IMPACT OF THE PRIOR IN BAYESIAN NONPARAMETRICS VIA OPTIMAL TRANSPORT ON RANDOM MEASURES

BY MARTA CATALANO[1,a] AND HUGO LAVENANT[2,b]

[1] *University of Warwick,* [a] *marta.catalano@warwick.ac.uk*

[2] *Bocconi University,* [b] *hugo.lavenant@bocconi.it*

The Dirichlet process has been a cornerstone in the development of Bayesian nonparametrics, allowing one to learn the law of the observations through closed-form expressions. Still, its learning rule is often too restrictive and many generalizations have been proposed to increase its flexibility, such as the class of normalized completely random measures. Here we investigate a simple yet fundamental matter: will a different prior *actually* guarantee a different learning outcome? To this end, we develop a new framework for assessing the merging rate of opinions with three leading pillars: i) investigate the identifiability of completely random measures; ii) measure their discrepancy through an optimal transport distance; iii) conduct an analysis a posteriori, with general techniques to unravel both finite-sample and asymptotic behaviour of the distance, as the number of observations grows. Our findings provide interesting insights on the impact of popular Bayesian nonparametric priors, with very mild assumptions on the data-generating process.

1. Introduction. In a Bayesian framework the distribution of the observations $(X_i)_{i\geq 1}$ is modelled through a random parameter θ. As the data arrives the initial distribution of θ (the *prior*) gets updated through conditional probability into the distribution of $\theta|X_1,\dots,X_n$ (the *posterior*). This allows one to make online inference on the parameter or on specific functionals of interest, as well as predictions on future observations. A long-standing topic in Bayesian statistics is to understand and quantify the impact of prior choices for θ on the final goal of the statistical analysis (see, e.g., [5, 19, 21]). The role of the impact is duplex: on the one hand one expects a small perturbation of the prior to have a small impact for robustness, on the other one wishes for large variations of the prior to have a considerable impact. Indeed, this ensures flexibility to the model, which, at least for a finite sample size, should be able to adapt to different prior opinions on the data-generating mechanism.

When the parameter space is finite-dimensional, a natural way to detect both robustness and flexibility is to compute, for two different priors $\mathcal{L}(\theta^1)$ and $\mathcal{L}(\theta^2)$, a distance between their posterior laws $d(\mathcal{L}(\theta^1|X_1,\dots,X_n),\mathcal{L}(\theta^2|X_1,\dots,X_n))$, which amounts to evaluating a distance between two probabilities in $\mathbb{R}^d$. For example, Ley, Reinert and Swan [28] have recently used the Wasserstein distance on the Euclidean space. This crucially leaves out Bayesian nonparametric (BNP) models, which use infinite-dimensional parameter spaces to avoid restrictive assumptions on the distribution of the data and increase the flexibility of Bayesian models. In this work we provide methodological and technical advances to measure the impact of the prior for a wide class of BNP models, which allow us to investigate three compelling questions that do not have a general answer in a nonparametric setting:

a. Given two different nonparametric priors, will the discrepancy between the induced inferences and predictions vanish as the number of observations goes to infinity, that is, will there be *merging of opinions*?

MSC2020 subject classifications: 60G55, 60G57, 49Q22, 62C10.

Keywords and phrases: Bayesian nonparametrics, Completely random measures, Cox process, Identifiability, Impact of the prior, Laplace method, Merging of opinions, Optimal transport, Wasserstein distance.

b. If there is asymptotic merging of opinions, how fast does the discrepancy vanish, that is, what is the *merging rate*?
c. If there is asymptotic merging of opinions, does the finite-sample discrepancy decrease steadily, as the observations grow? If not, when does the merging start?

For many BNP models the infinite-dimensional random parameter consists of a random measure $\tilde{\mu}$. The data on a Polish space $\mathbb{X}$ are modelled as conditionally independent and identically distributed (i.i.d.) given $\tilde{\mu}$, so that for every $n \geq 1$,

$$X_1, \ldots, X_n | \tilde{\mu} \overset{\text{iid}}{\sim} t(\tilde{\mu}), \tag{1}$$

for some transformation t. A popular choice for t is the normalization $t(\tilde{\mu}) = \tilde{\mu}/\tilde{\mu}(\mathbb{X})$, which allows one to define a random probability measure whenever $0 < \tilde{\mu}(\mathbb{X}) < +\infty$ almost surely (a.s.) [38, 23, 24]. Other common choices include kernel mixtures for densities and hazard rates [31, 12, 22], exponential transformations for survival functions [11] and cumulative transformations for cumulative hazards [20]; see [30] for a concise summary. All these works define a prior on $\tilde{\mu}$ through the notion of completely random measure (CRM), a rich class of distributions for random measures first defined by Kingman [26]. As a classical example, Ferguson [14] shows that by normalizing a gamma CRM one obtains the Dirichlet process, the cornerstone of Bayesian nonparametrics.

Consider two different priors $\mathcal{L}(\tilde{\mu}^1)$ and $\mathcal{L}(\tilde{\mu}^2)$ for model (1). Our motivating goal is to answer Questions a., b., and c. above for this class of models by introducing and studying a distance between posterior random measures,

$$d(\mathcal{L}(\tilde{\mu}^1 | X_1, \ldots, X_n), \mathcal{L}(\tilde{\mu}^2 | X_1, \ldots, X_n)). \tag{2}$$

To this end, we need to discuss the identifiability of $\tilde{\mu}$ and the choice of distance d. Identifiability of the random parameter $\tilde{\mu}$ is crucial to avoid evaluating a positive distance between random measures that induce the same model, that is, to erroneously detect a lack of merging of opinion. Identifiability depends on the choice of transformation t and the class of priors for $\tilde{\mu}$, and for this reason we focus the rest of this work on a specific setting, namely, normalization of CRMs. The task then amounts to understand which random measures induce the same normalization, divide the random measures in equivalence classes accordingly, and choose a representative for each equivalence class, both a priori and a posteriori. This will lead to the definition of scaled CRMs $\tilde{\mu}_{\mathcal{S}}$, such that $\mathbb{E}(\tilde{\mu}_{\mathcal{S}}(\mathbb{X})) = 1$, and their generalizations to random Lévy intensities.

Once the representatives for each class are established, one needs to define a suitable distance between them, which ideally conjugates computational efficiency with analytical tractability. We achieve this task by relying on optimal transport (OT). OT studies the mass transportation problem between (probability) measures, and it is conveniently used to define distances between measures in terms of the minimal cost of transporting mass from one measure to another; see [42, 40] for exhaustive references. This brings to a metric that stands out for its ability to take into account the geometry of the underlying space, which is especially useful with measures with atomic components, as it is usually the case in BNP. Thanks to the almost-conjugacy of models (1) with respect to a CRM prior [24], the bulk of the work is to define a distance between scaled CRMs. The law of a CRM may be decomposed as the sum of three components: a deterministic measure, a random measure with fixed atoms, and an a.s. discrete random measure

$$\tilde{\mu} = \sum_{(s,x) \in \mathcal{N}} s\, \delta_x, \tag{3}$$

where $\mathcal{N}$ is a Poisson point process on $(0, +\infty) \times \mathbb{X}$ with Lévy intensity ν. Prior specifications for a CRM usually restrict to the latter component and thus their law is specified

through a Lévy intensity. The first proposals to define a tractable distance between CRMs can be found in [7, 8] but the key progress of Catalano et al. [9] has been to define an OT distance directly at the level of the Lévy intensities, as follows. For a set A, one can focus on the corresponding Lévy measures $\nu_A(\cdot) = \nu(\cdot, A)$, which are typically measures with infinite mass. The extended Wasserstein distance $\mathcal{W}_*$ of Figalli and Gigli [15] remarkably allows for comparisons between such measures, so that

$$\sup_A \mathcal{W}_*(\nu_A^1, \nu_A^2) \tag{4}$$

defines a distance between CRMs [9] and leads to closed-form expressions whenever i) the CRMs do not have fixed atoms; ii) the CRMs are homogeneous, i.e., $\nu^i = \rho^i \otimes P^i$ is a product measure; iii) the CRMs share the same base measure, i.e., $P^1 = P^2$. These conditions are usually met in the setting of [9] but are very restrictive in our framework since i) and ii) typically hold *a priori* but not *a posteriori*, and iii) entails that the atoms of the CRM have the same distribution, which is likely to be the first feature one changes when choosing a different prior. We overcome these limitations of the distance in [9] by defining a new OT distance on the Lévy intensities that considers the joint distribution of the atoms and the jumps instead, thus avoiding the challenges of finding a supremum. It is worth underlining that when i), ii), iii) hold, and thus the distance in (4) is tractable, it coincides with our proposal.

This sets the ground for a principled framework to measure the impact of the prior in terms of a distance between (the representatives of) the posteriors, as in (2). We then apply this machinery to shed new light on the use of the Dirichlet process. The Dirichlet process [13] has been a cornerstone in the development of BNP, allowing one to make inference on the law of the observations through closed-form expressions. Still, its learning rule is often too restrictive, and many generalizations have been proposed to increase its flexibility, such as the class of normalized CRMs. From the practitioner's perspective, the price to pay for such flexibility is a higher level of mathematical sophistication and more demanding posterior computations, with additional cost in both the user's and the machine's time. We use our setup to investigate to which extent different parameters for the Dirichlet process or different nonparametric priors, such as the generalized gamma CRM [29], lead to different learning outcomes and can thus be useful in practice. This requires the development of analytical tools to treat both asymptotic (Questions a. and b.) and finite-sample properties (Quesion c.) of the OT distance, which bring to interesting results for this problem. Moreover, they may be adapted to further research questions, setting the basis for a general study of the impact of the prior in BNP.

Relation to previous work. The merging of opinions for model (1) could in principle be studied through the framework of Blackwell and Dubins [6]. However, this is restrictive for at least two reasons. First of all, merging is guaranteed for prior laws on $(X_n)_{n\geq 1}$ that are absolutely continuous with one another. Whereas this holds in most parametric settings, it is not usually the case in nonparametric non-dominated models, such as the one we are considering. As a classical example, two Dirichlet processes with different total base measure are mutually singular [27]. Moreover, [6] assumes that the data-generating mechanism coincides with one of the models under consideration, which may well not be the case in practice. Our aim is to overcome both restrictions and build a principled framework with very little assumptions on the data. Indeed, we propose one of the few asymptotic analysis that does not put strong assumptions on the distribution of the data (e.g., to come from the model or to be i.i.d. from a true distribution): this is possible because we focus our attention on the similarities between the learning mechanisms of different models rather than on the adherence of the learning mechanism to the unknown *truth*.

In the same spirit, our framework goes beyond consistency in that it does not assume the data to be i.i.d. from a true distribution. If, on the other hand, one were willing to make this assumption, consistency of the model with respect to each prior would entail merging of opinions (Question a.). However, this does not provide an answer to the merging rate (Question b.): indeed, if we manage to find the contraction rate for each prior to the truth, this could be (and usually is) considerably slower than the merging rate. As an extreme example, consider using the same prior twice: the distance between the posteriors is always zero, independently of its contraction rate to the truth. Thus, not only our framework has the advantage of not using strict assumptions on the data-generating process, it also provides new prospectives when one is willing to put them.

To our knowledge, this work provides the first theoretical set-up of *global sensitivity* analysis in Bayesian nonparametric non-dominated models, in the sense that it studies the performance of the model of interest with respect to any choice of prior. Previous studies on sensitivity usually consisted in varying the parameters of BNP priors or taking functional perurbations [32] and providing graphical comparisons between the posteriors. In a dominated setting, more classical distances between (posterior) densities can be considered, as in Saha and Kurtek [39]. A recent work of Giordano et al. [17] focuses on *local sensitivity*, which studies robustness with respect to small perturbation of the prior through differential approximations [19] and can be very useful when posterior distributions are not known in closed form. Since local sensitivity focuses on small perturbations, a 'good' impact is always a small impact. On the contrary, our framework does not limit to small perturbations, and thus a big impact of the prior can also be regarded as a benign sign of flexibility of the model. Though our statistical analysis focuses on the latter aspect, our results may be easily read or adapted to treat robustness as well.

Structure of the paper. Section 2 finds the class of CRMs that are identifiable for model (1) with respect to normalization and introduces the notion of scaled CRM, with some reference examples. Section 3 defines the OT distance between scaled CRMs, providing some intuition and key properties, which are then used to quantify the comparison between gamma and generalized gamma CRMs. Section 4 develops the ground to assess the impact of the prior in Bayesian nonparametrics. First, in Section 4.1, identifiability and OT are treated for a larger class of random measures, Cox CRMs, which naturally appear in the posterior characterization of CRMs. Then, in Section 4.2 we discuss the impact of the parameters of the Dirichlet process, whereas in Section 4.3 we shed light on the role of the concentration parameter of the generalized gamma CRM. All proofs are deferred to Section 5.

2. Identifiability of completely random measures. The goal of this section is to find the class of CRMs for which model (1) is identifiable with respect to normalization $t(\tilde{\mu}) = \tilde{\mu}/\tilde{\mu}(\mathbb{X})$. To this end, we first introduce the notion of CRM and provide a canonical decomposition of its Lévy intensity (Proposition 1). The section culminates with Theorem 2, which finds all *equivalent* CRMs, in the sense that they induce the same normalization. The choice of a representative for each equivalence class leads to the notion of scaled CRM. We conclude the section with a general formula to retrieve the Lévy intensity of a scaled CRM (Lemma 4) and provide some relevant examples.

We work on a Polish space $\mathbb{X}$ endowed with its Borel σ-algebra $\mathcal{B}(\mathbb{X})$, and we write $\mathcal{P}(\mathbb{X})$ for the space of probability measures over $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$. We recall that a random measure $\tilde{\mu}$ on a Polish space $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$ is a random element on the space of locally finite measures μ on $\mathbb{X}$, endowed with the σ-field generated by all projection maps $\mu \mapsto \mu(B)$, for $B \in \mathcal{B}(\mathbb{X})$. In the original definition by Kingman [26] completely random measures are random measures such

that, for every disjoint $B_1, \dots, B_n \in \mathcal{B}(\mathbb{X})$, the random variables $\tilde{\mu}(B_1), \dots, \tilde{\mu}(B_n)$ are mutually independent. As recalled in the introduction, this definition entails that any CRM may be decomposed in the sum of a deterministic measure, a discrete random measure with fixed atoms and a discrete random measure without fixed atoms, whose law is characterized by a Lévy intensity. Since it is the third component that plays the main role in Bayesian nonparametrics, we will give a slightly more restrictive definition of completely random measure, as follows. Here and after, $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution.

DEFINITION 1. A random measure $\tilde{\mu}$ is a completely random measure (CRM) if there exists ν a positive measure on $(0, +\infty) \times \mathbb{X}$ with $\iint \min(1, s)\, \mathrm{d}\nu(s, x) < +\infty$ such that the law of $\tilde{\mu}$ is defined by (3). We write $\tilde{\mu} \sim \mathrm{CRM}(\nu)$ and refer to ν as the Lévy intensity of $\tilde{\mu}$.

To exclude a pathological case we require the projection of ν onto $\mathbb{X}$ to not be reduced a Dirac mass, that is, the space $\mathbb{X}$ cannot be reduced to a singleton. In the latter case, we would run into identifiability issues. Note on the other hand that we have not imposed ν to be diffuse, so that $\tilde{\mu}$ can have fixed atoms corresponding to the atoms of the second marginal of ν. This will be useful when looking at the posterior distribution of CRMs in Section 4. Importantly, we assume that the CRM has *finite mean*, that is, $\mathbb{E}(\tilde{\mu}(\mathbb{X})) < +\infty$. Thanks to Campbell's theorem this condition can be stated in terms of the Lévy intensity, which must satisfy $\iint s\, \mathrm{d}\nu(s, x) < +\infty$.

PROPOSITION 1. *If $\tilde{\mu} \sim \mathrm{CRM}(\nu)$ has finite mean, then ν can be uniquely decomposed as $\mathrm{d}\nu(s, x) = \mathrm{d}\rho_x(s)\mathrm{d}P_0(x)$, where P_0 is a probability measure and ρ is a transition kernel on $\mathbb{X} \times \mathcal{B}(\mathbb{R}_+)$ that satisfy, for P_0-almost every (a.e.) $x \in \mathbb{X}$,*

$$\int_0^{+\infty} s\, \mathrm{d}\rho_x(s) = \mathbb{E}(\tilde{\mu}(\mathbb{X})); \qquad P_0(\cdot) = \frac{\mathbb{E}(\tilde{\mu}(\cdot))}{\mathbb{E}(\tilde{\mu}(\mathbb{X}))}.$$

We refer to $\mathrm{d}\nu(s, x) = \mathrm{d}\rho_x(s)\mathrm{d}P_0(x)$ as the canonical decomposition of ν.

REMARK 1. Intuitively, the measure P_0 describes the distribution of atoms, while ρ_x determines the conditional distribution of sizes of jumps at location x. This can be understood by approximating (3) with a compound distribution, so that for every $\epsilon > 0$

$$\sum_{\{(s,x) \in \mathcal{N} \,:\, s > \epsilon\}} s\, \delta_x \stackrel{\mathrm{d}}{=} \sum_{i=0}^{N} s_i\, \delta_{x_i}, \tag{5}$$

where by letting $r = \nu((\epsilon, +\infty) \times \mathbb{X})$, N is a Poisson random variable with mean r and $(s_i, x_i) \sim P$ independently, with $P = \nu|_{(\epsilon,+\infty)\times\mathbb{X}}/r$ being the normalized restriction of ν to $(\epsilon, +\infty) \times \mathbb{X}$.

For the normalization to be well-defined one needs $\tilde{\mu}(\mathbb{X}) > 0$ a.s., which at the level of the Lévy intensities corresponds to $\nu((0, +\infty) \times \mathbb{X}) = +\infty$ [38]. We will require something slightly stronger, that the CRM is *infintely active*. In terms of the canonical decomposition, this assumption reads

$$\text{for } P_0\text{-a.e. } x, \qquad \rho_x(0, +\infty) = +\infty.$$

This entails that there will always be an infinite number of jumps on a set A with $P_0(A) > 0$, and thus the compound Poisson approximation (5) can not be extended to $\epsilon = 0$. In particular, it guarantees that for any set A exactly one of the two alternatives holds: either $P_0(A) > 0$,

and in this case $\tilde{\mu}(A) > 0$ a.s., or $P_0(A) = 0$, which implies $\tilde{\mu}(A) = 0$ a.s. This will be useful for our identifiability results.

Quite clearly, different CRMs $\tilde{\mu}$ can lead to the same normalized random probability $\tilde{\mu}/\tilde{\mu}(\mathbb{X})$: for instance $\tilde{\mu}$ and $\alpha\tilde{\mu}$ leads to the same normalization for any $\alpha > 0$. In Theorem 2 below we build on a result on subordinators by David Aldous and Stevan Evans [36, Lemma 7.5] to prove that this can be the only source for the lack of identifiability.

THEOREM 2. *Let $\tilde{\mu}^i \sim \text{CRM}(\nu^i)$ be infinitely active with finite mean, for $i = 1, 2$. Then $\tilde{\mu}^1/\tilde{\mu}^1(\mathbb{X}) \stackrel{\text{d}}{=} \tilde{\mu}^2/\tilde{\mu}^2(\mathbb{X})$ if and only if $\tilde{\mu}^1 \stackrel{\text{d}}{=} \alpha\tilde{\mu}^2$, for some scalar $\alpha > 0$.*

Theorem 2 sheds new light on the equivalence relation $\sim$ that identifies CRMs leading to the same normalization, showing that $\tilde{\mu}^1 \sim \tilde{\mu}^2$ if and only if $\tilde{\mu}^1 \stackrel{\text{d}}{=} \alpha\tilde{\mu}^2$ for some $\alpha > 0$. We use it to choose one representative for each equivalence class, namely the only CRM in the class with total expected value equal to 1, as in the next definition.

DEFINITION 2. A CRM $\tilde{\mu}$ is a scaled CRM if $\mathbb{E}(\tilde{\mu}(\mathbb{X})) = 1$. If $\tilde{\mu}$ be a CRM with finite mean, we define its scaled CRM as

$$\tilde{\mu}_{\mathcal{S}} = \frac{\tilde{\mu}}{\mathbb{E}(\tilde{\mu}(\mathbb{X}))}.$$

COROLLARY 3. *Let $\tilde{\mu}^i \sim \text{CRM}(\nu^i)$ be infinitely active with finite mean, for $i = 1, 2$. Then $\tilde{\mu}^1/\tilde{\mu}^1(\mathbb{X}) \stackrel{\text{d}}{=} \tilde{\mu}^2/\tilde{\mu}^2(\mathbb{X})$ if and only if $\tilde{\mu}^1_{\mathcal{S}} \stackrel{\text{d}}{=} \tilde{\mu}^2_{\mathcal{S}}$.*

By Proposition 1 the mean measure $Q(\cdot) = \mathbb{E}(\tilde{\mu}(\cdot))$ of any scaled CRM with Lévy intensity $\text{d}\nu(s, x) = \text{d}\rho_x(s)\text{d}P_0(x)$ is the probability measure P_0 and, for P_0-a.e. $x \in \mathbb{X}$,

$$\int_0^{+\infty} s \, \text{d}\rho_x(s) = 1. \tag{6}$$

We highlight that, unlike normalized CRMs, scaled CRMs are CRMs. The following result provides a general way to find their Lévy intensities. We state it here for diffuse jumps components $\text{d}\rho_x(s) = \rho_x(s)\,\text{d}s$ for simplicity, though it can be extended to atomic jump components; see Lemma 22 in Section 5.

