

# Mass-gap extreme mass ratio inspirals

Zhen Pan,<sup>1,\*</sup> Zhenwei Lyu,<sup>1,2</sup> and Huan Yang<sup>1,2,†</sup>

<sup>1</sup>*Perimeter Institute for Theoretical Physics, Ontario, N2L 2Y5, Canada*

<sup>2</sup>*University of Guelph, Guelph, Ontario N1G 2W1, Canada*

In this work, we propose a new subclass of extreme-mass-ratio-inspirals (EMRIs): mass-gap EMRIs, consisting of a compact object in the lower mass gap  $\sim (2.5 - 5)M_\odot$  and a massive black hole (MBH). The mass-gap object (MGO) may be a primordial black hole or produced from a delayed supernova explosion. We calculate the formation rate of mass-gap EMRIs in both the (dry) loss-cone channel and the (wet) active galactic nucleus disk channel by solving Fokker-Planck-type equations for the phase-space distribution. In the dry channel, the mass-gap EMRI rate is strongly suppressed compared to the EMRI rate of stellar-mass black holes (sBHs) as a result of mass segregation effect. In the wet channel, the suppression is roughly equal to the mass ratio of sBHs over MGOs, because the migration speed of a compact object in an active galactic nucleus disk is proportional to its mass. We find that the wet channel is much more promising to produce mass-gap EMRIs observable by spaceborne gravitation wave detectors. (Non-)detection of mass-gap EMRIs may be used to distinguish different supernova explosion mechanisms and constrain the abundance of primordial black holes around MBHs.

## I. INTRODUCTION

Observations of Galactic X-ray binaries have indicated a dearth of compact objects around  $\sim (2.5 - 5)M_\odot$  in the mass spectrum [e.g., 1–3]. Whether this mass gap is a result of observational selection effects or underlying supernova (SN) explosion mechanism has been an open question for more than a decade [2, 4–6]. More recently new events detected with gravitational waves (GWs) and time-domain astronomy suggest the existence of mass-gap objects (MGOs) in compact object binaries and in Galactic non-interacting binaries - the compact objects of  $\sim (2.6 - 2.8)M_\odot$  in compact binary coalescence events GW190814 and GW200210 [7, 8], a compact object of  $\sim 3.3M_\odot$  as a non-interacting companion of a giant [9], and a compact object of  $\sim 3.0M_\odot$  as a non-interacting companion of a red giant [10] (see [11–13] for candidate MGOs in non-interacting binaries and as dark lens in the Milky Way). These identified MGOs provide evidences of a population of compact objects lying in the mass gap, or even a more extreme possibility that the mass gap itself does not exist. A natural question is that, if indeed a population of MGOs is present, what should be their origin? One viable option is MGOs are born in delayed SN explosions [6], which also provides an explanation to the merger rate of GW190814-like events [14, 15]. A more exotic possibility is that MGOs are primordial black holes (PBHs) [16–19], which have been intensively discussed in the context of compact binary mergers detected by LIGO/Virgo. With the upgrading of LIGO/Virgo and the coming era of third-generation detectors, more GW190814-like events are expected to be detected. However, as there is already a large number of proposed formation channels [e.g., 20–29], many of which are still subject to large theoretical uncertainties (see [30] for a recent review), it is unclear whether we will be able to nail down the the origin of MGOs with only the detection of stellar-mass binaries.

To better answer these questions, we investigate the possi-

bility that MGOs appear as components of some extreme mass ratio inspirals (EMRIs), to form mass-gap EMRIs. We propose that future detections of EMRIs and especially mass-gap EMRIs by spaceborne GW detectors, may be a useful probe for the origins of MGOs, thanks to the limited number of EMRI formation channels and distinct signatures of EMRI sources from different channels. There are two main channels of EMRI formation: the dry loss-cone channel [31–39] and the wet active galactic nucleus (AGN) disk channel [40–45] (other processes involving tidal disruption or tidal capture of binary sBHs, or tidal stripping of giant stars [46–49] may also contribute to EMRI formation and Naoz et al.[50] pointed out EMRI formation in MBH binaries is more efficient than around single MBHs). In the dry channel, a stellar-mass black hole (sBH) is scattered by stars in the nuclear stellar cluster and gravitationally captured by the MBH. There are two relevant timescales in the long-term dynamics: the GW emission timescale  $t_{\text{gw}}$  on which the sBH orbit shrink, and the diffusion timescale  $t_J$  on which the orbital angular momentum changes by successive scatterings [31, 34, 51–53]. If GW emission is more efficient with  $t_{\text{gw}} < t_J$ , the sBH becomes a stable EMRI with continually shrinking orbit until the final coalescence with the MBH. If scatterings are more efficient, i.e.,  $t_J < t_{\text{gw}}$ , the sBH is randomly scattered towards or away from the MBH. sBHs that are scattered into the MBH without losing much energy via GW emission are called prompt infalls [31]. As the GW emission timescale is much shorter for eccentric orbiters ( $e \rightarrow 1$ ), EMRIs in the dry channel are highly eccentric at formation.

A fraction  $\mathcal{O}(10^{-2} - 10^{-1})$  of MBHs in the universe (referred as AGNs) are actively accreting gas with an accretion disk [54, 55]. The presence of an accretion disk introduces new interactions affecting the motion of sBHs in the stellar cluster. For a sBH embedded in the AGN disk, its periodic motion produces density waves [56–59] that in turn drive the sBH to migrate inward, damp its orbital eccentricity and its inclination w.r.t. the disk plane. For a sBH on a highly inclined orbit, the effects of density waves becomes subdominant to dynamical friction [60, 61] arising from the relative motion of the sBH and the surrounding gas as it passes through the disk. As a result, sBHs are first captured onto the disk driven

\* zpan@perimeterinstitute.ca

† hyang@perimeterinstitute.ca

by dynamical fraction and density waves, and then migrate inward driven by the density waves upon reaching the vicinity of the MBH, where GW emission become dominant. We refer this type of EMRIs as wet EMRIs. Because density waves are very efficient at damping out the eccentricity, wet EMRIs are essentially circular. Dry and wet EMRIs can be easily distinguished from each other by measuring their eccentricities using spaceborne gravitational wave detectors [37–39]. In addition, the imprints of an accretion disk on the EMRI waveform may also be detectable [62, 63].

In the dry channel, sBHs are generally closer to the central MBH than MGOs due to the mass segregation, so that their EMRI rate should be larger. In the wet channel, the EMRI formation rates of both sBHs and MGOs mainly depend on the capture rate onto the AGN disk and the rate of migration along the disk. Both rates are proportional to the object mass, therefore the EMRI rate of MGOs is lower than that of sBHs by a factor of their mass ratio at most. We find the wet EMRI rate per AGN is generally higher than the dry EMRI rate per MBH by  $\mathcal{O}(10^1 - 10^3)$  for sBHs, and by  $\mathcal{O}(10^3 - 10^4)$  for MGOs. Taking into account the AGN fraction  $f_{\text{AGN}} = \mathcal{O}(10^{-2} - 10^{-1})$ , the wet channel turns out to be primary way in producing mass-gap EMRIs. If there is no gap in the mass spectrum of supernova remnants and roughly equal number of MGOs and sBHs are born in SN explosions, we expect LISA to detect  $\lesssim 1$  dry mass-gap EMRIs per year, and  $\mathcal{O}(1 - 10^2) \times (f_{\text{AGN}}/1\%)$  wet mass-gap EMRIs per year. In addition, future detections of dry and wet mass-gap EMRIs have interesting implications of MGO formation. An excess of mass-gap EMRI detection could be a signature of MGOs of more exotic origin, e.g., PBHs. The relative fraction of mass-gap EMRIs to sBH EMRIs is a sensitive probe for the mass spectrum of SN remnants and therefore the SN explosion mechanism.

The remaining part of this paper is organized as follows. In Sec. II and III, we introduce the two EMRI formation channels and the formation rates of MGOs and sBHs in these two channels. In Sec. IV, we forecast the detection prospects of these EMRIs by LISA. We summarize this work in Sec. V. Throughout this paper, we use geometrical units  $G = c = 1$ , and assume a flat  $\Lambda$ CDM cosmology with  $\Omega_m = 0.307$ ,  $\Omega_\Lambda = 1 - \Omega_m$  and  $H_0 = 67.7$  km/s/Mpc ( $h = 0.677$ ).

## II. DRY LOSS-CONE CHANNEL

In this section, we will first briefly introduce the loss-cone mechanism along with the Fokker-Planck equation governing the evolution of stars in a stellar cluster around a MBH. After that we calculate the EMRI formation rates of both sBHs and MGOs via the loss-cone mechanism.

### A. Loss-cone mechanism

Consider a star orbiting around a MBH, with specific binding energy  $E := \phi - v^2/2$  and specific angular momentum  $J$ , where  $\phi(r)$  is the (positive) gravitational potential and  $v^2/2$  is the kinetic energy. Its orbital motion is affected by two main

effects: GW emission which shrinks the orbit on a timescale  $t_{\text{gw}}$  and gravitational scatterings by other stars in the stellar cluster which changes the orbital angular momentum by order of unity on a timescale  $t_J$ . For a star on a tight and eccentric orbit where the GW emission is more efficient with  $t_{\text{gw}} < t_J$ , the orbit is stable against random scatterings and the star becomes an stable EMRI [31]. On the other hand, for a star on a wide and/or circular orbit where the GW emission is less efficient with  $t_{\text{gw}} > t_J$ , the star is expected to be scattered into a random direction: away from, towards or even directly into the central MBH.

In the phase space, a region of low angular momentum  $J < J_c(E)$  is usually referred as the loss cone, where a star usually promptly falls into the MBH within one orbital period  $P(E)$  if its angular momentum is not altered much by gravitational scatterings. As a result, the loss cone region satisfying  $P(E) < t_J$  is unpopulated (empty regime) and the loss cone region satisfying  $P(E) > t_J$  is populated (full regime). For the problem we are investigating, relevant orbits are of low energy ( $E \approx 0$ ) with semi-major axis length  $a \gg M_\bullet$  and the boundary of the loss cone is [64]

$$J_c(E \approx 0) = 4M_\bullet. \quad (1)$$

### B. Fokker-Planck equation

Statistical properties of stars can be described by their distribution functions  $f_i(t, \vec{r}, \vec{v})$  in the  $(\vec{r}, \vec{v})$  phase space, where  $i$  labels different star species. Following Refs. [65, 66], we approximate the distribution functions as  $f_i \approx f_i(t, E, R)$ , where  $R := J^2/J_c^2(E)$  is the normalized orbital angular momentum with  $J_c(E)$  being the maximum orbital angular momentum of a star with energy  $E$ . In order to relate the distribution function  $f(E, R)$  to the number density  $n(r)$ , and derive the Fokker-Planck equation, it is necessary to understand the properties of star orbits in given potential field  $\phi(r)$ , for which we summarize as follows [66]. The definition of energy suggests that

$$2(\phi - E) = v^2 = v_t^2 + v_r^2 = \frac{J^2}{r^2} + v_r^2, \quad (2)$$

where  $v_t$  and  $v_r$  are the tangential velocity and the radial velocity respectively. For a circular orbit of energy  $E$ , its orbit radius  $r_c(E)$  and angular momentum  $J_c(E)$  are determined by

$$\begin{aligned} J_c^2(E) &= -r_c^3 \phi'(r_c), \\ 2(\phi(r_c) - E) &= \frac{J_c^2}{r_c^2}. \end{aligned} \quad (3)$$

For a general non-circular orbit with parameters  $(E, R)$ , its turning points (apsis/periapsis)  $r_\pm$  are determined by

$$2(\phi(r_\pm) - E) = \frac{J^2}{r_\pm^2}, \quad (4)$$

and its orbit period  $P(E, R)$  is

$$P(E, R) = 2 \int_{r_-}^{r_+} \frac{dr}{v_r}. \quad (5)$$

Defining the particle number density in the  $(E, R)$  phase space as  $N(E, R)$ , with  $N(E, R)dEdR := \int_{r_-}^{r_+} d^3r d^3v f(E, R)$ , we have [65, 66]

$$\begin{aligned} N(E, R) &= 4\pi^2 P(E, R) J_c^2(E) f(E, R) \\ &:= C(E, R) f(E, R). \end{aligned} \quad (6)$$

The position-space particle number density  $n(r)$  can be expressed by the distribution function  $f(E, R)$  by [66]

$$n(r) = \frac{2\pi}{r^2} \int_0^{\phi(r)} dE J_c^2(E) \int_0^{R_{\max}} \frac{dR}{v_r} f(E, R), \quad (7)$$

where  $R_{\max}(r, E) = 2r^2(\phi(r) - E)/J_c^2(E)$ , and  $v_r(r, E, R) = 2(\phi - E) - J^2/r^2 = (R_{\max} - R)J_c^2(E)/r^2$ . In the case of thermal distribution  $f = f(E)$ , the above equation simplifies as [67]

$$n(r) = 4\pi \int_0^{\phi(r)} dE \sqrt{2(\phi(r) - E)} f(E). \quad (8)$$

With all these orbital properties, the Fokker-Planck equation governing the phase space evolution is written in the form of [33, 65, 66, 68, 69]

$$C \frac{\partial f}{\partial t} = -\frac{\partial}{\partial E} F_E - \frac{\partial}{\partial R} F_R, \quad (9)$$

with  $C = 4\pi^2 J_c^2 P$  being the weight function defined in Eq. (6) and  $F_{E,R}$  being the flux in the  $E/R$  direction:

$$\begin{aligned} -F_E &= C \left( D_{EE} \frac{\partial f}{\partial E} + D_{ER} \frac{\partial f}{\partial R} + D_E f \right), \\ -F_R &= C \left( D_{RR} \frac{\partial f}{\partial R} + D_{ER} \frac{\partial f}{\partial E} + D_R f \right). \end{aligned} \quad (10)$$

The diffusion and advection coefficients are functions of the distribution functions and are derived in Appendix A.

