The Impact of Recommendation Systems on Opinion Dynamics: Microscopic versus Macroscopic Effects

Conference Paper

Author(s):
Lanzetti, Nicolas ; Dörfler, Florian ; Pagan, Nicolo

The Impact of Recommendation Systems on Opinion Dynamics: Microscopic versus Macroscopic Effects

Nicolas Lanzetti, Florian Dörfler, and Nicolò Pagan

Abstract— **Recommendation systems are widely used in web services, such as social networks and e-commerce platforms, to serve personalized content to the users and, thus, enhance their experience. While personalization assists users in navigating through the available options, there have been growing concerns regarding its repercussions on the users and their opinions. Examples of negative impacts include the emergence of filter bubbles and the amplification of users' confirmation bias, which can cause opinion polarization and radicalization. In this paper, we study the impact of recommendation systems on users, both from a microscopic (i.e., at the level of individual users) and a macroscopic (i.e., at the level of a homogenous population) perspective. Specifically, we build on recent work on the interactions between opinion dynamics and recommendation systems to propose a model for this closed loop, which we then study both analytically and numerically. Among others, our analysis reveals that shifts in the opinions of individual users do not always align with shifts in the opinion distribution of the population. In particular, even in settings where the opinion distribution appears unaltered (e.g., measured via surveys across the population), the opinion of individual users might be significantly distorted by the recommendation system.**

I. Introduction

Over the past few years, recommendation systems have become an essential component of online services, including e-commerce platforms and social networking sites. Their primary objective is to filter through the vast amount of information available and guide users towards the most relevant content. Recommendation systems make use of diverse machine learning methods to assess the relevance of items and provide personalized content based on the recorded online behaviors of users. These techniques enable the systems to not only measure the absolute relevance of items but also tailor the recommendations to the users' expected tastes [1]. While recommendation systems have been a remarkable technological advancement, their impact on users' behavior has raised questions. Personalization is a key feature of these systems that improves the user experience but also poses concerns: Excessive personalization may limit the range of perspectives available to users, leading to "filter bubbles" [2]. These bubbles can induce opinion polarization and radicalization, which can be harmful [3]. Although later research [4] has downplayed concerns about the negative effects of personalization, evidence suggests that it has the potential to strengthen users' prejudices. For instance, numerous studies showed that it aggravates confirmation bias, which is the human propensity to seek and consider information that confirms their beliefs and ideas [5], [6]. This bias can lead to an unconscious one-sided argument-building process, reinforcing users' preconceived notions. Therefore, it is reasonable to conclude that personalization may exacerbate the confirmation bias phenomenon, potentially leading to further polarization and division among users.

Since empirical evidence supports the idea that confirmation bias is extensive, strong, and multiform, and its effects may be amplified by curation algorithms [3], a recent stream of literature [7]–[11] has started exploring the impact of the closed-loop dynamics between personalized recommendations and user preferences and opinions. For example, [9], [10] examined how this loop can reinforce user preferences and lead to polarization and filter bubbles, with [10] focusing mainly on the interaction between the user and the recommendation system, while [9] also included the effect of social influence by considering users being embedded in their social network. Differently, [8], [11] studied how to disentangle feedback loops in order to improve recommendation accuracy. The overall goal of this research field is to overcome the potential negative consequences of personalization by designing recommendation algorithms that can influence users' opinions and preferences in a more beneficial way [12], [13].

While these works generally show that the closed loop between opinion dynamics and recommendation systems bears the potential to steer individuals' opinions, it remains unclear (i) to what extent it may also steer the opinion distribution of a population (thus leading to significant concerns with regard to, e.g., political debates) and (ii) if seemingly similar opinion distributions (at the population level, e.g., surveys) can hide substantial individual shifts. To shed light on these questions, this paper builds on the recent model of [10] and examines the micro- and macroscopic impact of the closed loop between users and a recommendation system. This way, we can both study the impact of recommendation systems on the opinion distribution of a population and determine if shifts in opinions at the individual level can be concealed by their cumulative effect on the population level.

Contributions: With our work, we formalize the interaction between users and recommendation systems both from a microscopic (i.e., one user interacting with a recommendation system) and macroscopic (i.e., a homogeneous population interacting with a recommendation system) perspective. This way, we identify several tractable yet insightful instances,

N. Lanzetti and F. Dörfler are with the Automatic Control Laboratory, Department of Electrical Engineering and Information Technology, ETH Zürich, 8003 Zürich, Switzerland, `{lnicolas,doerfler}@ethz.ch`

N. Pagan is with the Social Computing Group, Department of Informatics, University of Zürich, 8003 Zürich, Switzerland, `nicolo.pagan@uzh.ch`

which we can investigate analytically and help us shed light on the impact of recommendation systems on users' opinions. Among others, our analytical analysis and numerical simulations uncover and explain a discrepancy between micro- and macroscopic behaviors, whereby the opinion of individual users is highly impacted by the recommendation system while, macroscopically, the opinion distribution remains unaffected. This insight reveals that, even when population surveys (e.g., exit polls) do not indicate opinion shifts, individuals' beliefs might be highly impacted by the recommendation systems.

Organization: This paper unfolds as follows. In Section II, we present our model of the closed loop between a user and a recommendation system. We study its properties in Section III and present numerical results in Section IV. Section V draws the conclusions of this paper. Proofs are deferred to the appendix.

A. Notation and background material

We denote by $\mathbb{R}_{\geq 0}$ the non-negative real numbers. The space of (Borel) probability distributions over $\mathbb{R}$ is $\mathcal{P}(\mathbb{R})$ and the space of probability distributions over $\mathbb{R}$ with finite second moment is $\mathcal{P}_2(\mathbb{R}) := \{\mu \in \mathcal{P}(\mathbb{R}) : \int_{\mathbb{R}} |x|^2 \mathrm{d}\mu(x) < +\infty\}$. The mean of a probability distribution $\mu \in \mathcal{P}(\mathbb{R})$ is $\mathbb{E}^\mu[x]$ and its variance is $\mathrm{Var}(\mu)$. The pushforward of a probability distribution $\mu \in \mathcal{P}(\mathbb{R})$ via a (Borel) map $f : \mathbb{R} \to \mathbb{R}$ is defined by $(f_\# \mu)(A) = \mu(f^{-1}(A))$ for each Borel set $A \subset \mathbb{R}$; if $X \sim \mu$, then $Y = f(X) \sim f_\# \mu$. The convolution of two probability distributions $\mu, \nu \in \mathcal{P}(\mathbb{R})$ is denoted by $\mu * \nu$; if $X \sim \mu$ and $Y \sim \nu$ are independent, then $X + Y \sim \mu * \nu$. The (type-p) Wasserstein distance between two probability distributions $\mu, \nu \in \mathcal{P}(\mathbb{R})$ is defined by

$$W_p(\mu, \nu) := \left(\min_{\gamma \in \Gamma(\mu, \nu)} \int_{\mathbb{R} \times \mathbb{R}} |x - y|^p \mathrm{d}\gamma(x, y) \right)^{\frac{1}{p}},$$

where $\Gamma(\mu, \nu) \subset \mathcal{P}(\mathbb{R} \times \mathbb{R})$ is the set all probability distributions over $\mathbb{R} \times \mathbb{R}$ with marginals μ and ν (referred to as transport plans) [14]. The Wasserstein distance is the minimum cost to transport μ onto ν when transporting a unit of mass from x to y costs $|x - y|^p$. Accordingly, a transport plan $\gamma \in \Gamma(\mu, \nu)$ encodes the allocation of probability mass: If (x, y) is in the support of γ, then some of the probability mass at x is displaced to y, or, equivalently, $\gamma(A \times B)$ is the mass transferred from the set $A \subset \mathbb{R}$ to the set $B \subset \mathbb{R}$. Finally, a sequence of probability distributions $(\mu_n)_{n \in \mathbb{N}} \subset \mathcal{P}_2(\mathbb{R})$ converges weakly in $\mathcal{P}_2(\mathbb{R})$ if $\int_{\mathbb{R}} \phi(x) \mathrm{d}\mu_n(x) \to \int_{\mathbb{R}} \phi(x) \mathrm{d}\mu(x)$ for all continuous functions $\phi : \mathbb{R} \to \mathbb{R}$ with at most quadratic growth (i.e., $\phi(x) \leq A(1 + |x|^2)$ for some $A > 0$).