LEMMA 4. *Let $\tilde{\mu} \sim \text{CRM}(\nu)$ such that $\text{d}\nu(s, x) = \rho_x(s)\,\text{d}s\,\text{d}P_0(x)$. Then $\tilde{\mu}_{\mathcal{S}}$ has Lévy intensity $\text{d}\nu_{\mathcal{S}}(s, x) = \mathbb{E}(\tilde{\mu}(\mathbb{X}))\rho_x(\mathbb{E}(\tilde{\mu}(\mathbb{X}))s)\,\text{d}s\,\text{d}P_0(x)$.*

EXAMPLE 1 (Gamma CRM). A gamma CRM with base probability $P_0 \in \mathcal{P}(\mathbb{X})$, total base measure $\alpha > 0$ and scale parameter $b > 0$ has Lévy intensity

$$\text{d}\nu(s, x) = \alpha\frac{e^{-bs}}{s}\text{d}s\,\text{d}P_0(x).$$

Thanks to Campbell's theorem one easily shows that $\mathbb{E}(\tilde{\mu}(\mathbb{X})) = \alpha/b$, so that by Lemma 4 the corresponding scaled gamma CRM has Lévy intensity

$$\text{d}\nu_{\mathcal{S}}(s, x) = \alpha\frac{e^{-\alpha s}}{s}\text{d}s\,\text{d}P_0(x). \tag{7}$$

Thus, $\tilde{\mu}_{\mathcal{S}}$ is a gamma CRM with base probability P_0, total base measure α and scale parameter α. This highlights the redundancy of the parameter $b > 0$, which will assumed to be equal to 1 in the rest of the paper, unless otherwise specified.

EXAMPLE 2 (Generalized gamma CRM). Another popular prior in BNP models is the normalized generalized gamma [29, 3]. The generalized gamma CRM with total base measure $\alpha > 0$, concentration parameter $\sigma \in (0,1)$, and base probability $P_0 \in \mathcal{P}(\mathbb{X})$ has Lévy intensity

$$\mathrm{d}\nu(s,x) = \frac{\alpha}{\Gamma(1-\sigma)} \frac{e^{-s}}{s^{1+\sigma}} \mathrm{d}s \, \mathrm{d}P_0(x),$$

where $\Gamma(z) = \int_0^{+\infty} x^{z-1} e^{-x} \, \mathrm{d}x$ is the gamma function. In particular, one recovers the gamma CRM as $\sigma \to 0$. It is easy to prove that $\mathbb{E}(\tilde{\mu}(\mathbb{X})) = \alpha$, so that by Lemma 4 the corresponding scaled CRM has Lévy intensity

$$\mathrm{d}\nu_{\mathcal{S}}(s,x) = \frac{\alpha^{1-\sigma}}{\Gamma(1-\sigma)} \frac{e^{-\alpha s}}{s^{1+\sigma}} \mathrm{d}s \, \mathrm{d}P_0(x). \tag{8}$$

Thus we can regard $\tilde{\mu}_{\mathcal{S}}$ as a generalized gamma CRM that also accounts for a scale parameter.

3. Optimal transport on completely random measures. In Section 2 we have shown that the class of scaled CRMs is identifiable for model (1) with respect to normalization. The goal of this section is to define an OT distance between scaled CRMs, which will be then used in Section 4 to measure the impact of the prior in Bayesian nonparametrics. We first recall some important notions that enable us to define our OT distance (Definition 5). After providing some useful computational properties (Lemma 6) and a foundational topological result (Theorem 7), we discuss two relevant examples. Their study is conducted through a general framework (Proposition 25, see Section 5) which could be useful in other settings.

As argued in the introduction, the plan is to use OT on the joint Lévy intensity $\mathrm{d}\nu(s,x) = \mathrm{d}\rho_x(s)\mathrm{d}P_0(x)$. There are two main challenges: (a) the classical theory of OT is developed for probability measures, whereas ν is a measure with typically infinite mass; (b) ν is a measure on $(0,+\infty) \times \mathbb{X}$, and its two components play a very different role, as highlighted in Remark 1. Indeed, P_0 is a probability measure on $\mathbb{X}$ describing the law of the atoms, whereas ρ_x is P_0-a.s. a measure with infinite mass on $(0,+\infty)$ that determines the conditional law of their corresponding jumps. We will thus preserve their different nature, in the spirit of the *nested Wasserstein distance* for probability measures, which was originally introduced by Pflug and Pichler [34] and it is often termed *adapted Wasserstein distance* in financial mathematics [2, 4]. Specifically we impose that, if an atom x is coupled to an atom y on the space $\mathbb{X}$, then whole measure ρ_x is also coupled to ρ_y. This leads to a distance that conjugates desirable topological properties (Theorem 7) with tractable computations.

First of all we recall the definition of Wasserstein distance $\mathcal{W}_p$ between probabilities and its extension $\mathcal{W}_{*,p}$ to positive measures. We refer to Section 5 for some classical properties that we will use throughout the present work. For a positive measure μ on a metric space $(\mathbb{Y}, d_{\mathbb{Y}})$, the p-th moment about a point $\bar{y} \in \mathbb{Y}$ is defined as $\int d_{\mathbb{Y}}(y,\bar{y})^p \mathrm{d}\mu(y)$. For finite measures, if the p-th moment about a point $\bar{y} \in \mathbb{Y}$ is finite, then it is also the case for the p-th moment about any another point $\bar{y}' \in \mathbb{Y}$: in this case we say that the measure has a finite p-th moment.

DEFINITION 3. For $p \geq 1$, the L^p-Wasserstein distance between two probability distributions P^1, P^2 on a metric space $(\mathbb{Y}, d_{\mathbb{Y}})$ with finite p-th moments is defined as

$$\mathcal{W}_p(P^1,P^2)^p = \min_{\pi \in \Pi(P^1,P^2)} \mathbb{E}_{(Y^1,Y^2)\sim\pi}(d_{\mathbb{Y}}(Y^1,Y^2)^p), \tag{9}$$

where $\Pi(P^1,P^2)$ corresponds to the set of couplings between P^1 and P^2, i.e., the joint laws of $\mathbb{Y}$-valued random variables whose marginal laws are P^1 and P^2.

Definition 3 has been extended to measures with different or infinite mass when $p = 2$ by Figalli and Gigli [15], through the notion of extended coupling. We refer to Catalano et al. [9] and Guillen, Mou and Święch [18] for a concise overview. We will only consider the case $\mathbb{Y} = (0, +\infty)$ in the present work, so we present directly a definition adapted to this case. We refer to [9, Proposition 4] for a link between the definition below and the general definition of extended Wasserstein distance at least in the case $p = 2$, and it can be easily extended to the general case $p \geq 1$.

DEFINITION 4. For $p \geq 1$, the L^p-extended Wasserstein distance between two positive Borel measures ρ^1, ρ^2 on $(0, +\infty)$ with finite p-th moments about 0 is defined as

$$\mathcal{W}_{*p}(\rho^1, \rho^2)^p = \int_0^{+\infty} |U_1^{-1}(s) - U_2^{-1}(s)|^p \mathrm{d}s, \tag{10}$$

where, for $i = 1, 2$, $U_i(s) = \rho^i(s, +\infty)$ is the tail integral of ρ^i and $U_i^{-1}(s) = \inf\{u \geq 0 : U_i(u) \leq s\}$ is its generalized inverse.

We are now ready to define an OT distance between scaled Lévy intensities $\mathrm{d}\nu(s, x) = \mathrm{d}\rho_x(s)\mathrm{d}P_0(x)$.

DEFINITION 5. For $p > 1$, we define an OT distance on the scaled Lévy intensities as

$$d_{\mathcal{W},p}(\nu^1, \nu^2)^p = \inf_{\pi \in \Pi(P_0^1, P_0^2)} \mathbb{E}_{(X,Y)\sim\pi} \left(d_{\mathbb{X}}(X, Y)^p + \mathcal{W}_{*p}(\rho_X^1, \rho_Y^2)^p \right), \tag{11}$$

provided P_0^i have finite first moment for $i = 1, 2$. In the case $p = 1$ we write simply $d_{\mathcal{W},1} = d_{\mathcal{W}}$.

PROPOSITION 5. *For any $p \geq 1$, the functions $\mathcal{W}_{*p}$ and $d_{\mathcal{W},p}$ define a distance over respectively the set of Borel measures on $(0, +\infty)$ with finite p-th moment about 0, and the set of scaled Lévy intensities such that the base measure P_0 has finite p-th moment.*

REMARK 2. The distance (11) is particularly convenient when at least one of the two Lévy intensities is homogeneous, i.e., $\nu^1 = \rho^1 \otimes P_0^1$. In such case, $\mathcal{W}_{*p}(\rho_X^1, \rho_Y^2)^p = \mathcal{W}_{*p}(\rho^1, \rho_Y^2)^p$ does not depend on the coupling and thus

$$d_{\mathcal{W},p}(\nu^1, \nu^2)^p = \mathcal{W}_p(P_0^1, P_0^2)^p + \mathbb{E}_{Y\sim P_0^2}(\mathcal{W}_{*p}(\rho^1, \rho_Y^2)^p).$$

When both Lévy intensities are homogeneous with the same base probability the formula simplifies even more to

$$d_{\mathcal{W},p}(\nu^1, \nu^2) = \mathcal{W}_{*p}(\rho^1, \rho^2).$$

We focus our attention on $d_{\mathcal{W},1} = d_{\mathcal{W}}$ because of two main advantages: (i) the assumption of bounded first moments is weaker than bounded p-th moments, for $p > 1$; (ii) to evaluate $\mathcal{W}_{*1} = \mathcal{W}_*$ one can avoid computing the generalized inverse of the tail integrals, thanks to Lemma 6 below, with great benefits in terms of computation time.

LEMMA 6. *Let ρ^1, ρ^2 be two measures on $(0, +\infty)$ with finite first moment about 0. Then*

$$\mathcal{W}_{*1}(\rho^1, \rho^2) = \int_0^{+\infty} |U_1(u) - U_2(u)| \mathrm{d}u,$$

where $U_i(u) = \rho^i(u, +\infty)$ is the tail integral of ρ^i. Moreover, if ρ^1, ρ^2 have first moment about 0 equal to 1, as in (6), $\mathcal{W}_*(\rho^1, \rho^2) \leq 2$.

We use the notation $d_{\mathcal{W}}(\tilde{\mu}^1, \tilde{\mu}^2) = d_{\mathcal{W}}(\nu^1, \nu^2)$ to indicate the distance between the laws of scaled CRMs $\tilde{\mu}^i \sim \text{CRM}(\nu^i)$ induced by their Lévy intensities. To further motivate its use as a distance between random measures, the next result shows that it controls the Wasserstein distance between (random) integrals. Note for a random measure $\tilde{\mu}$ on $\mathbb{X}$ and a measurable function $f : \mathbb{X} \to \mathbb{R}$, $\int f \mathrm{d}\tilde{\mu}$ is a real-valued random variable, and thus we can measure the distance between such random variables with the Wasserstein distance. For a function $f : \mathbb{X} \to \mathbb{R}$, we write $\text{Lip}(f) = \sup_{x \neq y} |f(x) - f(y)|/d_{\mathbb{X}}(x, y)$ for its best Lipschitz constant.

THEOREM 7. *Let $\tilde{\mu}^1$, $\tilde{\mu}^2$ be scaled* CRM*s on $\mathbb{X}$. Let $f : \mathbb{X} \to \mathbb{R}$ a function that is bounded and Lipschitz, and define $C = \max(\sup |f|, \text{Lip}(f))$. Then there holds*

$$\mathcal{W}_1\left(\int_{\mathbb{X}} f \mathrm{d}\tilde{\mu}^1, \int_{\mathbb{X}} f \mathrm{d}\tilde{\mu}^2\right) \leq C d_{\mathcal{W}}(\tilde{\mu}^1, \tilde{\mu}^2).$$

We conclude this section with two examples on the use of the OT distance in Definition 5 on popular priors in Bayesian nonparametrics. The general technical result that allows to study both examples, and possibly further ones, can be found in Section 5 as Proposition 25.

EXAMPLE 3 (OT on gamma CRMs). A question that is often of interest in BNP is to understand the role of the total base measure on the Dirichlet process, the normalization of a Gamma CRM (Example 1). We thus consider two gamma CRMs with total base measures α_i and base probabilities P_0^i having finite first moment, for $i = 1, 2$. Since the CRMs are homogeneous, thanks to Remark 2 and Lemma 6, one easily obtains

$$d_{\mathcal{W}}(\tilde{\mu}^1_{\mathcal{S}}, \tilde{\mu}^2_{\mathcal{S}}) = \int_0^{+\infty} |\alpha_1 \Gamma(0, \alpha_1 s) - \alpha_2 \Gamma(0, \alpha_2 s)| \mathrm{d}s + \mathcal{W}_1(P_0^1, P_0^2),$$

where $\Gamma(z, t) = \int_t^{+\infty} x^{1-z} e^{-x} \, \mathrm{d}x$ is the incomplete Gamma function. In particular, the distance is the sum of a jump component

$$\mathcal{J}(\alpha_1, \alpha_2) = \int_0^{+\infty} |\alpha_1 \Gamma(0, \alpha_1 s) - \alpha_2 \Gamma(0, \alpha_2 s)| \mathrm{d}s,$$

depending only on the total base measures, and of an atom component $\mathcal{W}_1(P_0^1, P_0^2)$, depending only on the base probability measures. The atom component involves a classical Wasserstein distance, which is well-studied in the literature. We thus focus our attention on $\mathcal{J}(\alpha_1, \alpha_2)$. First of all we observe that it may be easily computed through standard numerical integration. Not only, the next result shows that we may also unravel many analytical properties, which will be very useful when studying the impact of the prior in Section 4. Without loss of generality we assume $\alpha_1 \leq \alpha_2$ and study the behaviour of $\mathcal{J}(\alpha_1, \alpha_2)$ as α_2 increases.

PROPOSITION 8. *Let $0 < \alpha_1 \leq \alpha_2$. We define $C = \int_0^{+\infty} |\Gamma(0, t) - e^{-t}| \mathrm{d}t$. Then, $\mathcal{J}(\alpha_1, \alpha_2)$ satisfies the following:*

1. *$\alpha_2 \mapsto \mathcal{J}(\alpha_1, \alpha_2)$ is increasing.*
2. *We have the upper bound*

$$\mathcal{J}(\alpha_1, \alpha_2) \leq C \log\left(\frac{\alpha_2}{\alpha_1}\right). \tag{12}$$

3. *The asymptotic expansion of $\mathcal{J}(\alpha_1, \alpha_2)$ for α_2 close to α_1 is*

$$\mathcal{J}(\alpha_1, \alpha_2) = C\left(\frac{\alpha_2}{\alpha_1} - 1\right) - \frac{C}{2}\left(\frac{\alpha_2}{\alpha_1} - 1\right)^2 + o\left(\left(\frac{\alpha_2}{\alpha_1} - 1\right)^2\right).$$

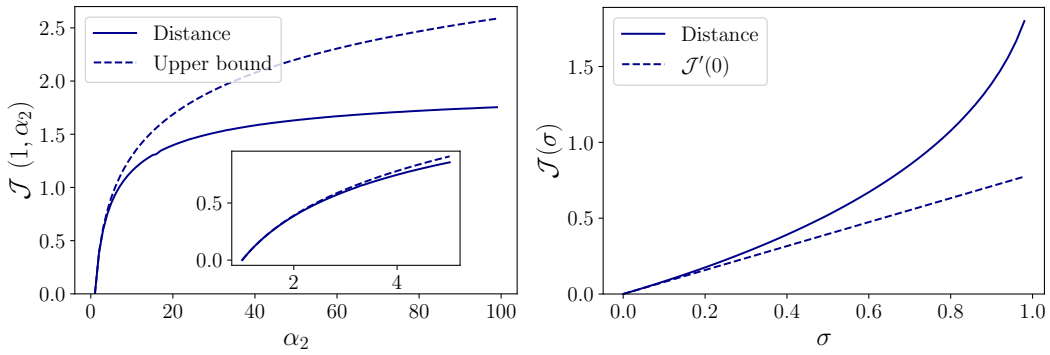

FIG 1. *Left: Jump component $\mathcal{J}(\alpha_1,\alpha_2)$ of $d_{\mathcal{W}}(\tilde{\mu}^1,\tilde{\mu}^2)$ when $\tilde{\mu}^1$ is a gamma CRM of parameters $(\alpha_1=1,P_0^1)$ and $\tilde{\mu}^2$ is a gamma CRM of parameters (α_2,P_0^2). The upper bound is the right hand side of (12). Right: Jump component $\mathcal{J}(\sigma)$ of $d_{\mathcal{W}}(\tilde{\mu}^1,\tilde{\mu}^2)$ when $\tilde{\mu}^1$ is a generalized gamma CRM of parameters (α,P_0,σ) and $\tilde{\mu}^2$ is a gamma CRM of parameters (α,P_0). The value of the derivative in zero is in Proposition 9.*

Note that the constant C can be evaluated numerically ($C\approx 0.56$). In particular, Proposition 8 provides an interpretable, computable and universal upper bound (12) for the distance, which is proved to be equal to the asymptotic expansion at order 2 of the distance for $\alpha_1\approx\alpha_2$. This will play a crucial role in Section 4.2. In the left plot of Figure 1 we fix $\alpha_1=1$ without loss of generality and we provide an empirical validation of the results of Proposition 8.

EXAMPLE 4 (OT on generalized gamma CRMs). Another important question in BNP is to understand the role of the concentration parameter $\sigma\in(0,1)$ on generalized gamma CRMs as in Example 2. We consider $\tilde{\mu}^1$ to be a generalized gamma CRM of parameters (α,σ,P_0) and $\tilde{\mu}^2$ to be a gamma CRM with same total base measure α and base probability P_0 having finite first moment. Then thanks to Remark 2 and Lemma 6, with a change of variable $t=\alpha s$,

$$d_{\mathcal{W}}(\tilde{\mu}^1_{\mathcal{S}},\tilde{\mu}^2_{\mathcal{S}})=\int_0^{+\infty}\left|\frac{\Gamma(-\sigma,t)}{\Gamma(1-\sigma)}-\Gamma(0,t)\right|\mathrm{d}t.$$

In particular, we observe that the discrepancy between the two models has only a jump component and it is not affected by the choice of α and P_0. We study the behaviour of $\mathcal{J}(\sigma)=d_{\mathcal{W}}(\tilde{\mu}^1_{\mathcal{S}},\tilde{\mu}^2_{\mathcal{S}})$ as σ increases.

PROPOSITION 9. *The function $\sigma\mapsto\mathcal{J}(\sigma)$ is increasing. Moreover,*

$$\mathcal{J}'(0)=\int_0^{+\infty}\left|\gamma\Gamma(0,t)+\int_t^{+\infty}\log(s)\frac{e^{-s}}{s}\,\mathrm{d}s\right|\mathrm{d}t,$$

with $\gamma=-\Gamma'(1)$ is the Euler–Mascheroni constant.

In the right plot of Figure 1 we plot $\mathcal{J}(\sigma)$ as σ increases. In particular, we observe that $\mathcal{J}'(0)\approx 0.79$ may be evaluated numerically.

4. Impact of the prior in Bayesian nonparametrics. The aim of this section is to measure the impact of the prior in Bayesian nonparametric models (1) through an OT distance between (identifiable) posterior laws of the random measures, that is, after a Bayesian update of the law of the prior $\tilde{\mu}$ with the data $X_{1:n}=(X_1,\dots,X_n)$. We denote the posterior random measure $\tilde{\mu}|X_{1:n}$ by $\tilde{\mu}^*$, omitting the dependence on n for notational convenience. In principle, we would like to use the framework developed in Section 2 and Section 3, and measure

the impact through the OT distance between scaled CRMs, $d_{\mathcal{W}}(\tilde{\mu}_{\mathcal{S}}^{1*}, \tilde{\mu}_{\mathcal{S}}^{2*})$. What makes it not possible at the present stage is that $\tilde{\mu}^{i*}$ is not a CRM, but rather maintains the independence of the increments only *conditionally* on a latent variable, as shown in James, Lijoi and Prünster [24, Theorem 1], which we report below as Theorem 10 for completeness. In Section 4.1 we thus introduce a larger class of random measures that accounts for random Lévy intensities, which we call doubly stochastic CRMs or Cox CRMs in analogy with the corresponding terminology for Poisson processes [10]. We establish a general criterion for identifiability of normalized Cox CRMs and use it to define a distance between representative Cox CRMs. In Section 4.2 we use this general framework to study the impact of the parameters of the Dirichlet processes, whereas in Section 4.3 we investigate the impact of the concentration parameter σ of the generalized gamma CRM. As far as the data $X_{1:n}$ are concerned, we do not make restrictive assumptions on the data-generating process, which is only required to satisfy a uniform integrability property (cf. Corollary 13 and Theorem 16). We work *conditionally* on $(X_n)_{n\geq 1}$, which is thus treated as a fixed deterministic sequence, and we will explain how our results only depend on the number of observations n and the number $k = k_n$ of distinct values among the n first observations.

In the rest of the section the base probability P_0 is assumed to be nonatomic, $(X_1^*, \ldots, X_k^*)$ are the $k \leq n$ distinct observations in $(X_1, \ldots, X_n)$ and $n_i = \#\{j : X_j = X_i^*\}$ is the number of observations equal to X_i^*.