Given initial condition  $f(t = 0, E, R)$ , we evolve the cluster according to the Fokker-Planck equation (9) subject to following boundary conditions. On the  $E \rightarrow 0$  boundary, we fix the distributions to their initial values, i.e.,

$$f(t, E, R)|_{E \rightarrow 0} = f(t = 0, E, R)|_{E \rightarrow 0}, \quad (11)$$

considering the long relaxation timescale there. On the  $R = 1$  boundary, the flux in the  $R$  direction should vanish,

$$F_R|_{R \rightarrow 1} = 0. \quad (12)$$

On the loss cone boundary  $R = R_{lc}(E) := J_{lc}^2/J_c^2(E)$ , there are two different regimes: full loss cone regime where

$$y_{lc} := \frac{R_{lc}}{(D_{RR}/R)_{R \rightarrow 0} P} < 1, \quad (13)$$

and empty regime where  $y_{lc} > 1$ . In the empty regime, stars are expected to fall into the MBH within one orbital period  $P$  (that's why the phase space is empty). In the full regime, stars are in general scattered into/out of the loss cone multiple times within one orbital period, therefore the phase space

is full of stars and the rate of stars falling into the MBH is low. Quantitatively, the flux in the  $R$  direction was obtained in Ref. [65] as

$$-\frac{F_R}{C} = \left( \frac{D_{RR}}{R} \right) \Big|_{R \rightarrow 0} \frac{f(R_0)}{\ln(R_0/R_{lc}) + \mathcal{F}(y_{lc})}, \quad (14)$$

where  $R_0$  is any small  $R$  in the range of  $R_{lc} \leq R \ll 1$ ,  $\mathcal{F}(y_{lc}) \sim 1/y_{lc}$  for  $y_{lc} \lesssim 1$  and  $\mathcal{F}(y_{lc}) \simeq 0.824 y_{lc}^{-1/2}$  for  $y_{lc} \gtrsim 1$ . At  $R_0 = R_{lc}$ , the above equation simplifies as  $F_R(R_{lc}) = 0$  in the full regime, and  $f(R_{lc}) = 0$  in the empty regime.

As a result, the EMRI rate and the prompt infall rate per MBH via loss cone is given by

$$\begin{aligned} \Gamma_{\text{emri,lc}} &= \int_{E_{\text{gw}}}^{+\infty} \vec{F} \cdot d\vec{l}, \\ \Gamma_{\text{infl,lc}} &= \int_{\sigma^2}^{E_{\text{gw}}} \vec{F} \cdot d\vec{l}, \end{aligned} \quad (15)$$

where  $\vec{F} = (F_E, F_R)$ ,  $d\vec{l} = (dE, dR)$  is the line element along the boundary of the loss cone, and  $E_{\text{gw}}$  is the critical energy where  $t_{\text{gw}} = t_J$ . To calculate the GW emission timescale  $t_{\text{gw}}$ , we use a recently corrected version of Peters' time-scale that accounts for eccentricity evolution and post-Newtonian corrections [51–53, 70], with

$$t_{\text{gw}} = \frac{5a^4}{256M_\star^2 m} \frac{(1 - e^2)^{7/2}}{1 + \frac{73}{24}e^2 + \frac{37}{96}e^4} 8^{1 - \sqrt{1-e}} e^{\frac{5M_\star}{a(1-e)}}, \quad (16)$$

where  $m$  is the mass of the star orbiting around the MBH,  $a$  and  $e$  are the orbital semi-major axis and the eccentricity, respectively. For calculating the diffusion timescale  $t_J$  in the  $J$ -direction, we use the approximation [31]

$$t_J \approx \frac{J^2}{J_c^2(E)} t_E(E, R) = \frac{J^2}{J_c^2(E)} \frac{E^2}{2D_{EE}(E, R = 0)}. \quad (17)$$

### C. EMRI rate and prompt infall rate

We initialize the system with Tremaine's MBH+stellar cluster model [71, 72], assuming 3 star species in the stellar cluster: stars with mass  $m_{\text{star}}$ , mass-gap objects with mass  $m_{\text{mgo}}$  and heavy sBHs with mass  $m_{\text{sbh}}$ . The total star/mgo/sBH mass in the cluster are  $M_{\text{star}}$ ,  $M_{\text{mgo}}$  and  $M_{\text{sbh}}$ , respectively. Their number densities in the Tremaine's cluster model are specified by

$$\begin{aligned} n_{\text{star}}(r) &= \frac{M_{\text{star}}}{m_{\text{star}}} \frac{3 - \gamma}{4\pi} \frac{r_a}{r^\gamma (r + r_a)^{4-\gamma}}, \\ n_i(r) &= \delta_i \times n_{\text{star}}(r), \end{aligned} \quad (18)$$

with  $i$  the index labelling different star species,  $r_a$  the density transition radius,  $\gamma$  the density scaling power index, and  $\delta_i$  the abundance of species  $i$  relative to stars.

As an example, we initialize a stellar cluster with three different star species with  $m_i = (1, 3, 10)M_\odot$ ,  $\delta_i = (1, 10^{-3}, 10^{-3})$  around a MBH with  $M_\bullet = 4 \times 10^6 M_\odot$ . Note if there is no

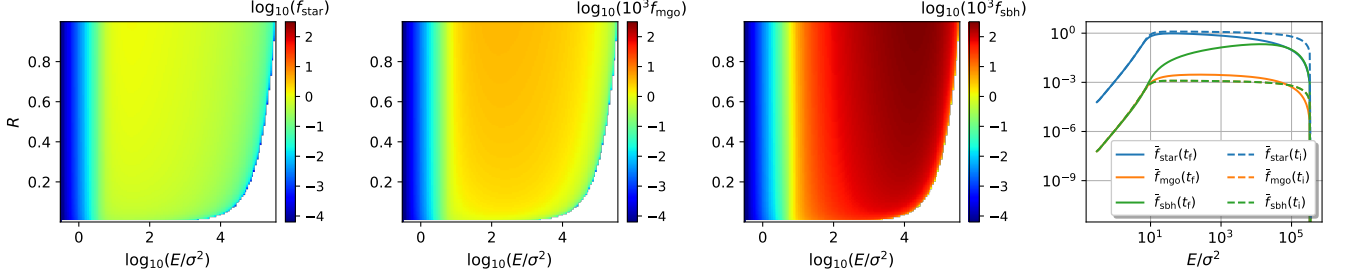


FIG. 1. In the fiducial stellar cluster around a MBH with  $M_\bullet = 4 \times 10^6 M_\odot$ , distribution functions  $f_i(E, R)$  ( $i = \text{star, mgo, sbh}$ ) at  $t = 2$  Gyr are shown in the first 3 panels. The 4th panel shows  $R$ -integrated distribution functions  $\bar{f}_i$  at  $t_i = 0$  (dashed lines) and at  $t_i = 2$  Gyr (solid lines). All the distribution functions are shown in units of  $10^5 \text{pc}^{-3} / (2\pi\sigma^2)^{3/2}$ .

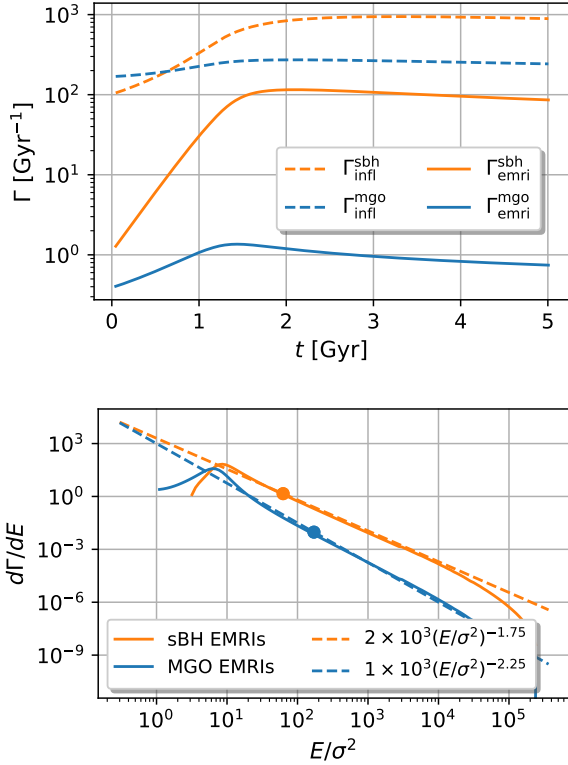


FIG. 2. Upper panel: EMRI rates  $\Gamma_{\text{emri,lc}}$  and prompt infall rates  $\Gamma_{\text{infl,lc}}$  of both sBHs and mass-gap objects in the loss-cone channel. Lower panel: differential rates  $d\Gamma/dE$  at  $t = 2$  Gyr, and the two dots are the critical energy  $E_{\text{gw}}^{\text{mgo,sbh}}$ .

gap in the mass spectrum of SN remnants with the power-law mass spectrum  $dN/dm \propto m^{-2.35}$  [73] holding in the whole mass range  $3M_\odot \leq m \leq 50M_\odot$ , we expect nearly equal number of MGOs and sBHs produced in SN explosions. The total star mass  $M_{\text{star}} = 20M_\bullet$ , the density transition radius  $r_a = 4r_h = 4M_\bullet/\sigma^2$  and the density power index  $\gamma = 1.5$ , where the star velocity dispersion  $\sigma$  satisfies the  $M_\bullet - \sigma$  rela-

tion [74, 75]

$$M_\bullet = 1.53 \times 10^6 \left( \frac{\sigma}{70 \text{ km/s}} \right)^{4.24}, \quad (19)$$

We evolve the cluster according to the Fokker-Planck equation (9) (see [44] for detailed numerical algorithm). In first 3 panels of Fig. 1, we show the distribution functions  $f_i(t, E, R)$  at  $t = 2$  Gyr, and in the 4th panel, we show the  $R$ -integrated functions  $\bar{f}_i(E) = \int_0^1 dR f_i(t, E, R)$  at  $t = 0$  and  $t = 2$  Gyr, respectively. From the 4th panel, we see sBHs (which are the most massive star component) concentrate around the MBH as a result of mass segregation, yielding a large increase in the distribution function for sBHs at small radii/large binding energy  $E$  with time [32, 34, 76], while little concentration is found for less massive MGOs or stars.

In the upper panel of Fig. 2, we show the EMRI rates and the prompt infall rates of both sBHs and MGOs, where  $\Gamma_{\text{emri}}^{\text{mgo}}$  is lower than  $\Gamma_{\text{emri}}^{\text{sbh}}$  by a factor of  $\mathcal{O}(10^2)$ , as a result of the stronger mass segregation and shorter GW emission timescale  $t_{\text{gw}}$  for sBHs. In contrast, the prompt infall rates are less affected by the mass segregation, because the prompt infall rate depends on the star density at lower energy ( $< E_{\text{gw}}$ ), while the EMRI rate depends on the star density at higher energy ( $> E_{\text{gw}}$ ) (Eq. 15), and the latter is more sensitive to the mass segregation (Fig. 1). As a result, we find the number of prompt infalls per EMRI  $N_p := \Gamma_{\text{infl}}/\Gamma_{\text{emri}}$  are  $N_p^{\text{sbh}} \approx 10$ ,  $N_p^{\text{mgo}} \approx 250$  for the fiducial model.