II. MODEL

In this section, we present our model. It consists of two interconnected parts: the user model and the recommendation system model; see Fig. 1. We start with a detailed description of each component. Then, we illustrate the behavior of our model in a numerical example. Our model is based on the mathematical framework from [10].
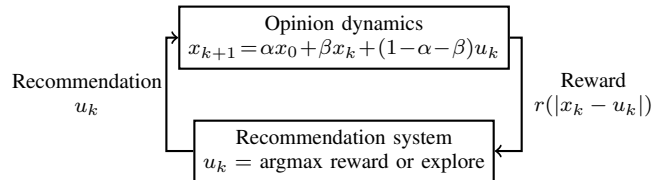

Fig. 1: Closed loop between a user (whose opinion dynamics are detailed in Section II-A) and a recommendation system (whose algorithm is detailed Section II-B).

A. Modeling of the users' opinion dynamics

We consider a large homogeneous population of users. The opinion of each user evolves according to the Friedkin-Johnson model [15]

$$x_{k+1} = \alpha x_0 + \beta x_k + (1 - \alpha - \beta)u_k, \tag{1}$$

where $x_k \in \mathbb{R}$ is the user's opinion at time k, x_0 is her/his opinion bias and initial opinion, and $u_k \in \mathbb{R}$ is the recommendation at time k. The parameters $\alpha \in [0, 1]$ and $\beta \in [0, 1)$ (with $\alpha + \beta \leq 1$) arbitrate between the impact of the user's bias, the current opinion, and the received recommendation on the future opinion. All users in the population share the same parameters α and β but have different biases. In particular, the bias/initial opinion distribution of the population is $\mu_0 \in \mathcal{P}_2(\mathbb{R})$. Similarly, we denote by μ_k the opinion distribution at time k or, equivalently, the probability distribution associated with the opinion of a generic user. Given the recommendation, the users produce a reward according to monotonically decreasing function $r : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$, so that the reward at time k is $r(|x_k - u_k|)$. Intuitively, the closer the recommendation is to the user's opinion, the more the user appreciates the content and thus the higher the benefit for the recommendation system (e.g., more clicks, more time on the platform). The reward is the only observable quantity; in particular, users' opinions are private and not revealed to the recommendation system.

B. Modeling of the recommendation system

The recommendation system aims at maximizing the reward. To do so, it outputs the recommendation that has generated the largest reward throughout the entire past until time k, with the exception of exploration steps, happening at the beginning of the horizon ($k = 0$) and every T steps ($k = nT$ for $n \in \mathbb{N}$). Thus, for $k \notin \{0, T, 2T, \ldots\}$

$$u_k = \underset{u_0, \ldots, u_{k-1}}{\mathrm{argmax}} \{r(|x_0 - u_0|), \ldots, r(|x_{k-1} - u_{k-1}|)\}.$$

When exploring, the recommendation system samples a recommendation from the recommendation distribution $\rho \in \mathcal{P}_2(\mathbb{R})$; i.e., $u_k \sim \rho$ for $k \in \{0, T, 2T, \ldots\}$. This strategy is in line with the classical ε-greedy action selection in multi-armed bandit problems in reinforcement learning, which also outputs the reward-maximizing action but explores with some probability ε (instead of at fixed time steps) [16, §2].

C. Discussion of the model

A few comments are in order. First, we consider a homogeneous population, whereby all users have the same α and β. This way, we can study the effect of the recommendation systems on *ensembles* of users, rather than on a specific user with given α and β, bias/initial opinion x_0, and realizations of the random recommendations. When presenting our results in Section IV, we consider various populations to illustrate the roles of α and β. Second, without loss of generality, we assume $\beta \neq 1$. If $\beta = 1$ (i.e., users are infinitely stubborn), we trivially conclude $x_{k+1} = x_k$ and $\mu_k = \mu_0$ for all k. Third, we do not include peer-to-peer interactions within the population, in line with our focus being the interaction between a recommendation system and its users. In any case, our model and analysis can be extended to include influences between the users (e.g., via graphons [17]). Nonetheless, we argue that, if interactions are to be considered, they do not happen *between* the users, but rather *through* the recommendation system. An example is collaborative filtering, where the system generates recommendations for a given user based on other users considered similar [18]. Fourth, there are of course many other models of recommendation systems based on more sophisticated algorithms, which include more general exploration strategies, collaborative filtering, priors on the users, machine learning techniques, etc. We leave their study to future research. Fifth, as we shall see below, the reward function r does not impact our analysis, as long as a mild (and, arguably, realistic) monotonicity assumption is satisfied. Finally, to ease the notation, we assume that α, β, and the recommendation distribution ρ are time-invariant. Provided this time-varying behavior is sufficiently regular, our model and analysis extend to the time-varying setting.

D. Illustrative example

To illustrate our model, we consider a population with $\alpha = 0.1$ and $\beta = 0.7$. For simulation purposes, we consider 5000 users, whose bias/initial opinion x_0 is distributed uniformly on $[0, 2]$ (blue, x-axis in Fig. 2). The recommendation distribution ρ is a zero-mean Gaussian with a standard deviation of 0.5 (red in Fig. 2) and exploration happens every 5th time step. We run the system for 50 time steps. Our results are in Fig. 2. At a macroscopic level, the opinion distribution shifts towards the recommendation distribution so that the final opinion distribution μ_N is a slightly asymmetric Gaussian (right plot in Fig. 2). At a microscopic level, the opinion of most (but not all) users is lower than their initial opinion (central plot in Fig. 2), which collectively contributes to the centralization of the opinion distribution observed at a macroscopic level.

III. ANALYSIS

In this section, we investigate some of the theoretic properties of our model. We start with a formulation of the dynamics in a tree structure, which immediately unveils the combinatorial nature of the problem. Thereafter, we study the more tractable yet insightful limit cases of $T = +\infty$ (no exploration) and $T = 1$ (exploration at every time step).
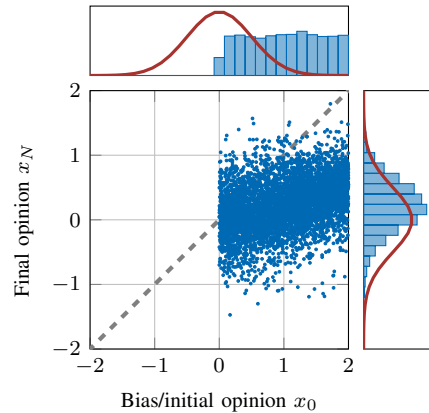

Fig. 2: Simulation of the closed-loop system over a horizon of 50 time steps. The central plot shows the final opinion of each user, as a function of their initial opinion. The top histogram shows the bias/initial opinion distribution; the one on the right is the final opinion distribution. For reference, we include, in solid red, the recommendation distribution.

