distribution associated with  $\hat{F}_{P_i}$ . Since the discrepancy  $\mathbb{E}[d_W^2(P_i, \mu_{\oplus}(T_i)) \mid T_i]$  between the random distributions  $P_i$  and the conditional Fréchet means  $\mu_{\oplus}(T_i)$  does not vanish as  $n \rightarrow \infty$ , difference quotients based on the estimated distributions  $\hat{P}_i$  are not directly suitable as an estimate of Wasserstein temporal gradients.

Accordingly, we utilize local Fréchet regression (Petersen and Müller, 2019a) to smooth the distributions  $\{\hat{P}_i\}$  over time, which yields consistent estimates of  $\mu_{\oplus}(t)$ , for any  $t \in \mathcal{T}$ . Following Petersen and Müller (2019a), we define the localized Fréchet mean by

$$\nu_{\oplus}(t) = \underset{p \in \mathcal{W}}{\text{argmin}} L_h(p, t), \quad \text{with } L_h(p, t) = \mathbb{E}[w(T, t, h)d_W^2(P, p)]. \quad (3.1)$$

Here,  $w(s, t, h) = K_h(s-t)[\kappa_2(t) - \kappa_1(t)(s-t)]/\sigma_0^2(t)$ , where  $\kappa_z(t) = \mathbb{E}[K_h(T-t)(T-t)^z]$ , for  $z = 0, 1, 2$ ,  $\sigma_0^2(t) = \kappa_0(t)\kappa_2(t) - \kappa_1(t)^2$ ,  $K_h(\cdot) = K(\cdot/h)/h$ ,  $K$  is a smoothing kernel, i.e., a density function symmetric around zero, and  $h = h(n) > 0$  is a bandwidth sequence. If assuming the distributions  $P_i$  are fully observed, setting  $\hat{w}(s, t, h) = K_h(s-t)[\hat{\kappa}_2(t) - \hat{\kappa}_1(t)(s-t)]/\hat{\sigma}_0^2(t)$ , where  $\hat{\kappa}_z(t) = n^{-1} \sum_{i=1}^n K_h(T_i-t)(T_i-t)^z$ , for  $z = 0, 1, 2$ , and  $\hat{\sigma}_0^2(t) = \hat{\kappa}_0(t)\hat{\kappa}_2(t) - \hat{\kappa}_1(t)^2$ , an oracle local Fréchet regression estimate is

$$\tilde{\nu}_\oplus(t) = \operatorname{argmin}_{p \in \mathcal{W}} \tilde{L}_n(p, t), \quad \text{with } \tilde{L}_n(p, t) = n^{-1} \sum_{i=1}^n \hat{w}(T_i, t, h) d_W^2(P_i, p). \quad (3.2)$$

In practice, we usually only observe random samples of measurements  $X_{ij}$  generated from  $P_i$ . Replacing  $P_i$  with the corresponding estimates  $\hat{P}_i$  as discussed in Section 3.1, a data-based local Fréchet regression estimate is

$$\hat{\nu}_\oplus(t) = \operatorname{argmin}_{p \in \mathcal{W}} \hat{L}_n(p, t), \quad \text{with } \hat{L}_n(p, t) = n^{-1} \sum_{i=1}^n \hat{w}(T_i, t, h) d_W^2(\hat{P}_i, p). \quad (3.3)$$

For simplicity, we assume for theoretical analysis that the support of the marginal density  $f_T$  of  $T$ , i.e.,  $\mathcal{T} = \{t \in \mathbb{R} : f_T(t) > 0\}$ , is connected. Let  $\mathcal{T}^\circ$  be the interior of  $\mathcal{T}$ . Furthermore, we require the following assumptions for the asymptotic analysis of the weights  $w(T, t, h)$  and  $\hat{w}(T_i, t, h)$  in  $\nu_\oplus(t)$  and  $\hat{\nu}_\oplus(t)$ , respectively.

- (R1) The kernel  $K$  is a probability density function, symmetric around zero and continuous on  $[-1, 1]$ , such that  $K(x) = 0$ , for all  $|x| > 1$ .
- (R2) The marginal density  $f_T$  of  $T$  exists and is continuous on  $\mathcal{T}$  and twice continuously differentiable on  $\mathcal{T}^\circ$ . The second-order derivative  $f_T''$  is bounded,  $\sup_{t \in \mathcal{T}^\circ} |f_T''(t)| < \infty$ .

For any  $t \in \mathcal{T}^\circ$ , with the local Fréchet regression estimate  $\hat{\nu}_\oplus(t)$  as per (3.3) and some small  $\Delta > 0$ , an estimate of the Wasserstein temporal gradient  $V_t$  in (2.2) is then given by

$$\hat{V}_{t,\Delta} = \frac{F_{\hat{\nu}_\oplus(t+\Delta)}^{-1} \circ F_{\hat{\nu}_\oplus(t)} - \operatorname{id}}{\Delta}, \quad (3.4)$$

where  $F_{\hat{\nu}_\oplus(s)}$  and  $F_{\hat{\nu}_\oplus(s)}^{-1}$  are the cumulative distribution function and quantile function of  $\hat{\nu}_\oplus(s)$  for  $s \in \mathcal{T}$ .

### 3.3. Parallel transport

Note that the true and estimated Wasserstein temporal gradients lie in different tangent spaces; specifically,  $V_t \in \mathcal{T}_{\mu_\oplus(t)}$  and  $\hat{V}_{t,\Delta} \in \mathcal{T}_{\hat{\nu}_\oplus(t)}$ . To quantify the estimation discrepancy of  $\hat{V}_{t,\Delta}$ , an expedient tool is parallel transport, which is commonly used for manifold-valued data (e.g., Yuan et al., 2012; Lin and

Yao, 2019; Petersen and Müller, 2019b; Chen et al., 2020). For two probability measures  $p_1, p_2 \in \mathcal{W}$ , a parallel transport operator  $\Gamma_{p_1, p_2} : \mathcal{L}_{p_1}^2 \rightarrow \mathcal{L}_{p_2}^2$  is defined by

$$\Gamma_{p_1, p_2} g = g \circ F_1^{-1} \circ F_2, \quad \text{for } g \in \mathcal{L}_{p_1}^2,$$

where  $F_2$  and  $F_1^{-1}$  are the cumulative distribution function of  $p_2$  and quantile function of  $p_1$ , respectively.

Note that since  $p_1$  and  $p_2$  are atomless, the tangent spaces satisfy  $\mathcal{T}_{p_k} \subset \mathcal{L}_{p_k}^2$  for  $k = 1, 2$ , and the parallel transport operator  $\Gamma_{p_1, p_2}|_{\mathcal{T}_{p_1}}$  restricted to the tangent space  $\mathcal{T}_{p_1}$  defines the parallel transport between tangent spaces  $\mathcal{T}_{p_1}$  and  $\mathcal{T}_{p_2}$ . Furthermore, the parallel transport operator  $\Gamma_{p_2, p_1}$  from  $\mathcal{L}_{p_2}^2$  to  $\mathcal{L}_{p_1}^2$  is the adjoint operator of  $\Gamma_{p_1, p_2}$ , i.e.,  $\langle \Gamma_{p_1, p_2} g_1, g_2 \rangle_{p_2} = \langle g_1, \Gamma_{p_2, p_1} g_2 \rangle_{p_1}$ . Thus, the discrepancy between functions  $g_1 \in \mathcal{T}_{p_1}$  and  $g_2 \in \mathcal{T}_{p_2}$  can be quantified by  $\|\Gamma_{p_2, p_1} g_2 - g_1\|_{p_1}$ .