For an analytic understanding of these results, we re-estimate the number of prompt infalls per EMRI  $N_p$  using previous analytic formula [Eq. (17) and (26) in [31]],

$$\begin{aligned} \Gamma_{\text{emri,lc}} &= \int_0^{a_{\text{gw}}} \frac{4\pi a^2 n(a)}{\ln(R_{\text{lc}}) t_{\text{rlx}}(a)} da \sim a_{\text{gw}}^{1.5-2p}, \\ \Gamma_{\text{infl,lc}} &= \int_{a_{\text{gw}}}^{a_{\text{max}}} \frac{4\pi a^2 n(a)}{\ln(R_{\text{lc}}) t_{\text{rlx}}(a)} da \sim a_{\text{max}}^{1.5-2p}, \end{aligned} \quad (20)$$

where  $n(a) \sim a^{-1.5-p}$  is the number density,  $t_{\text{rlx}}(a) \sim a^p$  is the local relaxation timescale,  $a_{\text{max}}$  is a characteristic radius of the stellar cluster, and  $a_{\text{gw}}$  is the critical radius where  $t_{\text{gw}} = t_J$  [Eqs.(16,17)]. For a single-species cluster filled with stars of mass  $m$ , the dependence of  $a_{\text{gw}}$  on mass  $m$  and on the MBH



mass  $M_\bullet$  is [31, 50],

$$\frac{a_{\text{gw}}}{M_\bullet/\sigma^2} \propto m^{\frac{2}{3-2p}} M_\bullet^0. \quad (21)$$

For comparison with numerical results, we formulate the above analytic results in the phase space, with the analytic differential rates

$$d\Gamma/dE = d\Gamma/da \times da/dE \sim E^{2p-2.5}, \quad (22)$$

which can be directly compared with the numerical results of the fiducial model at  $t = 2$  Gyr (lower panel of Fig. 2). From the comparison, we see the power laws  $d\Gamma_{\text{mgo}}/dE \sim E^{-2.25}$  and  $d\Gamma_{\text{sbbh}}/dE \sim E^{-1.75}$  are a good approximation at  $E \gtrsim 10\sigma^2$ , i.e.,  $2p_{\text{mgo}} \approx 0.25$ ,  $2p_{\text{sbbh}} \approx 0.75$ . In terms of the differential rates, the EMRI rate and the prompt infall rates are written as

$$\begin{aligned} \Gamma_{\text{emri,lc}} &= \int_{E_{\text{gw}}}^{\infty} d\Gamma/dE dE, \\ \Gamma_{\text{infl,lc}} &= \int_{\sigma^2}^{E_{\text{gw}}} d\Gamma/dE dE, \end{aligned} \quad (23)$$

where the critical energy are numerically found as  $E_{\text{gw}}^{\text{sbbh}} = 63\sigma^2$ ,  $E_{\text{gw}}^{\text{mgo}} = 171\sigma^2$  and they are consistent with the analytic expectation  $E_{\text{gw}}^{\text{sbbh}}/E_{\text{gw}}^{\text{mgo}} \approx (m_{\text{sbbh}}/m_{\text{mgo}})^{\frac{2}{3-2p}}$  [Eq. (21)]. Using the power-law approximations to the differential rates, we have

$$N_p \approx (E_{\text{gw}}/E_{\text{min}})^{1.5-2p}, \quad (24)$$

where  $E_{\text{min}} \in (1, 10)\sigma^2$  is an effective minimum energy (lower panel of Fig. 2). As a result, we obtain an analytic estimate  $N_p^{\text{sbbh}} \in (4, 22)$  and  $N_p^{\text{mgo}} \in (35, 600)$ .

Similar to the fiducial model, we initialize the stellar cluster around a MBH with mass in the range of  $(10^5, 10^7)M_\odot$ , then evolve the system for  $T_0 = 5$  Gyr, and summarize the time-averaged EMRI rates  $\bar{\Gamma}_{\text{emri}}^{\text{sbbh,mgo}}$  and prompt infall rates  $\bar{\Gamma}_{\text{infl}}^{\text{sbbh,mgo}}$  in Fig. 3. We find the EMRI rates peak around  $M_\bullet = 10^6 M_\odot$ , because the rates are limited by the longer relaxation timescale of the stellar cluster around a heavier MBH, while the rates are limited by the lower number of sBHBs and MGOs in the stellar cluster around a lighter MBH [44]. We find the time-averaged EMRI rates are  $\bar{\Gamma}_{\text{emri}}^{\text{sbbh}} = O(10 - 10^2)$  Gyr $^{-1}$  and  $\bar{\Gamma}_{\text{emri}}^{\text{mgo}} = O(1)$  Gyr $^{-1}$ . For longer evolution time  $T_0$ , the rates decrease further because of the depletion of sBHBs and MGOs in the stellar cluster. On average, the number of prompt infalls per EMRI  $N_p$  are similar to in the fiducial model with  $N_p^{\text{sbbh}} \approx 10$ ,  $N_p^{\text{mgo}} \approx 250$ , except  $N_p(M_\bullet = 10^5 M_\odot)$  is lower by a factor  $\sim 2$ . The nearly independence of  $N_p$  on the MBH mass  $M_\bullet$  comes from the independence of  $a_{\text{gw}}$  (or equivalently  $E_{\text{gw}}$ ) on  $M_\bullet$  [Eq. (21,24)], while  $N_p^{\text{mgo,sbbh}}(M_\bullet = 10^5 M_\odot)$  are lower simply because MGOs/sBHBs around the lighter MBH are of lower number and are quickly depleted via the loss cone, consequently the critical energy  $E_{\text{gw}}$  decreases and  $N_p$  is reduced.

### III. WET AGN DISK CHANNEL

In the presence of an accretion disk around a MBH, the distributions of all different orbiting object species are affected

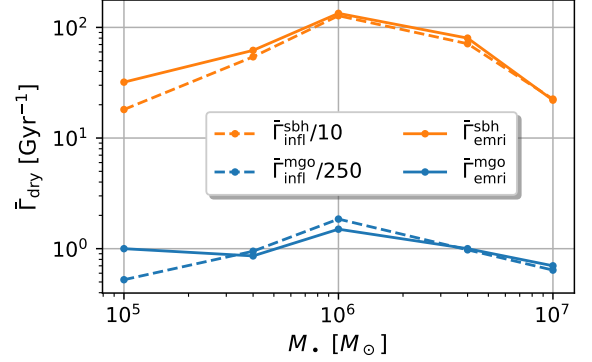


FIG. 3. Average EMRI rates  $\bar{\Gamma}_{\text{emri,dry}}$  and average prompt infall rates  $\bar{\Gamma}_{\text{infl,dry}}$  of both sBHBs and mass-gap objects in the loss-cone channel.

by the disk. As a result, the spherical symmetry is broken and the distribution function  $f(E, R, \mu)$  generally acquires dependence on the orbital inclination  $\iota$  w.r.t. the disk plane, where we have defined  $\mu := \cos \iota = \hat{\mathbf{J}} \cdot \hat{\mathbf{J}}_{\text{disk}}$ , with  $\hat{\mathbf{J}}$  and  $\hat{\mathbf{J}}_{\text{disk}}$  being the unit direction vectors of the star orbital angular momentum and the disk angular momentum, respectively.

For the problem we are considering, all stars can be conveniently decomposed as a cluster component and a disk component, i.e.,

$$f(E, R, \mu) \rightarrow f(E, \mu) + g(E)\delta(\mu - 1)\delta(R - 1), \quad (25)$$

where we have approximate the cluster-component distribution as  $R$ -independent and approximate the disk component as circular orbiters lying on the equator with  $\iota = 0$  ( $\mu = 1$ ), because the orbital eccentricity damping timescale is in general much shorter than the migration timescale (see subsection III A for details). With this decomposition, we have number density  $n(r, \theta)$  of the cluster-component stars and surface number density  $\Sigma(r)$  of the disk-component stars as

$$\begin{aligned} n(r, \theta) &= 4\pi \int_0^{\phi(r)} dE \sqrt{2(\phi(r) - E)} \bar{f}(E, \theta), \\ \Sigma(r) &= 2\pi^2 r E \sqrt{2(\phi(r) - E)} g(E) \Big|_{r=r_c(E)}, \end{aligned} \quad (26)$$

with

$$\bar{f}(E, \theta) = \frac{1}{2\pi} \int_0^{2\pi} d\eta f(E, \mu = \sin \theta \cos \eta), \quad (27)$$

where  $\theta$  is the polar angle w.r.t. to the  $\hat{\mathbf{J}}_{\text{disk}}$  direction.

In the remaining part of this section, we will first summarize the important interactions between stars and the accretion disk, then incorporate these interactions into the Fokker-Planck equation, and finally evolve the stellar cluster to calculate the wet EMRI rates.

#### A. Star-disk interactions

Interactions of an accretion disk with stars, MGOs and sBHBs are similar in aspects of density waves and dynamical

cal friction. In term of gas accretion onto stars, MGOs and sBHs within the AGN disk, the star size makes a difference. Compact objects of relatively small sizes only grow mildly within the AGN disk, while the stellar evolution is expected to be impacted by gas accretion onto stars of much larger sizes [77, 78]. In this work, we are not intended to model the star evolution in detail, and we simply assume no mass change of all the star species during the evolution period.

For illustration purpose, we take sBHs as an example. As a sBH orbits around the central MBH surrounded by an accretion disk, its periodic motion generates density waves, which in turn drive the star to migrate inward, damp its orbital eccentricity  $e$  and its orbital inclination  $\iota$ . For a highly inclined orbiter, the density wave effects become subdominant with respect to dynamical friction as it goes through the gas disk. The two effects (density waves and dynamical friction) together contribute to the advection in  $(E, \mu)$  space as [60, 61]

$$\begin{aligned} \langle \Delta \mu \rangle_t^{\text{disk}} &= (1 - \mu^2) \frac{\iota}{\sin \iota} \times \min. \left\{ \frac{0.544}{t_{\text{wav}}}, \frac{1.46}{t_{\text{wav}}} \frac{h^4}{\iota \sin^3(\iota/2)} \right\}, \\ \langle \Delta E \rangle_t^{\text{disk}} &= E \times \min. \left\{ \frac{2.7 + 1.1\alpha_s}{t_{\text{mig,I}}}, \frac{8.8}{t_{\text{mig,I}}} \frac{h^2}{\sin(\iota) \sin(\iota/2)} \right\}, \end{aligned} \quad (28)$$

with  $\langle \Delta X \rangle_t := \frac{\Delta X}{\Delta t} |_{\Delta t \rightarrow 0}$ ,  $\alpha_s := d \ln \Sigma / d \ln r$  and

$$t_{\text{mig,I}} = \frac{M}{m} \frac{M}{\Sigma r^2} h^2, t_{\text{wav}} = t_{\text{mig,I}} h^2, \quad (29)$$

where  $m$  is the mass of the orbiter,  $M$  is the total mass within radius  $r$ ,  $h$  and  $\Sigma$  are disk scale height and the disk surface density [79].

For a sBH captured into the disk, its orbital eccentricity will be damped by the eccentricity density waves on timescale  $t_{\text{wav}}$ , which is generally much shorter than all other relevant timescales, including the migration timescale  $t_{\text{mig,I}}$  and two diffusion timescales,  $E^2/D_{EE}$  and  $(1 - \mu^2)/D_{\mu\mu}$ . As a result, sBHs in the AGN disk are generally moving in circular orbits. For a sBH embedded in the gas disk, surrounding gas tends to flow towards it nearly in the radial direction at large distances, in the rest frame of the sBH. Due to the differential rotation of the gas disk, the inflowing gas generally carries non-zero angular momentum relative to the sBH, consequently circularizes and forms a disk around the sBH. The gas inflow rate at the outer boundary is usually super-Eddington and a strong outflow naturally emerges. As a result, a major part of the inflowing gas may escape as outflow and only the remaining part is accreted by the sBH (see Ref. [77] and references therein for detailed modeling). Because of the circularization process, it is reasonable to expect that the outflow carries minimal net momentum with respect to the sBH, and the momentum carried by the inflow eventually transfers to the sBH. The head wind contributes to the advection in the  $E$ -direction as

$$\langle \Delta E \rangle_t^{\text{wnd}} = \frac{2J}{\dot{J}_{\text{wnd}}} \quad (\text{for in-disk orbiters}), \quad (30)$$

where  $\dot{J}_{\text{wnd}}$  is sBH angular momentum loss rate due to the head wind (see Ref. [44] for calculation details).

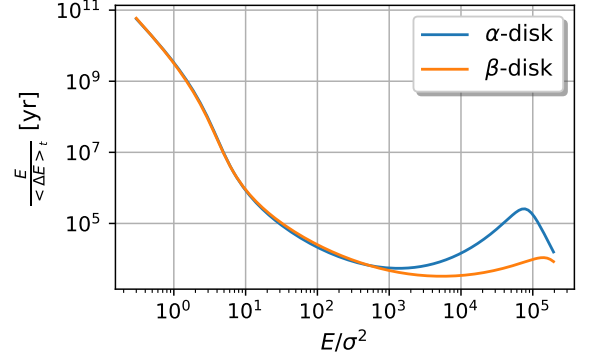


FIG. 4. The migration timescales of a  $10M_{\odot}$  BH embedded in the fiducial  $\alpha$  and  $\beta$  disks, respectively, where the GW emission becomes dominant at  $E/\sigma^2 \sim 10^5$  or equivalently  $a \sim 10^2 M_{\bullet}$ .