A. The general case: A combinatorial problem

A general analysis of the closed-loop system is combinatorial. After each exploration step, there are two possibilities: (i) the exploration input led to a higher reward and thus the recommendation system sticks to that recommendation for the subsequent steps, at least until the next exploration phase; (ii) the exploration input led to a lower reward and is therefore discarded, and the recommendation system "goes back" to the last recommendation. Accordingly, the probability of successful exploration coincides with the probability of sampling a recommendation (strictly) closer to the current opinion compared to the current recommendation. Formally:

Lemma 1 (Probability of successful exploration). *Let u_k be the current recommendation and x_k be the opinion at time k. Then, the probability of successful exploration is*

$$p(x_k, u_k) = F_\rho(x_k + |x_k - u_k|) - \rho(\{x_k + |x_k - u_k|\}) - F_\rho(x_k - |x_k - u_k|),$$

where $F_\rho : \mathbb{R} \to [0, 1]$ is the cumulative distribution function of the recommendation distribution ρ, and $\rho(\{a\})$ is the probability of sampling the recommendation $a \in \mathbb{R}$.

Lemma 1 predicates that the probability of successful exploration is controlled by the difference between the current opinion and recommendation, together with the properties of the recommendation distribution ρ. For instance, if ρ is uniform between -1 and $+1$, then $\rho(\{x\}) = 0$ for all $x \in \mathbb{R}$ and $F_\rho(x) = \frac{1}{2}(1 + \max\{-1, \min\{1, x\}\})$, so that $p(x_k, u_k) = \frac{1}{2}(1 + \max\{-1, \min\{1, x_k + |x_k - u_k|\}\}) - \frac{1}{2}(1 + \max\{-1, \min\{1, x_k - |x_k - u_k|\}\})$. If $x_k \pm |x_k - u_k| \in [-1, +1]$, then $p(x_k, u_k) = |x_k - u_k|$, showing that, at least in the case of a uniform recommendation distribution, the probability of successful exploration is precisely $|x_k - u_k|$. With Lemma 1, the dynamics are as in Fig. 3, which explains why the analysis of the closed-loop system becomes intractable after a few exploration steps.
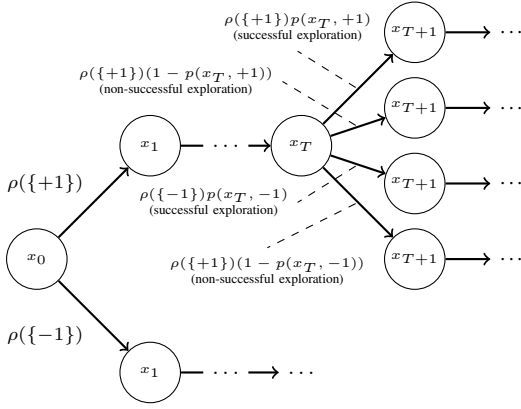
Fig. 3: Dynamics in the simple case of two possible recommendations (± 1, i.e., $\rho = p\delta_{-1} + (1-p)\delta_{+1}$ for $p \in [0,1]$) and deterministic bias/initial opinion $x_0 \in \mathbb{R}$ (i.e., $\mu_0 = \delta_{x_0}$). The transitions are $x_{k+1} = \alpha x_0 + \beta x_k + (1 - \alpha - \beta)u_k$; the quantity associated with each arrow is the probability of that transition.

Thus, in the sequel, we restrict our analysis to two tractable yet insightful limit cases: exploration only at the initial time and exploration at every time step.

B. Special case: No exploration

Suppose now that the recommendation system does not perform exploration; i.e., $T \to +\infty$. In this case, the initial recommendation, which is random, will be applied at all time steps, regardless of the reward returned by the user. As a result, each user's opinion converges to a convex combination of the bias and the received recommendation (at the initial time). Macroscopically, the opinion distribution approaches a "convex combination" of the bias distribution and the recommendation distribution:

Proposition 2 (No exploration). *Let $T = +\infty$. Let $\mu_0 \in \mathcal{P}_2(\mathbb{R})$ be the bias/initial opinion distribution, $\mu_k \in \mathcal{P}_2(\mathbb{R})$ be the opinion distribution at time k, and $\rho \in \mathcal{P}_2(\mathbb{R})$ be the recommendation distribution. Then, the opinion distribution μ_k converges weakly in $\mathcal{P}_2(\mathbb{R})$ to the opinion distribution*

$$\mu := \left(\frac{\alpha}{1-\beta}x\right)_{\#}\mu_0 * \left(\left(1 - \frac{\alpha}{1-\beta}\right)x\right)_{\#}\rho. \quad (2)$$

Proposition 2 proves weak convergence of the opinion distribution (i.e., convergence of the integral of each continuous function which grows at most quadratically) and does not prove strong convergence (i.e., $\mu_k(A) \to \mu(A)$ for every Borel set A). Namely, we prove that all macroscopic quantities converge (e.g., $\phi(x) = x$ in the definition of weak convergence in $\mathcal{P}_2(\mathbb{R})$ yields convergence of the expected value), and refrain from conducting a microscopic analysis for each and every (infinitesimal) user of the population. In particular, if $\alpha = 0$, then the opinion distribution asymptotically converges to the recommendation distribution. While simple, Proposition 2 unveils a fundamental discrepancy between a *macroscopic* analysis, aiming to study the opinion distribution of the population, and a *microscopic* analysis, which considers the change of opinions for individual users. We illustrate the phenomenon in the following analytic

example and discuss it more in detail when presenting our numerical results in Section IV.

Example 1 (Microscopic vs. macroscopic behavior). Suppose that the bias/initial opinion distribution μ_0 coincides with the recommendation distribution ρ and that all distributions are zero-mean Gaussian with standard deviation $\sigma > 0$. By Proposition 2 (together with the expressions for the pushforward and convolution of Gaussians), the opinion distribution converges to a zero-mean Gaussian with standard deviation $\sigma\sqrt{(\frac{\alpha}{1-\beta})^2 + (1 - \frac{\alpha}{1-\beta})^2}$, so that the (type-2) Wasserstein distance between the bias/initial opinion distribution μ_0 and the final distribution μ (defined in (2)) reads

$$W_2(\mu_0, \mu)^2 = \sigma^2 \left|1 - \sqrt{\left(\frac{\alpha}{1-\beta}\right)^2 + \left(1 - \frac{\alpha}{1-\beta}\right)^2}\right|^2,$$

where we used the closed-form expression for the Wasserstein distance between Gaussians. Conversely, from a microscopic perspective, the opinion of a user with bias x_0 who receives the recommendation u_0 converges to $\frac{\alpha}{1-\beta}x_0 + (1 - \frac{\alpha}{1-\beta})u_0$. Thus, the opinion shift is $(1 - \frac{\alpha}{1-\beta})(x_0 - u_0)$, and the expected squared opinion shift for a user is

$$\Delta = \left|1 - \frac{\alpha}{1-\beta}\right|^2 \int_{\mathbb{R}} \int_{\mathbb{R}} |x_0 - u_0|^2 \mathrm{d}\mu_0(x_0)\mathrm{d}\rho(u_0)$$
$$= 2\left|1 - \frac{\alpha}{1-\beta}\right|^2 \sigma^2.$$