### 3.4. Asymptotic theory

As discussed in Section 3.3, in order to justify  $\|\Gamma_{\hat{\nu}_{\oplus}(t), \mu_{\oplus}(t)} \hat{V}_{t, \Delta} - \bar{V}_t\|_{\mu_{\oplus}(t)}$  as a measure of estimation discrepancy of  $\hat{V}_{t, \Delta}$ , we require the atomlessness of  $\mu_{\oplus}(t)$  and  $\hat{\nu}_{\oplus}(t)$ . The former follows from (A1). However, the latter is not guaranteed in general. For theoretical derivations, we instead consider an atomless variant  $\check{\nu}_{\oplus}(t)$  of  $\hat{\nu}_{\oplus}(t)$ , which is defined as follows. Suppose  $\inf \mathcal{D} = x_0 < x_1 < \dots < x_B = \sup \mathcal{D}$  is an equidistant grid on  $\mathcal{D}$  with increment  $b$ . Then the cumulative distribution function of  $\check{\nu}_{\oplus}(t)$  is given by  $F_{\check{\nu}_{\oplus}(t)}(x) = F_{\hat{\nu}_{\oplus}(t)}(x_{l-1}) + b^{-1}(x - x_{l-1})[F_{\hat{\nu}_{\oplus}(t)}(x_l) - F_{\hat{\nu}_{\oplus}(t)}(x_{l-1})]$ , for  $x \in [x_{l-1}, x_l]$ ;  $F_{\check{\nu}_{\oplus}(t)}(x) = 0$  and  $1$  for  $x < x_0$  and  $x \geq x_B$ , respectively. We assume that  $b = b(n)$  is a positive sequence such that  $b \rightarrow 0$  as  $n \rightarrow \infty$ . Hence, an estimate of the Wasserstein temporal gradient at time  $t$  based on  $\check{\nu}_{\oplus}(s)$  with  $s \in \mathcal{T}$  is given by

$$\check{V}_{t, \Delta} = \frac{F_{\check{\nu}_{\oplus}(t+\Delta)}^{-1} \circ F_{\check{\nu}_{\oplus}(t)} - \text{id}}{\Delta}.$$

To obtain the convergence rate of  $\check{V}_{t, \Delta}$ , we also require the following assumption.

- (A2) The bivariate function  $(t, u) \mapsto F_{\mu_{\oplus}(t)}^{-1}(u)$  is twice differentiable;  $(t, u) \mapsto \partial^2 F_{\mu_{\oplus}(t)}^{-1}(u) / (\partial t \partial u)$  is continuous with respect to  $u$ . There exists a constant  $C > 0$  such that  $\sup_{x \in \mathcal{D}} f_{\mu_{\oplus}(t)}(x) \leq C$ ,  $\int_0^1 \sup_{t \in \mathcal{T}} |\partial^2 F_{\mu_{\oplus}(t)}^{-1}(u) / \partial t^2|^2 du \leq C$ , and  $\sup_{t \in \mathcal{T}, u \in (0, 1)} |\partial^2 F_{\mu_{\oplus}(t)}^{-1}(u) / (\partial t \partial u)| \leq C$ .

We take  $\Delta = h$ ; this choice, together with suitable values for  $h, b$  and  $\Delta$ , will lead to Wasserstein temporal gradient estimates  $\check{V}_{t, \Delta}$  with an asymptotic rate of convergence that matches the well-known optimal rate of derivative estimation for nonparametric regression for the case of real-valued responses assuming twice continuous differentiability of the regression function. This optimal rate is

for example achieved by derivative estimates based on local polynomial fitting (Müller, 1987; Fan and Gijbels, 1996).

**Theorem 1.** *Assume (A1)–(A2), (D1)–(D2) and (R1)–(R2). With  $\Delta = h$ , if  $h \rightarrow 0$ ,  $nh^3 \rightarrow \infty$ ,  $bh^{-1} \rightarrow \infty$ , and  $\alpha_m h^{-1} \rightarrow 0$ ,*

$$\begin{aligned} & \left\| \Gamma_{\check{\nu}_{\oplus}(t), \mu_{\oplus}(t)} \check{V}_{t, \Delta} - V_t \right\|_{\mu_{\oplus}(t)} \\ &= \left\| \frac{F_{\check{\nu}_{\oplus}(t+h)}^{-1} - F_{\check{\nu}_{\oplus}(t)}^{-1}}{h} - \frac{\partial F_{\mu_{\oplus}(t)}^{-1}}{\partial t} \right\| \\ &= O(h) + O_p\left((nh^3)^{-1/2}\right) + O(bh^{-1}) + O_p\left((\alpha_m h^{-1})^{1/2}\right). \end{aligned}$$

Furthermore, with  $h \sim n^{-1/5}$ ,  $b = O(n^{-2/5})$  and  $\alpha_m = O(n^{-3/5})$ ,

$$\left\| \Gamma_{\check{\nu}_{\oplus}(t), \mu_{\oplus}(t)} \check{V}_{t, \Delta} - V_t \right\|_{\mu_{\oplus}(t)} = O_p\left(n^{-1/5}\right). \quad (3.5)$$

Proofs are in the appendix.

#### 4. Implementation and simulations

There are two tuning parameters for implementation of Wasserstein temporal gradients, namely the bandwidth  $h$  involved in the local Fréchet regression as per (3.3) and the time increment  $\Delta$  used in the difference quotient estimator as per (3.4). As suggested by the theoretical analysis in Section 3.4, we take  $\Delta = h$  in practice. We choose the bandwidth  $h$  by leave-one-out cross validation, where the objective function to be minimized is the mean discrepancy between the local Fréchet regression estimates and the observed distributions; specifically,

$$h = \operatorname{argmin}_{h'} n^{-1} \sum_{i=1}^n d_W^2(\hat{\nu}_{\oplus h'}^{-i}(T_i), \hat{P}_i),$$

where  $\hat{\nu}_{\oplus h'}^{-i}(T_i)$  is the local Fréchet regression estimate of  $\mu_{\oplus}(T_i)$  obtained with bandwidth  $h'$  based on the sample excluding the  $i$ th pair  $(T_i, \hat{P}_i)$ , i.e.,

$$\hat{\nu}_{\oplus h'}^{-i}(T_i) = \operatorname{argmin}_{p \in \mathcal{W}} \frac{1}{n-1} \sum_{i' \neq i} \hat{w}(T_{i'}, T_i, h') d_W^2(\hat{P}_{i'}, p),$$

and  $\hat{P}_i$  is the estimate of  $P_i$  based on the observed measurements  $\{X_{ij}\}_{j=1}^{m_i}$  as discussed in Section 3.1. In practice, we replace leave-one-out cross validation by 10-fold cross validation when  $n > 30$ .

We generated data for simulations as follows:

Step 1: Set  $\mu_{\oplus}(t) = \mathcal{N}_{[0,1]}(\zeta_t, v_t^2)$ , a truncated Gaussian distribution on  $[0, 1]$  with  $\zeta_t = (t - 0.2)(t - 0.5)(t - 0.9) + 0.2$  and  $v_t = 0.15 + 0.03 \sin(2\pi t)$ .