GW emission only becomes important when the orbiter is very close to the MBH and it drives an advection in the  $E$  direction for a circular orbiter as [70]

$$\langle \Delta E \rangle_t^{\text{gw}} = \frac{64}{5} \frac{M^2 m}{a^4} E. \quad (31)$$

Accretion disk structure of AGNs has not been well understood especially in the outer parts, where both disk heating mechanism and the angular momentum transport mechanism are not clear. Three commonly used AGN disk models ( $\alpha/\beta$  disks [80], and TQM disk proposed by Thompson, Quataert, and Murray [81]) have been numerically solved and compared in our previous work [44]. In this work, we will use  $\alpha$  and  $\beta$  disks with accretion rate  $\dot{M}_{\bullet} = 0.1\dot{M}_{\bullet, \text{Edd}}$  as fiducial disk models (Fig. 5 in [44]). In Fig. 4, we show the migration timescales  $E/\langle \Delta E \rangle_t := E/(\langle \Delta E \rangle_t^{\text{disk}} + \langle \Delta E \rangle_t^{\text{wnd}} + \langle \Delta E \rangle_t^{\text{gw}})$  of a sBH with  $m_{\text{sBH}} = 10M_{\odot}$  embedded in the two fiducial accretion disks. The two disks only differs in inner parts where the radiation pressure dominates over the gas pressure and the two migration timescales only differs where  $E/\sigma^2 \gtrsim 10^3$ . We do not include the TQM disk model in this work because a much more efficient angular momentum transport mechanism is assumed in the TQM disk model, which is inconsistent with the turbulence viscosity driven by magnetorotational instability in inner parts of the accretion disk [82–84]. Due to the high efficiency of the angular momentum transport assumed in the TQM disk model, TQM disks are in general of lower surface density, therefore longer migration timescale  $t_{\text{mig,I}}$ , which hinders sBHs and MGOs from migrating to the vicinity of the central MBH and forming EMRIs if  $T_{\text{disk}} < t_{\text{mig,I}}$ . If  $T_{\text{disk}} > t_{\text{mig,I}}$ , the EMRI rates in TQM disks are similar to those in  $\alpha/\beta$ -disks (see [44] for details).

## B. Fokker-Planck equation

For cluster-component stars, the Fokker-Planck equation takes the form

$$C_\mu \frac{\partial f}{\partial t} = -\frac{\partial}{\partial E} F_E - \frac{\partial}{\partial \mu} F_\mu, \quad (32)$$

with flux

$$\begin{aligned} -F_E &= C_\mu \left( D_{EE} \frac{\partial f}{\partial E} + D_{E\mu} \frac{\partial f}{\partial \mu} + D_E f \right), \\ -F_\mu &= C_\mu \left( D_{\mu\mu} \frac{\partial f}{\partial \mu} + D_{E\mu} \frac{\partial f}{\partial E} + D_\mu f \right), \end{aligned} \quad (33)$$

and the weight function

$$C_\mu(E) := \frac{1}{2} \int_0^1 C(E, R) dR, \quad (34)$$

where the factor 2 comes from  $\int_{-1}^1 d\mu$ . All the coefficients of the Fokker-Planck equation (32) are contributed by star interactions with the accretion disk and by scatterings with both the cluster-component stars and the disk-component stars, where the first only contributes to the advection coefficients as

$$D_E = -(\langle \Delta E \rangle_t^{\text{dsk}} + \langle \Delta E \rangle_t^{\text{gw}}), \quad D_\mu = -\langle \Delta \mu \rangle_t^{\text{dsk}}. \quad (35)$$

and the latter two contributions are given in Eqs.(B16,B17), respectively.

With proper initial conditions, we evolve the system according to Eq. (32) with the following boundary conditions. On the  $E \rightarrow 0$  boundary, we again fix the distribution, i.e.,

$$f(t, E, R)|_{E \rightarrow 0} = f(t = 0, E, R)|_{E \rightarrow 0}. \quad (36)$$

On the  $E \rightarrow E_{\text{max}}$  boundary, where the evolution of distribution function is dominated by GW emission, we set

$$F_E|_{E \rightarrow E_{\text{max}}} = -C_\mu D_E f. \quad (37)$$

On the  $\mu = -1$  boundary, we use the zero-flux condition

$$F_\mu|_{\mu \rightarrow -1} = 0. \quad (38)$$

On the  $\mu = 1$  boundary, where the evolution of distribution function is dominated by the inclination damping arising from the normal density waves, we set

$$F_\mu|_{\mu \rightarrow 1} = -C_\mu D_\mu f. \quad (39)$$

For disk-component stars, the Fokker-Planck equation reduces to be 1-dimensional,

$$C_\mu \frac{\partial g}{\partial t} = \frac{\partial}{\partial E} \left[ C_\mu \left( D_{EE} \frac{\partial g}{\partial E} + D_{EG} \right) \right] + F_\mu(E, \mu = 1), \quad (40)$$

with the source term arising from disk-component stars captured by the disk. We also find the evolution of disk-component stars is dominated by the star-disk interactions and

GW emission [Eqs. (28,30,31)], so we simply neglect the contributions from scatterings, i.e.,

$$D_E = -(\langle \Delta E \rangle_t^{\text{dsk}} + \langle \Delta E \rangle_t^{\text{gw}} + \langle \Delta E \rangle_t^{\text{wnd}}), \quad D_{EE} = 0. \quad (41)$$

Now Eq. (40) is a first-order differential equation, which requires only one boundary condition, and we choose it as

$$g(t, E, R)|_{E \rightarrow 0} = g(t = 0, E, R)|_{E \rightarrow 0}. \quad (42)$$

The wet EMRI rate is determined by the flux in the  $E$  direction at the  $E_{\text{max}}$  boundary, i.e.,

$$\Gamma_{\text{emri,disk}} = -C_\mu D_E g|_{E=E_{\text{max}}}. \quad (43)$$

Strictly speaking, the cluster-component contribution should also be included in addition to the disk-component contribution. As we will see later, the disk-component contribution at the  $E_{\text{max}}$  boundary turns out to be dominant.

## C. Wet EMRI rates

Considering that the accretion history of MBHs is likely episodic [85, 86] and the active phase of an MBH is in general much shorter than its quiet phase [54, 55, 87], we simplify the duty cycle of an AGN as a long quiet phase of  $T_0 = 5$  Gyr followed by a short active phase of duration  $T_{\text{disk}} \in \{10^6, 10^7, 10^8\}$  yr. This simplified picture holds if the cluster evolution driven by purely two-body scatterings is negligible during a quiet phase between two active episodes, so that adjacent episodes can be effectively glued together as we understand the history of evolution.

For calculating the wet EMRI rate, we initialize the stellar cluster around a MBH and evolve the system for  $T_0$  in the same way as in the previous section, then turn on the accretion disk and initialize the cluster-component and disk-component distributions as

$$\begin{aligned} f_i(t = 0, E, \mu) &= 0.99 \int_0^1 f_i(t = T_0, E, R) dR, \\ g_i(t = 0, E) &= 0.01 \int_0^1 f_i(t = T_0, E, R) dR, \end{aligned} \quad (44)$$

where the integrand  $f_i(t = T_0, E, R)$  is the distribution function at the end of the quiet phase, and the initial fraction 0.01 is the typical disk aspect ratio [44], the exact value of which does not matter because the disk-component star densities are mostly determined by subsequent capture and migration processes. With this initialization, we continue the evolution of  $f_i(t, E, \mu)$  and  $g_i(t, E)$  according to Eqs.(32,40) for a duration  $T_{\text{disk}}$ .

As a fiducial model of the wet channel, we again consider a fiducial model with a MBH and a stellar cluster same to in the previous section, and a fiducial  $\alpha$ -disk with lifetime  $T_{\text{disk}} = 10^8$  yr (Fig. 4). In the first 3 panels of Fig. 5, we show the cluster-component distribution functions  $f_i(E, \mu)$  at  $t = 10^7$  yr, where we see low-inclination ( $\mu \rightarrow 1$ ) orbiters have been captured into the disk, therefore the distribution function  $f_i(E, \mu \rightarrow 1)$  is relatively lower. In the 4th panel, we show

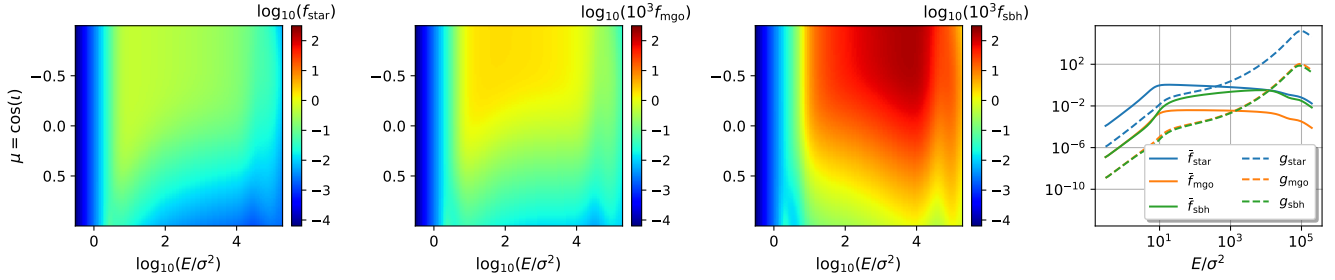


FIG. 5. In the fiducial stellar cluster around a MBH with  $M_\bullet = 4 \times 10^6 M_\odot$ , the cluster-component distribution functions  $f_i(E, \mu)$  ( $i = \text{star, mgo, sbh}$ ) at  $t = 10^7$  yr are shown in the first 3 panels. The 4th panel shows  $\mu$ -integrated cluster-component distribution functions  $\bar{f}_i(E)$  (solid lines) and disk-component  $g_i(E)$  (dashed lines). All the distribution functions are shown in units of  $10^5 \text{pc}^{-3} / (2\pi\sigma^2)^{3/2}$ .

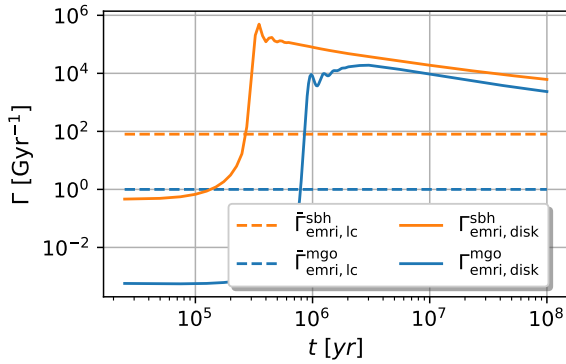


FIG. 6. Wet EMRI rates  $\Gamma_{\text{emri,disk}}(t)$  of sBHs and MGOs. For comparison, we also plot the average dry EMRI rates  $\bar{\Gamma}_{\text{emri,lc}}$  during the quiet phase as horizontal lines (Fig. 3).

the  $\mu$ -integrated distribution functions  $\bar{f}_i(E) = \int_{-1}^1 d\mu f_i(E, \mu)$  and the disk-component functions  $\bar{g}_i(E)$  at  $t = 10^7$  yr. The disk component dominates over the cluster component in the vicinity of the MBH (large  $E$ ), and the disk component peaks around  $E \sim 10^5 \sigma^2$  where the migration timescale peaks (Fig. 4). It is interesting to note that the disk-component distribution functions  $g_i(E)$  are proportional to their abundance  $\delta_i$  with no dependence on different star masses  $m_i$ , simply because both the rate of stars captured onto the disk and the migration rate of stars along the disk are proportional to their masses  $m_i$ , and the local density  $g_i(E)$  is determined by the ratio of the two rates in the equilibrium state.