For $\alpha \to 1 - \beta$, we recover the trivial case $1 - \alpha - \beta = 0$, whereby the recommendation has weight 0 in the opinion dynamics (1). Accordingly, both $W_2(\mu_0, \mu)^2$ and Δ converge to 0, and the micro- and macroscopic behaviors align. For $\alpha \to 0$ and fixed β, instead, $W_2(\mu_0, \mu)^2 \to 0$; i.e., the initial opinion distribution μ_0 (which, by assumption, equals ρ) coincides with the final opinion distribution μ (which, by Proposition 2, also equals ρ). However, $\Delta \to 2\sigma^2 > 0$. Thus, microscopically, each user's opinion is highly impacted by the recommendation system, while, macroscopically, the opinion distribution of the population is unaltered.

Finally, Proposition 2 considers the limit setting where the recommendation system explores only at the beginning and then sticks to the first, random, recommendation. Nonetheless, the intuition remains valid for sufficiently large values of exploration, as suggested by the numerical results in Fig. 4.

C. Special case: Continuous exploration

We consider now the opposite limit case, where the recommendation system explores at each time step; i.e., $T = 1$ and the recommendation received by a user is therefore random at each time step. In this setting, the opinion distribution converges to a "convex combination" of the bias/initial opinion distribution and a distribution "similar" to a Gaussian distribution. Perhaps surprisingly, this result is independent of the initial opinion distribution and, importantly, of the recommendation distribution. Formally:
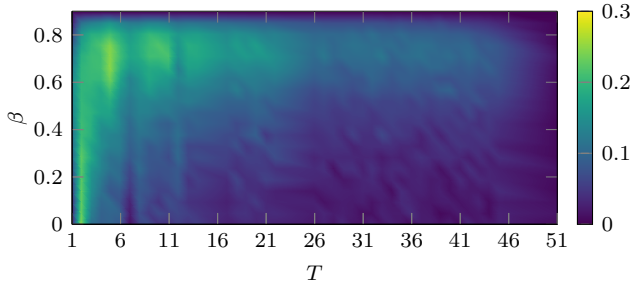
Fig. 4: The (type-1) Wasserstein distance between the final opinion distribution and the distribution (2) is small (dark blue in the colormap) already for relatively small values T, especially for β small (i.e., the distance is small). Thus, at least qualitatively, Proposition 2 remains valid also in non-asymptotic regimes. For this simulation, we used the same setting as in Section II-D.

Proposition 3 (Continuous exploration). *Let $T = 1$. Let $\mu_0 \in \mathcal{P}_2(\mathbb{R})$ be the bias/initial opinion distribution, $\mu_k \in \mathcal{P}_2(\mathbb{R})$ be the opinion distribution at time k, and $\rho \in \mathcal{P}_2(\mathbb{R})$ be the recommendation distribution. Then, the opinion distribution μ_k converges weakly in $\mathcal{P}_2(\mathbb{R})$ to some opinion distribution $\mu \in \mathcal{P}_2(\mathbb{R})$ with*

$$
\mathbb{E}^{\mu}[x] = \frac{\alpha}{1-\beta} \mathbb{E}^{\mu_0}[x] + \left(1 - \frac{\alpha}{1-\beta}\right) \mathbb{E}^{\rho}[x]
$$
$$
\mathrm{Var}(\mu) = \frac{\alpha^2}{(1-\beta)^2} \mathrm{Var}(\mu_0) + \frac{(1-\alpha-\beta)^2}{1-\beta^2} \mathrm{Var}(\rho). \tag{3}
$$

Moreover, regardless of the bias/initial opinion distribution μ_0 and the recommendation distribution ρ,

$$
\mu = \left(\frac{\alpha}{1-\beta}x\right)_{\#} \mu_0 * \left(\left(1 - \frac{\alpha}{1-\beta}\right)x\right)_{\#} \bar{\rho}, \tag{4}
$$

where $\bar{\rho}$ is "almost Gaussian", in the sense that its normalized distribution $\hat{\rho} := ((x - \mathbb{E}^{\bar{\rho}}[x])/\sqrt{\mathrm{Var}(\bar{\rho})})_{\#}\bar{\rho}$ satisfies

$$
W_1(\hat{\rho}, \Phi) \le \left(\frac{18}{\pi}\right)^{\frac{1}{3}} \left(\frac{1-\beta^2}{e\beta^2}\right)^{\frac{1}{12}} \sup_{\xi \ne 0} \frac{|C_\rho(\xi) - C_\Phi(\xi)|}{|\xi|^3}, \tag{5}
$$

where Φ is the zero-mean Gaussian probability distribution with unit variance and $C_\mu(\cdot)$ is the characteristic function of μ. Finally, if μ_0 and ρ are Gaussian distributions, then μ is Gaussian with the mean and variance in (3).

Proposition 3 predicates that the opinion distribution converges to a "convex combination" of the bias/initial condition distribution and a distribution which is "almost Gaussian", regardless of the recommendation distribution (that is, even for very "non-Gaussian" recommendation distributions). This effect is reminiscent of the central limit theorem in probability theory, which, however, does not apply here (indeed, we consider trajectories of a stochastic dynamic system and not the sum of i.i.d. random variables). In particular, if $\alpha = 0$, the opinion distribution will asymptotically be "almost Gaussian". The upper bound (5), which is valid for all α and β and is non-trivial (i.e., finite) for all distributions with finite third moment [19], indicates that the vicinity to a Gaussian distribution increases continuously as β approaches
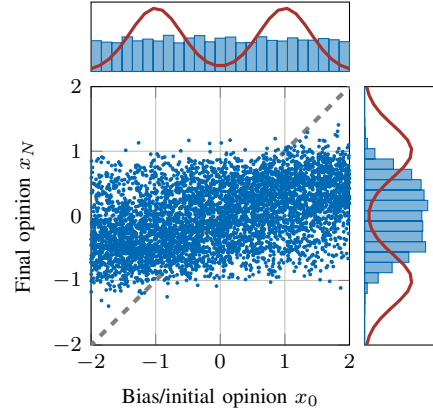

Fig. 5: Simulation of the closed-loop system over a horizon of 100. As suggested by our theoretic results in the case of infinitely frequent exploration (and with $\alpha = 0$), the opinion distribution approaches (but does not generally converge to) a Gaussian distribution, even though the bias/initial opinion distribution is uniform and the recommendation distribution (plotted in solid red) is bimodal (a mixture of Gaussians with mean ± 1 and standard deviation 0.4).

1. When all distributions are Gaussian, the result becomes exact. Among others, Proposition 3 explains why, for short exploration times, the final opinion distribution does not depend on the recommendation distribution and why, for small values of α, it resembles a Gaussian distribution, even if all underlying distributions are not Gaussian; see Fig. 5, where $\alpha = 0$, $\beta = 0.8$, and $T = 3$.