THEOREM 10 (James, Lijoi and Prünster [24]). *Let $\tilde{\mu}$ be a CRM with Lévy intensity $\mathrm{d}\nu(s,x) = \mathrm{d}\rho_x(s)\mathrm{d}P_0(x)$ and let $\tilde{\mu}^* = \tilde{\mu}|X_{1:n}$ be its posterior conditionally on the data $X_{1:n}$. Then there exists a latent variable $U = U_n$ with an explicit density function such that*

$$\tilde{\mu}^*|U \overset{\mathrm{d}}{=} \tilde{\mu}_U + \sum_{i=1}^{k} J_i^U \delta_{X_i^*},$$

where $\tilde{\mu}_U$ is a CRM with Lévy intensity $\mathrm{d}\nu_U(s,x) = \mathrm{e}^{-Us}\mathrm{d}\rho_x(s)\mathrm{d}P_0(x)$, the non-negative jumps J_i^U are mutually independent and independent from $\tilde{\mu}_U$ with density function $f_i(s) \propto s^{n_i}\mathrm{e}^{-Us}\mathrm{d}\rho_{X_i^*}(s)$.

4.1. *Identifiability and optimal transport on Cox completely random measures.* In this section we introduce Cox CRMs and discuss their identifiability (Theorem 11), which leads to Cox scaled CRMs and to a natural OT distance to measure their discrepancy.

DEFINITION 6. A random measure $\tilde{\mu}$ on $\mathbb{X}$ is a Cox CRM or a doubly stochastic CRM if there exists a random measure $\tilde{\nu}$ on $(0, +\infty) \times \mathbb{X}$ such that $\tilde{\mu}|\tilde{\nu} \sim \mathrm{CRM}(\tilde{\nu})$.

The distribution of a Cox CRM is thus determined by the one of a Cox process, whose law is *uniquely* determined by the one of its Lévy intensity [25, Theorem 3.3]. For any Cox CRM we defined its scaled version conditionally on the Lévy intensity.

DEFINITION 7. Let $\tilde{\mu}$ be a Cox CRM such that $\tilde{\mu}|\tilde{\nu} \sim \mathrm{CRM}(\tilde{\nu})$ has a.s. finite mean and infinite activity. Then we define its Cox scaled CRM $\tilde{\mu}_{\mathcal{S}}$ as

$$\tilde{\mu}_{\mathcal{S}}|\tilde{\nu} = \frac{\tilde{\mu}}{\mathbb{E}(\tilde{\mu}(\mathbb{X})|\tilde{\nu})}.$$

We observe that when the Lévy intensity is deterministic, a Cox CRM is a CRM and its Cox scaled CRM coincides with the scaled CRM in Definition 2. The next result motivates the use of Cox scaled CRMs by determining their identifiability with respect to normalization, an analogue to Corollary 3 for CRMs. We recall that two measures are said equivalent if they share the same sets of measure zero.

THEOREM 11. *Let $\tilde{\mu}^i$ be a Cox CRM such that $\tilde{\mu}^i|\tilde{\nu}^i \sim \mathrm{CRM}(\tilde{\nu}^i)$ is a.s. an infinitely active CRM with finite mean, for $i = 1, 2$. In addition, for $\mathrm{d}\tilde{\nu}^i(s,x) = \mathrm{d}\tilde{\rho}^i(s)\,\mathrm{d}\tilde{P}_0^i(x)$ we assume that there exists a deterministic and not purely atomic measure Q^i on $\mathbb{X}$ such that $\tilde{P}_0^i$ is equivalent to Q^i a.s.*

Then $\tilde{\mu}^1/\tilde{\mu}^1(\mathbb{X}) \overset{\mathrm{d}}{=} \tilde{\mu}^2/\tilde{\mu}^2(\mathbb{X})$ if and only if $\tilde{\mu}^1_{\mathcal{S}} \overset{\mathrm{d}}{=} \tilde{\mu}^2_{\mathcal{S}}$.

The assumption that $\tilde{P}_0^i$ is equivalent to a fixed measure Q^i is satisfied for Cox CRM having the structure defined in Theorem 10 by taking $Q^i = P_0^i + \sum_{i=1}^k \delta_{X_i^*}$. As in Section 2 this enforces the following alternative for any set A: either $\tilde{\mu}^i(A) = 0$ a.s., or $\tilde{\mu}^i(A) > 0$ a.s.

We define a distance between the laws of Cox scaled CRMs $\tilde{\mu}$ through a distance between the laws of their random scaled Lévy intensities $\tilde{\nu}$, by relying once again on OT. Indeed, the space of scaled Lévy intensities is a metric space with respect to the OT distance in Definition 5. Following Definition 3, we can thus consider the L^1-Wasserstein distance on this metric space, which leads to

$$d_{\mathcal{W}}(\tilde{\mu}^1, \tilde{\mu}^2) = \inf_{(\tilde{\nu}_1, \tilde{\nu}_2)} \mathbb{E}(d_{\mathcal{W}}(\tilde{\nu}^1, \tilde{\nu}^2)), \tag{13}$$

where the infimum is taken over all joint laws having for marginals the random Lévy intensities of $\tilde{\mu}^1$ and $\tilde{\mu}^2$. When restricted to CRMs, $d_{\mathcal{W}}$ coincides with (11).

4.2. *Impact for the Dirichlet process.* In this section we study the impact of the parameters of the Dirichlet process, corresponding to a normalized gamma CRM, by studying the OT distance between scaled posteriors. To this end, we first show that the posterior is a Cox CRM and find its corresponding scaled version (Lemma 12). Then, we prove a number of finite-sample and asymptotic properties for the OT distance (Proposition 14), making strong use of the analysis *a priori* in Proposition 8. This will allow to us to state unprecedented conclusions about the merging rate of opinions (Corollary 15) and study its finite-sample behaviour.

First of all we show that the posterior is a Cox CRM. This result builds on Theorem 10, by further showing that the jumps J_i^U are infinitely divisible, and can thus be included in the random Lévy intensity, as shown in Lemma 27 in Section 5.4.

LEMMA 12. *Let $\tilde{\mu}$ be a gamma CRM of parameters (α, P_0). Then $\tilde{\mu}^*$ is a Cox CRM and $\tilde{\mu}^*_{\mathcal{S}}$ is a scaled gamma CRM with canonical Lévy intensity $\mathrm{d}\nu^*(s,x) = \rho^*(s)\mathrm{d}s\,\mathrm{d}P_0^*(x)$, where*

$$\rho^*(s) = (\alpha + n)\frac{e^{-(\alpha+n)s}}{s}; \qquad P_0^* = \frac{\alpha}{\alpha+n}P_0 + \frac{1}{\alpha+n}\sum_{i=1}^n \delta_{X_i}.$$

Lemma 12 shows that the posterior of a gamma CRM is a Cox CRM with a peculiar feature: all the randomness of its Lévy intensity is contained in a scale parameter, so that the corresponding Cox scaled CRM reduces to a scaled CRM.

COROLLARY 13. *Let $\tilde{\mu}^i$ be a gamma CRM of parameters (α^i, P_0^i) with P_0^i having finite first moment, for $i = 1, 2$. Then, $d_{\mathcal{W}}(\tilde{\mu}^{1*}_{\mathcal{S}}, \tilde{\mu}^{2*}_{\mathcal{S}}) = \mathcal{J} + \mathcal{A}$, with $\mathcal{J} = \mathcal{J}(\alpha_1, \alpha_2, n)$ and $\mathcal{A} = \mathcal{A}(\alpha_1, P_0^1, \alpha_2, P_0^2, n)$ defined as*

$$\mathcal{J} = \int_0^{+\infty} |(\alpha_1 + n)\Gamma(0, (\alpha_1 + n)s) - (\alpha_2 + n)\Gamma(0, (\alpha_2 + n)s)|\mathrm{d}s;$$

$$\mathcal{A} = \mathcal{W}_1\left(\frac{\alpha_1}{\alpha_1+n}P_0^1 + \frac{1}{\alpha_1+n}\sum_{i=1}^n \delta_{X_i}, \frac{\alpha_2}{\alpha_2+n}P_0^2 + \frac{1}{\alpha_2+n}\sum_{i=1}^n \delta_{X_i}\right).$$

We observe that $\mathcal{J}$ measures the discrepancy between the jumps of the posterior random measures, which does not depend on the base probability measures and depends on the observations only through their cardinality n. As for $\mathcal{A}$, it measures the discrepancy between the atoms a posteriori and it coincides with the Wasserstein distance between the predictive distributions $\mathcal{L}(X_{n+1}|X_{1:n})$ of the corresponding models.

PROPOSITION 14. *Consider the framework of Corollary 13 where, without loss of generality, $\alpha_1 \leq \alpha_2$, and let $C = \int_0^{+\infty} |\Gamma(0,t) - e^{-t}| \mathrm{d}t$. Then, there holds*

1. *$n \mapsto \mathcal{J}(\alpha_1, \alpha_2, n)$ is decreasing in n.*
2. *For every $n \geq 0$,*

$$\mathcal{J}(\alpha_1, \alpha_2, n) \leq C \log\left(\frac{\alpha_2 + n}{\alpha_1 + n}\right). \tag{14}$$

3. *As n goes to $+\infty$,*

$$\mathcal{J}(\alpha_1, \alpha_2, n) = \frac{C(\alpha_2 - \alpha_1)}{n} + O\left(\frac{1}{n^2}\right).$$

4. *For every $n \geq 1$,*

$$\mathcal{A} \leq \frac{\alpha_1}{\alpha_1 + n} \mathcal{W}_1(P_0^1, P_0^2) + \frac{(\alpha_2 - \alpha_1)n}{(\alpha_1 + n)(\alpha_2 + n)} \mathcal{W}_1\left(\frac{1}{n}\sum_{i=1}^n X_i, P_0^2\right),$$

with equality if $P_0^1 = P_0^2$ or $\alpha^1 = \alpha^2$.

Proposition 14 gives us the means to unravel both finite-sample and asymptotic properties of the distance between posterior random measures. From the asymptotic point of view, the next result is an immediate yet remarkable consequence. For sequences $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ we write $a_n \asymp b_n$ when the quotient $(a_n/b_n)_{n\in\mathbb{N}}$ stays bounded from above and from below by a strictly positive constant.

COROLLARY 15. *Let $\tilde{\mu}^i$ be a gamma CRM of parameters (α^i, P_0^i) with P_0^i having finite first moment, for $i = 1, 2$, and let $(X_n)_{n\geq 1}$ satisfy $\sup_n \left|\frac{1}{n}\sum_{i=1}^n X_i\right| < +\infty$. Then,*

$$d_{\mathcal{W}}(\tilde{\mu}_{\mathcal{S}}^{1*}, \tilde{\mu}_{\mathcal{S}}^{2*}) \asymp \frac{1}{n}.$$

Corollary 15 states that there will always be merging of opinions, and provides us with the merging rate of $1/n$, which holds independently from the choice of parameters (α_i, P_0^i). Moreover, thanks to Proposition 14 we can also study the *starting time* of the merging, which highlights a different behaviour for the jump and the atom component, as detailed below.

Point 1. of Proposition 14 states that the jump component $\mathcal{J}$ decreases steadily in n, that is, the merging starts with the first observation (time 1). We illustrate this behavior in Figure 2, where without loss of generality we fix $\alpha = 1$ and plot the value of $\mathcal{J}$ as a function of n, for different values of α_2, together with the upper bound 14, which becomes asymptotically exact as $n \to +\infty$, as stated in point 3. of Proposition 14.

As for the atom component $\mathcal{A}$, this can show different behaviours depending on the parameters (α_i, P_0^i), as showcased by the study of the upper bounds in point 4. of Proposition 14. In particular, if $\alpha_1 = \alpha_2 = \alpha$,

$$\mathcal{A} = \frac{\alpha}{\alpha + n} \mathcal{W}_1(P_0^1, P_0^2).$$

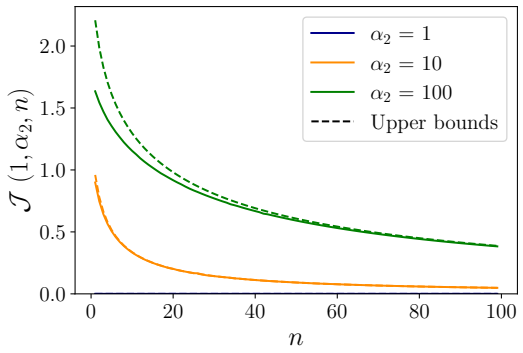

FIG 2. *Jump component $\mathcal{J}$ of $d_{\mathcal{W}}(\tilde{\mu}^{1*}_{\mathcal{S}}, \tilde{\mu}^{2*}_{\mathcal{S}})$ when $\tilde{\mu}^1$ is a gamma$(\alpha_1 = 1, P^1_0)$ CRM and $\tilde{\mu}^2$ is a gamma(α_2, P^2_0) CRM, as the number of observations n increases. The upper bounds are the right hand side of (14).*

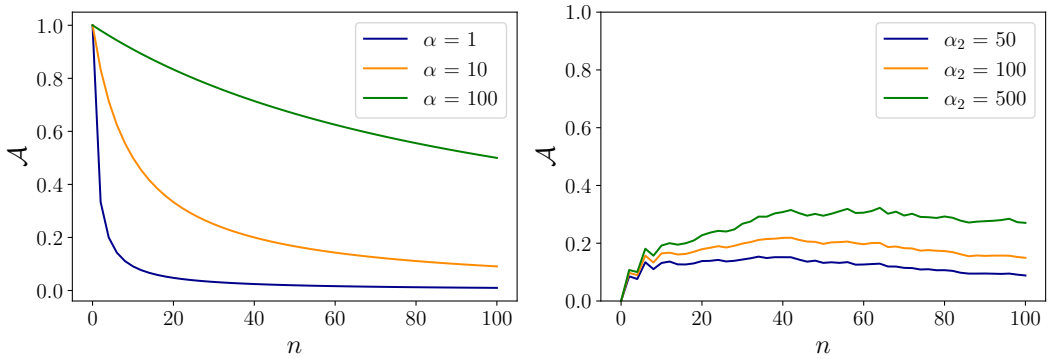

FIG 3. *Atom component $\mathcal{A}$ of $d_{\mathcal{W}}(\tilde{\mu}^{1*}_{\mathcal{S}}, \tilde{\mu}^{2*}_{\mathcal{S}})$, as the number of observations n increases, in two different scenarios. Left: $\tilde{\mu}^1$ is a gamma$(\alpha_1 = \alpha, P^1_0 = \mathcal{N}(1,1))$ CRM and $\tilde{\mu}^2$ is a gamma$(\alpha_2 = \alpha, P^2_0 = \mathcal{N}(2,1))$ CRM, where $\mathcal{N}(m, v)$ is a Gaussian distribution of mean m and variance v. Right: $\tilde{\mu}^1$ is a gamma$(\alpha_1 = 10, P^1_0 = \mathcal{N}(1,1))$ CRM and $\tilde{\mu}^2$ is a gamma$(\alpha_2, P^2_0 = \mathcal{N}(1,1))$ CRM. In both scenarios the observations are an i.i.d. sample from a Poisson distribution of mean 1.*

Thus in this scenario $\mathcal{A}$ steadily decreases as n increases, that is, the merging of the atoms starts at time 1. The left plot of Figure 3 validates it numerically by taking $\alpha_1 = \alpha_2 = \alpha$ for different values of α, P^1_0 and P^2_0 Gaussian distributions with different mean and same variance, and i.i.d. observations from a Poisson distribution.

If on the other hand $P^1_0 = P^2_0$,

$$\mathcal{A} = \frac{(\alpha_2 - \alpha_1)n}{(\alpha_1 + n)(\alpha_2 + n)} \mathcal{W}_1\left(P_0, \frac{1}{n}\sum_{i=1}^{n} X_i\right).$$

In many settings $\mathcal{W}_1(P_0, \frac{1}{n}\sum_{i=1}^{n} X_i)$ will converge to a strictly positive number (e.g., i.i.d. samples from $P \neq P_0$), so that for n large enough, the behaviour of $\mathcal{A}$ with respect to n reduces to the behaviour of the prefactor

$$\omega(n) = \frac{(\alpha_2 - \alpha_1)n}{(\alpha_1 + n)(\alpha_2 + n)}.$$

This quantity can be studied analytically: it first increases and reaches a maximum at $n \approx \sqrt{\alpha_1\alpha_2}$ before decreasing to 0 at rate $1/n$, hinting that this may well be the case for $\mathcal{A}$ as well, at least when $\sqrt{\alpha_1\alpha_2}$ is large enough. We tested this behavior numerically. In Figure 3 on the right we fix $P_0^1 = P_0^2 = P_0$ to be a Gaussian distribution, $\alpha_1 = 10$ and α_2 to vary, whereas the observations are i.i.d. from P a Poisson distribution: in this setting we know that $\mathcal{W}_1(P_0, \frac{1}{n}\sum_{i=1}^n X_i)$ converges to $\mathcal{W}_1(P_0, P)$ with rate $1/n$ [16]. We see that the distance is not decreasing in n, the maximum being achieved for $\bar{n} \approx 34$ when $\alpha_2 = 50$, $\bar{n} \approx 42$ when $\alpha_2 = 100$ and $\bar{n} \approx 64$ when $\alpha_2 = 500$. As expected, the maximum is better approximated by $\sqrt{\alpha_1\alpha_2}$ (equal to approximately 21, 31, and 72, respectively) as $\sqrt{\alpha_1\alpha_2}$ increases.

In conclusion, depending on the total base measure and the base probabilities, we can have different behaviours: either a steady decrease, or an increase phase followed by a decrease, but ultimately the asymptotic rate of convergence is $1/n$.

From a statistical perspective this finding is quite remarkable: even if two Bayesians have the same prior on a phenomenon, the difference between their opinions may increase if they see a finite amount of data, according to how much confidence they are willing to put on their prior guesses.

4.3. *Impact for the generalized gamma prior.* In this section we study the impact of the concentration parameter σ of a generalized gamma CRM, that is, the effect of using a normalized generalized gamma CRM instead of a Dirichlet process in model (1). As in the case of Gamma CRMs, we first prove that the posterior is a Cox CRM and find the expression of its Cox scaled CRM (Proposition 16), which is used for the expression of the OT distance between the posteriors in Corollary 17. We then develop an in-depth asymptotic study of the OT distance as the number of observations diverge. Since the Lévy intensities depend on a latent variable $U = U_n$, we first need to unravel its asymptotic behaviour, which crucially depends on the number of distinct observations $k = k_n$ (Theorem 18). This gives us the means to study its implications on the behaviour on the merging of opinions (Theorem 19 and Proposition 20).

We show that the posterior is a Cox CRM with the same technique used in Section 4.2 for the gamma CRM. We define

$$c = c(\alpha, \sigma, n, k, U) = \alpha(1+U)^\sigma + n - k\sigma.$$

PROPOSITION 16. *Let $\tilde{\mu}$ be a generalized gamma CRM of parameters (α, σ, P_0). Then $\tilde{\mu}^*$ is a Cox* CRM *and $\tilde{\mu}^*_{\mathcal{S}}$ is a Cox scaled CRM whose random Lévy intensity has canonical form $\mathrm{d}\tilde{\nu}^*_{\mathcal{S}}(s,x) = \tilde{\rho}^*_x(s)\mathrm{d}s\,\mathrm{d}\tilde{P}^*_0(x)$, where*

$$\tilde{\rho}^*_x(s) = \frac{c^{1-\sigma}}{\Gamma(1-\sigma)}\frac{e^{-cs}}{s^{1+\sigma}}\mathbb{1}_{\mathbb{X}\setminus\{X_1^*,\dots,X_k^*\}}(x) + c\frac{e^{-cs}}{s}\mathbb{1}_{\{X_1^*,\dots,X_k^*\}}(x),$$

$$\tilde{P}^*_0 = \frac{\alpha(1+U)^\sigma}{c}P_0 + \frac{1}{c}\sum_{i=1}^k (n_i - \sigma)\delta_{X_i^*},$$

and $U = U_n$ has probability density proportional to $u^{n-1}(1+u)^{k\sigma-n}e^{-\frac{\alpha}{\sigma}(1+u)^\sigma}$.

We then use this expression to compute the OT distance a posteriori between a generalized gamma and a gamma CRM.

COROLLARY 17. *Let $\tilde{\mu}^1$ be a generalized gamma CRM of parameters (α, σ, P_0) with P_0 having finite first moment, and let $\tilde{\mu}^2$ be a gamma CRM with same total base measure α and base probability P_0. Then,*

$$d_{\mathcal{W}}(\tilde{\mu}^{1*}, \tilde{\mu}^{2*}) = \mathbb{E}_U(\mathcal{J}_1^U + \mathcal{J}_2^U + \mathcal{A}^U),$$

where

$$\mathcal{J}_1^U = \frac{\alpha(1+U)^\sigma}{c}\mathcal{W}_{*1}\left(\frac{c^{1-\sigma}}{\Gamma(1-\sigma)}\frac{e^{-cs}}{s^{1+\sigma}}\mathrm{d}s, (\alpha+n)\frac{e^{-(\alpha+n)s}}{s}\mathrm{d}s\right),$$

$$\mathcal{J}_2^U = \frac{n-k\sigma}{c}W_{*1}\left(c\frac{e^{-cs}}{s}\mathrm{d}s, (\alpha+n)\frac{e^{-(\alpha+n)s}}{s}\mathrm{d}s\right),$$

$$\mathcal{A}^U = \mathcal{W}_1\left(\frac{\alpha(U+1)^\sigma}{c}P_0 + \frac{1}{c}\sum_{i=1}^k(n_i-\sigma)\delta_{X_i^*}, \frac{\alpha}{\alpha+n}P_0 + \frac{1}{\alpha+n}\sum_{i=1}^k n_i\delta_{X_i^*}\right),$$

with U and c are defined as in Proposition 16.