In Fig. 6, we show the wet EMRI rates for both sBHs and MGOs as functions of time. Initially both of them are low because it takes sometime for sBHs and MGOs that are captured by the disk to migrate to the MBH; and the two rates peaks around  $t = 3 \times 10^5$  yr and  $t = 10^6$  yr, respectively, because the migration timescale is inversely proportional to the star mass  $m_i$ . After the peak time, the rate of stars captured by the disk is in equilibrium with the corresponding EMRI rate, and both of them steadily decrease with time  $\propto t^{-0.5}$ , which is the typical behavior of diffusion processes with absorbing boundary conditions. We find the wet EMRI rate of sBHs in the equilib-

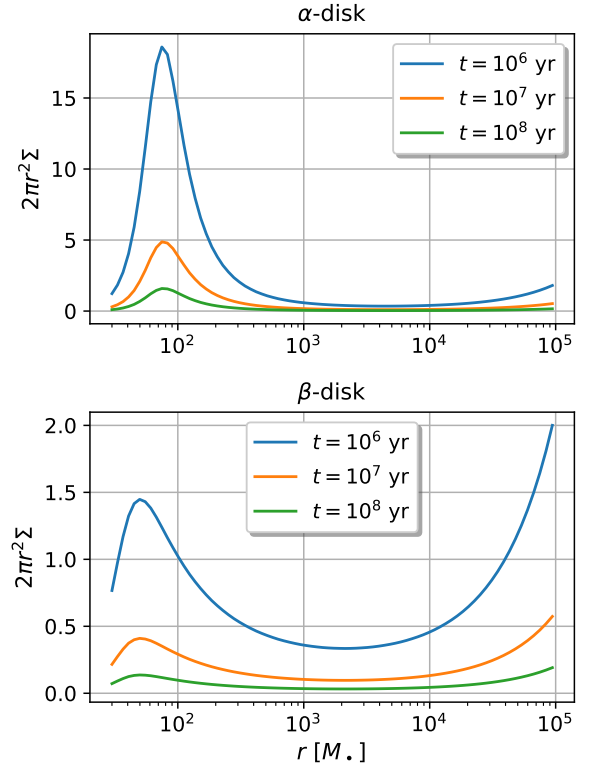


FIG. 7. The surface number density  $\Sigma(r)$  of disk-component sBHs at different times for the two different AGN disk models.

rium state is higher than the dry EMRI rate by  $\mathcal{O}(10^2 - 10^3)$  for the fiducial model, which is consistent with the result of the slow disk capture scenario in our previous work (Fig. 10 of Ref. [44]), though the detailed time dependence does not match exactly because we approximated the effect of density waves as inclination-independent and neglected the contribution from dynamical friction in the previous work, while we have used more refined prescription of star-disk interaction [Eq. (28)] and self-consistent calculation of the disk capture rate [Eq. (40)] in this work. In the loss-cone channel, the EMRI rate of MGOs is largely suppressed compared to sBHs



because MGOs benefit little from the mass segregation (see Figs. 1 and 2), while in the disk channel, the EMRI rate of MGOs is lower than that of sBHs by a factor of  $\sim m_{\text{sBH}}/m_{\text{MGO}}$  in the equilibrium state (Fig. 6). As a result, we find the wet EMRI rate of MGOs in the equilibrium state is higher than the dry rate by  $O(10^3 - 10^4)$  for the fiducial model.

As shown in Fig. 4, the migration timescale of a sBH in an AGN disk peaks around  $r \sim 10^2 M_\bullet$ , where the GW emission start to dominate over other processes. As a result, a number of sBHs aggregate around this radius due to a traffic jam (4th panel of Fig. 5). For clarity, we show the surface number density of disk-component sBHs  $\Sigma$  in Fig. 7 at different times. Compared with the  $\beta$ -disk, more sBHs aggregate around  $r \sim 10^2 M_\bullet$  in the fiducial  $\alpha$ -disk because of slower migration speed. A similar result (with slightly higher surface number densities) was obtained in our previous work [77] though the disk capture rate was not calculated from first principles.

Similar to the fiducial model, we initialize a stellar cluster of a MBH with mass in the range of  $(10^5 - 10^7) M_\odot$ , evolve the MBH+cluster system for  $T_0 = 5$  Gyr, then turn on an accretion disk around the MBH and initialize the cluster-component and disk-component distributions, and continue the evolution for  $T_{\text{disk}}$ . The average wet EMRI rates of sBHs and MGOs

$$\bar{\Gamma}_{\text{wet}} = \frac{1}{T_{\text{disk}}} \int_0^{T_{\text{disk}}} \Gamma_{\text{emri,disk}} dt, \quad (45)$$

are summarized in Table I. For long disk lifetime  $T_{\text{disk}} \gtrsim 10^7$  yr, the wet EMRI rates of sBHs and MGOs are higher for heavier MBHs because more sBHs and MGOs are available around heavier MBHs. For short disk lifetime  $T_{\text{disk}} = 10^6$  yr, the trend reverses for  $M_\bullet \gtrsim 10^6 M_\odot$  because longer migration timescale in AGN disks ( $\gtrsim T_{\text{disk}}$ ) around heavier MBHs hinders the sBHs and MGOs captured by the AGN disk from migrating to the vicinity of the MBH and becoming EMRIs. In comparison with the dry channel, we find the wet EMRI rate of sBHs is usually higher by  $O(10^1 - 10^3)$  and the wet EMRI rate of MGOs is usually higher by  $O(10^3 - 10^4)$ .

#### IV. DETECTION PROSPECTS

In addition to the generic EMRI rates per MBH/AGN obtained in the previous two sections, a few extra pieces of information are needed for calculating the LISA detectable EMRI rate: the mass function of MBHs  $dN_\bullet/dM_\bullet$  and the fraction of MBHs living a stellar cusp which is supposed to be destroyed during a MBH merger following a previous galaxy merger.

Following Ref. [37], we consider two MBH mass functions in the range of  $(10^4, 10^7) M_\odot$ ,

$$\begin{aligned} f_{\bullet,-0.3} : \frac{dN_\bullet}{d \log M_\bullet} &= 0.01 \left( \frac{M_\bullet}{3 \times 10^6 M_\odot} \right)^{-0.3} \text{ Mpc}^{-3}, \\ f_{\bullet,+0.3} : \frac{dN_\bullet}{d \log M_\bullet} &= 0.002 \left( \frac{M_\bullet}{3 \times 10^6 M_\odot} \right)^{+0.3} \text{ Mpc}^{-3}, \end{aligned} \quad (46)$$

where  $f_{\bullet,-0.3}$  is an approximation to the mass function in the model assuming MBHs are seeded by Pop-III stars and grow via accretion and mergers [88], and  $f_{\bullet,+0.3}$  is purely a phenomenological model [89].

In the frame of observers on the earth, the differential dry and wet EMRI rates are written as

$$\begin{aligned} \frac{d^2 \mathcal{R}_{\text{dry}}}{dM_\bullet dz} &= \frac{1}{1+z} \frac{dN_\bullet}{dM_\bullet} \frac{dV_c(z)}{dz} C_{\text{cusp}}(M_\bullet, z) \bar{\Gamma}_{\text{dry}}(M_\bullet; N_p), \\ \frac{d^2 \mathcal{R}_{\text{wet}}}{dM_\bullet dz} &= \frac{f_{\text{AGN}}}{1+z} \frac{dN_\bullet}{dM_\bullet} \frac{dV_c(z)}{dz} C_{\text{cusp}}(M_\bullet, z) \bar{\Gamma}_{\text{wet}}(M_\bullet; \mathbb{M}), \end{aligned} \quad (47)$$

where  $z$  is cosmological redshift,  $V_c(z)$  is the comoving volume of the universe up to redshift  $z$ ,  $C_{\text{cusp}}(M_\bullet, z)$  is the fraction of MBHs embedded in a stellar cusp, where we use the same  $C_{\text{cusp}}(M_\bullet, z)$  function as in [37] for models with MBH mass function  $f_{\bullet,-0.3}$  and take  $C_{\text{cusp}}(M_\bullet, z) = 1$  for models with MBH mass function  $f_{\bullet,+0.3}$ .  $\bar{\Gamma}_{\text{wet}}(M_\bullet; \mathbb{M})$  is the average wet EMRI rate (Table I) and we conservatively take the AGN fraction as  $f_{\text{AGN}} = 1\%$ .

In consistent with Ref. [37], we parametrize the average dry EMRI rate of sBHs as

$$\bar{\Gamma}_{\text{dry}}^{\text{sBH}}(M_\bullet; N_p) = C_{\text{dep}}(M_\bullet; N_p) C_{\text{grow}}(M_\bullet; N_p) \Gamma_{\text{lc}}^{\text{sBH}}(M_\bullet), \quad (48)$$

where we take the number of prompt infalls per EMRI as  $N_p = 10$  (Fig. 2), the generic rate

$$\Gamma_{\text{lc}}^{\text{sBH}}(M_\bullet) = \Gamma_0 \left( \frac{M_\bullet}{10^6 M_\odot} \right)^{-0.19}, \quad (49)$$

with  $\Gamma_0 \in (30, 300) \text{ Gyr}^{-1}$  [37, 45], and two correction factors are correction from possible depletion of sBHs available  $C_{\text{dep}}(M_\bullet; N_p)$  and correction capping the MBH mass growth via accreting sBHs (from both prompt infalls and EMRIs)  $C_{\text{grow}}(M_\bullet; N_p)$ , respectively. Though there is uncertainty of a factor of  $O(10)$  in the generic rate  $\Gamma_{\text{lc}}^{\text{sBH}}$ , we will see that the uncertainty does not propagate to the average rate  $\bar{\Gamma}_{\text{dry}}^{\text{sBH}}$  with the two corrections. The depletion correction is formulated as

$$C_{\text{dep}} = \min. \left\{ \frac{T_{\text{dep}}}{T_{\text{rlx}}}, 1 \right\}, \quad (50)$$

where  $T_{\text{dep}}$  is the depletion timescale of sBHs residing in the MBH influence sphere ( $r_c = 2M_\bullet/\sigma^2$ )

$$T_{\text{dep}} = \frac{\Sigma m_{\text{sBH}}}{(1 + N_p) \Gamma_{\text{lc}}^{\text{sBH}} m_{\text{sBH}}}, \quad (51)$$

assuming the total mass of sBHs in the influence sphere is  $\Sigma m_{\text{sBH}} = 0.06 M_\bullet$ , and the relaxation timescale at  $r = r_c$  is

$$T_{\text{rlx}} = \left( \frac{\sigma}{20 \text{ km/s}} \right) \left( \frac{r_c}{1 \text{ pc}} \right)^2 \text{ Gyr}. \quad (52)$$

The MBH growth correction comes from requiring the MBH mass grows no more than  $1/e$  via accreting sBHs,

$$C_{\text{grow}} = \min. \left\{ e^{-1} \frac{M_\bullet}{\Delta M_\bullet}, 1 \right\}, \quad (53)$$

TABLE I. Average EMRI rates ( $\text{Gyr}^{-1}$ ) of sBHs and MGOs per AGN, assuming sBHs and MGOs are produced from SN explosion with equal numbers.

AGN disk	$M_{\bullet}/M_{\odot}$	$(\bar{\Gamma}_{\text{wet}}^{\text{mgo}}, \bar{\Gamma}_{\text{wet}}^{\text{sbh}})_{T_{\text{disk}}=10^6\text{yr}}$	$(\bar{\Gamma}_{\text{wet}}^{\text{mgo}}, \bar{\Gamma}_{\text{wet}}^{\text{sbh}})_{T_{\text{disk}}=10^7\text{yr}}$	$(\bar{\Gamma}_{\text{wet}}^{\text{mgo}}, \bar{\Gamma}_{\text{wet}}^{\text{sbh}})_{T_{\text{disk}}=10^8\text{yr}}$
$\alpha$ -disk	$1 \times 10^7$	$(0.1 \times 10^3, 4.1 \times 10^4)$	$(1.2 \times 10^4, 4.9 \times 10^4)$	$(6.7 \times 10^3, 1.7 \times 10^4)$
	$4 \times 10^6$	$(0.7 \times 10^3, 9.0 \times 10^4)$	$(1.2 \times 10^4, 3.7 \times 10^4)$	$(4.7 \times 10^3, 1.2 \times 10^4)$
	$1 \times 10^6$	$(1.0 \times 10^4, 4.4 \times 10^4)$	$(7.4 \times 10^3, 1.3 \times 10^4)$	$(2.5 \times 10^3, 4.3 \times 10^3)$
	$4 \times 10^5$	$(1.1 \times 10^4, 1.7 \times 10^4)$	$(4.3 \times 10^3, 4.5 \times 10^3)$	$(1.5 \times 10^3, 1.4 \times 10^3)$
	$1 \times 10^5$	$(2.7 \times 10^3, 1.6 \times 10^3)$	$(1.1 \times 10^3, 0.6 \times 10^3)$	$(0.5 \times 10^3, 0.3 \times 10^3)$
$\beta$ -disk	$1 \times 10^7$	$(4.6 \times 10^3, 9.3 \times 10^4)$	$(1.3 \times 10^4, 5.1 \times 10^4)$	$(6.7 \times 10^3, 1.7 \times 10^4)$
	$4 \times 10^6$	$(7.3 \times 10^3, 1.1 \times 10^5)$	$(1.3 \times 10^4, 3.8 \times 10^4)$	$(4.7 \times 10^3, 1.2 \times 10^4)$
	$1 \times 10^6$	$(1.5 \times 10^4, 4.8 \times 10^4)$	$(7.6 \times 10^3, 1.4 \times 10^4)$	$(2.5 \times 10^3, 4.3 \times 10^3)$
	$4 \times 10^5$	$(1.2 \times 10^4, 1.8 \times 10^4)$	$(4.4 \times 10^3, 4.9 \times 10^3)$	$(1.5 \times 10^3, 1.4 \times 10^3)$
	$1 \times 10^5$	$(3.7 \times 10^3, 3.2 \times 10^3)$	$(1.1 \times 10^3, 0.8 \times 10^3)$	$(0.5 \times 10^3, 0.3 \times 10^3)$

TABLE II. Forecasted Total and LISA detectable (with  $\text{SNR} \geq 20$ ) EMRI rates of sBHs and MGOs in the redshift range  $0 < z < 4.5$  assuming sBHs and MGOs are produced from SN explosions with equal numbers. For the wet channel, we have assumed a conservative AGN fraction  $f_{\text{AGN}} = 1\%$  throughout the universe.