- Step 2: Sample  $a_i \sim \text{Unif}\{\pm 10\pi, \pm 11\pi, \dots, \pm 14\pi\}$  and  $T_i \sim \text{Unif}[0, 1]$  independently, for  $i = 1, \dots, n$ . Set  $P_i = g_{a_i} \# \mu_{\oplus}(T_i)$ , where  $g_a(x) = x - |a|^{-1} \sin(ax)$  with  $a \in \mathbb{R} \setminus \{0\}$  and  $x \in \mathbb{R}$ .
- Step 3: Draw an independently and identically distributed sample  $\{X_{ij}\}_{j=1}^m$  of size  $m$  from each of the distributions  $\{P_i\}_{i=1}^n$ .

Four cases were considered with  $n \in \{50, 200\}$  and  $m \in \{25, 500\}$ . We simulated 500 runs for each pair  $(n, m)$ . To evaluate the performance of the Wasserstein temporal gradient estimate based on the local Fréchet regression as per (3.4), we computed the integrated error (IE) for given  $t \in [0, 1]$ ; specifically,

$$\text{IE}(n, m, t) = \left\| \frac{F_{\hat{\nu}_{\oplus}(t+\Delta)}^{-1} - F_{\hat{\nu}_{\oplus}(t)}^{-1}}{\Delta} - \frac{\partial}{\partial t} F_{\mu_{\oplus}(t)}^{-1} \right\|. \tag{4.1}$$

The results are summarized in the boxplots of IEs in Figure 3. It can be seen that the estimation error decreases as  $n$  or  $m$  increases.

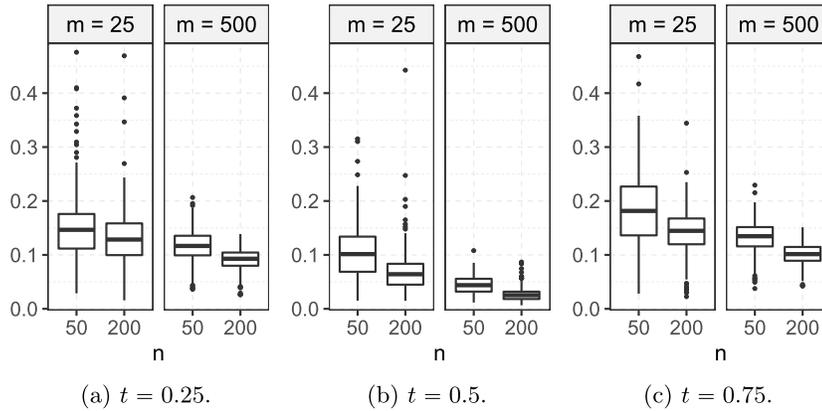


FIG 3. Boxplots of the integrated errors (IEs) as per (4.1) of the 500 runs for  $t \in \{0.25, 0.5, 0.75\}$  and each  $(n, m) \in \{50, 200\} \times \{25, 500\}$ .

### 5. Applications

In this section, we will demonstrate the proposed Wasserstein gradients for time-dependent household income and human mortality data. As mentioned before, the underlying densities are practically never known and need to be estimated from data that they generate. In the household income and mortality examples, the data are reported in the form of histograms, respectively life tables. Our methods can be applied in a straightforward way to histogram data; specifically we estimate the densities by applying a smoothing step, e.g., using local linear regression. For local Fréchet regression, we use the Epanechnikov kernel function  $K(t) = 0.75(1 - t^2)\mathbf{1}_{[-1,1]}(t)$  and choose smoothing bandwidths  $h$  by cross validation.

### 5.1. Household income data

Many studies have been conducted on income distribution and inequality (Jones, 1997; Heathcote et al., 2010), since this is a major measure of economic equality/inequality. The evolution of income distributions over time is of particular interest as it provides quantification of the directions in which income inequality is evolving. The US Census Bureau provides histogram data of US household income over calendar years from 1994 to 2016, available at [www.census.gov](http://www.census.gov). To make incomes of different years comparable, adjustments for inflation have been made, using the year 2000 as baseline for constant dollars.

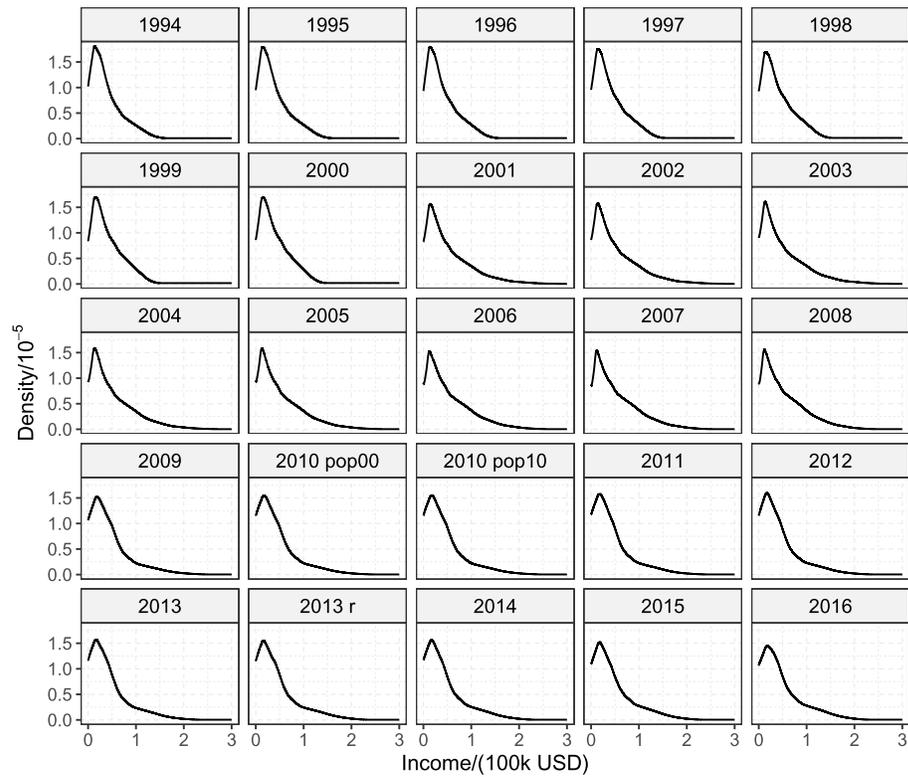


FIG 4. Densities of US household incomes for 1994–2016, where “2010 pop00” and “2010 pop10” represent the distribution of 2010 based on the population census of 2000 and 2010, respectively, and “2013” and “2013 r” represent the distributions on previous and redesigned questionnaires, respectively.

We focus on incomes less than \$300,000. The data require some preprocessing, as the width of the histogram bins changed between 2000 and 2001; for 2010, due to census changes, two datasets are available based on both census 2010 and 2000 populations; for 2013, two sets of data are also available and one of them is based on a redesigned questionnaire which has been used since

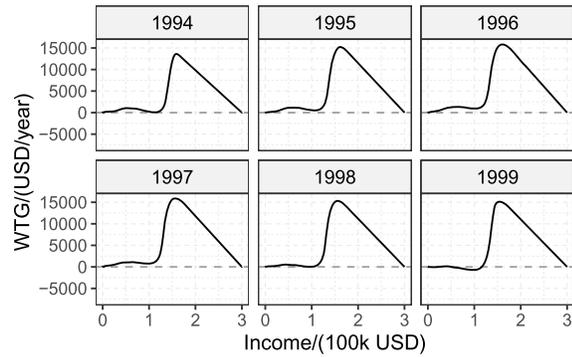
then. To mitigate against these changes, which potentially introduce artificial variation, we divided the whole period into four parts: 1994–2000, 2001–2010, 2010–2013 and 2013–2016. Although another change of bin width occurred between 2008 and 2009, we keep the entire period 2001–2010 in order to cover the financial crisis of 2008 well within the time interval. The densities constructed by smoothing the histogram data are shown in Figure 4, where “2010 pop00”, “2010 pop10”, “2013” and “2013 r” represents the income distribution of 2010 based on the population census of 2000 and 2010, and of 2013 based the previous and redesigned questionnaires, respectively.