IV. NUMERICAL RESULTS

Our numerical analysis[1] concerns the discrepancy between the micro- and macroscopic behavior of the opinion distribution, as a function of α, β and the exploration time T. We consider homogeneous populations with $\alpha \in \{0, 0.1, 0.2\}$, $\beta \in \{0, 0.1, \ldots, 1 - \alpha\}$, and bias/initial opinion x_0 uniformly distributed between -2 and 2, and recommendation systems with exploration time $T \in \{1, \ldots, 21\}$ and recommendations distributed according to the standard zero-mean Gaussian distribution with unit variance. For each setting, we perform 20 exploration cycles. To quantify *microscopic* opinion shifts, we average the difference between users' initial and final opinions (i.e., $\frac{1}{M}\sum_{i=1}^{M} |x_{0,i} - x_{N,i}|$ with $x_{0,i}$ and $x_{N,i}$ being the initial and final opinion of user $i \in \{1, \ldots, M\}$). To quantify *macroscopic* opinion shifts, instead, we use the (type-1) Wasserstein distance between bias/initial opinion distribution μ_0 and final opinion distribution μ_N.

Our results are summarized in Fig. 6. First, for small values of α, our simulations suggest a qualitative discrepancy between micro- and macroscopic changes: When the microscopic change is largest (yellow), the macroscopic change is lowest (blue), and vice versa. The trend becomes less marked and eventually disappears as α increases, as already observed in Example 1 for the limit case $T = +\infty$. This discrepancy suggests that, even if the opinion distribution can be proven to be relatively stable (e.g., through surveys), the opinions

[1]Python implementation: https://gitlab.ethz.ch/lnicolas/impact-of-recommendation-systems-on-opinion-dynamics.
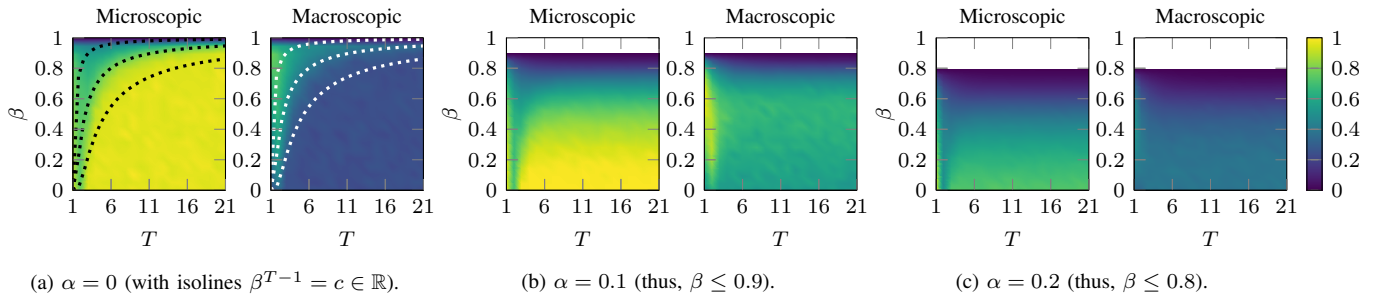
(a) $\alpha = 0$ (with isolines $\beta^{T-1} = c \in \mathbb{R}$).   (b) $\alpha = 0.1$ (thus, $\beta \leq 0.9$).   (c) $\alpha = 0.2$ (thus, $\beta \leq 0.8$).

Fig. 6: In each figure, the left plot shows the microscopic opinion shift as a function of β and T, measured as the average difference between the initial and the final opinion across the population. The right plot shows the macroscopic opinion shift, captured by the (type-1) Wasserstein distance between the initial opinion distribution and the final one. For comparison purposes, we normalize each value by the maximum value across all β and T, so that all values are between 0 and 1. Remarkably, our result highlights a marked discrepancy between the micro- and macroscopic behavior: Even if the opinion distribution can be shown to remain stable (e.g., via surveys), the opinion of individual users might be significantly impacted by the recommendation system, especially for low α.

of individual users may be significantly impacted by the recommendation system. Second, for $\alpha \ll 1 - \beta$, we observe similar behavior along isolines of $\beta^{T-1} = c \in \mathbb{R}_{\geq 0}$ in the sense that settings with similar β^{T-1} yield similar micro- and macroscopic behaviors (see Fig. 6a). Intuitively, this is a consequence of the probability of successful exploration being modulated by $|x_k - u_k|$ (cf. Lemma 1), since between two exploration steps k and $k - T$:

$$|x_k - u_k| \overset{(1)}{=} \left| \beta^{T-1}(x_{k+1-T} - u_k) + \frac{\alpha - \alpha\beta^{T-1}}{1 - \beta}(x_0 - u_k) \right|$$
$$\overset{\alpha \ll 1-\beta}{\approx} \beta^{T-1}|x_{k+1-T} - u_k|.$$

Thus, settings with similar β^{T-1} result in similar probabilities of successful exploration and thus in similar outcomes.

V. CONCLUSIONS

We studied the impact of recommendation systems on opinion dynamics from a microscopic (i.e., at the individual level) and macroscopic (i.e., at the population level) perspective. We analyzed theoretically and numerically the interaction between users and a recommendation system. Among others, our work explains why and in which circumstances we observe a fundamental discrepancy between micro- and macroscopic effects, whereby the opinions of individual users are drastically affected by the recommendation system whereas the opinion distribution of the population is unchanged.

In future research, we would like to (i) further investigate the properties of our model, (ii) consider more sophisticated recommendation systems (e.g., collaborative filtering [18]), and (iii) model the dynamics of opinion distribution directly in the probability space (e.g., see [20]).

REFERENCES

[1] J. A. Konstan and J. Riedl, "Recommender systems: from algorithms to user experience," *User modeling and user-adapted interaction*, vol. 22, pp. 101–123, 2012.

[2] E. Pariser, *The filter bubble: What the Internet is hiding from you.* penguin UK, 2011.

[3] C. R. Sunstein, *# Republic: Divided democracy in the age of social media.* Princeton University Press, 2018.

[4] E. Bakshy, S. Messing, and L. A. Adamic, "Exposure to ideologically diverse news and opinion on Facebook," *Science*, vol. 348, no. 6239, pp. 1130–1132, 2015.

[5] R. S. Nickerson, "Confirmation bias: A ubiquitous phenomenon in many guises," *Review of general psychology*, vol. 2, no. 2, pp. 175–220, 1998.

[6] J. Klayman and Y.-W. Ha, "Confirmation, disconfirmation, and information in hypothesis testing." *Psychological review*, vol. 94, no. 2, p. 211, 1987.

[7] N. Pagan, J. Baumann, E. Elokda, G. De Pasquale, S. Bolognani, and A. Hannák, "A classification of feedback loops and their relation to biases in automated decision-making systems," *arXiv preprint arXiv:2305.06055*, 2023.

[8] A. Sinha, D. F. Gleich, and K. Ramani, "Deconvolving feedback loops in recommender systems," *Advances in neural information processing systems*, vol. 29, 2016.

[9] N. Perra and L. E. Rocha, "Modelling opinion dynamics in the age of algorithmic personalisation," *Scientific reports*, vol. 9, no. 1, pp. 1–11, 2019.

[10] W. S. Rossi, J. W. Polderman, and P. Frasca, "The closed loop between opinion formation and personalized recommendations," *IEEE Transactions on Control of Network Systems*, vol. 9, no. 3, pp. 1092–1103, 2021.