Wet EMRIs	$f_{\bullet}$	AGN disk	$T_{\text{disk}}$ [yr]	Total rates of (MGOs, sBHs) [ $\text{yr}^{-1}$ ]	LISA detectable rates of (MGOs, sBHs) [ $\text{yr}^{-1}$ ]
	$f_{\bullet,-0.3}$	$\alpha$ -disk	$10^6$	(1900, 6400)	(50, 480)
			$10^7$	(1400, 2500)	(24, 130)
			$10^8$	(540, 860)	(10, 54)
		$\beta$ -disk	$10^6$	(2700, 8200)	(65, 530)
			$10^7$	(1400, 2600)	(24, 150)
			$10^8$	(540, 860)	(10, 54)
	$f_{\bullet,+0.3}$	$\alpha$ -disk	$10^6$	(110, 1000)	(3, 34)
			$10^7$	(180, 470)	(1, 10)
			$10^8$	(71, 160)	(< 1, 3)
		$\beta$ -disk	$10^6$	(200, 1300)	(5, 38)
			$10^7$	(190, 500)	(2, 11)
			$10^8$	(71, 160)	(< 1, 3)
Dry EMRIs	$f_{\bullet}$			Total rates of (MGOs, sBHs) [ $\text{yr}^{-1}$ ]	LISA detectable rates of (MGOs, sBHs) [ $\text{yr}^{-1}$ ]
	$f_{\bullet,-0.3}$			(79, 1300)	(1, 120)
	$f_{\bullet,+0.3}$			(3, 130)	(< 1, 10)

with

$$\Delta M_{\bullet} = m_{\text{sBH}}(1 + N_{\text{p}})C_{\text{dep}}(M_{\bullet}, N_{\text{p}})\Gamma_{\text{lc}}^{\text{sBH}}(M_{\bullet}, N_{\text{p}})T_{\text{emri}}(M_{\bullet}), \quad (54)$$

and

$$T_{\text{emri}} = \int dt \frac{dt}{dz} C_{\text{cusp}}(M_{\bullet}, z) \quad (55)$$

is the total duration when a MBH lives in a stellar cusp.

With these two corrections, we find the average EMRI rate of sBHs is well fitted by

$$\begin{aligned} & \bar{\Gamma}_{\text{dry}}^{\text{sBH}}(M_{\bullet}; N_{\text{p}} = 10) \\ & = \min. \left\{ 26 \left( \frac{M_{\bullet}}{10^5 M_{\odot}} \right), 30 \left( \frac{M_{\bullet}}{10^6 M_{\odot}} \right)^{-0.19} \right\} \text{Gyr}^{-1}. \end{aligned} \quad (56)$$

with little dependence on the generic rate  $\Gamma_0$  as long as it is higher than  $30 \text{ Gyr}^{-1}$ , because the average rate is in fact determined by the number of sBHs available around MBHs and the MBH growth limit via accreting sBHs. For the average EMRI rate of MGOs, we simply take it as  $\bar{\Gamma}_{\text{dry}}^{\text{mgo}}(M_{\bullet}) \approx 1 \text{ Gyr}^{-1}$  (Fig. 3).

With all the elements for calculating the differential EMRI rates ready [Eq. (47)], we calculate the total EMRI rates of

MGOs and sBHs from the two channels and the LISA detectable EMRI rates. We first sample the EMRI sources according to the differential rates [Eq. (47)], then compute the EMRI waveform using the Augment Analytic Kludge [90–92] and the expected signal-to-ratio (SNR) by the LISA detector (see all the source sampling and SNR computation details in the previous work [45]). The forecast results are listed in Table II. For the well-motivated MBH mass function  $f_{\bullet,-0.3}$ , we expect LISA to detect  $\sim 1$  mass-gap EMRIs,  $\mathcal{O}(10^2)$  sBH EMRIs from the dry channel,  $\mathcal{O}(10 - 10^2) \times (f_{\text{AGN}}/1\%)$  mass-gap EMRIs, and  $\mathcal{O}(10^2 - 10^3) \times (f_{\text{AGN}}/1\%)$  sBH EMRIs from the wet channel per year. For the less optimistic MBH mass function  $f_{\bullet,+0.3}$ , the expected detection numbers are overall lower by a factor of  $\mathcal{O}(10)$ .

## V. SUMMARY AND DISCUSSION

### A. Summary

In the dry EMRI formation channel, the formation rate of mass-gap EMRIs is strongly suppressed compared with EMRIs of sBHs, because sBHs are heavier and accumulate closer

to MBH due to the mass segregation effect and therefore easier to form EMRIs (Fig. 2). In the wet channel, the EMRI formation turns out to be much more efficient than in the dry channel because the capture of compact objects onto the accretion disk and subsequent inward migration along the disk are highly efficient in transporting compact objects [44, 45]. Both the capture rate onto to disk and the migration speed along the disk are linearly proportional to the compact object mass, so that the formation rate of wet mass-gap EMRIs in the equilibrium state is suppressed by a factor of  $\sim m_{\text{mgo}}/m_{\text{sbb}}$  assuming their abundances are equal ( $\delta_{\text{sbb}} = \delta_{\text{mgo}}$ ). As a result, we find the wet EMRI rate of sBHs per AGN is higher than the dry rate per MBH by  $\mathcal{O}(10^1 - 10^3)$ , and the wet EMRI rate of MGOs per AGN is higher than the dry rate per MBH by  $\mathcal{O}(10^3 - 10^4)$ . Accounting for the AGN fraction  $f_{\text{AGN}} = \mathcal{O}(10^{-2} - 10^{-1})$ , the wet channel turns out to be the dominant channel of mass-gap EMRI formation. As for the LISA detection prospects, we expect LISA to detect no more than  $\sim 1$  dry mass-gap EMRIs, and  $\mathcal{O}(10 - 10^2) \times (f_{\text{AGN}}/1\%)$  wet mass-gap EMRIs per year for the physically motivated MBH mass function  $f_{\bullet,-0.3}$ . For the less optimistic MBH mass function  $f_{\bullet,+0.3}$ , the expected detection numbers are lower by  $\mathcal{O}(10)$  (Table II).

## B. Discussion

As shown above, the expected number of EMRI detections (denoted as  $D_{\text{sbb}}$  and  $D_{\text{mgo}}$ ) are sensitive to the unknown MBH mass function, while the ratio  $D_{\text{sbb}}/D_{\text{mgo}}$  is not, which can be used as a more robust probe to the MGO abundance and origin.

If LISA detects  $D_{\text{mgo}}$  dry mass-gap EMRIs and  $D_{\text{sbb}}$  dry EMRIs of sBHs per year, we can infer the relative abundance of MGOs and sBHs within nuclear stellar clusters as

$$\frac{\delta_{\text{mgo}}}{\delta_{\text{sbb}}} \approx \frac{D_{\text{mgo}}/D_{\text{sbb}}}{R_{\text{sbb}}^{\text{mgo}}}, \quad (57)$$

where  $R_{\text{sbb}}^{\text{mgo}} \approx 1/120$  is the ratio of expected detection numbers of two different EMRIs assuming MGOs and sBHs are of the same abundance (Table II). In a similar way making use of Eq. (57), one can again infer the relative abundance of MGOs  $\delta_{\text{mgo}}/\delta_{\text{sbb}}$  from detections of wet EMRIs, where  $R_{\text{sbb}}^{\text{mgo}} = (1/10 - 1/5)$  (Table II) varies little over all different model parameters for the parameter space we considered. The inferred relative abundance  $\delta_{\text{mgo}}/\delta_{\text{sbb}}$  can be used to constrain the SN explosion mechanisms, where the delayed SN explosion mechanism predicts  $\delta_{\text{mgo}}/\delta_{\text{sbb}} \rightarrow 1$  while the rapid explosion mechanism predicts  $\delta_{\text{mgo}}/\delta_{\text{sbb}} \rightarrow 0$  [6].

An excess of mass-gap EMRI detection by LISA is a signature of MGOs of exotic origins (e.g., PBHs [16, 18, 19, 93, 94]). If these MGOs are of primordial origin, their abundance around MBHs may be further used to constrain the mass fraction of PBHs in dark matter (DM)  $f_{\text{mgo}} := \Omega_{\text{mgo}}/\Omega_{\text{DM}}$ . This constraint sensitively depends on the DM distribution around MBHs, which is poorly understood theoretically. We consider two extremal cases: (1) DM around MBHs traces baryons with the DM to baryon ratio  $\Omega_{\text{DM}}/\Omega_{\text{B}}$ ; (2) the DM density around MBHs follows the NFW profile [95].

In case (1), the abundance of MGOs relative to stars is

$$\frac{\delta_{\text{mgo}}}{\delta_{\text{star}}} = \frac{f_{\text{mgo}}\Omega_{\text{DM}}}{(1 - f_{\text{gas}})\Omega_{\text{B}}} \frac{m_{\text{star}}}{m_{\text{mgo}}}, \quad (58)$$

where  $f_{\text{gas}}$  is the mass fraction of gas in baryons and  $f_{\text{mgo}}$  is formulated as

$$\begin{aligned} f_{\text{mgo}} &= (1 - f_{\text{gas}}) \frac{\delta_{\text{sbb}}}{\delta_{\text{star}}} \frac{D_{\text{mgo}}/D_{\text{sbb}}}{R_{\text{sbb}}^{\text{mgo}}} \frac{\Omega_{\text{B}}}{\Omega_{\text{DM}}} \frac{m_{\text{mgo}}}{m_{\text{star}}} \\ &= 6 \times 10^{-4} (1 - f_{\text{gas}}) \frac{\delta_{\text{sbb}}/\delta_{\text{star}}}{10^{-3}} \frac{D_{\text{mgo}}/D_{\text{sbb}}}{R_{\text{sbb}}^{\text{mgo}}} \frac{\Omega_{\text{B}}/\Omega_{\text{DM}}}{0.2} \frac{m_{\text{mgo}}/m_{\text{star}}}{3} \end{aligned} \quad (59)$$

where we have used Eq.(57).

In case (2), the DM abundance around MBHs is usually much lower, with the total DM mass with the MBH influence radius  $M_{\text{DM}}(< r_c) \approx 0.3\%M_{\bullet}$  (see Appendix C for details). For comparison, the total star mass is  $\Sigma m_{\text{star}}(< r_c) \approx 2M_{\bullet}$  and the total mass of astrophysical MGOs is  $\Sigma m_{\text{mgo}}(r < r_c) \approx 0.6\%M_{\bullet} \times (\delta_{\text{mgo}}/10^{-3})$ , i.e., the DM abundance around MBHs is comparable with that of astrophysical MGOs. In this case, it is unlikely to observe excess of mass-gap EMRIs. Therefore, an excess of mass-gap EMRI detection by LISA would disfavor the NFW distribution of MGOs as DM.

Many other proposals to explain GW190814-like events involve hierarchical mergers, e.g., in young stellar clusters [23], triple systems [26], AGN disks [28], etc. If the abundance of MGO production can be reliably estimated in these scenarios, the rate of mass gap EMRIs may also be used to test these models. For example, as young stellar clusters are not expected to host massive black holes, if they are the only places that MGOs are produced, we should expect the mass-gap EMRI rate to be minimized.

One working assumption we used is no mass change of MGOs in AGN disks. As shown in Fig. 4, the typical migration timescale of MGOs is  $\sim 10^6$  yr, which is much shorter than the Salpeter timescale  $5 \times 10^7$  yr (mass doubling timescale with the Eddington accretion rate). But the accretion rate of MGOs/sBHs in AGN disks is uncertain [77, 96]. If gas accretion onto MGOs was highly super-Eddington and largely increased their masses, identifying the distorted mass gap is less straightforward, and the commonly used techniques of searching for mass-gap features in the mass spectrum of LIGO/Virgo events [8] would also be valuable for our purpose.

## ACKNOWLEDGMENTS

Z. P. and H. Y. are supported by the Natural Sciences and Engineering Research Council of Canada and in part by Perimeter Institute for Theoretical Physics. Research at Perimeter Institute is supported in part by the Government of Canada through the Department of Innovation, Science and Economic Development Canada and by the Province of Ontario through the Ministry of Colleges and Universities.