Figure 4 reveals not much variation in the income distributions over time except around 2008. The estimated Wasserstein temporal gradients as per (3.4) (with bandwidths  $h = 1.99, 1.55, 1.50$  and  $1.50$  years for the four periods, respectively, chosen by cross validation, see Section 4 and time increment  $\Delta = h$ ) demonstrate how the income of poor, middle-class and rich households evolved for the four periods in Figure 5. The Wasserstein temporal gradients for the ending years of each period cannot be well estimated due to the relative large value of  $\Delta$ , and hence the results for 2000, “2010 pop00”, “2013” and 2016 are not displayed. Since the local Fréchet regression has increased variance near endpoints, the estimated Wasserstein temporal gradients on the two ends of each period are somewhat unreliable.

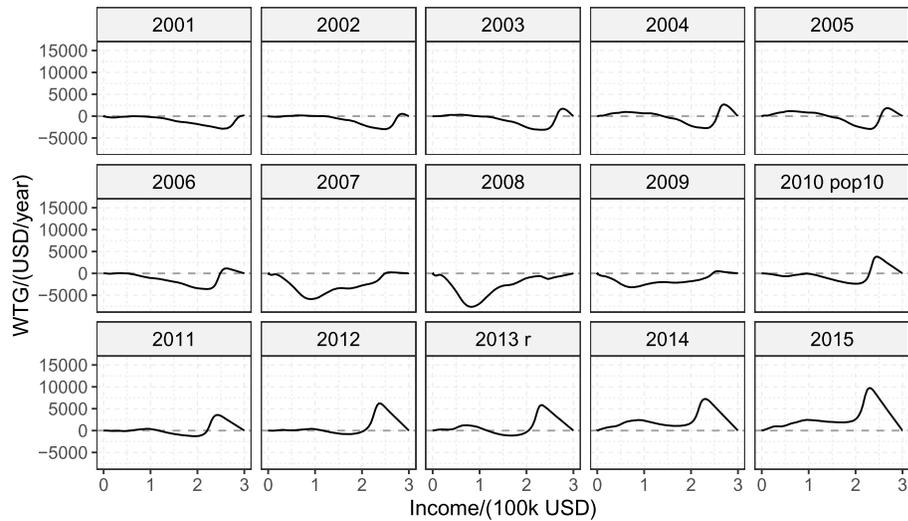
It can be seen in Figure 5 that for the period 1994–1999, incomes of households at the same percentile levels increased almost throughout, except for relatively poor households whose incomes tended to decrease in 1999. Incomes of households earning more than \$150,000 per year increased much faster than the other incomes. For the second period 2001–2010, the economic status of the lower and middle earners was stable in the first three years, rose in 2004–2006, and then declined starting in 2007. Higher incomes declined until 2002, and beginning in 2003, a divide manifested itself in the higher income levels: The lower tier of higher incomes was associated with declining income, whereas the higher tier was associated with increasing income, except for 2007 and 2008. Note that in 2007 and 2008, all household incomes tended to decrease, coinciding with the financial crisis. For the last two periods, it can be seen that household incomes gradually recovered from the crisis. While top incomes above 240,000 US dollars always gained, households with relatively low incomes did not recover until around 2014.

## 5.2. Human mortality data

The analysis of mortality data across countries and species has found interest in demography and statistics (Carey et al., 1992; Chiou and Müller, 2009; Ouellette and Bourbeau, 2011; Hyndman et al., 2013; Shang and Hyndman, 2017). Of particular interest is how the distribution of age-of-death evolves over time. The Human Mortality Database ([www.mortality.org](http://www.mortality.org)) provides data of yearly life tables for 37 countries or areas, from which the distributions of ages-at-death in terms of histograms can be extracted.



(a) 1994–1999.



(b) 2001–2015.

FIG 5. Estimates of the Wasserstein temporal gradients (solid curves) as per (3.4) for US household income distribution flows for 1994–1999 and 2001–2015. Positive values indicate an increasing trend; negative values indicate a decreasing trend.

We focus on ages-at-death in the age interval  $[0, 100]$  (in years) and take Russia, Sweden and the United States as three examples. The densities obtained by smoothing the histogram data for females and males separately are shown in Figures 6, 7 and 1, respectively. It can be seen that densities of mortality and their changes vary across these three countries, which is partly due to the different domains in terms of calendar years during which country-specific mortality has been recorded, which goes much further into the past for Sweden than for the other countries. Estimates of the Wasserstein temporal gradients have been obtained with bandwidths  $h$  chosen by cross validation as discussed in Section 4 per gender and country (see Table 1 for details) and  $\Delta = h$ .

TABLE 1  
Bandwidths used in the local Fréchet regression for the age-at-death distributions.

	Russia	Sweden	USA
Females	1.87	2.09	2.40
Males	1.66	1.78	2.52

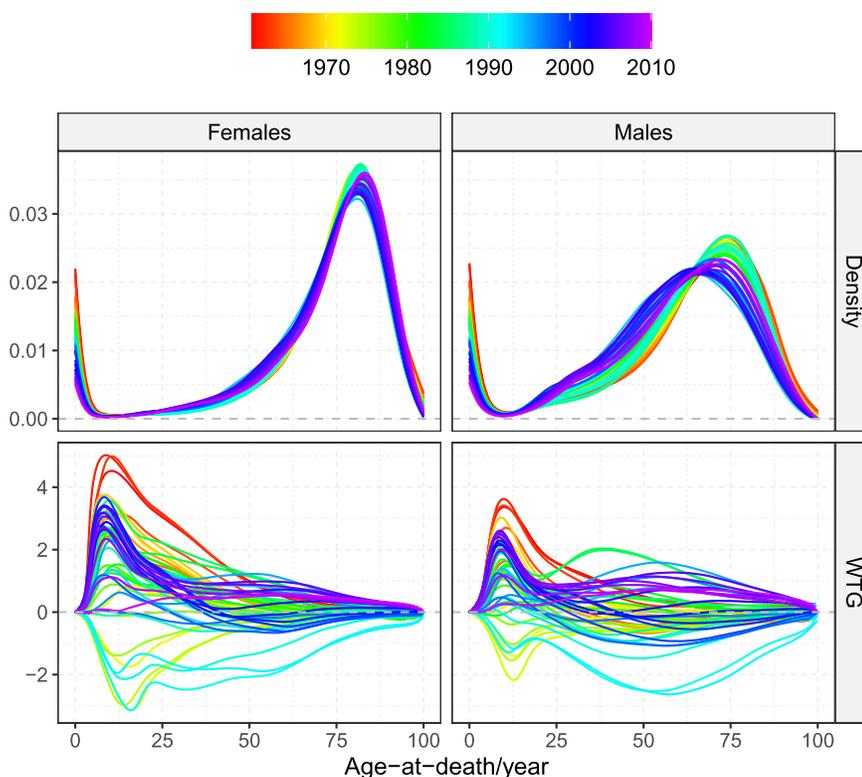


FIG 6. Top: Time-varying densities of age-at-death (in years) with females in the left column and males in the right column for Russia from 1961 to 2010. Bottom: Estimates of the (unit-free) Wasserstein temporal gradients of the age-at-death (in years) distributions from 1961 to 2010, where positive values indicate increasing trend and negative values indicate decreasing trend.