[11] S. Dean and J. Morgenstern, "Preference dynamics under personalized recommendations," in *Proceedings of the 23rd ACM Conference on Economics and Computation*, 2022, pp. 795–816.

[12] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 4, no. 1, 2010, pp. 90–97.

[13] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Steering user behavior with badges," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 95–106.

[14] C. Villani, *Optimal transport: old and new.* Springer, 2009, vol. 338.

[15] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," *Journal of Mathematical Sociology*, vol. 15, no. 3-4, pp. 193–206, 1990.

[16] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction.* MIT press, 2018.

[17] L. Lovász, *Large networks and graph limits.* American Mathematical Soc., 2012, vol. 60.

[18] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Advances in artificial intelligence*, vol. 2009, 2009.

[19] G. Katriel, "A quantitative discounted central limit theorem using the Fourier metric," *Statistics & Probability Letters*, vol. 145, pp. 321–326, 2019.

[20] N. Lanzetti, J. Hajar, and F. Dörfler, "Modeling of political systems using Wasserstein gradient flows," in *2022 IEEE 61st Conference on Decision and Control (CDC).* IEEE, 2022, pp. 364–369.

[21] L. Aolaritei, N. Lanzetti, H. Chen, and F. Dörfler, "Distributional uncertainty propagation via optimal transport," *arXiv preprint arXiv:2205.00343*, 2023.

[22] Y. S. Chow and H. Teicher, *Probability theory: independence, interchangeability, martingales.* Springer Science & Business Media, 2003.

[23] J. A. Carrillo, G. Toscani *et al.*, "Contractive probability metrics and

asymptotic behavior of dissipative kinetic equations," *Riv. Mat. Univ. Parma*, vol. 6, no. 7, pp. 75–198, 2007.

[24] M. Gelbrich, "On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces," *Mathematische Nachrichten*, vol. 147, no. 1, pp. 185–203, 1990.

APPENDIX

A. Preliminaries

We start with two properties of the Wasserstein distance:

Lemma 4 (Pushforward and W_2 [21, Proposition 3]). *Let $\mu, \nu \in \mathcal{P}(\mathbb{R})$ and $f, g : \mathbb{R} \to \mathbb{R}$ be Borel maps. Then,*

$$W_2(f_{\#}\mu, g_{\#}\nu) = \sqrt{\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R} \times \mathbb{R}} |f(x) - g(y)|^2 \mathrm{d}\gamma(x,y)}.$$

Moreover, $W_2(f_{\#}\mu, g_{\#}\mu) \leq \sqrt{\int_{\mathbb{R}} |f(x) - g(x)|^2 \mathrm{d}\mu(x)}$, and, if $f(x) = ax$, $W_2(f_{\#}\mu, f_{\#}\nu) \leq a W_2(\mu, \nu)$.

Lemma 5 (Convolution and W_2 [21, Proposition 10]). *Let $\mu, \nu, \rho \in \mathcal{P}(\mathbb{R})$. Then, $W_2(\mu * \rho, \nu * \rho) \leq W_2(\mu, \nu)$.*

Second, we study the discounted sum of i.i.d. random variables. Here, $\mathcal{B}([0,1])$ denotes the Borel sigma algebra on $[0,1]$ and λ the Lebesgue measure.

Lemma 6 ("Central limit theorem" for discounted sums). *Consider i.i.d. real random variables $(X_n)_{n \in \mathbb{N}}$ with mean m, variance σ^2, and law $\nu \in \mathcal{P}(\mathbb{R})$ on a probability space $([0,1], \mathcal{B}([0,1]), \lambda)$. Let $\beta \in (0,1)$ and define the random variable S_β as the almost sure limit of $S_{\beta,k} := \sum_{l=0}^{k-1} \beta^{k-1-l}(1-\beta)X_l$. Let $\hat{\mu}_\beta \in \mathcal{P}(\mathbb{R})$ be the law of the "normalized" S_β (i.e., $\hat{S}_\beta = \sqrt{\frac{1+\beta}{(1-\beta)\sigma^2}}(S_\beta - \mu)$). Then,*

$$W_1(\hat{\mu}_\beta, \Phi) \leq \left(\frac{18}{\pi}\right)^{\frac{1}{3}} \left(\frac{1-\beta^2}{e\beta^2}\right)^{\frac{1}{12}} \sup_{\xi \neq 0} \frac{|C_\nu(\xi) - C_\Phi(\xi)|}{|\xi|^3},$$

where $C_\mu(\cdot)$ is the characteristic function of μ and Φ is the zero-mean Gaussian probability measure with unit variance.

Proof. Without loss of generality, assume X_l is zero-mean; indeed, if $\sum_{l=0}^{k} \beta^{k-1-l}(1-\beta)(X_l - m)$ converges to S_β, then $\sum_{l=0}^{k} \beta^{k-1-l}(1-\beta)X_l$ converges to $S_\beta + m$. Then, S_β is well-defined: Standard arguments of the proof of the Khintchine-Kolmogorov Convergence Theorem [22, Theorem 1, §5] yield almost sure convergence. The rest of the proof is analogous to the proof of [19] (with β instead of a and for simplicity $s = 3$), together with [23, Theorem 2.21] to derive the bound for the Wasserstein distance. $\square$

B. Proofs

Proof of Lemma 1. By definition of successful exploration,

$$p = \rho(\{u \in \mathbb{R} : r(|x_k - u|) < r(|x_k - u_k|)\})$$
$$\overset{\heartsuit}{=} \rho(\{u \in \mathbb{R} : |x_k - u| < |x_k - u_k|\})$$
$$= \rho(\{u \in \mathbb{R} : x_k - |x_k - u_k| < u < x_k + |x_k - u_k|\})$$
$$= \rho(\{u \in \mathbb{R} : u \leq x_k + |x_k - u_k|\})$$
$$\quad - \rho(\{x_k + |x_k - u_k|\})$$
$$\quad - \rho(\{u \in \mathbb{R} : u \leq x_k - |x_k - u_k|\}),$$

where $\heartsuit$ follows from strict monotonicity of the reward. With the definition of F_ρ, we conclude the proof. $\square$

Proof of Proposition 2. To start, consider a user with bias/initial opinion $x_0 \in \mathbb{R}$ and let $u_0 \in \mathbb{R}$ be the recommendation at the initial time. The opinion at time k is

$$x_k = \beta^k x_0 + \sum_{l=0}^{k-1} \beta^{k-1-l}(\alpha x_0 + (1-\alpha-\beta)u_0)$$
$$= \frac{\alpha + \beta^k(1-\alpha-\beta)}{1-\beta}x_0 + \frac{(1-\beta^k)(1-\alpha-\beta)}{1-\beta}u_0 \quad (6)$$
$$= (\eta + \beta^k(1-\eta))x_0 + (1-\beta^k)(1-\eta)u_0,$$

where $\eta := \alpha/(1-\beta) \in [0,1]$. Without loss of generality, assume $\eta \neq 1$; else $\alpha = 1 - \beta$, and we directly conclude $x_k = x_0$ and so $\mu_k = \mu_0$ for all $k \in \mathbb{N}$. By (6), the opinion distribution at time k is $\mu_k = \mu_{0,k} * \rho_k$ with $\mu_k := ((\eta + \beta^k(1-\eta))x)_{\#}\mu_0$ and $\rho_k := ((1-\beta^k)(1-\eta)x)_{\#}\rho$. We now prove that the (type-2) Wasserstein distance between μ_k and μ decays exponentially to 0. To ease the notation, we write $\mu = \bar{\mu}_0 * \bar{\rho}$ with $\bar{\mu}_0 = (\eta x)_{\#}\mu_0$ and $\bar{\rho} = ((1-\eta)x)_{\#}\rho$. Then,