REMARK 3. The expression in Corollary 17 can be approximated through Monte Carlo integration, that is, by taking $u_1, \ldots, u_N \overset{\text{iid}}{\sim} \mathcal{L}(U)$ and then evaluating the unbiased estimator

$$\hat{d}_N = \frac{1}{N}\sum_{i=1}^N(\mathcal{J}_1^{u_i} + \mathcal{J}_2^{u_i} + \mathcal{A}^{u_i}),$$

where $\mathcal{J}_1^{u_i}$ and $\mathcal{J}_2^{u_i}$ can be evaluated through numerical integration and $\mathcal{A}^{u_i}$ only involves the (classical) Wasserstein distance, with many existing fast algorithms to approximate it [33].

If, on the other hand, one is interested in understanding some general features of $d_{\mathcal{W}}$ as the sample size n grows, one needs to understand the asymptotic behaviour of the latent variables $U = U_n$. The next result finds the asymptotic distribution of $(1+U)^\sigma$ as the number of observations diverge. We focus on $(1+U)^\sigma$ instead of U because the former is featured in the expressions of Corollary 17. The core of the proof shows that $(1+U)^\sigma$ has a log-concave density, and concentrates around a point of maximum of its log-density. Indeed, with a change of variable one finds that the density of $(1+U)^\sigma$ is proportional to $\exp(-f_n(x))$, where

$$f_n(x) = -(k_n-1)\log(x) + \frac{\alpha}{\sigma}x - (n+1)\log\left(1 - \frac{1}{x^{1/\sigma}}\right)$$

is a convex function. If f_n were proportional to n, say $f_n(x) = ng(x)$ for some smooth function g, then the result would come directly from Laplace's method. Here the situation is more delicate as there are different scales (namely, the constant α/σ, k_n and n) in the function f_n. We extend Laplace's method to this setting in Lemma 28 in Section 5, which yields a general result of independent interest. Note on the other hand that the density of U itself is *not* a log-concave.

Before stating the result let us first introduce the following standard notations for positive deterministic sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$:

- $a_n \ll b_n$ if $(a_n/b_n)_{n\in\mathbb{N}}$ converges to 0. Equivalently, $a_n = o(b_n)$.
- $a_n \lesssim b_n$ if $(a_n/b_n)_{n\in\mathbb{N}}$ stays bounded from above.
- $a_n \sim b_n$ if $(a_n/b_n)_{n\in\mathbb{N}}$ converges to 1.
- $a_n \asymp b_n$ if $(a_n/b_n)_{n\in\mathbb{N}}$ stays bounded from above, and from below by a strictly positive constant. Equivalently, $a_n \lesssim b_n$ and $b_n \lesssim a_n$.

THEOREM 18. *As $n \to +\infty$, for the L^1 convergence,*

$$\lim_{n\to+\infty}\frac{(1+U)^\sigma}{r_n} = 1,$$

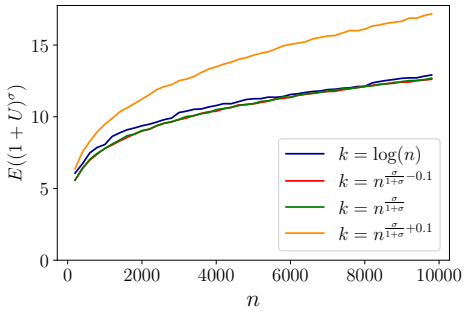

FIG 4. *Expected value of $(1+U)^\sigma$ for different numbers of unique values $k = k_n$, as the number of observations n increases. The concentration parameter is fixed to $\sigma = 0.3$.*

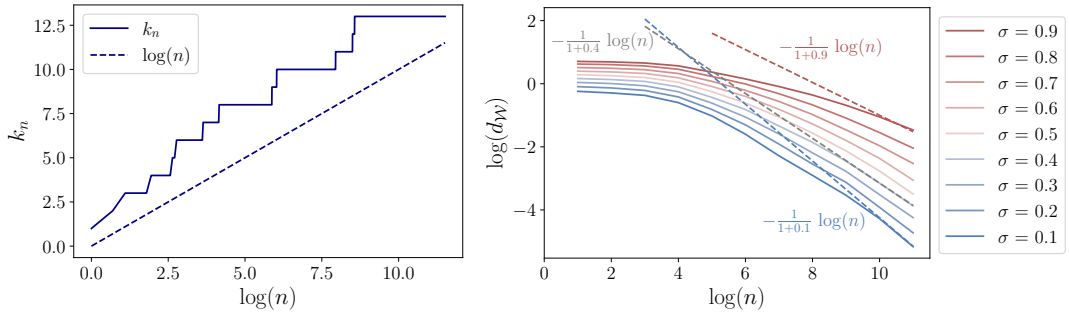

FIG 5. *Left: Unique values $k = k_n$ for a conditionally i.i.d. sample from a Dirichlet process of parameters $(\alpha = 1, P_0 = \mathcal{N}(0,1))$, as the number of observations n increases. Right: $d_{\mathcal{W}}(\tilde{\mu}^{1*}_{\mathcal{S}}, \tilde{\mu}^{2*}_{\mathcal{S}})$ when $\tilde{\mu}^1$ is a generalized gamma$(\alpha_1 = 100, P^1_0 = \mathcal{N}(0,1), \sigma)$ CRM and $\tilde{\mu}^2$ is a gamma$(\alpha_2 = 100, P^2_0 = \mathcal{N}(0,1))$ CRM, as the number of observations n increases exponentially. The observations are conditionally i.i.d. sample from a Dirichlet process of parameters $(\alpha = 1, P_0 = \mathcal{N}(0,1))$.*

where the sequence $(r_n)_{n\in\mathbb{N}}$ is defined by

$$r_n = \begin{cases} \alpha^{-\frac{\sigma}{1+\sigma}}\, n^{\frac{\sigma}{1+\sigma}} & \text{if } k \ll n^{\frac{\sigma}{1+\sigma}}, \\ \gamma\, n^{\frac{\sigma}{1+\sigma}} & \text{if } k \sim \lambda n^{\frac{\sigma}{1+\sigma}}, \\ \sigma/\alpha\, k & \text{if } k \gg n^{\frac{\sigma}{1+\sigma}}, \end{cases}$$

being here γ is the unique solution of $\sigma\lambda x^{-1} + x^{-(1+\sigma)/\sigma} = \alpha$.

Theorem 18 highlights a transition in the asymptotic behaviour of $(1+U)^\sigma$ as the number of unique values k increases. Indeed, one can easily derive that

$$\mathbb{E}((1+U)^\sigma) \asymp \max\left(n^{\frac{\sigma}{1+\sigma}}, k\right).$$

This implies that the growth rate of $(1+U)^\sigma$ is not affected by k if k is not diverging or diverging sufficiently slow, otherwise it coincides with k itself, as pictured in Figure 4.

We now use Theorem 18 to study the asymptotic behaviour of the distance in Corollary 16.

THEOREM 19. *Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ be as in Corollary 17 and let $(X_n)_{n\geq 1}$ satisfy (i) $k_n \ll n$; (ii) $\sup_n \left|\frac{1}{n}\sum_{i=1}^n X_i\right| < +\infty$; (iii) $\sup_k \left|\frac{1}{k}\sum_{i=1}^k X^*_i\right| < +\infty$. Then as $n \to +\infty$,*

$$d_{\mathcal{W}}(\tilde{\mu}^{1*}_{\mathcal{S}}, \tilde{\mu}^{2*}_{\mathcal{S}}) \asymp \max(n^{-1/(1+\sigma)}, k/n).$$

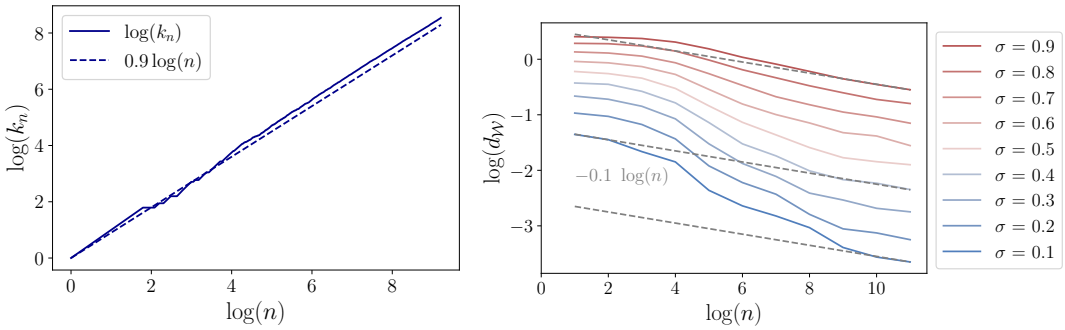

FIG 6. *Left: Unique values $k = k_n$ for a conditionally i.i.d. sample from a Pitman-Yor process of parameters $(\alpha = 1, \sigma = 0.9, P_0 = \mathcal{N}(0,1))$, as the number of observations n increases. Right: $d_{\mathcal{W}}(\tilde{\mu}^{1*}_{\mathcal{S}}, \tilde{\mu}^{2*}_{\mathcal{S}})$ when $\tilde{\mu}^1$ is a generalized gamma$(\alpha_1 = 100, P^1_0 = \mathcal{N}(0,1), \sigma)$ CRM and $\tilde{\mu}^2$ is a gamma$(\alpha_2 = 100, P^2_0 = \mathcal{N}(0,1))$ CRM, as the number of observations n increases exponentially. The observations are conditionally i.i.d. sample from a Pitman-Yor process of parameters $(\alpha = 1, \sigma = 0.9, P_0 = \mathcal{N}(0,1))$.*

Theorem 19 shows many interesting properties of the merging between a gamma and a generalized gamma CRM. First of all, recall that when comparing two gamma CRMs as in Section 4.2, the change in the parameters (α, P_0) did not change the merging rate of $1/n$. On the contrary, Theorem 19 shows that when the number of distinct values k diverges sufficiently slowly, a change in σ has a heavy impact on the merging, whose rate of convergence slows down to $n^{-\frac{1}{1+\sigma}} \in (1/n, 1/\sqrt{n})$. To illustrate this situation, in Figure 5 we sampled conditionally i.i.d. observations from a Dirichlet process of parameters $(\bar{\alpha}, \bar{P}_0)$, since it is well known that $k_n \sim \bar{\alpha} \log(n)$ a.s. [27] and thus $k_n \lesssim n^{\frac{\sigma}{1+\sigma}}$ a.s. for every σ.

Moreover, Theorem 19 also shows that as the number of unique values increases, the rate drastically deteriorates and does not depend on σ anymore. This is pictured in Figure 6, where we sampled conditionally i.i.d. observations from a Pitman-Yor process of parameters $(\bar{\alpha}, \bar{\sigma}, \bar{P}_0)$ [37], since in such case $\mathbb{E}(k_n) \asymp n^{\bar{\sigma}}$ [35], and thus for $\bar{\sigma} \in [0.5, 1)$ we expect $k_n \gtrsim n^{\frac{\sigma}{1+\sigma}}$ for every σ.

Eventually, if $k \asymp n$ there will not be merging of opinions. As a notable example, this happens a.s. when the observations are sampled i.i.d. from a continuous distribution, as described in the next result. It is worth underlining that the lack of merging should not worry the practitioner too much: since model (1) with normalization gives positive probability to ties among the observations, i.e., $\mathbb{P}(X_i = X_j) > 0$, when dealing with independent and continuous data other models for random densities are to be preferred, such as kernel mixtures over a random probability measure.

PROPOSITION 20. *Let $\tilde{\mu}_1$ and $\tilde{\mu}_2$ be as in Corollary 17 and let $X_n \overset{\text{iid}}{\sim} P$, where P is a continuous distribution with finite first moment. Then as $n \to +\infty$, for a.e. realization of $(X_n)_{n \geq 1}$,*

$$d_{\mathcal{W}}(\tilde{\mu}^{1*}_{\mathcal{S}}, \tilde{\mu}^{2*}_{\mathcal{S}}) \to \sigma \left(\int_0^{+\infty} \left| \frac{\Gamma(-\sigma, t)}{\Gamma(1-\sigma)} - \Gamma(0,t) \right| \mathrm{d}t + \mathcal{W}_1(P_0, P) \right).$$

We observe that the limit in Proposition 20 does not depend on α and it is composed by two terms. The first term coincides with the distance *a priori* between a scaled gamma and scaled generalized gamma with same total base measure, and can thus be seen as a measure of discrepancy between the priors for the jumps. The second term coincides with the distance between the base measure and the law of the observations, and can thus be seen

as a measure of discrepancy between the (common) prior for the atoms and the data. Thanks to Proposition 9 we know that both terms are increasing in σ, which thus acts as an overall measure of confidence in the prior.

From a statistical perspective, Theorem 19 unravels the regimes where using a normalized generalized gamma CRM will actually guarantee a different learning outcome, with the following take-home message: either the number of observations n is sufficiently small or the number of distinct values k should be sufficiently large, and how large it must be depends on the choice of concentration σ, namely, $k > n^{\frac{\sigma}{1+\sigma}}$. As notable examples, this happens when the observations are i.i.d. from a continuous distribution or when they are conditionally i.i.d. given a normalized generalized gamma CRM or a Pitman-Yor process, but it does not happen when the observations are i.i.d. from a discrete distribution with a finite number of atoms or when they are conditionally i.i.d. given a Dirichlet process or a hierarchical Dirichlet process.

5. Proofs. We here present all the proofs of our results, divided by section. First, we recall some classical results about the L_1-Wasserstein distance that will be repeatedly used in the following. We refer the interested reader to, e.g., [42, 40, 1].

The L_1-Wasserstein distance defined in (3) is finite if P^1, P^2 have finite first moments and, in this case,

$$\mathcal{W}_1(P^1, P^2) \leq \mathbb{E}_{P^1}(d_{\mathbb{Y}}(y_0, X)) + \mathbb{E}_{P^2}(d_{\mathbb{Y}}(y_0, Y)), \tag{15}$$

where y_0 is any point in $\mathbb{Y}$. Moreover, the following topological property holds.

PROPOSITION 21. *The function $\mathcal{W}_1$ is a distance over the set of probability distributions over Y with finite first moments, and it metrizes the topology of weak convergence together with convergence of the first moments.*

A useful property that we often use is the convexity inequality

$$\mathcal{W}_1(\lambda P^1 + (1-\lambda)Q^1, \lambda P^2 + (1-\lambda)Q^2) \leq \lambda \mathcal{W}_1(P^1, P^2) + (1-\lambda)\mathcal{W}_1(Q^1, Q^2), \tag{16}$$

which holds for all $\lambda \in (0,1)$ and P^1, P^2, Q^1, Q^2 probability distributions with finite first moment. Moreover, if $P^1 = P^2 = P$ then there is equality, that is,

$$\mathcal{W}_1(\lambda P + (1-\lambda)Q^1, \lambda P + (1-\lambda)Q^2) = (1-\lambda)\mathcal{W}_1(Q^1, Q^2). \tag{17}$$

This last equality case is specific to the Wasserstein distance of order and can be proved using the dual formulation of optimal transport, see for instance Equation (3.1) in [40].

We conclude by recalling the notion of pushforward of a measure. If $f : \mathbb{X} \to \mathbb{Y}$ is a measurable map between two measurable spaces and ρ is a positive measure on $\mathbb{X}$, then $f\#\rho$ is the measure on $\mathbb{Y}$ defined as $(f\#\rho)(B) = \rho(f^{-1}(B))$ for any measurable set $B \subseteq \mathbb{Y}$.

5.1. *Proofs of Section 2.* In this section we prove Proposition 1, Theorem 2 and Lemma 4, which reduces to an auxiliary Lemma 22.

PROOF OF PROPOSITION 1. Campbell's formula yields, for any Borel set $A \subseteq \mathbb{X}$,

$$\iint_{(0,+\infty)\times A} s \, \mathrm{d}\nu(s,x) = \mathbb{E}(\tilde{\mu}(A)) < +\infty.$$

By applying it for $A = \mathbb{X}$, we see that the measure $\bar{\nu} = s\,\mathrm{d}\nu(s,x)$ is finite on the product space $(0,+\infty) \times \mathbb{X}$. Thus its disintegration $\mathrm{d}\bar{\nu}(s,x) = \mathrm{d}\varrho_x(s)\mathrm{d}Q(x)$ is well-defined, with

Q the second marginal measure of $\bar{\nu}$ on $\mathbb{X}$ and ϱ a probability kernel. Again by Campbell's formula, it is easy to see that $Q(A) = \mathbb{E}(\tilde{\mu}(A))$ for any Borel set A. We define

$$P_0 = \frac{Q}{Q(\mathbb{X})}, \qquad \mathrm{d}\rho_x(s) = Q(\mathbb{X})\frac{\mathrm{d}\varrho_x(s)}{s},$$

which entails the sought decomposition: P_0 is a probability measure and since ϱ is a probability kernel, it is clear that for P_0-a.e. x there holds

$$\int_0^{+\infty} s\,\mathrm{d}\rho_x(s) = Q(\mathbb{X})\int_0^{+\infty}\mathrm{d}\varrho_x(s) = Q(\mathbb{X}) = \mathbb{E}(\tilde{\mu}(\mathbb{X})). \qquad \square$$

Before proving Theorem 2 we first state a preliminary result that can be seen as a special case of [7, Lemma 7]. We report its proof for completeness.

LEMMA 22. *Let $\tilde{\mu}$ be a CRM with Lévy intensity $\rho_x(\mathrm{d}s)\mathrm{d}P_0(x)$ and let $k > 0$ be a constant. Then $\tilde{\mu}_k(A) := \tilde{\mu}(A)/k$ is a CRM with Lévy intensity $\mathrm{d}(S_k\#\rho_x)(s)\,\mathrm{d}P_0(x)$, being $S_k : s \in (0,+\infty) \to s/k \in (0,+\infty)$ the scaling by $1/k$.*

In particular, if $\mathrm{d}\rho_x(s) = \rho_x(s)\mathrm{d}s$ has a density with respect to the Lebesgue measure, $\tilde{\mu}_k$ has Lévy intensity $k\rho_x(ks)\,\mathrm{d}s\,\mathrm{d}P_0(x)$.

PROOF. Clearly, $\tilde{\mu}_k$ has independent evaluations on disjoint sets. Its Laplace functional satisfies

$$\mathbb{E}\left(\exp\left(-\int_{\mathbb{X}} g(x)\,\mathrm{d}\tilde{\mu}_k(x)\right)\right) = \exp\left\{-\int_{(0,+\infty)\times\mathbb{X}}(1-e^{-s\frac{g(x)}{k}})\mathrm{d}\rho_x(s)\mathrm{d}P_0(x))\right\}$$

which shows the conclusion in the general case by definition of the pushforward measure. For the case where $\rho_x(\mathrm{d}s) = \rho_x(s)\mathrm{d}s$ we do the change of variables $t = s/k$. $\square$

PROOF OF THEOREM 2. As written above, the proof is inspired by a result on subordinators by David Aldous and Stevan Evans [36, Lemma 7.5]. The first and second steps below can be seen as adaptation of their ideas to our context.

By independence of the evaluations on disjoint sets it suffices to prove that $\tilde{\mu}^1(A) \overset{\mathrm{d}}{=} \alpha\tilde{\mu}^2(A)$ for every Borel set A. We write $\mathrm{d}\nu^i(s,x) = \mathrm{d}\rho_x^i(s)\mathrm{d}P_0^i(x)$ for the Lévy intensity of $\tilde{\mu}^i$, as in Proposition 1.

Preliminary step: a note on null-sets. By the infinite activity assumption it follows that $\tilde{\mu}^i(A) = 0$ a.s. if and only if $P_0^i(A) = 0$. In particular $\tilde{\mu}^i(A)/\tilde{\mu}^i(\mathbb{X}) = 0$ a.s. if and only if $P_0^i(A) = 0$. Since $\tilde{\mu}^1/\tilde{\mu}^1(\mathbb{X}) \overset{\mathrm{d}}{=} \tilde{\mu}^2/\tilde{\mu}^2(\mathbb{X})$, we conclude that P_0^1, P_0^2 are equivalent, that is, they have the same null set. Moreover, still by the infinite activity assumption, if $P_0^1(A) > 0$, then $0 < \tilde{\mu}^i(A) < +\infty$ a.s., for $i = 1, 2$.

First step: defining auxiliary Cox CRMs. Let $A \subseteq \mathbb{X}$ a subset with $P_0^1(A) > 0$ and A^c its complement. For every $i = 1, 2$ we can define a random measure on A^c as

$$\tilde{\xi}^i(\cdot) = \frac{\tilde{\mu}^i(\cdot)}{\tilde{\mu}^i(A)}.$$

We will prove that $\tilde{\xi}^1 \overset{\mathrm{d}}{=} \tilde{\xi}^2$ and then leverage this equality to get information on the Lévy intensities. Because of the independence of the evaluations on disjoint sets, conditionally on $\tilde{\mu}^i(A)$, $\tilde{\xi}^i$ is a CRM on A^c. Thus, $\tilde{\xi}^i$ is a Cox CRM as in Definition 6 and to prove that $\tilde{\xi}^1 \overset{\mathrm{d}}{=} \tilde{\xi}^2$

it suffices to prove that $\tilde{\xi}^1(B) \overset{\mathrm{d}}{=} \tilde{\xi}^2(B)$ for every $B \subset A^c$. By dividing both numerator and denominator by $\tilde{\mu}^i(\mathbb{X})$, we observe that $\tilde{\xi}^i(B) = \tilde{p}^i(B)/\tilde{p}^i(A)$, where $\tilde{p}^i = \tilde{\mu}^i/\tilde{\mu}^i(\mathbb{X})$. Since $\tilde{p}^1 \overset{\mathrm{d}}{=} \tilde{p}^2$, then $(\tilde{p}^1(B), \tilde{p}^1(A)) \overset{\mathrm{d}}{=} (\tilde{p}^2(B), \tilde{p}^2(A))$ and thus $\tilde{\xi}^1(B) \overset{\mathrm{d}}{=} \tilde{\xi}^2(B)$.