For Russia, the densities of ages-at-death from 1961 to 2010 are shown in the top two panels in Figure 6. The age-at-death distributions are quite different between females and males; female adults tend to live longer than males. The estimates of the Wasserstein temporal gradients for 1961–2010 for Russia are shown in the bottom two panels in Figure 6, and were obtained based on data from 1959 to 2014; the estimated gradients for the first and last two years were excluded due to boundary effects. Between 1970 and 2000, the movement of mortality to higher ages and thus longer life was interrupted, with a lot of

variation during this period, and resumed only in the 2000s, where substantial improvement occurs in children's mortality. In the 1990s, there was a remarkable reversal in the trend of longevity increase, as the estimates of the Wasserstein temporal gradients were negative for those years, especially for young females and mid-age males.

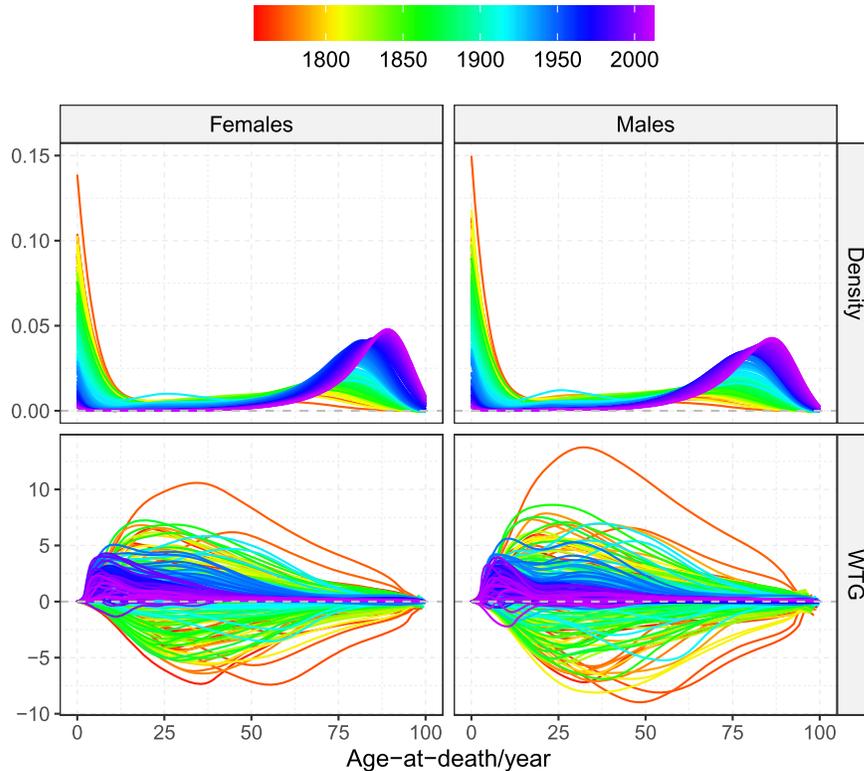


FIG 7. *Top: Time-varying densities of age-at-death (in years) with females in the left column and males in the right column for Sweden from 1754 to 2012. Bottom: Estimates of the (unit-free) Wasserstein temporal gradients for the distributions of age-at-death (in years) distributions from 1754 to 2012, where positive values indicate increasing trend; negative values indicate decreasing trend.*

For Sweden, as shown in Figure 7, the densities of ages-at-death of females and males are quite similar, indicating a general increase in longevity over the years. The estimated Wasserstein temporal gradients for Sweden in Figure 7 from 1754 to 2012, which are obtained based on data from 1751 to 2016, show some volatility in the age-at-death distributions for both females and males, especially before 1950. Compared to Russia, the evolution of the age-at-death distributions in Sweden is more balanced—years where the distribution moves to the left (right) are followed by years with a rightward (leftward) movement in the distribution. The Wasserstein temporal gradients for Sweden vary in a

much larger range than Russia, which is partly due to the inclusion of early calendar years, where the variation of mortality from year to year was much larger, compared to more recent calendar years. For example, the top orange curve for 1773 demonstrates a massive increasing trend in life span for both females and males while the bottom orange curves for 1770–1771 demonstrate a strongly decreasing trend.

For the US, the age-at-death distributions are somewhat similar across genders. The estimates of the Wasserstein temporal gradients from 1936 to 2010 obtained based on data from 1933 to 2015 for the US in Figure 8 indicate that age-at-death distributions tend to move to the right in almost all years, suggesting increasing longevity. However, for several of the years since the 1980s, reversals can be found for both females and males. A major reversal can be found for the males from young adults to middle age during 1983–1987. This puzzling reversal has been attributed to drug use (e.g., Case and Deaton, 2015).

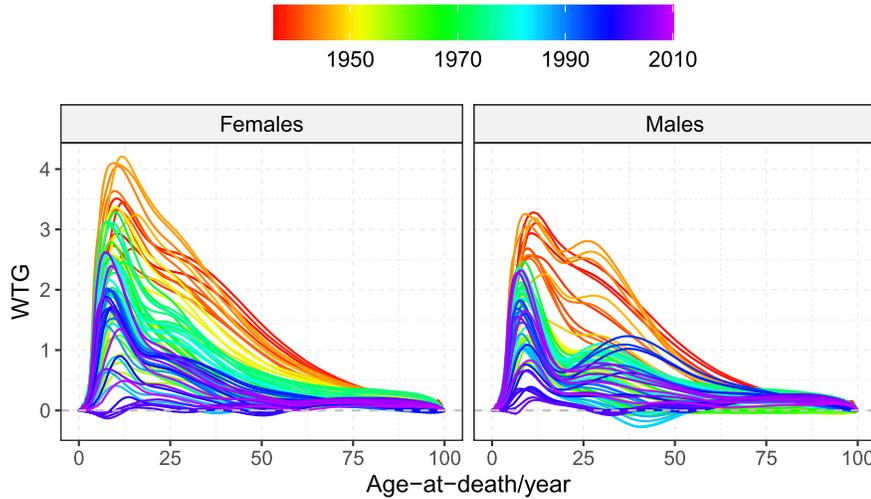


FIG 8. Estimates of the (unit-free) Wasserstein temporal gradients of the age-at-death (in years) distributions for the US from 1936 to 2010, with females on the left and males on the right. Positive values indicate increasing trend; negative values indicate decreasing trend.

### Appendix A: Derivation of Theorem 1

For  $t \in \mathcal{T}$ , we define

$$\begin{aligned}
 q_t(\cdot) &= \mathbb{E} [w(T, t, h)F_P^{-1}(\cdot)], \\
 \bar{q}_t(\cdot) &= n^{-1} \sum_{i=1}^n w(T_i, t, h)F_{P_i}^{-1}(\cdot), \\
 \tilde{q}_t(\cdot) &= n^{-1} \sum_{i=1}^n \hat{w}(T_i, t, h)F_{P_i}^{-1}(\cdot),
 \end{aligned}$$

$$\hat{q}_t(\cdot) = n^{-1} \sum_{i=1}^n \hat{w}(T_i, t, h) \hat{F}_{P_i}^{-1}(\cdot),$$

where  $F_P^{-1}$ ,  $F_{P_i}^{-1}$  and  $\hat{F}_{P_i}^{-1}$  are the quantile functions of  $P$ ,  $P_i$  and  $\hat{P}_i$ , respectively. Considering any fixed  $t \in \mathcal{T}^\circ$ , we will show that

$$\left\| \frac{q_{t+h} - q_t}{h} - \frac{\partial F_{\mu_\oplus}^{-1}}{\partial t} \right\| = O(h). \quad (\text{A.1})$$