$$W_2(\mu_k, \mu) = W_2(\mu_{0,k} * \rho_k, \bar{\mu}_0 * \bar{\rho})$$
$$\overset{\heartsuit}{\leq} W_2(\mu_{0,k} * \rho_k, \bar{\mu}_0 * \rho_k) + W_2(\bar{\mu}_0 * \rho_k, \bar{\mu}_0 * \bar{\rho})$$
$$\overset{\spadesuit}{\leq} W_2(\mu_{0,k}, \bar{\mu}_0) + W_2(\rho_k, \bar{\rho})$$
$$\overset{\clubsuit}{\leq} \sqrt{\int_{\mathbb{R}} |(\eta + \beta^k(1-\eta))x - \eta x|^2 \mathrm{d}\mu_0(x)}$$
$$\quad + \sqrt{\int_{\mathbb{R}} |(1-\beta^k)(1-\eta)u - (1-\eta)u|^2 \mathrm{d}\rho(u)}$$
$$= \beta^k(1-\eta)\left(\sqrt{\int_{\mathbb{R}} |x|^2 \mathrm{d}\mu_0(x)} + \sqrt{\int_{\mathbb{R}} |u|^2 \mathrm{d}\rho(u)}\right),$$

where in $\heartsuit$ we used triangle inequality; in $\spadesuit$ we used Lemma 5 and $\mu * \nu = \nu * \mu$; and $\clubsuit$ leverages Lemma 4. Since $\mu_0, \rho \in \mathcal{P}_2(\mathbb{R})$ and $\beta \in (0,1)$, we have $W_2(\mu_k, \mu) \to 0$ and thus weak convergence in $\mathcal{P}_2(\mathbb{R})$ [14, Theorem 6.9]. $\square$

Proof of Proposition 3. At time $k \in \mathbb{N}$, the opinion of a user with bias (and thus initial opinion) $x_0 \in \mathbb{R}$ reads

$$x_k = (\eta + \beta^k(1-\eta))x_0 + (1-\eta)\sum_{l=0}^{k-1} \beta^{k-1-l}(1-\beta)u_l, \quad (7)$$

where we proceed as in (6) and $\eta := \alpha/(1-\beta) \in [0,1)$ (again, the case $\eta = 1$ follows separately). Thus, the opinion distribution at time k is $\mu_k = \mu_{0,k} * ((1-\eta)x)_{\#}\rho_k$ with

$$\mu_{0,k} := ((\eta + \beta^k(1-\eta))x)_{\#}\mu_0$$
$$\rho_k := *_{l=0}^{k-1}(\beta^{k-1-l}(1-\beta)x)_{\#}\rho,$$

where $*_{l=0}^{k-1}(\beta^{k-1-l}(1-\beta)x)_{\#}\rho$ is the short-hand notation for $(\beta^{k-1}(1-\beta)x)_{\#}\rho * \ldots * ((1-\beta)x)_{\#}\rho$. We now show that $(\mu_k)_{k \in \mathbb{N}}$, $(\mu_{0,k})_{k \in \mathbb{N}}$, and $(\rho_k)_{k \in \mathbb{N}}$ converge weakly in $\mathcal{P}_2(\mathbb{R})$ to some $\bar{\mu}, \bar{\mu}_0, \bar{\rho} \in \mathcal{P}_2(\mathbb{R})$. It suffices that all sequences are Cauchy (w.r.t. the Wasserstein distance); convergence then follows from completeness of the Wasserstein space [14,

Theorem 6.18]. We start with $(\rho_k)_{k\in\mathbb{N}}$. Let $\varepsilon > 0$ (w.l.o.g., $\varepsilon < \beta$) and take $N_\rho > \log_\beta(\varepsilon(2\frac{1-\beta}{1+\beta}\int_\mathbb{R}|u|^2\mathrm{d}\rho(u))^{-1/2})$, well-defined since $\rho \in \mathcal{P}_2(\mathbb{R})$. Then, for $m > n > N_\rho$

$$W_2(\rho_n,\rho_m)^2$$
$$= W_2\big(*_{l=0}^{n-1}\left(\beta^{n-1-l}(1-\beta)x\right)_\#\rho,$$
$$*_{l=0}^{m-1}\left(\beta^{m-1-l}(1-\beta)x\right)_\#\rho\big)^2$$
$$= W_2\big(\delta_0 * \left(*_{l=0}^{n-1}\left(\beta^{n-1-l}(1-\beta)x\right)_\#\rho\right),$$
$$\left(*_{l=0}^{m-n-1}\left(\beta^{m-1-l}(1-\beta)x\right)_\#\rho\right)$$
$$* \left(*_{l=0}^{n-1}\left(\beta^{n-1-l}(1-\beta)x\right)_\#\rho\right)\big)^2$$
$$\overset{\heartsuit}{\leq} W_2\left(\delta_0, *_{l=0}^{m-n-1}\left(\beta^{m-1-l}(1-\beta)x\right)_\#\rho\right)^2$$
$$= \int_\mathbb{R}|x|^2\mathrm{d}\left(*_{l=0}^{m-n-1}\left(\beta^{m-1-l}(1-\beta)x\right)_\#\rho\right)(x)$$
$$\overset{\spadesuit}{=} \int_{\mathbb{R}^{m-n-1}}|u_0+\ldots+u_{m-n-1}|^2$$
$$\mathrm{d}\left(\left(\beta^{m-1}(1-\beta)x\right)_\#\rho\right)(u_0)\cdots$$
$$\mathrm{d}\left(\left(\beta^n(1-\beta)x\right)_\#\rho\right)(u_{m-n-1})$$
$$\overset{\clubsuit}{=} \int_{\mathbb{R}^{m-n-1}}|\beta^{m-1}(1-\beta)u_0+\ldots$$
$$+\beta^n(1-\beta)u_{m-n-1}|^2\mathrm{d}\rho(u_0)\cdots\mathrm{d}\rho(u_{m-n-1})$$
$$\overset{\diamondsuit}{\leq} 2(1-\beta)^2\int_{\mathbb{R}^{m-n-1}}\beta^{2m-2}|u_0|^2+\ldots$$
$$+\beta^{2n}|u_{m-n-1}|^2\mathrm{d}\rho(u_0)\cdots\mathrm{d}\rho(u_{m-n-1})$$
$$= 2(1-\beta)^2\Big(\beta^{2m-2}\int_\mathbb{R}|u_0|^2\mathrm{d}\rho(u_0)+\ldots$$
$$+\beta^{2n}\int_\mathbb{R}|u_{m-n-1}|^2\mathrm{d}\rho(u_{m-n-1})\Big)$$
$$= 2(1-\beta)^2\left(\beta^{2m-2}+\ldots+\beta^{2n}\right)\int_\mathbb{R}|u|^2\mathrm{d}\rho(u)$$
$$= 2(1-\beta)^2\frac{\beta^{2n+1}-\beta^{2m-1}}{1-\beta^2}\int_\mathbb{R}|u|^2\mathrm{d}\rho(u)$$
$$\leq 2\frac{1-\beta}{1+\beta}\beta^{2N_\rho}\int_\mathbb{R}|u|^2\mathrm{d}\rho(u),$$