Second step: deducing information on Lévy intensities. By Lemma 22, conditionally on $\tilde{\mu}^i(A)$, ξ^i is a CRM with Lévy intensity $\mathrm{d}(S_{\tilde{\mu}^i(A)}\#\rho^i_x)(s)\,\mathrm{d}P^i_0(x)$. Here $S_k : (0,+\infty) \to (0,+\infty)$ is defined by $S_k(s) = s/k$, it is the scaling by $1/k$. Thus $\mathrm{d}\tilde{\nu}^i(s,x) = \mathrm{d}(S_{\tilde{\mu}^i(A)}\#\rho^i_x)(s)\,\mathrm{d}P^i_0(x)$ is the random Lévy intensity of the Cox CRM ξ^i. By standard results on Cox processes [25, Theorem 3.3], the law of the random Lévy intensity characterizes the process, which implies

$$\mathrm{d}(S_{\tilde{\mu}^1(A)}\#\rho^1_x)(s)\,\mathrm{d}P^1_0(x) \overset{\mathrm{d}}{=} \mathrm{d}(S_{\tilde{\mu}^2(A)}\#\rho^2_x)(s)\,\mathrm{d}P^2_0(x).$$

We then multiply both sides by s and integrate over $(0,+\infty) \times A^c$. By definition of the pushforward, and since ρ^i_x have first moment equal to $\mathbb{E}(\tilde{\mu}^i(\mathbb{X}))$ (see Proposition 1), we deduce that

$$\frac{P^1_0(A^c)\mathbb{E}(\tilde{\mu}^1(\mathbb{X}))}{\tilde{\mu}^1(A)} \overset{\mathrm{d}}{=} \frac{P^2_0(A^c)\mathbb{E}(\tilde{\mu}^2(\mathbb{X}))}{\tilde{\mu}^2(A)}.$$

So we conclude $\tilde{\mu}^1(A) \overset{\mathrm{d}}{=} \alpha_A \tilde{\mu}^2(A)$, at least if $P^1_0(A)$ and $P^1_0(A^c)$ do not vanish (cf. preliminary step). To conclude the proof, we need to show that α_A does not depend on A and treat the case where $P^1_0(A) = 0$ or $P^1_0(A^c) = 0$.

Third step: showing that α_A does not depend on A. First we take A such that α_A and α_{A^c} are well defined, that is, both $P^1_0(A)$ and $P^1_0(A^c)$ are strictly positive, and we show that $\alpha_A = \alpha_{A^c}$. By the discussion of the first step:

$$\frac{\tilde{\mu}^1(A^c)}{\tilde{\mu}^1(A)} = \tilde{\xi}^1(A^c) \overset{\mathrm{d}}{=} \tilde{\xi}^2(A^c) = \frac{\tilde{\mu}^2(A^c)}{\tilde{\mu}^2(A)}.$$

Moreover since $\tilde{\mu}^i$ is a CRM, $(\tilde{\mu}^1(A), \tilde{\mu}^1(A^c)) \overset{\mathrm{d}}{=} (\alpha_A\tilde{\mu}^2(A), \alpha_{A^c}\tilde{\mu}^2(A^c))$, and both correspond to a random vector with independent components. We deduce that

$$\frac{\tilde{\mu}^1(A^c)}{\tilde{\mu}^1(A)} \overset{\mathrm{d}}{=} \frac{\alpha_{A^c}}{\alpha_A}\frac{\tilde{\mu}^2(A^c)}{\tilde{\mu}^2(A)},$$

and this necessarily yields $\alpha_{A^c} = \alpha_A$. It follows that

$$\tilde{\mu}^1(\mathbb{X}) = \tilde{\mu}^1(A) + \tilde{\mu}^1(A^c) \overset{\mathrm{d}}{=} \alpha_A\left(\tilde{\mu}^2(A) + \tilde{\mu}^2(A^c)\right) = \alpha_A\tilde{\mu}^2(\mathbb{X}).$$

This implies $\alpha_A = \alpha_{\mathbb{X}}$, which thus does not depend on A. We denote it by α.

Eventually we have to consider the case $P^1_0(A) = 0$ or $P^1_0(A^c) = 0$. If $P^1_0(A) = 0$, by the preliminary step also $P^2_0(A) = 0$, so that both $\tilde{\mu}^1(A)$ and $\tilde{\mu}^2(A)$ are a.s. equal to 0. In particular, $\tilde{\mu}^1(A) = \alpha\tilde{\mu}^2(A) = 0$ a.s. On the other hand, if $P^1_0(A^c) = 0$, then $P^2_0(A^c) = 0$, and thus $\tilde{\mu}^1(A) = \tilde{\mu}^1(\mathbb{X}) \overset{\mathrm{d}}{=} \alpha\tilde{\mu}^2(\mathbb{X})$ while $\tilde{\mu}^2(A) = \tilde{\mu}^2(\mathbb{X})$ a.s., which completes the proof. □

PROOF OF LEMMA 4. The proof follows by Lemma 22 by taking $k = \mathbb{E}(\tilde{\mu}(\mathbb{X}))$. □

5.2. *Proofs of Section 3.* In this section we prove Proposition 5, Lemma 6, Theorem 7, Proposition 8 and Proposition 9, together with some auxiliary results: Lemma 23, Lemma 24, Proposition 25, Lemma 26. Before heading to the proof of Proposition 5, we introduce some notations and auxilliary results. We denote by $\mathscr{L}$ the Lesbegue measure on $(0,+\infty)$ and $\mathscr{L}^2$ the Lebesgue measure on $(0,+\infty)\times(0,+\infty)$. We start with a preliminary lemma which mimics [40, Proposition 2.2].

LEMMA 23. *Let ρ a positive measure on $(0,+\infty)$ with tail integral U_ρ and let U_ρ^{-1} denote its generalized inverse, as in Definition 5. Then the restriction to $(0,+\infty)$ of $U_\rho^{-1}\#\mathscr{L}$ is equal to ρ.*

PROOF. We only need to evaluate $U_\rho^{-1}\#\mathscr{L}$ on sets $(a,+\infty)$ for $a>0$ and show it coincides with $U_\rho(a)$. Indeed, $(U_\rho^{-1}\#\mathscr{L})(a,+\infty)$ is equal to

$$\mathscr{L}(\{s\in(0,+\infty)\ :\ U_\rho^{-1}(s)>a\})=\mathscr{L}(\{s\in(0,+\infty)\ :\ s<U_\rho(a)\})=U_\rho(a),$$

where we used that $U_\rho^{-1}(s)>a$ if and only if $s<U_\rho(a)$, which follows from the definition of U_ρ^{-1}. □

LEMMA 24. *Let $\nu^i=\mathrm{d}\rho_x^i(s)\mathrm{d}P_0^i(x)$ for $i=1,2$ two scaled Lévy intensities and $\pi\in\Pi(P_0^1,P_0^2)$. Then there exists a measure $\overline{\pi}=\overline{\pi}(s_1,x_1,s_2,x_2)$ on $([0,+\infty)\times\mathbb{X})^2$ with three properties: i) the projection $\overline{\pi}_{12}$ on the first two marginals, when restricted to $(0,+\infty)\times\mathbb{X}$, coincide with ν^1; ii) the projection $\overline{\pi}_{34}$ on the last two marginals, when restricted to $(0,+\infty)\times\mathbb{X}$, coincide with ν^2; iii) the coupling $\overline{\pi}$ satisfies*

$$\mathbb{E}_{(X,Y)\sim\pi}\left(d_{\mathbb{X}}(X,Y)^p+\mathcal{W}_{*p}(\rho_X^1,\rho_Y^2)^p\right)$$
$$=\int_{([0,+\infty)\times\mathbb{X})^2}\left(|s_1-s_2|^p+d_{\mathbb{X}}(x_1,x_2)^p\right)\,\mathrm{d}\overline{\pi}(s_1,x_1,s_2,x_2)$$

PROOF. We define $\overline{\pi}$ as follows. Consider $(X,Y)\sim\pi$. Then for any positive measurable function $f:([0,+\infty)\times\mathbb{X})^2\to\mathbb{R}$ we define

$$\int f\,\mathrm{d}\overline{\pi}=\mathbb{E}_{X,Y}\left(\int_0^{+\infty}f\left(U_{\rho_X^1}^{-1}(s),X,U_{\rho_Y^2}^{-1}(s),Y\right)\,\mathrm{d}s\right).\tag{18}$$

This defines a positive measure on $([0,+\infty)\times\mathbb{X})^2$. We use (18) to prove the three properties. First, for every $f(s_1,x_1,s_2,x_2)=g(s_1,x_1)$ such that $g(0,x_1)=0$, by Lemma 23 and since the first marginal of π is P_0^1,

$$\int f\mathrm{d}\overline{\pi}=\mathbb{E}\left(\int g(U_{\rho_X^1}^{-1}(s),X)\,\mathrm{d}s\right)=\iint g(s,x)\,\mathrm{d}\rho_x(s)\mathrm{d}P_0^1(x)=\iint g\,\mathrm{d}\nu^1.$$

This proves i) and similar techniques may be used for ii). As for iii), we consider $f(s_1,x_1,s_2,x_2)=|s_1-s_2|^p+d_{\mathbb{X}}(x_1,x_2)^p$.

$$\int f\,\mathrm{d}\overline{\pi}=\mathbb{E}_{X,Y}\left(\int_0^{+\infty}\left|U_{\rho_X^1}^{-1}(s)-U_{\rho_Y^2}^{-1}(s)\right|^p\mathrm{d}s\right)+\mathbb{E}(d_{\mathbb{X}}(X,Y)^p)$$
$$=\mathbb{E}_{X,Y}\left(\mathcal{W}_{*p}(\rho_X^1,\rho_Y^2)+d_{\mathbb{X}}(X,Y)^p\right).\quad\square$$

With the help of the key Lemma 24 we can now move to the fundamental results about our OT distance. First, we prove that the distances we consider are indeed distances.

PROOF OF PROPOSITION 5. We start with $\mathcal{W}_{*p}$. It clearly satisfies the triangle inequality as it corresponds to a L^p distance between the inverse tail integrals U^{-1}. Identifiability comes from the fact that $U_1^{-1} = U_2^{-1}$ a.e. implies $U_1 = U_2$ a.e., and thus $\rho^1 = \rho^2$.

Next we move to $d_{\mathcal{W},p}$. For the triangle inequality, consider ν^1, ν^2 and ν^3. If π^{12} and π^{23} are couplings respectively in $\Pi(P_0^1, P_0^2)$ and $\Pi(P_0^2, P_0^3)$, by the gluing Lemma [42, Chapter 1], we can find (X, Y, Z) such that $(X, Y) \sim \pi^{12}$ and $(Y, Z) \sim \pi^{23}$. Then we see that

$$d_{\mathbb{X}}(X,Z)^p + \mathcal{W}_{*p}(\rho_X^1, \rho_Z^3)^p \leq (d_{\mathbb{X}}(X,Y) + d_{\mathbb{X}}(Y,Z))^p + (\mathcal{W}_{*p}(\rho_X^1, \rho_y^2) + \mathcal{W}_{*p}(\rho_Y^2, \rho_Z^3))^p$$

thanks to the triangle inequality for $d_{\mathbb{X}}$ and $\mathcal{W}_{*p}$. We then take the expectation, and use Minkowski's inequality to see that

$$d_{\mathcal{W},p}(\nu^1, \nu^3) \leq \mathbb{E}\left(d_{\mathbb{X}}(X,Y)^p + \mathcal{W}_{*p}(\rho_X^1, \rho_Y^2)^p\right)^{1/p} + \mathbb{E}\left(d_{\mathbb{X}}(Y,Z)^p + \mathcal{W}_{*p}(\rho_Y^2, \rho_Z^3)^p\right)^{1/p}.$$

Taking the infimum in π^{12} and π^{23} in the right hand side yields the triangle inequality.

Eventually, we have to check identifiability, that is, what happens if $d_{\mathcal{W},p}(\nu^1, \nu^2) = 0$. Let $(\pi^n)_n$ a minimizing sequence in (11). With the help of Lemma 24 we construct $\overline{\pi}^n$ which are couplings between ν^1 and ν^2. Let f any function on $(0, +\infty) \times \mathbb{X}$ which is Lipschitz continuous and compactly supported, say with Lipschitz constant C. Then

$$\int f(s,x)\,\mathrm{d}\nu^1(s,x) - \int f(s,x)\,\mathrm{d}\nu^2(s,x) = \iint (f(s_1,x_1) - f(s_2,x_2))\,\mathrm{d}\overline{\pi}^n(s_1,x_1,s_2,x_2)$$

thus we can estimate

$$\left|\int f(s,x)\,\mathrm{d}\nu^1(s,x) - \int f(s,x)\,\mathrm{d}\nu^2(s,x)\right| \leq C \iint |s_1 - s_2| + d_{\mathbb{X}}(x_1,x_2)\,\mathrm{d}\overline{\pi}^n(s_1,x_1,s_2,x_2).$$

Sending n to $+\infty$, the right hand side converges to 0 thanks to Lemma 24 and the assumption $d_{\mathcal{W},p}(\nu^1, \nu^2) = 0$. Thus ν^1 and ν^2 have the same integrals against Lipschitz continuous and compactly supported, they coincide as measures. □

PROOF OF LEMMA 6. The proof is similar to [40, Proposition 2.17], which focuses on an analogous result for the (classic) $\mathcal{W}_1$ on probabilities. First we observe that

$$\int_0^{+\infty} |U_1^{-1}(s) - U_2^{-1}(s)|\,\mathrm{d}s = \\ = \mathscr{L}^2\{(s,u) : U_1^{-1}(s) \leq u < U_2^{-1}(s)\} + \mathscr{L}^2\{(s,u) : U_2^{-1}(s) \leq u < U_1^{-1}(s)\}$$

Thanks to the properties of the generalized inverse of a non-decreasing function, $U_i^{-1}(s) \leq u < U_j^{-1}(s)$ if and only if $U_i^{-1}(s) > u \geq U_j^{-1}(s)$. Thus the last expression is equal to

$$\mathscr{L}^2\{(s,u) : U_1(u) \leq s < U_2(u)\} + \mathscr{L}^2\{(s,u) : U_2(s) \leq s < U_1^{-1}(u)\} = \\ = \int_0^{+\infty} |U_1(u) - U_2(u)|\,\mathrm{d}u.$$

To prove the second part of the Lemma, we start by the triangle inequality, which yields

$$\int_0^{+\infty} |U_1(u) - U_2(u)|\mathrm{d}u \leq \int_0^{+\infty} U_1(u)\mathrm{d}u + \int_0^{+\infty} U_2(u)\mathrm{d}u,$$

and then use Fubini's theorem to notice that, for $i = 1, 2$,

$$\int_0^{+\infty} U_i(u)\mathrm{d}u = \int_0^{+\infty} \int_u^{+\infty} \mathrm{d}\rho^i(s)\,\mathrm{d}u = \int_0^{+\infty} s\,\mathrm{d}\rho^i(s) = 1. \qquad \square$$

PROOF OF THEOREM 7. The main idea of the proof is to use the coupling between Lévy intensities given by Lemma 24 to build a coupling between the corresponding CRMs.

We fix $\varepsilon > 0$. Consider $\pi \in \Pi(P_0^1, P_0^2)$ an ε -optimal coupling with respect to the infimum in the definition of $d_{\mathcal{W}}$ (11). We look at $\overline{\pi}$ the measure on $([0, +\infty) \times \mathbb{X})^2$ given by Lemma 24 in the case $p = 1$. In particular

$$\int_{([0,+\infty)\times\mathbb{X})^2} (|s_1 - s_2| + d_{\mathbb{X}}(x_1, x_2))\ \mathrm{d}\overline{\pi}(s_1, x_1, s_2, x_2) \leq d_{\mathcal{W},p}(\nu^1, \nu^2) + \varepsilon.$$

We consider $\tilde{\mathcal{N}}$ a Poisson point process on $([0, +\infty) \times \mathbb{X})^2$ with intensity $\overline{\pi}$. Then, because of the marginal properties of $\overline{\pi}$,

$$\tilde{\mu}^1 \overset{\mathrm{d}}{=} \sum_{(s_1,x_1,s_2,x_2)\in\tilde{\mathcal{N}}} s_1\delta_{x_1}, \qquad \tilde{\mu}^2 \overset{\mathrm{d}}{=} \sum_{(s_1,x_1,s_2,x_2)\in\tilde{\mathcal{N}}} s_2\delta_{x_2}.$$

We have built a coupling between the two CRMs $\tilde{\mu}^1$ and $\tilde{\mu}^2$. In particular, for a test function $f : \mathbb{X} \to \mathbb{R}$, it yields a coupling between the two random variables $\int_{\mathbb{X}} f\,\mathrm{d}\tilde{\mu}^1$ and $\int_{\mathbb{X}} f\,\mathrm{d}\tilde{\mu}^2$, and we have

$$\int_{\mathbb{X}} f\,\mathrm{d}\tilde{\mu}^1 - \int_{\mathbb{X}} f\,\mathrm{d}\tilde{\mu}^2 = \sum_{(s_1,x_1,s_2,x_2)\in\tilde{\mathcal{N}}} s_1 f(x_1) - s_2 f(x_2).$$

We want to bound the right hand side of the equation above in expectation. Note that by the triangle inequality

$$\begin{aligned} |s_1 f(x_1) - s_2 f(x_2)| &= |s_1 f(x_1) - s_2 f(x_1) + s_2 f(x_1) - s_2 f(x_2)| \\ &\leq C(|s_1 - s_2| + s_2 d_{\mathbb{X}}(x_1, x_2)) \end{aligned}$$

where $C = \max(\sup|f|, \mathrm{Lip}(f))$. Using Campbell's formula, we find that

$$\mathbb{E}\left(\left|\int_{\mathbb{X}} f\,\mathrm{d}\tilde{\mu}^1 - \int_{\mathbb{X}} f\,\mathrm{d}\tilde{\mu}^2\right|\right) \leq C \int |s_1 - s_2| + s_2 d_{\mathbb{X}}(x_1, x_2)\,\mathrm{d}\overline{\pi}.$$

The left hand side is always an upper bound for $\mathcal{W}_1(\int_{\mathbb{X}} f\,\mathrm{d}\tilde{\mu}^1, \int_{\mathbb{X}} f\,\mathrm{d}\tilde{\mu}^2)$, so to conclude the proof we only need to estimate the right hand side. For the second term in the integral, thanks the precise structure of the coupling given by Lemma 24, see (18), we have

$$\int s_2\, d_{\mathbb{X}}(x_1, x_2)\,\mathrm{d}\overline{\pi} = \mathbb{E}_{X,Y}\left(d_{\mathbb{X}}(X, Y) \int_0^{+\infty} U^{-1}_{\rho_Y^2}(s)\,\mathrm{d}s\right).$$

Thanks to Lemma 23, $\int_0^{+\infty} U^{-1}_{\rho_Y^2}(s)\,\mathrm{d}s = \int_0^{+\infty} s\rho_Y^2(s)\mathrm{d}s = 1$ a.s., so that

$$\int s_2\, d_{\mathbb{X}}(x_1, x_2)\,\mathrm{d}\overline{\pi} = \mathbb{E}_{X,Y}\left(d_{\mathbb{X}}(X, Y)\right) = \int d_{\mathbb{X}}(x_1, x_2)\,\mathrm{d}\overline{\pi}.$$

Summing up,

$$\mathcal{W}_1\left(\int_{\mathbb{X}} f\mathrm{d}\tilde{\mu}^1, \int_{\mathbb{X}} f\mathrm{d}\tilde{\mu}^2\right) \leq C \int |s_1 - s_2| + d_{\mathbb{X}}(x^1, x^2)\,\mathrm{d}\overline{\pi} \leq C(d_{\mathcal{W}}(\nu^1, \nu^2) + \varepsilon),$$

and we can conclude as ε is arbitrary. $\square$

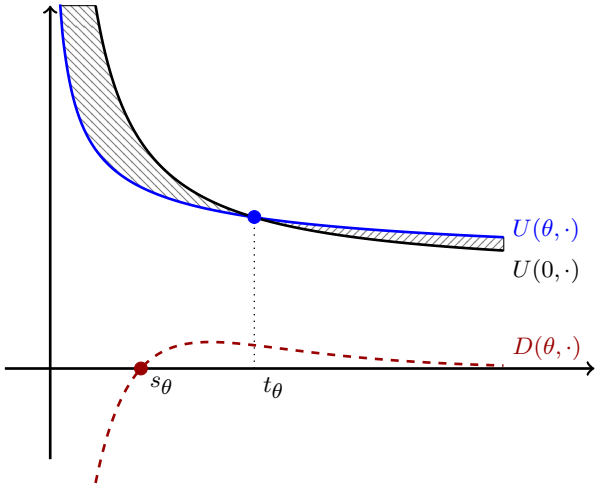

FIG 7. *Illustration of the geometric situation depicted in Proposition 25 with the functions $U(\theta,\cdot)$, $U(\theta,0)$ and the derivative $D(\theta,\cdot)=\frac{\partial U}{\partial\theta}(\theta,\cdot)$. The function $F(\theta)$ corresponds to the (unsigned) area between the two curves. By Assumption 1 the areas before and after t_θ are the same, so $F(\theta)$ is twice the area before t_θ, and also twice the area after t_θ.*

We now develop a general framework for the study of the jump component of the distance $d_{\mathcal{W}}$ on CRMs. This will be the main ingredient to prove Proposition 8 and Proposition 9.