### Appendix A: Diffusion and advection coefficients in the Fokker-Planck equation (9)

In Refs. [65, 66, 68], the diffusion and the advection coefficients of a single-species cluster have been derived in detail. Following Refs. [33, 69], we extend them to multi-species cases. We first define auxiliary functions:

$$\begin{aligned} F_0^{(j)}(E, r) &= (4\pi m_j)^2 \ln \Lambda \int_{-\infty}^E dE' \bar{f}_j(E'), \\ F_n^{(j)}(E, r) &= (4\pi m_j)^2 \ln \Lambda \int_E^{\phi(r)} dE' \left( \frac{\phi - E'}{\phi - E} \right)^{n/2} \bar{f}_j(E'), \end{aligned} \quad (\text{A1})$$

where  $n \geq 1$ ,  $j$  is the index labelling different star species,  $\ln \Lambda$  the Coulomb's logarithm which we take as  $\ln \Lambda = 10$ , and

$$\bar{f}_j(E) := \int_0^1 f_j(E, R) dR. \quad (\text{A2})$$

With these auxiliary functions, the coefficients are written as

$$\begin{aligned} D_{EE}^{(i)} &= \sum_j \frac{2}{3P} \int_{r_-}^{r_+} \frac{dr}{v_r} v^2 (F_0^{(j)} + F_3^{(j)}), \\ D_E^{(i)} &= \sum_j -\frac{2}{P} \int_{r_-}^{r_+} \frac{dr}{v_r} F_1^{(j)} \times \frac{m_i}{m_j}, \\ D_{ER}^{(i)} &= \sum_j \frac{4}{3P} R \int_{r_-}^{r_+} \frac{dr}{v_r} \left( \frac{v^2}{v_c^2} - 1 \right) (F_0^{(j)} + F_3^{(j)}), \\ D_{RR}^{(i)} &= \sum_j \frac{4}{3P} \frac{R}{J_c^2} \int_{r_-}^{r_+} \frac{dr}{v_r} \left\{ 2 \frac{r^2}{v^2} \left[ v_t^2 \left( \frac{v^2}{v_c^2} - 1 \right)^2 + v_r^2 \right] F_0^{(j)} \right. \\ &\quad \left. + 3 \frac{r^2}{v^2} v_r^2 F_1^{(j)} + \frac{r^2}{v^2} \left[ 2v_t^2 \left( \frac{v^2}{v_c^2} - 1 \right)^2 - v_r^2 \right] F_3^{(j)} \right\}, \\ D_R^{(i)} &= \sum_j -\frac{4}{P} \frac{R}{v_c^2} \int_{r_-}^{r_+} \frac{dr}{v_r} \left( 1 - \frac{v_c^2}{v^2} \right) F_1^{(j)} \times \frac{m_i}{m_j}, \end{aligned} \quad (\text{A3})$$

where  $i$  (similar to  $j$ ) is also the star species index,  $v_t = J/r$  is the tangential velocity, and  $v_c(E)$  is the velocity of a circular orbit with energy  $E$ .

### Appendix B: Diffusion and advection coefficients in the Fokker-Planck equation (32)

Following Refs. [33, 65, 66, 69], we derive the diffusion and advection coefficients in the Fokker-Planck equation (32) in this section. Consider a star  $m$  with orbital energy  $E$  and velocity  $\mathbf{v}$  changes its energy by  $\Delta E$  and velocity by  $\Delta \mathbf{v}$  due to a scattering with a field star  $m_a$ . In the following orthonormal basis,

$$\begin{aligned} \hat{\mathbf{v}} &= \mathbf{v}/v, \\ \hat{\mathbf{J}} &= \mathbf{J}/J = \mathbf{r} \times \mathbf{v}/J, \\ \hat{\mathbf{w}} &= \mathbf{v} \times \mathbf{J}/|\mathbf{v} \times \mathbf{J}|, \end{aligned} \quad (\text{B1})$$

the velocity change is written as

$$\Delta \mathbf{v} = \Delta v_{\parallel} \hat{\mathbf{v}} + \Delta \mathbf{v}_{\perp} = \Delta v_{\parallel} \hat{\mathbf{v}} + \Delta v_J \hat{\mathbf{J}} + \Delta v_w \hat{\mathbf{w}}. \quad (\text{B2})$$

As a result, the changes in energy and angular momentum are

$$\Delta E = -\frac{1}{2} (\Delta v_{\parallel})^2 - \frac{1}{2} (\Delta v_{\perp})^2 - v \Delta v_{\parallel}, \quad (\text{B3})$$

and

$$\begin{aligned} \Delta \mathbf{J} &= \mathbf{r} \times \Delta \mathbf{v} \\ &= J \left( \frac{\Delta v_{\parallel}}{v} - \frac{v_r}{v_t} \frac{\Delta v_w}{v} \right) \hat{\mathbf{J}} + J \frac{\Delta v_J}{v} \left( \frac{v_r}{v_t} \hat{\mathbf{w}} - \hat{\mathbf{v}} \right) \\ &:= \Delta \mathbf{J}_{\parallel} + \Delta \mathbf{J}_{\perp}, \end{aligned} \quad (\text{B4})$$

with  $\Delta \mathbf{J}_{\perp} = J(\Delta v_J/v) \sqrt{1 + (v_r/v_t)^2} = J(\Delta v_J/v_t) = r \Delta v_J$  and

$$\Delta J := |\mathbf{J} + \Delta \mathbf{J}| - J = J \frac{\Delta v_{\parallel}}{v} - r v_r \frac{\Delta v_w}{v} + \frac{1}{2} \frac{r^2 (\Delta v_J)^2}{J}. \quad (\text{B5})$$

For  $\mu = \cos \iota = J_z/J$  with  $J_z$  the  $z$ -component angular momentum, its change is

$$\Delta \mu = \frac{\Delta J_z}{J} - \mu \frac{\Delta J}{J} = \frac{\Delta v_J}{v} \left( \frac{v_r}{v_t} \hat{\mathbf{w}} - \hat{\mathbf{v}} \right) \cdot \hat{\mathbf{z}} - \frac{\mu}{2} \frac{r^2 (\Delta v_J)^2}{J^2}. \quad (\text{B6})$$

Defining  $\langle \Delta X \rangle_t := \frac{\langle \Delta X \rangle}{\Delta t} |_{\Delta t \rightarrow 0}$  and  $\langle \Delta X \Delta Y \rangle_t := \frac{\langle \Delta X \Delta Y \rangle}{\Delta t} |_{\Delta t \rightarrow 0}$  (where  $\langle \rangle$  is ensemble average over scatterings a star has experienced), it is straightforward to see

$$\begin{aligned} \langle \Delta E \rangle_t &= -\frac{1}{2} \langle (\Delta v_{\parallel})^2 \rangle_t - \frac{1}{2} \langle (\Delta v_{\perp})^2 \rangle_t - v \langle \Delta v_{\parallel} \rangle_t, \\ \langle (\Delta E)^2 \rangle_t &= v^2 \langle (\Delta v_{\parallel})^2 \rangle_t, \\ \langle \Delta \mu \rangle_t &= \langle \frac{\Delta v_J}{v} \rangle_t \left( \frac{v_r}{v_t} \hat{\mathbf{w}} - \hat{\mathbf{v}} \right) \cdot \hat{\mathbf{z}} - \frac{\mu}{2} \frac{r^2}{J^2} \langle (\Delta v_J)^2 \rangle_t, \\ \langle (\Delta \mu)^2 \rangle_t &= \left\langle \left( \frac{\Delta v_J}{v} \left( \frac{v_r}{v_t} \hat{\mathbf{w}} - \hat{\mathbf{v}} \right) \cdot \hat{\mathbf{z}} \right)^2 \right\rangle_t \\ &= \left\langle \frac{(\Delta v_J)^2}{v^2} \left( \frac{v_r^2}{v_t^2} + 1 \right) \frac{\sin^2 \iota}{2} \right\rangle_t \\ &= \frac{1 - \mu^2}{2} \frac{r^2}{J^2} \langle (\Delta v_J)^2 \rangle_t, \\ \langle \Delta E \Delta \mu \rangle_t &= - \left( \frac{v_r}{v_t} \hat{\mathbf{w}} - \hat{\mathbf{v}} \right) \cdot \hat{\mathbf{z}} \langle \Delta v_{\parallel} \Delta v_J \rangle_t, \end{aligned} \quad (\text{B7})$$

accurate to quadratic order in  $\Delta v$ .

In the case of spherical symmetry, coefficients  $\langle \Delta \mathbf{v} \rangle_t$  and  $\langle \Delta \mathbf{v} \Delta \mathbf{v} \rangle_t$  has been derived by Binney and Tremaine [69] (assuming the field stars  $m_a$  are symmetrically distributed in the azimuthal direction in the rest frame of particle  $m$ ) as

$$\begin{aligned} \langle \Delta v_{\parallel} \rangle_t &= -\kappa \frac{m + m_a}{m_a} \int_0^v dv_a \frac{v_a^2}{v^2} f_a(v_a), \\ \langle (\Delta v_{\parallel})^2 \rangle_t &= \frac{2}{3} \kappa \left( \int_0^v dv_a \frac{v_a^4}{v^3} f_a(v_a) + \int_v^{\infty} dv_a v_a f_a(v_a) \right) \\ \langle (\Delta v_{\perp})^2 \rangle_t &= \frac{2}{3} \kappa \left( \int_0^v dv_a \left( \frac{3v_a^2}{v} - \frac{v_a^4}{v^3} \right) f_a(v_a) + 2 \int_v^{\infty} dv_a v_a f_a(v_a) \right) \end{aligned} \quad (\text{B8})$$

$\langle \Delta v_{\perp} \rangle_t = 0$  and  $\langle \Delta v_{\parallel} \Delta v_{\perp} \rangle_t = 0$ , with  $\kappa = (4\pi m_a)^2 \ln \Lambda$ . It is straightforward to extend the above results to the non-spherical symmetry case, with the replacement  $f_a(v_a) \rightarrow \bar{f}_a(v_a, \theta)$ , where

$$\bar{f}_a(v_a, \theta) := \frac{1}{2\pi} \int_0^{2\pi} d\eta f_a(v_a, \mu = \sin \theta \cos \eta) d\eta, \quad (\text{B9})$$

and  $\theta$  is the polar angle w.r.t. the  $z$ -axis.

With  $\langle \Delta v \rangle_t$  and  $\langle \Delta v \Delta v \rangle_t$  ready, the derivation of coefficients in the Fokker-Planck equation (32) is parallel to the previous section and we outline it as follows. Define functions

$$\begin{aligned} F_0^{(j)}(E, \theta) &= (4\pi m_j)^2 \ln \Lambda \int_{-\infty}^E dE' \bar{f}_j(E', \theta), \\ F_n^{(j)}(E, r, \theta) &= (4\pi m_j)^2 \ln \Lambda \int_E^{\phi(r)} dE' \bar{f}_j(E', \theta) \left( \frac{\phi - E'}{\phi - E} \right)^{n/2}, \end{aligned} \quad (\text{B10})$$

with  $n \geq 1$  and

$$\bar{f}_j(E, \theta) := \frac{1}{2\pi} \int_0^{2\pi} d\eta f_j(E, \mu = \sin \theta \cos \eta) d\eta. \quad (\text{B11})$$

With these auxiliary functions, the local diffusion/advection coefficients  $\hat{D}(E, R, r, \theta)$  (which depend on energy  $E$ , normalized angular momentum  $R$ , and spatial coordinates  $r$  and  $\theta$ ) are

$$\begin{aligned} \hat{D}_{EE}^{(i)} &= \sum_j \frac{1}{3} v^2 (F_0^{(j)} + F_3^{(j)}), \\ \hat{D}_E^{(i)} &= \sum_j -F_1^{(j)} \times \frac{m_i}{m_j}, \\ D_{\mu\mu}^{(i)} &= \sum_j \frac{1-\mu^2}{4} \frac{1}{3} v_t^{-2} (2F_0^{(j)} + 3F_2^{(j)} - F_4^{(j)}), \\ D_{\mu}^{(i)} &= \sum_j \frac{1-\mu^2}{4} \frac{1}{3} \frac{\partial}{\partial \mu} v_t^{-2} (2F_0^{(j)} + 3F_2^{(j)} - F_4^{(j)}), \\ D_{\mu E}^{(i)} &= 0. \end{aligned} \quad (\text{B12})$$

In general, the orbital coordinates  $(r, \theta)$  of a star orbit with semi-major axis  $a(E)$  and eccentricity  $e$  are specified by

$$r = \frac{a(1-e^2)}{1+e \cos \psi'}, \quad \cos \theta = \cos(\psi' + \phi) \sin \iota, \quad (\text{B13})$$

with  $\phi \in [0, 2\pi]$  specifying the pericenter location in the azimuthal direction, and  $\psi' \in [0, 2\pi]$  is the orbital phase w.r.t the pericenter. With orbit average, we obtain

$$\bar{D}(E, R, \mu, \phi) = \frac{2}{P} \int_{r_-}^{r_+} \frac{dr}{v_r} \hat{D}(E, R, r, \theta). \quad (\text{B14})$$

After ensemble average over  $R$  and  $\phi$ , we arrive at the final form  $D(E, \mu) = \langle \bar{D}(E, R, \mu, \phi) \rangle_{R, \phi}$ .