For  $q_t$ , we will show that with sufficiently small  $h$ ,

$$\partial q_t(u)/\partial u > 0, \quad \text{for all } u \in (0, 1), \quad (\text{A.2})$$

i.e.,  $q_t$  is a quantile function, whence we will show that

$$\left\| F_{\hat{\nu}_\oplus(t)}^{-1} - q_t \right\| \leq \|\hat{q}_t - q_t\|. \quad (\text{A.3})$$

Hence,

$$\begin{aligned} & \left\| \frac{F_{\hat{\nu}_\oplus(t+h)}^{-1} - F_{\hat{\nu}_\oplus(t)}^{-1}}{h} - \frac{q_{t+h} - q_t}{h} \right\| \\ & \leq h^{-1} \left( \left\| F_{\hat{\nu}_\oplus(t+h)}^{-1} - q_{t+h} \right\| + \left\| F_{\hat{\nu}_\oplus(t)}^{-1} - q_t \right\| \right) \\ & \leq h^{-1} (\|\hat{q}_{t+h} - q_{t+h}\| + \|\hat{q}_t - q_t\|). \end{aligned}$$

Furthermore, we will show that

$$\begin{aligned} \|\hat{q}_{t+h} - \tilde{q}_{t+h}\| &= O_p\left((\alpha_m h)^{1/2}\right), \\ \|\tilde{q}_{t+h} - \bar{q}_{t+h}\| &= O_p\left((nh)^{-1/2}\right), \\ \|\bar{q}_{t+h} - q_{t+h}\| &= O_p\left((nh)^{-1/2}\right). \end{aligned} \quad (\text{A.4})$$

Similar results hold when replacing  $t+h$  with  $t-h$ . We note that for all  $t \in \mathcal{T}$ ,  $\|F_{\check{\nu}_\oplus(t)}^{-1} - F_{\hat{\nu}_\oplus(t)}^{-1}\| \leq b$ , a.s. In conjunction with the atomlessness of  $\mu_\oplus(t)$  and  $\check{\nu}_\oplus(t)$ , taking  $\Delta = h$  yields

$$\begin{aligned} & \left\| \Gamma_{\check{\nu}_\oplus(t), \mu_\oplus(t)} \check{V}_{t, \Delta} - V_t \right\|_{\mu_\oplus(t)} \\ &= \left\| \frac{F_{\check{\nu}_\oplus(t+h)}^{-1} - F_{\check{\nu}_\oplus(t)}^{-1}}{h} - \frac{\partial F_{\mu_\oplus}^{-1}}{\partial t} \right\| \\ &= O(h) + O_p\left((nh^3)^{-1/2}\right) + O(bh^{-1}) + O_p\left((\alpha_m h^{-1})^{1/2}\right), \end{aligned}$$

whence (3.5) follows with  $h \sim n^{-1/5}$ ,  $b = O(n^{-2/5})$  and  $\alpha_m = O(n^{-3/5})$ . Next, we will prove (A.1)–(A.4), respectively.

For any given  $t_0 \in \mathcal{T}^\circ$ , there exists  $\rho > 0$  such that  $[t_0 - \rho, t_0 + \rho] \subset \mathcal{T}$ . We note that under (R1) and (R2), as  $h \rightarrow 0$ , it holds for  $z \in \{0, 1\}$  that

$$\begin{aligned} \mathbb{E} [K_h(T-t)(T-t)^{2z}] &= h^{2z} [f_T(t)\mathcal{K}_{1,2z} + O(h^2)], \\ \mathbb{E} [K_h(T-t)(T-t)^{2z+1}] &= h^{2z+2} [f'_T(t)\mathcal{K}_{1,2z+2} + o(1)], \\ \mathbb{E} [K_h(T-t)|T-t|^{2z+1}] &= O(h^{2z+1}), \\ \mathbb{E} [K_h(T-t)(T-t)^z]^2 &= O(h^{2z-1}), \end{aligned} \tag{A.5}$$

where  $\mathcal{K}_{k,l} = \int_{-1}^1 K(s)^k s^l ds$ , for  $k, l \in \mathbb{N}$  and the  $O$  and  $o$  terms are uniform in  $t \in [t_0 - \rho, t_0 + \rho]$ . For the following proofs, we define  $\kappa_z^+(t) = \mathbb{E}[K_h(T_i-t)|T_i-t|^z]$  and  $\hat{\kappa}_z^+(t) = n^{-1} \sum_{i=1}^n [K_h(T_i-t)|T_i-t|^z]$ , for  $z = 0, 1, 2, 3$ .

*Proof of (A.1).* Applying a Taylor expansion yields

$$\begin{aligned} &\left\| \frac{q_{t+h} - q_t}{h} - \frac{\partial F_{\mu_{\oplus}(t)}^{-1}}{\partial t} \right\| \\ &\leq \frac{1}{2h} [\mathbb{E} |w(T, t+h, h)(T-t)^2| + \mathbb{E} |w(T, t, h)(T-t)^2|] \\ &\quad \times \left[ \int_0^1 \sup_{t' \in \mathcal{T}} \left| \frac{\partial^2 F_{\mu_{\oplus}(t')}^{-1}(u)}{\partial t'^2} \right|^2 du \right]^{1/2}. \end{aligned}$$

Under (R1) and (R2), for  $z = 0, 1, 2, 3$ , it follows from similar arguments to (A.5) that as  $h \rightarrow 0$ ,

$$\begin{aligned} &\mathbb{E} |w(T, t+h, h)(T-(t+h))^z| \\ &\leq \frac{\kappa_2(t+h)\kappa_z^+(t+h) + |\kappa_1(t+h)|\kappa_{z+1}^+(t+h)}{\sigma_0^2(t+h)} = O(h^z). \end{aligned}$$

Hence,

$$\begin{aligned} &\mathbb{E} |w(T, t+h, h)(T-t)^2| \\ &= \mathbb{E} |w(T, t+h, h)(T-(t+h))^2| + 2h\mathbb{E} |w(T, t+h, h)(T-(t+h))| \\ &\quad + h^2\mathbb{E} |w(T, t+h, h)| \\ &= O(h^2). \end{aligned}$$

Similarly,  $\mathbb{E} |w(T, t, h)(T-t)^2| = O(h^2)$ . In conjunction with (A2), (A.1) follows.  $\square$

*Proof of (A.2).* We note that by (A.5) and (A2),

$$\begin{aligned} &\sup_{u \in (0,1)} \left| \frac{\partial q_t(u)}{\partial u} - \frac{\partial F_{\mu_{\oplus}(t)}^{-1}(u)}{\partial u} \right| \\ &\leq \mathbb{E} |w(T, t, h)(T-t)| \sup_{t' \in \mathcal{T}, u \in (0,1)} \left| \frac{\partial^2 F_{\mu_{\oplus}(t')}^{-1}(u)}{\partial t' \partial u} \right| \\ &= O(h), \end{aligned}$$

and that

$$\inf_{u \in (0,1)} \frac{\partial F_{\mu_{\oplus}(t)}^{-1}(u)}{\partial u} = \left( \sup_{x \in \mathcal{D}} f_{\mu_{\oplus}(t)}(x) \right)^{-1} \geq C^{-1} > 0.$$

Thus, with sufficiently small  $h$ ,  $\partial q_t(u)/\partial u > 0$ , for all  $u \in (0, 1)$ .  $\square$