PROPOSITION 25. *Let $U(\theta,t)$ a continuous function $[0,+\infty)\times(0,+\infty)\to\mathbb{R}$ such that the following assumptions hold:*

1. *$\mathcal{I}(t)=\int_0^{+\infty}U(\theta,t)\,\mathrm{d}t$ does not depend on θ.*
2. *For any $\theta>0$, there exists one single point t_θ such that $U(0,t_\theta)=U(\theta,t_\theta)$.*
3. *The function $D(\theta,t)=\frac{\partial U}{\partial\theta}(\theta,t)$ is defined everywhere, continuous and satisfies the integrability condition:*

$$\forall\Theta>0,\quad \int_0^{+\infty}\sup_{0\le\theta\le\Theta}|D(\theta,t)|\mathrm{d}t<+\infty.$$

   Moreover, for any $\theta\ge 0$ the function $D(\theta,\cdot)$ vanishes at a single point s_θ.
4. *For $t\ge\max(t_\theta,s_\theta)$ (or equivalently for $t\le\min(t_\theta,s_\theta)$) the functions $U(\theta,t)-U(0,t)$ and $D(\theta,t)$ have the same sign.*

Then the function

$$F:\theta\mapsto\int_0^{+\infty}|U(\theta,t)-U(0,t)|\mathrm{d}t$$

is strictly increasing, differentiable everywhere and

$$F'(\theta)\le\int_0^{+\infty}|D(\theta,t)|\mathrm{d}t, \tag{19}$$

with equality when $\theta=0$.

REMARK 4. In our examples $U(\theta,\cdot)=\rho_\theta(\cdot,+\infty)$ is the tail integral of a measure ρ_θ. Assumption 1 can then be verified through the identity $\int_0^{+\infty}U(\theta,t)\mathrm{d}t=\int_0^{+\infty}s\,\mathrm{d}\rho_\theta(s)$, together with condition (6).

PROOF. Up to changing U into $-U$, we can assume that $U(\theta,t)-U(0,t)$ is negative on $(0,t_\theta)$ and positive on $(t_\theta,+\infty)$. We refer to Figure 7 for an illustration of the situation. We can write

$$F(\theta)=\int_0^{t_\theta}(U(0,t)-U(\theta,t))\mathrm{d}t+\int_{t_\theta}^{+\infty}(U(\theta,t)-U(0,t))\mathrm{d}t.$$

As moreover $\int_0^{+\infty}U(0,t)\mathrm{d}t=\int_0^{+\infty}U(\theta,t)\mathrm{d}t$ thanks to Assumption 1 we can rewrite

$$F(\theta)=2\int_0^{t_\theta}(U(0,t)-U(\theta,t))\mathrm{d}t=2\int_{t_\theta}^{+\infty}(U(\theta,t)-U(0,t))\mathrm{d}t.$$

When we take the derivative (thanks to the integrability Assumption 3 we can differentiate under the $\int$ sign), using that the integrand vanishes exactly at t_θ, we find three different expressions of the derivative:

$$F'(\theta)=-\int_0^{t_\theta}D(\theta,t)\mathrm{d}t+\int_{t_\theta}^{+\infty}D(\theta,t)\mathrm{d}t=-2\int_0^{t_\theta}D(\theta,t)\mathrm{d}t=2\int_{t_\theta}^{+\infty}D(\theta,t)\mathrm{d}t.$$

Recall that s_θ is the point where $D(\theta,t)$ vanishes. By Assumption 4, we know that $D(\theta,t)$ is negative on $(0,s_\theta)$ and positive on $(s_\theta,+\infty)$. If $s_\theta\le t_\theta$, using the last expression for F' yields $F'(\theta)>0$. On the other hand, if $t_\theta\le s_\theta$, then it is by using the second expression that we find $F'(\theta)>0$. This concludes the proof that F is strictly increasing.

Eventually the first expression of the derivative yields our bound (19) when we take the absolute value with the help of the triangle equality.

For the equality in (19) when $\theta=0$, we first justify that t_θ converges to s_0 as $\theta\to 0$. Fix $t_1<s_0<t_2$. As $D(0,\cdot)$ changes sign around s_0, then $U(\theta,t_1)-U(0,t_1)\sim\theta D(0,t_1)$ and $U(\theta,t_2)-U(0,t_2)\sim\theta D(0,t_2)$ take different signs for θ small enough. By continuity of $U(\theta,\cdot)$ we deduce that $t_1<t_\theta<t_2$ if θ is small enough. This is enough to show that t_θ converges to s_0 as $\theta\to 0$. Thus, taking the limit in the first expression of F' we find

$$\lim_{\theta\to 0}F'(\theta)=-\int_0^{s_0}D(0,t)\mathrm{d}t+\int_{s_0}^{+\infty}D(0,t)\mathrm{d}t=\int_0^{+\infty}|D(0,t)|\mathrm{d}t$$

as $D(0,\cdot)$ is negative on $(0,s_0)$ and positive on $(s_0,+\infty)$. This is enough to justify that $F'(0)$ exists and is equal to the value on the right hand side. □

Proposition 25 is at the core of the proof of Proposition 8. Before moving to its proof, we state and prove a technical lemma.

LEMMA 26. *For any fixed $\alpha>1$, the equation $\Gamma(0,t)=\alpha\Gamma(0,\alpha t)$ has exactly two solutions, $t=0$ and $t_\alpha>0$. Moreover, denoting $\bar t$ the point of maximum of the function $t\mapsto t\Gamma(0,t)$, there holds*

$$\frac{\bar t}{\alpha}\le t_\alpha\le\bar t.$$

PROOF. Let us call $f:t\mapsto t\Gamma(0,t)$, so that the equation can be rewritten as $f(t)=f(\alpha t)$. Standard analysis yields that f is strictly increasing between 0 and $\bar t$, and strictly decreasing after. Quite easily, this yields that $f(t)<f(\alpha t)$ when $0<t\le\bar t/\alpha$, while $f(t)>f(\alpha t)$ when $t\ge\bar t$. Eventually, on $(\bar t/\alpha,\bar t)$ the function $f(\alpha t)-f(t)$ is strictly decreasing and continuous, and thus it vanishes at exactly one point. □

PROOF OF PROPOSITION 8. First of all we observe that by a simple change of variable $\mathcal{J}(\alpha_1, \alpha_2) = \mathcal{J}(1, \alpha_2/\alpha_1)$. Thus, by a slight abuse of notations, we just need to study $\mathcal{J}(\alpha) = \mathcal{J}(1, \alpha)$. Moreover, by symmetry between α_1 and α_2, we can restrict to the case $\alpha_2 \geq \alpha_1$, that is, $\alpha \geq 1$. The reader can easily check that Proposition 8 is nothing else than the result below. □

PROPOSITION. *Let $\mathcal{J} : [1, +\infty) \to \mathbb{R}$ defined by*

$$\mathcal{J}(\alpha) = \int_0^{+\infty} |\Gamma(0,t) - \alpha\Gamma(0,\alpha t)| \mathrm{d}t.$$

The function $\mathcal{J}$ is increasing and

$$\mathcal{J}(\alpha) \leq C \log(\alpha),$$

for $C = \int_0^{+\infty} |\Gamma(0,t) - e^{-t}| \mathrm{d}t$. Moreover the asymptotic expansion of $\mathcal{J}$ for α close to 1 is

$$\mathcal{J}(\alpha) = C(\alpha - 1) - \frac{C}{2}(\alpha - 1)^2 + o((\alpha - 1)^2).$$

PROOF. We will use Proposition 25 with $\theta = \alpha - 1$ and $U(\theta, t) = \alpha\Gamma(0, \alpha t)$.

By Lemma 26, $\Gamma(0,t) - \alpha\Gamma(0,\alpha t)$ is negative on $(0, t_\alpha)$ and positive on $(t_\alpha, +\infty)$, where t_α is the strictly positive solution of the equation $\Gamma(0,t) = \alpha\Gamma(0,\alpha t)$.

Moreover,

$$\frac{\partial}{\partial\alpha}(\alpha\Gamma(0,\alpha t)) = \Gamma(0,\alpha t) - \exp(-\alpha t),$$

and one can check that this function also vanishes at a single point, is negative after that point and positive before. For the integrability Assumption 3 note that for any $\alpha \geq 1$

$$|\Gamma(0,\alpha t) - \exp(-\alpha t)| \leq \Gamma(0,t) + \exp(-t)$$

and the right hand side integrates to $2 < +\infty$ as can be checked with Fubini's theorem.

Thus by Proposition 25 we directly see that $\mathcal{J}$ is increasing with

$$\mathcal{J}'(\alpha) \leq \int_0^{+\infty} |\Gamma(0,\alpha t) - \exp(-\alpha t)| \mathrm{d}t = \frac{1}{\alpha} \int_0^{+\infty} |\Gamma(0,t) - e^{-t}| \mathrm{d}t.$$

Integrating this inequality between 1 and α yields the upper bound.

As for the asymptotic expansion, it is enough to prove that $C\log(\alpha) - \mathcal{J} \leq o((\alpha-1)^2)$ and use the Taylor expansion of the logarithm. As in the proof of Proposition 25, let s_α the point where $\frac{\partial}{\partial\alpha}(\alpha\Gamma(0,\alpha t))$ vanishes. It clearly satisfies $s_\alpha = \bar{t}/\alpha$, where $\bar{t}$ is as in Lemma 26. The latter guarantees that $s_\alpha \leq t_\alpha$. Then,

$$\begin{aligned}
&\int_0^{+\infty} |\Gamma(0,\alpha t) - \exp(-\alpha t)| \mathrm{d}t - \mathcal{J}'(\alpha) \\
&= \int_0^{s_\alpha} (\Gamma(0,\alpha t) - \exp(-\alpha t)) \mathrm{d}t - \int_{s_\alpha}^{+\infty} (\Gamma(0,\alpha t) - \exp(-\alpha t)) \mathrm{d}t \\
&\quad - \int_0^{t_\alpha} (\Gamma(0,\alpha t) - \exp(-\alpha t)) \mathrm{d}t + \int_{t_\alpha}^{+\infty} (\Gamma(0,\alpha t) - \exp(-\alpha t)) \mathrm{d}t \\
&= 2 \int_{s_\alpha}^{t_\alpha} |\Gamma(0,\alpha t) - \exp(-\alpha t)| \mathrm{d}t = \frac{1}{\alpha} \int_{\alpha s_\alpha}^{\alpha t_\alpha} |\Gamma(0,t) - e^{-t}| \mathrm{d}t \\
&\leq \frac{1}{\alpha} \int_{\bar{t}}^{\alpha\bar{t}} |\Gamma(0,t) - e^{-t}| \mathrm{d}t,
\end{aligned}$$

where for the last inequality we used $\alpha s_\alpha = \bar{t}$ and $\alpha t_\alpha \leq \alpha \bar{t}$, thanks to Lemma 26. As $\bar{t}$ is a point where $\Gamma(0,t) - e^{-t}$ vanishes, for a constant $\bar{C}$ and at least for t close to $\bar{t}$, we have

$$|\Gamma(0,t) - e^{-t}| \leq \bar{C}|t - \bar{t}|.$$

By integrating this inequality we find that, for α close to 1,

$$\int_0^{+\infty} |\Gamma(0,\alpha t) - \exp(-\alpha t)| \mathrm{d}t - \mathcal{J}'(\alpha) \leq \frac{\bar{C}\bar{t}^2(\alpha-1)^2}{\alpha}.$$

By integrating now in α, we can estimate the difference between $\mathcal{J}$ and its upper bound by

$$C\log(\alpha) - \mathcal{J}(\alpha) \leq \int_1^\alpha \frac{\bar{C}\bar{t}^2(\alpha-1)^2}{\alpha} \mathrm{d}\alpha \leq \bar{C}\frac{\bar{t}^2}{6}(\alpha-1)^3,$$

which yields our result. □

PROOF OF PROPOSITION 9. We study the function

$$\mathcal{J}(\sigma) := \int_0^{+\infty} \left| \Gamma(0,t) - \frac{\Gamma(-\sigma,t)}{\Gamma(1-\sigma)} \right| \mathrm{d}t$$

by using Proposition 25 with $\theta = \sigma$ and $U(\theta,t) = \Gamma(-\sigma,t)/\Gamma(1-\sigma)$.

We first claim that there exists a unique point t_σ where the integrand in $\mathcal{J}$ vanishes, and that the integrand is positive before t_σ and negative after. Indeed, let us introduce the integrand

$$f_\sigma(t) = \Gamma(0,t) - \frac{\Gamma(-\sigma,t)}{\Gamma(1-\sigma)}.$$

It is easy to check that $\lim_{t\to 0} f_\sigma(t) = -\infty$ while $\lim_{t\to+\infty} f_\sigma(t) = 0$. Moreover, since

$$f'_\sigma(t) = -\frac{e^{-t}}{t} + \frac{e^{-t}}{t^{1+\sigma}\Gamma(1-\sigma)},$$

we see that f_σ is first increasing and then decreasing. Thus it admits a unique zero, and it is positive before its zero and negative after.

Then we study the derivative by introducing

$$D(\sigma,t) = \frac{\partial}{\partial\sigma}\left[\frac{\Gamma(-\sigma,t)}{\Gamma(1-\sigma)}\right].$$

We recall that the derivative of the Gamma incomplete function satisfies

$$\frac{\partial\Gamma(z,t)}{\partial z} = -\int_t^{+\infty} \log(x)\, x^{1-z} e^{-x}\, \mathrm{d}x$$

A straightforward computation then leads to

$$D(\sigma,t) = \frac{1}{\Gamma(1-\sigma)}\left\{\frac{\Gamma'(1-\sigma)}{\Gamma(1-\sigma)}\Gamma(-\sigma,t) - \int_t^{+\infty} \log(s)\frac{e^{-s}}{s^{1+\sigma}}\,\mathrm{d}s\right\}.$$

From this expression we see that

$$\lim_{t\to 0} D(\sigma,t) = +\infty, \qquad \lim_{t\to+\infty} D(\sigma,t) = 0.$$

Moreover, $D(\sigma,\cdot)$ is decreasing until $a_\sigma = \exp(\Gamma'(1-\sigma)/\Gamma(1-\sigma)) < 1$ and increasing after. Thus it vanishes at a single point $s_\sigma \in (0,a_\sigma)$, it is positive on $(0,s_\sigma)$ and negative after.

We also need to check the integrability Assumption 3. Fix $\Sigma \in (0,1)$. Notice first that $\Gamma(1-\sigma)$ and $\Gamma'(1-\sigma)$ are uniformly bounded for $\sigma \in [0,\Sigma]$. Up to a constant C which depends only on Σ we deduce that for $\sigma \in [0,\Sigma]$,

$$|D(\sigma,t)| \leq C\left(|\Gamma(-\sigma,t)| + \int_t^{+\infty} |\log(s)| \frac{e^{-s}}{s^{1+\sigma}}\,\mathrm{d}s\right) \leq C\int_t^{+\infty} (1+|\log(s)|)\frac{e^{-s}}{\min(s,s^{1+\Sigma})}\,\mathrm{d}s.$$

We integrate the upper in t and find, with the help of Fubini's theorem:

$$\begin{aligned}\int_0^{+\infty} \sup_{\sigma\in[0,\Sigma]} |D(t,\sigma)|\,\mathrm{d}t &\leq C\int_0^{+\infty}\left(\int_t^{+\infty}(1+|\log(s)|)\frac{e^{-s}}{\min\left(s,s^{1+\Sigma}\right)}\,\mathrm{d}s\right)\mathrm{d}t \\ &= C\int_0^{+\infty}(1+|\log(s)|)\max\left(1,s^{-\Sigma}\right)e^{-s}\,\mathrm{d}s < +\infty.\end{aligned}$$

Eventually we can conclude by Proposition 25 that $\mathcal{J}$ is increasing in σ and

$$\mathcal{J}'(0) = \int_0^{+\infty} |D(0,t)|\mathrm{d}t = \int_0^{+\infty}\left|\Gamma'(1)\Gamma(0,t) - \int_t^{+\infty}\log(s)\frac{e^{-s}}{s}\,\mathrm{d}s\right|\mathrm{d}t.$$

Noting that $\Gamma'(1) = -\gamma$ is the Euler–Mascheroni constant, we get our result. □

5.3. *Proofs of Section 4.1.* This section is devoted to the proof of Theorem 11.

PROOF OF THEOREM 11. We follows the same line of proof as Theorem 2, but with additional steps to handle the new layers of randomness.

We write $\mathrm{d}\tilde{\nu}^i(s,x) = \mathrm{d}\tilde{\rho}^i_x(s)\mathrm{d}\tilde{P}^i_0(x)$ for the random Lévy intensity of $\tilde{\mu}^i$, as in Proposition 1. We define the real-valued random variable $\tilde{q}^i = \mathbb{E}(\tilde{\mu}^i(\mathbb{X})|\tilde{\nu}^i)$, in such a way that

$$\int_0^{+\infty} s\,\mathrm{d}\tilde{\rho}^i_x(s) = \tilde{q}^i,$$

as written in Proposition 1. We recall that $S_k : (0,+\infty) \to (0,+\infty)$ is defined by $S_k(s) = s/k$, it is the scaling by $1/k$.

We only need to prove that the Lévy intensities of the scaled Cox CRM are the same, that is, given Lemma 22, we want to prove the following equality in distribution of random measures:

$$\mathrm{d}(S_{\tilde{q}^1}\#\tilde{\rho}^1_x)(s)\mathrm{d}\tilde{P}^1_0(x) \overset{\mathrm{d}}{=} \mathrm{d}(S_{\tilde{q}^2}\#\tilde{\rho}^2_x)(s)\mathrm{d}\tilde{P}^2_0(x).$$

Preliminary step: a note on null-sets. By the infinite activity assumption and the assumption that all measures $\tilde{P}^i_0$ are equivalent to a fixed measure Q^i it follows that $\tilde{\mu}^i(A) = 0$ a.s. if and only if $Q^i(A) = 0$. In particular $\tilde{\mu}^i(A)/\tilde{\mu}^i(\mathbb{X}) = 0$ a.s. if and only if $Q^i(A) = 0$. Since $\tilde{\mu}^1/\tilde{\mu}^1(\mathbb{X}) \overset{\mathrm{d}}{=} \tilde{\mu}^2/\tilde{\mu}^2(\mathbb{X})$, we conclude that Q^1, Q^2 are equivalent, that is, they have the same null set. Now we call $Q = Q^1$ and we know that the following alternative holds: either $Q(A) > 0$, and then $0 < \tilde{\mu}^i(A) < +\infty$ a.s., for $i = 1,2$; or $Q(A) = 0$, and in such a case $\tilde{\mu}^i(A) = 0$ a.s., for $i = 1,2$.

First step: starting as in Theorem 2. Let $A \subseteq \mathbb{X}$ a subset with $Q(A) > 0$, it implies $\tilde{P}^i_0(A) > 0$ a.s. for $i = 1,2$ and in particular $0 < \tilde{\mu}^i(A) < +\infty$ a.s. for $i = 1,2$. We define the randoms measures $\tilde{\xi}^i$ on A^c as

$$\tilde{\xi}^i(\cdot) = \frac{\tilde{\mu}(\cdot)}{\tilde{\mu}^i(A)}.$$

As in the proof of Theorem 2, we know that $\tilde{\xi}^1(B) = \tilde{\xi}^2(B)$ for any set $B \subseteq A^c$. Moreover, conditionally to $(\tilde{\nu}^i, \tilde{\mu}^i(A))$, the random measure $\tilde{\xi}^i$ is a Cox CRM, and its (random) Lévy intensity is $\mathrm{d}(S_{\tilde{\mu}^i(A)}\#\tilde{\rho}^i_x)(s)\,\mathrm{d}\tilde{P}^i_0(x)$. As in the proof of Theorem 2, invoking [25, Theorem 3.3], we have equality in law of the random Lévy intensities, that is,

$$\mathrm{d}(S_{\tilde{\mu}^1(A)}\#\tilde{\rho}^1_x)(s)\,\mathrm{d}\tilde{P}^1_0(x) \overset{\mathrm{d}}{=} \mathrm{d}(S_{\tilde{\mu}^2(A)}\#\tilde{\rho}^2_x)(s)\,\mathrm{d}\tilde{P}^2_0(x).$$

We take this equality and integrate it over $(0,+\infty)\times B$ where $B \subseteq A^c$. We use the normalization of $\tilde{\rho}^i_x$ recalled above and the definition of the pushforward measure to conclude

$$\frac{\tilde{q}^1\tilde{P}^1_0(B)}{\tilde{\mu}^1(A)} \overset{\mathrm{d}}{=} \frac{\tilde{q}^2\tilde{P}^2_0(B)}{\tilde{\mu}^2(A)}.$$

Note that this equality is valid for every sets A, B with $Q(A) > 0$ and $A \cap B = \emptyset$.