where in $\heartsuit$ we leveraged Lemma 5, with $\delta_0 * \mu = \mu$ and $\mu*\nu = \nu*\mu$; in $\spadesuit$ we used the definition of convolution; $\clubsuit$ follows from $\int_\mathbb{R}\phi(x)\mathrm{d}(f_\#\mu)(x) = \int_\mathbb{R}\phi(f(x))\mathrm{d}\mu(x)$; and in $\diamondsuit$ we used that $(a+b)^2 \leq 2(a^2+b^2)$ for all $a,b\in\mathbb{R}$. By definition of N_ρ we therefore conclude

$$W_2(\rho_n,\rho_m) \leq \beta^{N_\rho}\sqrt{2\frac{1-\beta}{1+\beta}\int_\mathbb{R}|u|^2\mathrm{d}\rho(u)} < \varepsilon. \quad (8)$$

Thus, $(\rho_k)_{k\in\mathbb{N}}$ is Cauchy. For $(\mu_k)_{k\in\mathbb{N}}$, let $N_\mu > \log_\beta(\varepsilon(1-\eta)^{-1}((\int_\mathbb{R}|x|^2\mathrm{d}\mu_0(x))^{1/2}+(2\frac{1-\beta}{1+\beta}\int_\mathbb{R}|u|^2\mathrm{d}\rho(u))^{1/2})^{-1})$ and $m > n > N_\mu$. The distance between $\mu_{0,m}$ and $\mu_{0,n}$ is

$$W_2(\mu_{0,m},\mu_{0,n})^2$$
$$\overset{\heartsuit}{\leq} \int_\mathbb{R}|(\eta+\beta^m(1-\eta))x-(\eta+\beta^n(1-\eta))x|^2\mathrm{d}\mu_0(x)$$
$$= (1-\eta)^2(\beta^n-\beta^m)^2\int_\mathbb{R}|x|^2\mathrm{d}\mu_0(x)$$

$$\leq (1-\eta)^2\beta^{2N_\mu}\int_\mathbb{R}|x|^2\mathrm{d}\mu_0(x) \quad (9)$$

where in $\heartsuit$ we used Lemma 4. Then,

$$W_2(\mu_m,\mu_n)$$
$$= W_2(\mu_{0,m}*((1-\eta)x)_\#\rho_m, \mu_{0,n}*((1-\eta)x)_\#\rho_n)$$
$$\leq W_2(\mu_{0,m}*((1-\eta)x)_\#\rho_m, \mu_{0,n}*((1-\eta)x)_\#\rho_m)$$
$$+ W_2(\mu_{0,n}*((1-\eta)x)_\#\rho_m, \mu_{0,n}*((1-\eta)x)_\#\rho_n)$$
$$\overset{\heartsuit}{\leq} W_2(\mu_{0,m},\mu_{0,n}) + W_2(((1-\eta)x)_\#\rho_m, ((1-\eta)x)_\#\rho_n)$$
$$\overset{\spadesuit}{\leq} W_2(\mu_{0,m},\mu_{0,n}) + (1-\eta)W_2(\rho_m,\rho_n)$$
$$\overset{(8),(9)}{\leq} \beta^N(1-\eta)\left(\sqrt{\int_\mathbb{R}|x|^2\mathrm{d}\mu_0(x)}+\sqrt{2\frac{1-\beta}{1+\beta}\int_\mathbb{R}|u|^2\mathrm{d}\rho(u)}\right)$$
$$\overset{\clubsuit}{<} \varepsilon,$$

where in $\heartsuit$ we used Lemma 5 with $\mu*\nu = \nu*\mu$, in $\spadesuit$ we used Lemma 4, and in $\clubsuit$ follows from the definition of N_μ. Thus, $(\mu_k)_{k\in\mathbb{N}}$ is Cauchy too. By completeness of $\mathcal{P}_2(\mathbb{R})$ endowed with the Wasserstein distance [14, Theorem 6.18], $(\mu_k)_{k\in\mathbb{N}}, (\rho_k)_{k\in\mathbb{N}}$ converge to some $\bar\mu, \bar\rho \in \mathcal{P}_2(\mathbb{R})$, respectively. Finally, by (9), $(\mu_{0,k})_{k\in\mathbb{N}}$ is Cauchy too, and it converges to $\bar\mu_0 = (\eta x)_\#\mu_0$: By Lemma 4

$$W_2(\mu_{0,k},\bar\mu_0) \leq \sqrt{\int_\mathbb{R}|(\eta+\beta^k(1-\eta))x-\eta x|^2\mathrm{d}\mu_0(x)}$$
$$= \beta^k(1-\eta)\sqrt{\int_\mathbb{R}|x|^2\mathrm{d}\mu_0(x)} \to 0$$

since μ_0 has finite second moment. Then,

$$W_2(\mu,\bar\mu)$$
$$= W_2(\bar\mu_0*((1-\eta)x)_\#\bar\rho, \bar\mu)$$
$$\overset{\heartsuit}{\leq} W_2(\bar\mu_0*((1-\eta)x)_\#\bar\rho, \mu_{0,k}*((1-\eta)x)_\#\bar\rho)$$
$$+ W_2(\mu_{0,k}*((1-\eta)x)_\#\bar\rho, \mu_{0,k}*((1-\eta)x)_\#\rho_k)$$
$$+ W_2(\mu_{0,k}*((1-\eta)x)_\#\rho_k, \bar\mu)$$
$$\overset{\spadesuit}{\leq} W_2(\bar\mu_0,\mu_{0,k}) + W_2(((1-\eta)x)_\#\bar\rho, ((1-\eta)x)_\#\rho_k)$$
$$+ W_2(\mu_k,\mu)$$
$$= W_2(\bar\mu_0,\mu_{0,k}) + (1-\eta)W_2(\bar\rho,\rho_k) + W_2(\mu_k,\mu) \to 0.$$

where in $\heartsuit$ follows from triangle inequality and in $\heartsuit$ we used Lemma 5 with $\mu*\nu = \nu*\mu$ and the definition of μ_k. Thus, $W_2(\mu,\bar\mu) = 0$ and so $\bar\mu = \mu$. The definition of weak convergence in $\mathcal{P}_2(\mathbb{R})$ directly yields convergence of the first two moments and, thus, of the variance, which establishes (3) (together with the expressions for the pushforward/convolution of the expected value/variance). To prove (5), it now suffices to deploy Lemma 6. Finally, if μ_0 and ρ are Gaussian, then μ_k is Gaussian for all k. Let μ be Gaussian with the mean and variance in (3). By Gelbrich [24],

$$W_2(\mu_k,\mu)^2 = |\mathbb{E}^{\mu_k}[x]-\mathbb{E}^\mu[x]|^2 + \left|\sqrt{\mathrm{Var}(\mu_k)}-\sqrt{\mathrm{Var}(\mu)}\right|^2$$

which converges to 0 as $k \to \infty$. So, $(\mu_k)_{k\in\mathbb{N}}$ converges weakly in $\mathcal{P}_2(\mathbb{R})$ to μ. This concludes the proof. $\square$