*Proof of (A.3).*

$$\|\hat{q}_t - q_t\|^2 - \|F_{\hat{\nu}_{\oplus}(t)}^{-1} - q_t\|^2 = \|\hat{q}_t - F_{\hat{\nu}_{\oplus}(t)}^{-1}\|^2 + 2 \langle \hat{q}_t - F_{\hat{\nu}_{\oplus}(t)}^{-1}, F_{\hat{\nu}_{\oplus}(t)}^{-1} - q_t \rangle.$$

If  $\langle \hat{q}_t - F_{\hat{\nu}_{\oplus}(t)}^{-1}, F_{\hat{\nu}_{\oplus}(t)}^{-1} - q_t \rangle < 0$ , then there exists  $u \in (0, 1)$  such that

$$\begin{aligned} & \|uq_t + (1-u)F_{\hat{\nu}_{\oplus}(t)}^{-1} - \hat{q}_t\|^2 - \|F_{\hat{\nu}_{\oplus}(t)}^{-1} - \hat{q}_t\|^2 \\ &= u^2 \|F_{\hat{\nu}_{\oplus}(t)}^{-1} - q_t\|^2 + 2u \langle \hat{q}_t - F_{\hat{\nu}_{\oplus}(t)}^{-1}, F_{\hat{\nu}_{\oplus}(t)}^{-1} - q_t \rangle < 0, \end{aligned}$$

which contradicts the fact that  $F_{\hat{\nu}_{\oplus}(t)}^{-1} = \operatorname{argmin}_{F^{-1}} n^{-1} \sum_{i=1}^n \hat{w}(T_i, t, h) \|\hat{F}_{P_i}^{-1} - F^{-1}\|^2 = \operatorname{argmin}_{F^{-1}} \|\hat{q}_t - F^{-1}\|^2$ , where the minimization is over all the quantile functions of distributions in  $\mathcal{W}$ . Therefore,  $\langle \hat{q}_t - F_{\hat{\nu}_{\oplus}(t)}^{-1}, F_{\hat{\nu}_{\oplus}(t)}^{-1} - q_t \rangle \geq 0$ , whence (A.3) follows.  $\square$

*Proof of (A.4).* Under (R1)–(R2) and (D1)–(D2),

$$\begin{aligned} & \mathbb{E} \left( \|\hat{q}_{t+h} - \tilde{q}_{t+h}\|^2 \right) \\ & \leq \mathbb{E} \left( \frac{1}{n} \sum_{i=1}^n \hat{w}(T_i, t+h, h)^2 \|\hat{F}_{P_i}^{-1} - F_{P_i}^{-1}\|^2 \right) \\ & = \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \hat{w}(T_i, t+h, h)^2 \mathbb{E} \left( \|\hat{F}_{P_i}^{-1} - F_{P_i}^{-1}\|^2 \mid T_i, P_i \right) \right] \\ & \leq C_1 \alpha_m \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \hat{w}(T_i, t+h, h)^2 \right] \\ & \leq C_2 \frac{\alpha_m}{h} \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n \left( \frac{\hat{\kappa}_2(t+h) - \hat{\kappa}_1(t+h)(T_i - (t+h))}{\hat{\sigma}_0^2(t+h)} \right)^2 K_h(T_i - (t+h)) \right] \\ & = C_2 \alpha_m h^{-1} \mathbb{E} \left[ \frac{\hat{\kappa}_2(t+h)^2 \hat{\kappa}_0(t+h) - \hat{\kappa}_2(t+h) \hat{\kappa}_1(t+h)^2}{\hat{\kappa}_2(t+h) \hat{\kappa}_0(t+h) - \hat{\kappa}_1(t+h)^2} \right] \\ & = C_2 \alpha_m h^{-1} \mathbb{E} [\hat{\kappa}_2(t+h)] \\ & = C_2 \alpha_m h^{-1} \kappa_2(t+h) \\ & \leq C_3 \alpha_m h, \end{aligned}$$

where  $C_1, C_2, C_3$  are constants. Hence,  $\|\hat{q}_{t+h} - \tilde{q}_{t+h}\| = O_p((\alpha_m h)^{1/2})$ .

Next, we observe that

$$\begin{aligned} \|\tilde{q}_{t+h} - \bar{q}_{t+h}\| &\leq \sup_{u \in (0,1)} \frac{1}{n} \sum_{i=1}^n |(\hat{w}(T_i, t+h, h) - w(T_i, t+h, h)) F_{P_i}^{-1}(u)| \\ &\leq \left( \sup_{x \in \mathcal{D}} |x| \right) \frac{1}{n} \sum_{i=1}^n |\hat{w}(T_i, t+h, h) - w(T_i, t+h, h)| \\ &\leq \left( \sup_{x \in \mathcal{D}} |x| \right) \left( |\hat{\kappa}_0(t+h)| \left| \frac{\hat{\kappa}_2(t+h)}{\hat{\sigma}_0^2(t+h)} - \frac{\kappa_2(t+h)}{\sigma_0^2(t+h)} \right| \right. \\ &\quad \left. + |\hat{\kappa}_1^+(t+h)| \left| \frac{\hat{\kappa}_1(t+h)}{\hat{\sigma}_0^2(t+h)} - \frac{\kappa_1(t+h)}{\sigma_0^2(t+h)} \right| \right). \end{aligned}$$

We note that

$$\mathbb{E} [K_h(T - (t+h))(T - (t+h))^z]^2 = O(h^{2z-1}). \tag{A.6}$$

In conjunction with (A.5) and Taylor expansion,

$$\begin{aligned} \hat{\kappa}_0(t+h) &= \kappa_0(t+h) + O_p((nh)^{-1/2}), \\ \hat{\kappa}_1^+(t+h) &= \kappa_1^+(t+h) + O_p(h(nh)^{-1/2}), \\ \left| \frac{\hat{\kappa}_2(t+h)}{\hat{\sigma}_0^2(t+h)} - \frac{\kappa_2(t+h)}{\sigma_0^2(t+h)} \right| &= O_p((nh)^{-1/2}), \\ \left| \frac{\hat{\kappa}_1^+(t+h)}{\hat{\sigma}_0^2(t+h)} - \frac{\kappa_1^+(t+h)}{\sigma_0^2(t+h)} \right| &= O_p(h^{-1}(nh)^{-1/2}), \end{aligned}$$

whence we have  $\|\tilde{q}_{t+h} - \bar{q}_{t+h}\| = O_p((nh)^{-1/2})$ .

Lastly, we observe that

$$\begin{aligned} &\|\bar{q}_{t+h} - q_{t+h}\| \\ &\leq \frac{\kappa_2(t+h)}{\sigma_0^2(t+h)} \frac{1}{n} \sum_{i=1}^n \|K_h(T_i - (t+h)) F_{P_i}^{-1} - \mathbb{E} [K_h(T - (t+h)) F_P^{-1}]\| \\ &\quad + \frac{h|\kappa_1(t+h)|}{\sigma_0^2(t+h)} \frac{1}{n} \sum_{i=1}^n \left\| K_h(T_i - (t+h)) \frac{T_i - (t+h)}{h} F_{P_i}^{-1} \right. \\ &\quad \left. - \mathbb{E} \left[ K_h(T - (t+h)) \frac{T - (t+h)}{h} F_P^{-1} \right] \right\| \\ &= O_p((nh)^{-1/2}), \end{aligned}$$

which follows from (A.6) and the boundedness of  $\mathcal{D}$ . □

### References

AGGARWAL, O. P. (1955). Some minimax invariant procedures for estimating a cumulative distribution function. *Annals of Mathematical Statistics* **26** 450–463. [MR0070914](#)