Second step: doing the same reasoning on a new Cox CRM. Fix a set B with $Q(B) > 0$, it guarantees that $\tilde{P}^i_0(B) > 0$ a.s. for $i = 1, 2$. We define the random measures $\tilde{\zeta}^i$ on B^c

$$\tilde{\zeta}^i(\cdot) = \frac{\tilde{\mu}^i(\cdot)}{\tilde{q}^i\tilde{P}^i_0(B)}.$$

Conditionally to $\tilde{\nu}^i$ these are CRM, thus they are Cox CRM and their random Lévy intensity is $\mathrm{d}(S_{\tilde{q}^i\tilde{P}^i_0(B)}\#\tilde{\rho}^i_x)(s)\,\mathrm{d}\tilde{P}^i_0(x)$. Moreover, the previous step showed that $\tilde{\zeta}^1(A) \overset{\mathrm{d}}{=} \tilde{\zeta}^2(A)$ for any set $A \subseteq B^c$ with $Q(A) > 0$. This equality in distribution easily extends to the case $Q(A) = 0$, as it implies $\tilde{\mu}^i(A) = 0$ a.s. for $i = 1, 2$. Thus, invoking again [25, Theorem 3.3]:

$$\mathrm{d}(S_{\tilde{q}^1\tilde{P}^1_0(B)}\#\tilde{\rho}^1_x)(s)\,\mathrm{d}\tilde{P}^1_0(x) \overset{\mathrm{d}}{=} \mathrm{d}(S_{\tilde{q}^2\tilde{P}^2_0(B)}\#\tilde{\rho}^2_x)(s)\,\mathrm{d}\tilde{P}^2_0(x). \tag{20}$$

Third step: deducing equality in law of the scaled random Lévy intensities. We take the previous equality in distribution, multiply by s, and integrate it over $(0,+\infty)\times B^c$. Using again the normalization of $\tilde{\rho}^i_x$, we know conclude

$$\frac{\tilde{P}^1_0(B^c)}{\tilde{P}^1_0(B)} \overset{\mathrm{d}}{=} \frac{\tilde{P}^2_0(B^c)}{\tilde{P}^2_0(B)}.$$

As $\tilde{P}^i_0(B) + \tilde{P}^i_0(B^c) = 1$, we deduce that $\tilde{P}^1_0(B) \overset{\mathrm{d}}{=} \tilde{P}^2_0(B)$, at least if $Q(B) > 0$. We actually deduce more, as we can say that, for $i = 1, 2$,

$$\tilde{P}^i_0(B) \quad \text{is a deterministic function of} \quad \mathrm{d}(S_{\tilde{q}^i\tilde{P}^i_0(B)}\#\tilde{\rho}^i_x)(s)\,\mathrm{d}\tilde{P}^i_0(x).$$

Note that $\mathrm{d}(S_{\tilde{q}^i}\#\tilde{\rho}^i_x)(s)\mathrm{d}\tilde{P}^i_0(x)$ is obtained from $\mathrm{d}(S_{\tilde{q}^i\tilde{P}^i_0(B)}\#\tilde{\rho}^i_x)(s)\mathrm{d}\tilde{P}^i_0(x)$ from a scaling of the s variable by $\tilde{P}^i_0(B)$. Thus

$$\mathrm{d}(S_{\tilde{q}^i}\#\tilde{\rho}^i_x)(s)\mathrm{d}\tilde{P}^i_0(x) \quad \text{is a deterministic function of} \quad \mathrm{d}(S_{\tilde{q}^i\tilde{P}^i_0(B)}\#\tilde{\rho}^i_x)(s)\,\mathrm{d}\tilde{P}^i_0(x).$$

Combined with (20) we deduce

$$\mathrm{d}(S_{\tilde{q}^1}\#\tilde{\rho}^1_x)(s)\mathrm{d}\tilde{P}^1_0(x) \overset{\mathrm{d}}{=} \mathrm{d}(S_{\tilde{q}^2}\#\tilde{\rho}^2_x)(s)\mathrm{d}\tilde{P}^2_0(x)$$

at least on $(0,+\infty)\times B^c$, where B is a set satisfying $Q(B) > 0$. To conclude the proof we only need to prove that equality in distribution is valid on $(0,+\infty)\times\mathbb{X}$, that is, take $B = \emptyset$.

Fourth step: removing the assumption $Q(B) > 0$. Recall that we assumed that Q is not purely atomic. By a theorem of Sierpinski we can find a non-increasing sequence of sets

(B_n) with $Q(B_n) > 0$ but $Q(\cap_n B_n) = 0$. The measures $\mathrm{d}(S_{\tilde{q}^i} \# \tilde{\rho}^i_x)(s)\mathrm{d}\tilde{P}^i_0(x)$ have the same law for $i = 1, 2$ on $(0, +\infty) \times B_n^c$ for any n. By a monotone convergence argument, it implies that it is also the case on $(0, +\infty) \times (\cup_n B_n^c)$. Eventually, as $\mathrm{d}(S_{\tilde{q}^i} \# \tilde{\rho}^i_x)(s)\mathrm{d}\tilde{P}^i_0(x)$ gives a.s. measure zero to $(0, +\infty) \times \cap_n B_n$ for $i = 1, 2$ we conclude that the law is the same on the whole space $(0, +\infty) \times \mathbb{X}$ and we are done. □

5.4. *Proofs of Section 4.2.* In this section we prove Lemma 12 and Proposition 14, together with an auxiliary results, Lemma 27, which will also be used in the next section.

Informally, we recall that a random variable is infinitely divisible if its probability distribution can be expressed of the sum of an arbitrary number of i.i.d. random variables. One can prove that for every measure ρ on $(0, +\infty)$ such that $\int (s^2 \wedge 1)\mathrm{d}\rho(s) < +\infty$, a random variable J such that

$$\log\left(\mathbb{E}\left(e^{-\lambda J}\right)\right) = -\int (1 - e^{-\lambda s})\,\mathrm{d}\rho(s).$$

is infinitely divisible. We call J an infinitely divisible with Lévy measure ρ and refer to [41] for a deep account on infinite divisibility.

LEMMA 27. *Let $\tilde{\mu} \sim \mathrm{CRM}(\nu)$ on a Polish space $\mathbb{X}$ and let J be an infinitely divisible random variable on $(0, +\infty)$ independent of $\tilde{\mu}$ with Lévy measure ρ. Then for every $y \in \mathbb{X}$, $\tilde{\mu} + J\delta_y \sim \mathrm{CRM}(\nu + \rho \otimes \delta_y)$.*

PROOF. For any measurable function f on $\mathbb{X}$, since J and $\tilde{\mu}$ are independent,

$$\begin{aligned}
\mathbb{E}\left(\exp\left(-\int f(x)\mathrm{d}(\tilde{\mu} + J\delta_y)(x)\right)\right) &= \mathbb{E}\left(\exp\left(-\int f(x)\mathrm{d}\tilde{\mu}(x)\right)\right)\mathbb{E}\left(\exp\left(-f(y)J\right)\right) \\
&= \exp\left(-\int_0^\infty \int_{\mathbb{X}} (1 - e^{-sf(x)})\mathrm{d}\nu(s,x)\right)\exp\left(-\int_0^\infty (1 - e^{-sf(y)})\mathrm{d}\rho(s)\right) \\
&= \exp\left(-\int_0^\infty \int_{\mathbb{X}} (1 - e^{-sf(x)})\mathrm{d}(\nu(s,x) + \rho(s)\delta_y(x))\right).
\end{aligned}$$

Since CRMs are characterized by their Laplace functional this proves the result. □

PROOF OF LEMMA 12. By specializing Theorem 10 to a gamma CRM, we know that there exists a latent random variable U such that

$$\tilde{\mu}^* | U \overset{\mathrm{d}}{=} \tilde{\mu}_U + \sum_{i=1}^k J_i^U \delta_{X_i^*},$$

where $\tilde{\mu}_U$ is a gamma CRM with Lévy intensity $s^{-1}e^{-(U+1)s}\alpha P_0$ and the jumps $J_i^U \sim \mathrm{gamma}(n_i, U + 1)$, where $n_i = \#\{j : X_j = X_i^*\}$ is the number of observations equal to X_i^* and everything is independent. Since the gamma distribution is infinitely divisible, thanks to Lemma 27 this implies that $\tilde{\mu}^*|U$ is a gamma CRM with canonical Lévy intensity

$$\mathrm{d}\nu_U(s,x) = (\alpha + n)\frac{e^{-(U+1)s}}{s}\mathrm{d}s\left(\frac{\alpha}{\alpha + n}\mathrm{d}P_0(x) + \frac{1}{\alpha + n}\sum_{i=1}^n \delta_{X_i}(x)\right).$$

Thus $\tilde{\mu}^*$ is a Cox CRM that is conditionally a gamma CRM. By Example 1 its corresponding Cox scaled CRM $\tilde{\mu}^*_{\mathcal{S}}$ has Lévy intensity

$$\mathrm{d}\tilde{\nu}^i_{\mathcal{S}}(s,x)|U = (\alpha + n)\frac{e^{(\alpha+n)s}}{s}\left(\frac{\alpha}{\alpha + n}\mathrm{d}P_0(x) + \frac{1}{\alpha + n}\sum_{i=1}^n \delta_{X_i}(x)\right).$$

In particular, we observe that by scaling the conditional Lévy intensity all the randomness disappears and we are left with a deterministic Lévy intensity. □

PROOF OF PROPOSITION 14. To prove 1., 2. and 3. we can use Proposition 8 on the comparison between Gamma CRMs, since $\mathcal{J} = \mathcal{J}(\alpha_1, \alpha_2, n)$ coincides with $\mathcal{J}(\alpha_1 + n, \alpha_2 + n)$ studied in Proposition 8. Note that we use only the leading term in the asymptotic expansion of $\mathcal{J}$ to simplify the exposition.

The proof of 4. relies on the behavior of the Wasserstein with respect to convex combination, which we recalled in (16) in Section 5. Let $\lambda_i = \alpha_i(\alpha_i + n)^{-1}$, so that $0 < \lambda_1 < \lambda_2 < 1$ for every n. Then by using (16) twice,

$$\begin{aligned}\mathcal{A} &\leq \lambda_1 \mathcal{W}_1(P_0^1, P_0^2) + (1-\lambda_1)\mathcal{W}_1\left(\frac{1}{n}\sum_{i=1}^n X_i, \frac{\lambda_2 - \lambda_1}{1-\lambda_1}P_0^2 + \frac{1-\lambda_2}{1-\lambda_1}\frac{1}{n}\sum_{i=1}^n X_i\right)\\ &\leq \lambda_1 \mathcal{W}_1(P_0^1, P_0^2) + (\lambda_2 - \lambda_1)\mathcal{W}_1\left(\frac{1}{n}\sum_{i=1}^n X_i, P_0^2\right).\end{aligned}$$

The proof of the upper bound follows by substituting the values of λ_1 and λ_2. Moreover, if $\alpha_1 = \alpha_2$ or $P_0^1 = P_0^2$, we can reduce it to an equality by using (17) in Section 5. □

5.5. *Proofs of Section 4.3.* In this section we prove Proposition 16, Theorem 18, Theorem 19 and Proposition 20, together with the auxiliary Lemma 28, which is both relevant for our analysis and of independent interest.

PROOF OF PROPOSITION 16. The proof follows the same lines as Lemma 12. By specializing the formula for the jumps J_i^U in Theorem 10 we see that they follow a gamma distribution. Since the gamma distribution is infinitely divisible, thanks to Lemma 27 $\tilde{\mu}^*|U$ is characterized by the Lévy intensity

$$\nu_U^*(s,x) = \frac{\alpha}{\Gamma(1-\sigma)}\frac{e^{-(U+1)s}}{s^{1+\sigma}}\mathrm{d}s\,\mathrm{d}P_0(x) + \frac{e^{-(U+1)s}}{s}\sum_{i=1}^k (n_i - \sigma)\,\mathrm{d}s\delta_{X_i^*}(x).$$

With simple calculations we derive

$$\mathbb{E}(\tilde{\mu}^*(\mathbb{X})|U) = \frac{\alpha(U+1)^\sigma + n - k\sigma}{U+1}.$$

Denote $c = \alpha(U+1)^\sigma + n - k\sigma$. Then by Lemma 4 the corresponding Cox scaled CRM $\tilde{\mu}_{\mathcal{S}}^*$ has random Lévy intensity

$$\tilde{\nu}_{\mathcal{S}}^*(\mathrm{d}s, \mathrm{d}x) = \frac{\alpha}{\Gamma(1-\sigma)}\left(\frac{U+1}{c}\right)^\sigma \frac{e^{-cs}}{s^{1+\sigma}}\mathrm{d}s\,\mathrm{d}P_0(x) + \frac{e^{-cs}}{s}\sum_{i=1}^k (n_i - \sigma)\,\mathrm{d}s\delta_{X_i^*}(x).$$

We now find its canonical expression. Starting from the previous formula, by integrating out $f(s) = s$ we find the mean measure of $\tilde{\mu}_{\mathcal{S}}^*$ to be

$$P_0^* = \frac{\alpha(U+1)^\sigma}{c}P_0 + \frac{1}{c}\sum_{i=1}^k (n_i - \sigma)\delta_{X_i^*}.$$

In general it seems not trivial to find the canonical decomposition of a sum of homogeneous CRMs, which overall is a non-homogeneous CRM. In this case, however, we can use the fact

that P_0 is nonatomic, so that every component of the sum has different support. Thus, the disintegration of $\tilde{\nu}^*_S$ with respect to P^*_0 is

$$\frac{c^{1-\sigma}}{\Gamma(1-\sigma)}\frac{e^{-cs}}{s^{1+\sigma}}\mathbb{1}_{\mathbb{X}\setminus\{X^*_1,\dots,X^*_k\}}(x)\,\mathrm{d}s + c\frac{e^{-cs}}{s}\mathbb{1}_{\{X^*_1,\dots,X^*_k\}}(x)\,\mathrm{d}s.$$

Moreover, the expression of the distribution of U can be easily derived by specializing the formula above James, Lijoi and Prünster [24, Theorem 1], which we do not report here for brevity. □

The proof of Theorem 18 relies on the following lemma.

LEMMA 28. *Let $(Y_n)_{n\in\mathbb{N}}$ a sequence of random variables on $I=(a,+\infty)$ with densities proportional to $\exp(-f_n)$, for some some sequence $(f_n)_{n\in\mathbb{N}}$ of functions from $I\to\mathbb{R}$. We assume:*

1. *Each f_n is smooth, convex, and minimized at a single point $r_n\in I$, with the sequence $(r_n)_{n\in\mathbb{N}}$ diverging to $+\infty$.*
2. *We can find $0<\alpha<\beta<+\infty$ and $\gamma>1/2$ such that, for every n large enough, for $x\in[r_n-r_n^\gamma, r_n+r_n^\gamma]$, $\alpha\le r_n f_n''(x)\le\beta$.*

Then, for the topology of L^1 convergence there holds

$$\lim_{n\to+\infty}\frac{Y_n}{r_n}=1.$$

REMARK 5. The second condition quantifies $f_n''\asymp 1/r_n$ on a large enough neighborhood of r_n. This inequality, together with the convexity of f_n, enables to show that Y_n indeed concentrates around r_n. Such condition will be satisfied in our case, and the scaling $f_n''(r_n)\asymp 1/r_n$ comes from the logarithms in the expression of f_n.

PROOF. Without loss of generality, we can assume $\gamma<1$ so that $\gamma\in(1/2,1)$.

First step: simplifying the setting. We need to prove that

$$A_n=\frac{1}{r_n}\left(\int_I|x-r_n|\exp(-f_n(x))\,\mathrm{d}x\right)\left(\int_I\exp(-f_n(x))\,\mathrm{d}x\right)^{-1}$$

converges to 0 as $n\to+\infty$. We first shift the function f_n by defining $g_n(x)=f_n(x-r_n)-f_n(x_n)$, which is still convex but now minimized at $g_n(0)=0$. The bound bound in 2. now reads, for $x\in[-r_n^\gamma,r_n^\gamma]$,

$$\alpha\le r_n g_n''(x)\le\beta.$$

Moreover g_n is defined on $I_n=(a-r_n,+\infty)$. Thus we can rewrite

$$A_n=\frac{1}{r_n}\left(\int_{I_n}|x|\exp(-g_n(x))\,\mathrm{d}x\right)\left(\int_{I_n}\exp(-g_n(x))\,\mathrm{d}x\right)^{-1}$$

We set $J_n=[-r_n^\gamma,r_n^\gamma]$ and introduce

$$B_n=\int_{J_n}\exp(-g_n(x))\,\mathrm{d}x,\qquad C_n=\int_{I_n\setminus J_n}|x|\exp(-g_n(x))\,\mathrm{d}x.$$

We see clearly that $\int_{J_n} |x| \exp(-g_n(x))\,\mathrm{d}x \leq r_n^\gamma B_n$. Thus, bounding from below the denominator of A_n by B_n,

$$\begin{aligned} A_n &\leq \frac{\int_{I_n} |x| \exp(-g_n(x))\,\mathrm{d}x}{r_n B_n} = \frac{\int_{J_n} |x| \exp(-g_n(x))\,\mathrm{d}x + \int_{I_n \setminus J_n} |x| \exp(-g_n(x))\,\mathrm{d}x}{r_n B_n} \\ &\leq \frac{r_n^\gamma}{r_n} + \frac{C_n}{r_n B_n}. \end{aligned}$$

The first term converges to 0 because $\gamma < 1$, while we have more work to do to prove the convergence to zero of the second term.

Second step. Lower bound on B_n. On $J_n = [-r_n^\gamma, r_n^\gamma]$, using $g_n(0) = 0$ together with the bound $g_n'' \leq \beta / r_n$, we find

$$g_n(x) \leq \frac{\beta}{2r_n} x^2.$$

Plugging this back in the expression for B_n,

$$B_n \geq \int_{-r_n^\gamma}^{r_n^\gamma} \exp\left(-\frac{\beta}{2r_n} x^2\right) \mathrm{d}x = \sqrt{\frac{2\pi r_n}{\beta}} \operatorname{erf}\left(\sqrt{\frac{\beta}{2}} r_n^{\gamma - 1/2}\right),$$

being erf the error function. Sending n to $+\infty$, as $r_n^{\gamma-1/2} \to +\infty$, we see that $B_n \gtrsim \sqrt{r_n}$, and precisely that

$$\liminf_{n \to +\infty} \frac{B_n}{\sqrt{r_n}} \geq \sqrt{\frac{2\pi}{\beta}}.$$

Third step. Upper bound on C_n. We need to analyze C_n. We will focus on the integral over $[r_n^\gamma, +\infty)$, as the other case is symmetric. We first notice that, again thanks to the bounds on g_n'',

$$g_n'(r_n^\gamma) \geq \alpha r_n^{\gamma - 1}, \qquad g_n(r_n^\gamma) \geq \frac{\alpha}{2} r_n^{2\gamma - 1}.$$

As the function g_n is convex, thus above its tangents, we deduce that for $x \geq r_n^\gamma$,

$$g_n(x) \geq \frac{\alpha}{2} r_n^{2\gamma-1} + \alpha r_n^{\gamma-1}(x - r_n^\gamma).$$

We plug this bound in the integral we want to compute to obtain, using the change of variables $y = \alpha r_n^{\gamma-1}(x - r_n^\gamma)$,

$$\begin{aligned} \int_{r_n^\gamma}^{+\infty} x \exp(-g_n(x))\,\mathrm{d}x &\leq \int_{r_n^\gamma}^{+\infty} x \exp\left(-\frac{\alpha}{2} r_n^{2\gamma-1} - \alpha r_n^{\gamma-1}(x - r_n^\gamma)\right) \mathrm{d}x \\ &= \exp\left(-\frac{\alpha}{2} r_n^{2\gamma-1}\right) \int_{r_n^\gamma}^{+\infty} x \exp\left(-\alpha r_n^{\gamma-1}(x - r_n^\gamma)\right) \mathrm{d}x \\ &= \frac{\exp\left(-\frac{\alpha}{2} r_n^{2\gamma-1}\right)}{\alpha r_n^{\gamma-1}} \int_0^{+\infty} \left(\frac{y}{\alpha r_n^{\gamma-1}} + r_n^\gamma\right) \exp(-y)\,\mathrm{d}y \\ &= \frac{\exp\left(-\frac{\alpha}{2} r_n^{2\gamma-1}\right)}{\alpha r_n^{\gamma-1}} \left(\frac{1}{\alpha r_n^{\gamma-1}} + r_n^\gamma\right). \end{aligned}$$

By symmetry, C_n is bounded by twice the quantity above.

Fourth step. Putting all estimates together. Plugging this back together we obtain

$$A_n \lesssim r_n^{\gamma-1} + \frac{1}{r_n^{3/2}} \frac{\exp\left(-\frac{\alpha}{2} r_n^{2\gamma-1}\right)}{\alpha r_n^{\gamma-1}} \left(\frac{1}{\alpha r_n^{\gamma-1}} + r_n^{\gamma}\right).$$

Sending $n \to +\infty$, the first term in the sum converges to 0, as well as the second one as the exponential factor $\exp\left(-\frac{\alpha}{2} r_n^{2\gamma-1}\right)$ dominates everything as $2\gamma - 1 > 0$. □

PROOF OF THEOREM 18. Let $h(U) = (1+U)^{\sigma}$. With a change of variable we obtain that the density of $h(U)$ is proportional to $\exp(-f_n(x))$, with

$$f_n(x) = -(k_n - 1)\log(x) + \frac{\alpha}{\sigma} x - (n+1)\log\left(1 - \frac{1}{x^{1/\sigma}}\right),$$

for $x > 1$, where the subscript $k_n = k$ is introduced to underline the dependence on n. We will show that $h(U)$ concentrates around the point of minimum of f_n through Lemma 28. A direct computation yields

$$f_n'(x) = -\frac{k_n - 1}{x} + \frac{\alpha}{\sigma} - \frac{n+1}{\sigma(x^{1+1/\sigma} - x)};$$

$$f_n''(x) = \frac{k_n - 1}{x^2} + (n+1)\frac{(1+1/\sigma)x^{1/\sigma} - 1}{\sigma(x^{1+1/\sigma} - x)^2}.$$

The last expression tells us that the function f_n is indeed convex as $f_n'' \geq 0$. Moreover, f_n is minimized at a single point which is a zero of f_n'. Thus, it is minimized at r_n a solution of the equation

$$\frac{k_n - 1}{x} + \frac{n+1}{\sigma(x^{1+1/\sigma} - x)} = \frac{\alpha}{\sigma}. \tag{21}$$

As the left hand side is the sum of two non-negative terms decreasing in n, we see that $r_n \to +\infty$ and

$$r_n \gtrsim \max\left(k_n, n^{\sigma/(1+\sigma)}\right). \tag{22}$$

We find the asymptotic behaviour of r_n depending on the behavious of k_n. We distinguish three cases.

First case. If $k_n \ll n^{\sigma/(1+\sigma)}$, from (22) then we deduce

$$\frac{k_n - 1}{r_n} \ll 1.$$

Thus we can neglect the first term in (21) and conclude that

$$\frac{n+1}{\sigma(r_n^{1+1/\sigma} - r_n)} \sim \frac{\alpha}{\sigma} \qquad \Leftrightarrow \qquad r_n \sim \left(\frac{n}{\alpha}\right)^{\sigma/(1+\sigma)}.$$

Second case. On the other hand if $k_n \gg n^{\sigma/(1+\sigma)}$, then from (22) we deduce

$$\frac{k_n - 1}{r_n} \gg \frac{n+1}{\sigma(r_n^{1+1/\sigma} - r_n)},$$

and thus we can neglect the second term in (21) and conclude that

$$\frac{k_n - 1}{r_n} \sim \frac{\alpha}{\sigma} \qquad \Leftrightarrow \qquad r_n \sim \frac{\sigma k_n}{\alpha}.$$

Third case. If $k_n \sim \lambda n^{\sigma/(1+\sigma)}$, then both terms in the left hand side of (21) are of the same order. Specifically, one can see that the reduced variable $\bar{r}_n = x_n/n^{\sigma/(1+\sigma)}$ will converge to a solution $\bar{r}$ of

$$\frac{\lambda}{\bar{r}} + \frac{1}{\sigma \bar{r}^{1+1/\sigma}} = \frac{\alpha}{\sigma}.$$

To apply Lemma 28 we need to control of the derivative f_n'' in each of the three cases. As $r_n \to +\infty$, f_n'' is asymptotically equivalent, say for $x \in [1/2r_n, 2r_n]$, to

$$\frac{k_n}{x^2} + \frac{(1+1/\sigma)n}{\sigma x^{2+1/\sigma}}. \tag{23}$$

We show that the last expression is asymptotically equivalent to $1/r_n$, say for $x \in [1/2r_n, 2r_n]$. This will allow us to use Lemma 28 for any value of $\gamma \in (1/2, 1)$.

First case. If $k_n \ll n^{\sigma/(1+\sigma)}$ the the first term in (23) is negligible compared to the second one which behaves like

$$\frac{n}{r_n^{2+1/\sigma}} \asymp \frac{1}{n^{\sigma/(1+\sigma)}} \asymp \frac{1}{r_n}.$$

Second case. If $k_n \gg n^{\sigma/(1+\sigma)}$ the the second term in (23) is negligible compared to the first one which behaves like $1/r_n$.

Third case. If $k_n \sim \lambda n^{\sigma/(1+\sigma)}$ the both terms are of the same order of magnitude, and once again they both behave like $1/r_n$. □

PROOF OF THEOREM 19. We study each term in the sum $\mathbb{E}(\mathcal{J}_1^U) + \mathbb{E}(\mathcal{J}_2^U) + \mathbb{E}(\mathcal{A}^U)$ sequentially, where all expectations are taken with respect to the random variable U. We will show

$$\max(\mathbb{E}(\mathcal{J}_1^U), \mathbb{E}(\mathcal{A}^U)) \lesssim \max(n^{-1/(1+\sigma)}, k/n), \quad \mathbb{E}(\mathcal{J}_2^U) \asymp \max(n^{-1/(1+\sigma)}, k/n).$$

First term. We recall that

$$\mathcal{J}_1^U = \frac{\alpha(1+U)^\sigma}{c} \mathcal{W}_{*1}\left(\frac{c^{1-\sigma}}{\Gamma(1-\sigma)} \frac{e^{-cs}}{s^{1+\sigma}} \mathrm{d}s, (\alpha+n) \frac{e^{-(\alpha+n)s}}{s} \mathrm{d}s \right).$$

Using Proposition 18 we obtain

$$\mathbb{E}\left(\frac{\alpha(1+U)^\sigma}{c} \right) \leq \mathbb{E}\left(\frac{\alpha(1+U)^\sigma}{n - k\sigma} \right) \lesssim \max\left(\frac{1}{n^{1/(1+\sigma)}}, \frac{k}{n} \right).$$

Since the extended Wasserstein distance is bounded by 2 (see Lemma 6), as $n \to +\infty$, $\mathbb{E}(\mathcal{J}_1^U) \lesssim \max(n^{-1/(1+\sigma)}, k/n)$.

Second term. We recall that

$$\mathcal{J}_2^U = \frac{n - k\sigma}{c} \mathcal{W}_{*1}\left(c \frac{e^{-cs}}{s} \mathrm{d}s, (\alpha+n) \frac{e^{-(\alpha+n)s}}{s} \mathrm{d}s \right),$$

and with Lemma 6 we know it is bounded by 2. Using the same argument as before, the prefactor satisfies

$$\mathbb{E}\left(\frac{n - k\sigma}{c} \right) = 1 - \mathbb{E}\left(\frac{\alpha(1+U)^\sigma}{c} \right) \sim 1,$$

and also convergence in probability holds. Then we deal with the extended Wasserstein distance through Proposition 8, which entails

$$\mathcal{W}_{*1}\left(c\frac{e^{-cs}}{s}\mathrm{d}s, (\alpha+n)\frac{e^{-(\alpha+n)s}}{s}\mathrm{d}s\right) \sim C\left|\frac{c}{\alpha+n}-1\right|,$$

and it is bounded from above by 2 (see Lemma 6). The asymptotic analysis of Proposition 18 guarantees that, at least in probability

$$\frac{c}{\alpha+n}-1 = \frac{\alpha(1+U)^{\sigma}/n - k/n\sigma - \alpha/n}{1+\alpha/n} \asymp \max\left(\frac{1}{n^{1/(1+\sigma)}}, \frac{k}{n}\right).$$

Thus the same asymptotic holds in probability for $\mathcal{J}_2^U$ thus we can deduce (by boundedness) that it also hold in expectation.

Last term. We recall that

$$\mathcal{A}^U = \mathcal{W}_1\left(\frac{\alpha(U+1)^{\sigma}}{c}P_0 + \frac{1}{c}\sum_{i=1}^{k}(n_i-\sigma)\delta_{X_i^*}, \frac{\alpha}{\alpha+n}P_0 + \frac{1}{\alpha+n}\sum_{i=1}^{k} n_i\delta_{X_i^*}\right).$$

We use the convexity estimate (16) of the Wasserstein distance. By applying it twice we obtain

$$\mathcal{A}^U \leq \lambda_1\mathcal{W}_1\left(P_0, \frac{1}{n}\sum_{i=1}^{k} n_i\delta_{X_i^*}\right) + \lambda_2\mathcal{W}_1\left(\frac{1}{n-k\sigma}\sum_{i=1}^{k}(n_i-\sigma)\delta_{X_i^*}, \frac{1}{n}\sum_{i=1}^{k} n_i\delta_{X_i^*}\right),$$

where

$$\lambda_1 = \frac{\alpha(1+U)^{\sigma}}{c} - \frac{\alpha}{\alpha+n}, \qquad \lambda_2 = 1 - \frac{\alpha(1+U)^{\sigma}}{c}.$$

From the previous discussion we know that $\mathbb{E}(\lambda_1) \lesssim \max(n^{-1/(1+\sigma)}, k/n)$ while $\mathbb{E}(\lambda_2) \sim 1$. We then use the upper bound (15) for the Wasserstein distance, relying also on the assumption we made that the first moment of the (deterministic) sequence $(X_n)_{n\geq 1}$, as well as the one of $(X_k^*)_{k\geq 1}$, stays bounded. This yields that the first term of the sum is bounded by $\max(n^{-1/(1+\sigma)}, k/n)$ in expectation. We then use one last time the convexity argument by noting that

$$\frac{1}{n}\sum_{i=1}^{k} n_i\delta_{X_i^*} = \frac{n-k\sigma}{n}\frac{1}{n-k\sigma}\sum_{i=1}^{k}(n_i-\sigma)\delta_{X_i^*} + \frac{k\sigma}{n}\frac{1}{k}\sum_{i=1}^{k}\delta_{X_i^*}$$

which yields

$$\mathcal{W}_1\left(\frac{1}{n-k\sigma}\sum_{i=1}^{k}(n_i-\sigma)\delta_{X_i^*}, \frac{1}{n}\sum_{i=1}^{k} n_i\delta_{X_i^*}\right) \leq \frac{k\sigma}{n}\mathcal{W}_1\left(\frac{1}{n-k\sigma}\sum_{i=1}^{k}(n_i-\sigma)\delta_{X_i^*}, \frac{1}{k}\sum_{i=1}^{k}\delta_{X_i^*}\right).$$

We conclude as for the first term with (15) and the assumption that the first moments of $(X_n)_{n\geq 1}$ and $(X_k^*)_{k\geq 1}$ stay bounded. Summing up,

$$\mathbb{E}(\mathcal{A}^U) \lesssim \max\left(\frac{1}{n^{1/(1+\sigma)}}, \frac{k}{n}\right) + \frac{k}{n} \lesssim \max\left(\frac{1}{n^{1/(1+\sigma)}}, \frac{k}{n}\right).$$

□

PROOF OF PROPOSITION 20. Since P is continuous $k = n$ almost surely. We consider a realization of X_n with $k = n$, and thus the only source of randomness now comes from U. With a similar reasoning to Proposition 19, we will show

$$\mathbb{E}(\mathcal{J}_1^U) \sim \sigma \int_0^{+\infty} \left| \frac{\Gamma(-\sigma, t)}{\Gamma(1-\sigma)} - \Gamma(0,t) \right| \mathrm{d}t, \qquad \mathbb{E}(\mathcal{J}_2^U) \ll 1, \qquad \mathbb{E}(\mathcal{A}^U) \sim \sigma \mathcal{W}_1(P_0, P).$$

First of all we observe that from Proposition 18 and the definition of c,

$$\lim_{n\to+\infty} \frac{\alpha(1+U)^\sigma}{c} = \sigma, \quad \lim_{n\to+\infty} \frac{n - k\sigma}{c} = 1-\sigma, \quad \lim_{n\to+\infty} \frac{c}{n} = 1,$$

and all the limits hold in L^1.

First term. We recall that

$$\mathcal{J}_1^U = \frac{\alpha(1+U)^\sigma}{c} \mathcal{W}_* \left(\frac{c^{1-\sigma}}{\Gamma(1-\sigma)} \frac{e^{-cs}}{s^{1+\sigma}} \mathrm{d}s, (\alpha+n) \frac{e^{-(\alpha+n)s}}{s} \mathrm{d}s \right).$$

The prefactor converges in L^1 to σ while staying bounded by 1. For the extended Wasserstein distance, with a change of variable $t = cs$,

$$\begin{aligned}
&\mathcal{W}_{*1} \left(\frac{c^{1-\sigma}}{\Gamma(1-\sigma)} \frac{e^{-cs}}{s^{1+\sigma}} \mathrm{d}s, (\alpha+n) \frac{e^{-(\alpha+n)s}}{s} \mathrm{d}s \right) \\
&= \int_0^{+\infty} \left| \frac{c\Gamma(-\sigma, cs)}{\Gamma(1-\sigma)} - (\alpha+n)\Gamma(0, (\alpha+n)s) \right| \mathrm{d}s \\
&= \int_0^{+\infty} \left| \frac{\Gamma(-\sigma,t)}{\Gamma(1-\sigma)} - \frac{\alpha+n}{c} \Gamma\left(0, \frac{\alpha+n}{c} t\right) \right| \mathrm{d}t.
\end{aligned}$$

From this we claim that

$$\lim_{n\to+\infty} \mathbb{E}\left(\mathcal{W}_{*1} \left(\frac{c^{1-\sigma}}{\Gamma(1-\sigma)} \frac{e^{-cs}}{s^{1+\sigma}} \mathrm{d}s, (\alpha+n) \frac{e^{-(\alpha+n)s}}{s} \mathrm{d}s \right) \right) = \int_0^{+\infty} \left| \frac{\Gamma(-\sigma,t)}{\Gamma(1-\sigma)} - \Gamma(0,t) \right| \mathrm{d}t,$$

which can be proved by using the dominated convergence twice: for the inner integral using that $(\alpha+n)/c$ converges (e.g., in probability) to 1, for the expectation using that each integral is is bounded by 2 (see Lemma 6).

Second term. We recall that

$$\mathcal{J}_2^U = \frac{n - k\sigma}{c} \mathcal{W}_{*1} \left(c \frac{e^{-cs}}{s} \mathrm{d}s, (\alpha+n) \frac{e^{-(\alpha+n)s}}{s} \mathrm{d}s \right).$$

The prefactor stays bounded by 1. For the extended Wasserstein distance we use the bound of Proposition 8 to achieve

$$\mathcal{W}_{*1} \left(c \frac{e^{-cs}}{s} \mathrm{d}s, (\alpha+n) \frac{e^{-(\alpha+n)s}}{s} \mathrm{d}s \right) \lesssim \log\left(\left| \frac{c}{\alpha+n} \right| \right).$$

As $c/(\alpha+n)$ converges in L^1 to 1, while the logarithm is sublinear at $+\infty$, we easily conclude that $\mathbb{E}(\mathcal{J}_2^U)$ converges to $\log(1) = 0$.

Last term. We recall that

$$\mathcal{A}^U = \mathcal{W}_1 \left(\frac{\alpha(U+1)^\sigma}{c} P_0 + \frac{(1-\sigma)}{c} \sum_{i=1}^n \delta_{X_i}, \frac{\alpha}{\alpha+n} P_0 + \frac{1}{\alpha+n} \sum_{i=1}^n \delta_{X_i} \right),$$

where we have used that here the sequences $(X_n)_{n\geq 1}$ and $(X_k^*)_{k\geq 1}$ coincide. The empirical distribution of $(X_n)_{n\geq 1}$ converges to P for the topology of weak convergence together with convergence of the first moment, as P has finite first moment. Thus,

$$\frac{\alpha(U+1)^\sigma}{c}P_0 + \frac{(1-\sigma)}{c}\sum_{i=1}^{n}\delta_{X_i} \to \sigma P_0 + (1-\sigma)P;$$

$$\frac{\alpha}{\alpha+n}P_0 + \frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{X_i} \to P.$$

Since the Wasserstein distance metrizes this convergence, as recalled in Proposition 21 in Section 5, we conclude that

$$\mathbb{E}(\mathcal{A}^U) \to \mathcal{W}_1(\sigma P_0 + (1-\sigma)P, P) = \sigma\mathcal{W}_1(P_0, P),$$

where the last equality follows by (17). □

REFERENCES

[1] AMBROSIO, L., BRUÉ, E. and SEMOLA, D. (2021). *Lectures on optimal transport*. Springer.

[2] BACKHOFF, J., BARTL, D., BEIGLBÖCK, M. and WIESEL, J. (2022). Estimating processes in adapted Wasserstein distance. *The Annals of Applied Probability* **32** 529–550.

[3] BARRIOS, E., LIJOI, A., NIETO-BARAJAS, L. E. and PRÜNSTER, I. (2013). Modeling with Normalized Random Measure Mixture Models. *Statistical Science* **28** 313 – 334. https://doi.org/10.1214/13-STS416

[4] BARTL, D., BEIGLBÖCK, M. and PAMMER, G. (2021). The Wasserstein space of stochastic processes. *arXiv preprint arXiv:2104.14245*.

[5] BERGER, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. *Journal of Statistical Planning and Inference* **25** 303-328. https://doi.org/10.1016/0378-3758(90)90079-A

[6] BLACKWELL, D. and DUBINS, L. (1962). Merging of Opinions with Increasing Information. *The Annals of Mathematical Statistics* **33** 882 – 886. https://doi.org/10.1214/aoms/1177704456

[7] CATALANO, M., LIJOI, A. and PRÜNSTER, I. (2020). Approximation of Bayesian models for time-to-event data. *Electron. J. Statist.* **14** 3366–3395. https://doi.org/10.1214/20-EJS1746

[8] CATALANO, M., LIJOI, A. and PRÜNSTER, I. (2021). Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *Ann. Statist.* **49** 2916-2947.

[9] CATALANO, M., LAVENANT, H., LIJOI, A. and PRÜNSTER, I. (2022+). A Wasserstein index of dependence for random measures. *Arxiv*.

[10] COX, D. R. (1955). Some Statistical Methods Connected with Series of Events. *Journal of the Royal Statistical Society: Series B (Methodological)* **17** 129-157. https://doi.org/10.1111/j.2517-6161.1955.tb00188.x

[11] DOKSUM, K. (1974). Tailfree and Neutral Random Probabilities and Their Posterior Distributions. *The Annals of Probability* **2** 183–201.

[12] DYKSTRA, R. L. and LAUD, P. (1981). A Bayesian Nonparametric Approach to Reliability. *The Annals of Statistics* **9** 356–367.

[13] FERGUSON, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics* **1** 209 – 230. https://doi.org/10.1214/aos/1176342360

[14] FERGUSON, T. S. (1974). Prior Distributions on Spaces of Probability Measures. *The Annals of Statistics* **2** 615 – 629. https://doi.org/10.1214/aos/1176342752

[15] FIGALLI, A. and GIGLI, N. (2010). A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions. *Journal de Mathèmatiques Pures et Appliquèes* **94** 107-130.

[16] FOURNIER, N. and GUILLIN, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields* **162** 707–738.

[17] GIORDANO, R., LIU, R., JORDAN, M. I. and BRODERICK, T. (2022). Evaluating Sensitivity to the Stick-Breaking Prior in Bayesian Nonparametrics. *Bayesian Analysis* 1 – 53. https://doi.org/10.1214/22-BA1309

[18] GUILLEN, N., MOU, C. and ŚWIĘCH, A. (2019). Coupling Lévy measures and comparison principles for viscosity solutions. *Transactions of the American Mathematical Society* **372** 7327–70.

[19] GUSTAFSON, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics* **24** 174 – 195. https://doi.org/10.1214/aos/1033066205

[20] HJORT, N. L. (1990). Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics* **18** 1259–1294.

[21] INSUA, D. R. and RUGGERI, F. (2000). *Robust Bayesian Analysis* **152**. Springer.

[22] JAMES, L. F. (2005). Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. *The Annals of Statistics* **33** 1771–1799.

[23] JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2006). Conjugacy as a distinctive feature of the Dirichlet process. *Scandinavian Journal of Statistics* **33** 105–120.

[24] JAMES, L. F., LIJOI, A. and PRÜNSTER, I. (2009). Posterior Analysis for Normalized Random Measures with Independent Increments. *Scandinavian Journal of Statistics* **36** 76-97. https://doi.org/10.1111/j.1467-9469.2008.00609.x

[25] KALLENBERG, O. (2017). *Random Measures, Theory and Applications*. *Probability Theory and Stochastic Modelling*. Springer International Publishing.

[26] KINGMAN, J. F. C. (1967). Completely Random Measures. *Pacific J. Math.* **21** 59–78.

[27] KORWAR, R. M. and HOLLANDER, M. (1973). Contributions to the Theory of Dirichlet Processes. *The Annals of Probability* **1** 705 – 711. https://doi.org/10.1214/aop/1176996898

[28] LEY, C., REINERT, G. and SWAN, Y. (2017). Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. *The Annals of Applied Probability* **27** 216–241.

[29] LIJOI, A., MENA, R. H. and PRÜNSTER, I. (2007). Controlling the reinforcement in Bayesian nonparametric mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **69** 715-740. https://doi.org/10.1111/j.1467-9868.2007.00609.x

[30] LIJOI, A. and PRÜNSTER, I. (2010). *Models beyond the Dirichlet process* In *Bayesian Nonparametrics* 80–136. Cambridge University Press. https://doi.org/10.1017/CBO9780511802478.004

[31] LO, A. Y. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics* **12** 351–357.

[32] NIETO-BARAJAS, L. E. and PRÜNSTER, I. (2009). A sensitivity analysis for Bayesian Nonparametric density estimators. *Statistica Sinica* **19** 685–705.

[33] PEYRÉ, G. and CUTURI, M. (2019). Computational optimal transport – with applications to data science. *Foundations and Trends® in Machine Learning* **11** 355–607.

[34] PFLUG, G. C. and PICHLER, A. (2012). A distance for multistage stochastic optimization models. *SIAM Journal on Optimization* **22** 1–23.

[35] PITMAN, J. (2006). *Combinatorial stochastic processes*. *Lecture Notes in Mathematics* **1875**. Springer-Verlag, Berlin. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard. https://doi.org/10.1007/b11601500 MR2245368

[36] PITMAN, J. and YOR, M. (1992). Arcsine Laws and Interval Partitions Derived from a Stable Subordinator. *Proceedings of the London Mathematical Society* **s3-65** 326-356. https://doi.org/10.1112/plms/s3-65.2.326

[37] PITMAN, J. and YOR, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25** 855–900.

[38] REGAZZINI, E., LIJOI, A. and PRÜNSTER, I. (2003). Distributional results for means of normalized random measures with independent increments. *The Annals of Statistics* **31** 560–585.

[39] SAHA, A. and KURTEK, S. (2019). Geometric Sensitivity Measures for Bayesian Nonparametric Density Estimation Models. *Sankhya Ser A.* **81(1)**.

[40] SANTAMBROGIO, F. (2015). Optimal transport for applied mathematicians. *Birkhäuser, NY* **55** 94.

[41] SATO, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. *Cambridge Studies in Advanced Mathematics*. Cambridge University Press.

[42] VILLANI, C. (2009). *Optimal transport: old and new* **338**. Springer.