- AMBROSIO, L. (2003). Optimal transport maps in Monge–Kantorovich problem. *arXiv preprint math/0304389*. [MR1957525](#)
- AMBROSIO, L., GIGLI, N. and SAVARÉ, G. (2004). Gradient flows with metric and differentiable structures, and applications to the Wasserstein space. *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl* **15** 327–343. [MR2148889](#)
- AMBROSIO, L., GIGLI, N. and SAVARÉ, G. (2008). *Gradient Flows: in Metric Spaces and in the Space of Probability Measures*. Springer. [MR2401600](#)
- BIGOT, J., CAZELLES, E. and PAPADAKIS, N. (2019). Penalization of barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis* **51** 2261–2285. [MR3955234](#)
- BIGOT, J., GOUET, R., KLEIN, T. and LÓPEZ, A. (2017). Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques* **53** 1–26. [MR3606732](#)
- BOLSTAD, B. M., IRIZARRY, R. A., ÅSTRAND, M. and SPEED, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19** 185–193.
- CAMBANIS, S., SIMONS, G. and STOUT, W. (1976). Inequalities for  $Ek(X, Y)$  when the marginals are fixed. *Probability Theory and Related Fields* **36** 285–294. [MR0420778](#)
- CAREY, J. R., LIEDO, P., OROZCO, D. and VAUPEL, J. (1992). Slowing of mortality rates at older ages in large Medfly cohorts. *Science* **258** 457–461.
- CASE, A. and DEATON, A. (2015). Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. *Proceedings of the National Academy of Sciences* **112** 15078–15083.
- CAZELLES, E., SEGUY, V., BIGOT, J., CUTURI, M. and PAPADAKIS, N. (2018). Geodesic PCA versus log-PCA of histograms in the Wasserstein space. *SIAM Journal on Scientific Computing* **40** B429–B456. [MR3780753](#)
- CHEN, Y., LIN, Z. and MÜLLER, H.-G. (2020). Wasserstein regression. *arXiv preprint arXiv:2006.09660* (to appear in the *Journal of the American Statistical Association*).
- CHENG, C. and PARZEN, E. (1997). Unified estimators of smooth quantile and quantile density functions. *Journal of Statistical Planning and Inference* **59** 291–307. [MR1450503](#)
- CHIOU, J.-M. and MÜLLER, H.-G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *Journal of the American Statistical Association* **104** 572–585. [MR2751439](#)
- CHOWDHURY, J. and CHAUDHURI, P. (2019). Nonparametric depth and quantile regression for functional data. *Bernoulli* **25** 395–423. [MR3892324](#)
- DELICADO, P. and VIEU, P. (2017). Choosing the most relevant level sets for depicting a sample of densities. *Computational Statistics* **32** 1083–1113. [MR3703579](#)
- FALK, M. (1983). Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. *Statistica Neerlandica* **37** 73–83. [MR0711744](#)
- FALK, M. (1984). Relative deficiency of kernel type estimators of quantiles. *Annals of Statistics* **12** 261–268. [MR0733512](#)

- FAN, J. and GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman & Hall, London. [MR1383587](#)
- GALICHON, A. (2017). A survey of some recent applications of optimal transport methods to econometrics. *The Econometrics Journal* **20** C1–C11. [MR3685644](#)
- HEATHCOTE, J., PERRI, F. and VIOLANTE, G. L. (2010). Unequal we stand: An empirical analysis of economic inequality in the United States, 1967–2006. *Review of Economic dynamics* **13** 15–51.
- HOEFFDING, W. (1940). Masstabinvariante Korrelationstheorie. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin* **5** 181–233. [MR0004426](#)
- HYNDMAN, R. J., BOOTH, H. and YASMEEN, F. (2013). Coherent mortality forecasting: the product-ratio method with functional time series models. *Demography* **50** 261–283.
- JONES, C. I. (1997). On the evolution of the world income distribution. *The Journal of Economic Perspectives* **11** 19–36.
- KLOECKNER, B. (2010). A geometric study of Wasserstein spaces: Euclidean spaces. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze* **9** 297–323. [MR2731158](#)
- LEBLANC, A. (2012). On estimating distribution functions using Bernstein polynomials. *Annals of the Institute of Statistical Mathematics* **64** 919–943. [MR2960952](#)
- LIN, Z. and YAO, F. (2019). Intrinsic Riemannian functional data analysis. *Annals of Statistics* **47** 3533–3577. [MR4025751](#)
- MALLOWS, C. L. (1972). A note on asymptotic joint normality. *Annals of Statistics* **43** 508–515. [MR0298812](#)
- MÜLLER, H.-G. (1987). Weighted local regression and kernel methods for non-parametric curve fitting. *Journal of the American Statistical Association* **82** 231–238. [MR0883351](#)
- OUELLETTE, N. and BOURBEAU, R. (2011). Changes in the age-at-death distribution in four low mortality countries: A nonparametric approach. *Demographic Research* **25** 595–628.
- PANARETOS, V. M. and ZEMEL, Y. (2016). Amplitude and phase variation of point processes. *Annals of Statistics* **44** 771–812. [MR3476617](#)
- PARZEN, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association* **74** 105–121. [MR0529528](#)
- PETERSEN, A. and MÜLLER, H.-G. (2016). Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics* **44** 183–218. [MR3449766](#)
- PETERSEN, A. and MÜLLER, H.-G. (2019a). Fréchet regression for random objects with Euclidean predictors. *Annals of Statistics* **47** 691–719. [MR3909947](#)
- PETERSEN, A. and MÜLLER, H.-G. (2019b). Wasserstein covariance for multiple random densities. *Biometrika* **106** 339–351. [MR3949307](#)
- READ, R. (1972). The asymptotic inadmissibility of the sample distribution function. *Annals of Mathematical Statistics* **43** 89–95. [MR0300378](#)
- RUBNER, Y., GUIBAS, L. J. and TOMASI, C. (1997). The earth mover’s distance, multi-dimensional scaling, and color-based image retrieval. In *Proceed-*

- ings of the ARPA image understanding workshop*, vol. 661. 668.
- SANTAMBROGIO, F. (2017). {Euclidean, metric, and Wasserstein} gradient flows: An overview. *Bulletin of Mathematical Sciences* **7** 87–154. [MR3625852](#)
- SHANG, H. L. and HYNDMAN, R. J. (2017). Grouped functional time series forecasting: An application to age-specific mortality rates. *Journal of Computational and Graphical Statistics* **26** 330–343. [MR3640190](#)
- STURM, K.-T. (2003). Probability measures on metric spaces of nonpositive curvature. *Heat Kernels and Analysis on Manifolds, Graphs, and Metric Spaces (Paris, 2002)* **338** 357–390. [MR2039961](#)
- VILLANI, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society. [MR1964483](#)
- VILLANI, C. (2008). *Optimal Transport: Old and New*, vol. 338. Springer Science & Business Media. [MR2459454](#)
- YANG, S.-S. (1985). A smooth nonparametric estimator of a quantile function. *Journal of the American Statistical Association* **80** 1004–1011. [MR0819607](#)
- YUAN, Y., ZHU, H., LIN, W. and MARRON, J. (2012). Local polynomial regression for symmetric positive definite matrices. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74** 697–719. [MR2965956](#)
- ZEMEL, Y. and PANARETOS, V. M. (2019). Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli* **25** 932–976. [MR3920362](#)
- ZHANG, Z. and MÜLLER, H.-G. (2011). Functional density synchronization. *Computational Statistics and Data Analysis* **55** 2234–2249. [MR2786984](#)