In fact, we find circular orbits is a good approximation in calculating the coefficients. For circular orbits, the orbital equation is simplified as

$$r = r_c(E), \quad \cos \theta = \cos \psi' \sin \iota \quad (\text{B15})$$

and the diffusion/advection coefficients arising from scatterings with the cluster-component stars

$$\begin{aligned} D_{EE}^{(i)} &= \sum_j \frac{1}{3\pi} \int_0^{\pi} d\psi' v^2 (F_0^{(j)} + F_3^{(j)}), \\ D_E^{(i)} &= \sum_j -\frac{1}{\pi} \int_0^{\pi} d\psi' F_1^{(j)} \times \frac{m_i}{m_j}, \\ D_{\mu\mu}^{(i)} &= \sum_j \frac{1-\mu^2}{4} \frac{1}{3\pi} \int_0^{\pi} d\psi' v_t^{-2} (2F_0^{(j)} + 3F_2^{(j)} - F_4^{(j)}), \\ D_{\mu}^{(i)} &= \sum_j \frac{1-\mu^2}{4} \frac{1}{3\pi} \frac{\partial}{\partial \mu} \int_0^{\pi} d\psi' v_t^{-2} (2F_0^{(j)} + 3F_2^{(j)} - F_4^{(j)}), \\ D_{\mu E}^{(i)} &= 0. \end{aligned} \quad (\text{B16})$$

With the same approximation, the diffusion/advection coefficients arising from scatterings with the disk-component stars are

$$\begin{aligned} D_{EE}^{(i)} &= \sum_j \frac{1}{3} \epsilon v^2 (G_0^{(j)} + G_3^{(j)}), \\ D_E^{(i)} &= \sum_j -\epsilon G_1^{(j)} \times \frac{m_i}{m_j}, \\ D_{\mu\mu}^{(i)} &= \sum_j \frac{1-\mu^2}{4} \frac{1}{3} \epsilon v_t^{-2} (2G_0^{(j)} + 3G_2^{(j)} - G_4^{(j)}), \\ D_{\mu}^{(i)} &= \sum_j \frac{1-\mu^2}{4} \frac{1}{3} \frac{\partial}{\partial \mu} \epsilon v_t^{-2} (2G_0^{(j)} + 3G_2^{(j)} - G_4^{(j)}), \\ D_{\mu E}^{(i)} &= 0. \end{aligned} \quad (\text{B17})$$

where  $\epsilon(h, \iota) = \min\{1, \frac{2}{\pi} \arcsin(\frac{h}{\iota})\} \approx \frac{2}{\pi} \arcsin(\frac{h}{h+\iota})$  is fraction of the orbit lying inside the disk component, and

$$\begin{aligned} G_0^{(j)}(E) &= (4\pi m_j)^2 \ln \Lambda \int_{-\infty}^E dE' g_j(E'), \\ G_n^{(j)}(E, r) &= (4\pi m_j)^2 \ln \Lambda \int_E^{\phi(r)} dE' g_j(E') \left( \frac{\phi - E'}{\phi - E} \right)^{n/2}, \end{aligned} \quad (\text{B18})$$

with  $n \geq 1$ .

### Appendix C: DM NFW profile

If DM density around a MBH follows the NFW profile [95]

$$\rho_{\text{DM}}(r) = \frac{\rho_s}{\frac{r}{R_s} \left(1 + \frac{r}{R_s}\right)^2}, \quad (\text{C1})$$

the total DM mass within radius  $r$  is written as  $M_{\text{DM}}(< r) = 4\pi \rho_s R_s^3 \mathcal{G}(c)$ , with concentration  $c := r/R_s$  and  $\mathcal{G}(c) = \ln(1+c) - c/(1+c)$ , where  $\rho_s$  and  $R_s$  are the characteristic density and radius, respectively. For relating  $\rho_s$  and  $R_s$  to the MBH mass  $M_{\bullet}$ , we need the aid of a commonly used cutoff radius within which the average DM density is 200 times the critical



density of the universe  $\rho_{\text{crit}}$ , i.e.,

$$M_{200} = 4\pi\rho_s R_s^3 \mathcal{G}(c_{200}), \quad \frac{\rho_s}{\rho_{\text{crit}}} = \frac{200}{3} \frac{c_{200}^3}{\mathcal{G}(c_{200})}, \quad (\text{C2})$$

As found in Ref. [97],  $M_{200}$  and  $M_\bullet$  are correlated with

$$\frac{M_\bullet}{10^7 M_\odot} \approx \left( \frac{M_{200}}{10^{12} M_\odot} \right)^{1.65}. \quad (\text{C3})$$

For low redshift, the concentration  $c_{200}$  has a weak dependence on the mass  $M_{200}$  [98]

$$\log_{10} c_{200} = 0.905 - 0.101 \log_{10}(M_{200}/10^{12} h^{-1} M_\odot). \quad (\text{C4})$$

Combining the above three equations, we find the total mass of DM within the influence radius  $r_c$  is  $M_{\text{DM}}(< r_c) \approx 0.3\% M_\bullet$  for  $M_\bullet \in (10^5, 10^7) M_\odot$ .

- 
- [1] C. D. Bailyn, R. K. Jain, P. Coppi, and J. A. Orosz, *Astrophys. J.* **499**, 367 (1998), arXiv:astro-ph/9708032 [astro-ph].
- [2] F. Özel, D. Psaltis, R. Narayan, and J. E. McClintock, *Astrophys. J.* **725**, 1918 (2010), arXiv:1006.2834 [astro-ph.GA].
- [3] F. Özel, D. Psaltis, R. Narayan, and A. Santos Villarreal, *Astrophys. J.* **757**, 55 (2012), arXiv:1201.1006 [astro-ph.HE].
- [4] L. Kreidberg, C. D. Bailyn, W. M. Farr, and V. Kalogera, *Astrophys. J.* **757**, 36 (2012), arXiv:1205.1805 [astro-ph.HE].
- [5] K. Belczynski, G. Wiktorowicz, C. L. Fryer, D. E. Holz, and V. Kalogera, *Astrophys. J.* **757**, 91 (2012), arXiv:1110.1635 [astro-ph.GA].
- [6] C. L. Fryer, K. Belczynski, G. Wiktorowicz, M. Dominik, V. Kalogera, and D. E. Holz, *Astrophys. J.* **749**, 91 (2012), arXiv:1110.1726 [astro-ph.SR].
- [7] LIGO Scientific Collaboration and Virgo Collaboration, *Astrophys.J.Lett.* **896**, L44 (2020), arXiv:2006.12611 [astro-ph.HE].
- [8] LIGO Scientific Collaboration and Virgo Collaboration, arXiv e-prints, arXiv:2111.03606 (2021), arXiv:2111.03606 [gr-qc].
- [9] T. A. Thompson, C. S. Kochanek, K. Z. Stanek, C. Badenes, R. S. Post, T. Jayasinghe, D. W. Latham, A. Bieryla, G. A. Esquerdo, P. Berlind, M. L. Calkins, J. Tayar, L. Lindegren, J. A. Johnson, T. W. S. Holoiien, K. Auchtettl, and K. Covey, *Science* **366**, 637 (2019), arXiv:1806.02751 [astro-ph.HE].
- [10] T. Jayasinghe, K. Z. Stanek, T. A. Thompson, C. S. Kochanek, D. M. Rowan, P. J. Vallely, K. G. Strassmeier, M. Weber, J. T. Hinkle, F. J. Hamsch, D. V. Martin, J. L. Prieto, T. Pessi, D. Huber, K. Auchtettl, L. A. Lopez, I. Ilyin, C. Badenes, A. W. Howard, H. Isaacson, and S. J. Murphy, *MNRAS* **504**, 2577 (2021), arXiv:2101.02212 [astro-ph.SR].
- [11] B. Giesers, S. Dreizler, T.-O. Husser, S. Kamann, G. Anglada Escudé, J. Brinchmann, C. M. Carollo, M. M. Roth, P. M. Weilbacher, and L. Wisotzki, *MNRAS* **475**, L15 (2018), arXiv:1801.05642 [astro-ph.SR].
- [12] B. Giesers, S. Kamann, S. Dreizler, T.-O. Husser, A. Askar, F. Göttgens, J. Brinchmann, M. Latour, P. M. Weilbacher, M. Wendt, and M. M. Roth, *A&A* **632**, A3 (2019), arXiv:1909.04050 [astro-ph.SR].
- [13] Ł. Wyrzykowski and I. Mandel, *A&A* **636**, A20 (2020), arXiv:1904.07789 [astro-ph.SR].
- [14] M. Zevin, M. Spera, C. P. L. Berry, and V. Kalogera, *Astrophys.J.Lett.* **899**, L1 (2020), arXiv:2006.14573 [astro-ph.HE].
- [15] P. Drozda, K. Belczynski, R. O’Shaughnessy, T. Bulik, and C. L. Fryer, arXiv e-prints, arXiv:2009.06655 (2020), arXiv:2009.06655 [astro-ph.HE].
- [16] B. Carr, S. Clesse, J. Garcia-Bellido, and F. Kuhnel, arXiv e-prints, arXiv:1906.08217 (2019), arXiv:1906.08217 [astro-ph.CO].
- [17] H. Yang, W. E. East, and L. Lehner, *Astrophys. J.* **856**, 110 (2018), [Erratum: *Astrophys.J.* 870, 139 (2019)], arXiv:1710.05891 [gr-qc].
- [18] S. Clesse and J. Garcia-Bellido, arXiv e-prints, arXiv:2007.06481 (2020), arXiv:2007.06481 [astro-ph.CO].
- [19] K. Jedamzik, *Phys. Rev. Lett.* **126**, 051302 (2021), arXiv:2007.03565 [astro-ph.CO].
- [20] M. Safarzadeh, A. S. Hamers, A. Loeb, and E. Berger, *Astrophys.J.Lett.* **888**, L3 (2020), arXiv:1911.04495 [astro-ph.HE].
- [21] G. Fragione, A. Loeb, and F. A. Rasio, *Astrophys.J.Lett.* **895**, L15 (2020), arXiv:2002.11278 [astro-ph.GA].
- [22] Y. Yang, V. Gayathri, I. Bartos, Z. Haiman, M. Safarzadeh, and H. Tagawa, *Astrophys.J.Lett.* **901**, L34 (2020), arXiv:2007.04781 [astro-ph.HE].
- [23] S. Rastello, M. Mapelli, U. N. Di Carlo, N. Giacobbo, F. Santoliquido, M. Spera, A. Ballone, and G. Iorio, *MNRAS* **497**, 1563 (2020), arXiv:2003.02277 [astro-ph.HE].
- [24] T. Broadhurst, J. M. Diego, and G. F. Smoot, arXiv e-prints, arXiv:2006.13219 (2020), arXiv:2006.13219 [astro-ph.CO].
- [25] A. S. Hamers, G. Fragione, P. Neunteufel, and B. Kocsis, *MNRAS* **506**, 5345 (2021), arXiv:2103.03782 [astro-ph.HE].
- [26] W. Lu, P. Beniamini, and C. Bonnerot, *MNRAS* **500**, 1817 (2021), arXiv:2009.10082 [astro-ph.HE].
- [27] B. Liu and D. Lai, *MNRAS* **502**, 2049 (2021), arXiv:2009.10068 [astro-ph.HE].
- [28] H. Tagawa, B. Kocsis, Z. Haiman, I. Bartos, K. Omukai, and J. Samsing, *Astrophys. J.* **908**, 194 (2021), arXiv:2012.00011 [astro-ph.HE].
- [29] M. Arca Sedda, *Astrophys.J.Lett.* **908**, L38 (2021), arXiv:2102.03364 [astro-ph.HE].
- [30] I. Mandel and F. S. Broekgaarden, arXiv e-prints, arXiv:2107.14239 (2021), arXiv:2107.14239 [astro-ph.HE].
- [31] C. Hopman and T. Alexander, *Astrophys. J.* **629**, 362 (2005), arXiv:astro-ph/0503672 [astro-ph].
- [32] M. Preto and P. Amaro-Seoane, *Astrophys.J.Lett.* **708**, L42 (2010), arXiv:0910.3206 [astro-ph.GA].
- [33] B. Bar-Or and T. Alexander, *Astrophys. J.* **820**, 129 (2016), arXiv:1508.01390 [astro-ph.GA].
- [34] P. Amaro-Seoane and M. Preto, *Classical and Quantum Gravity* **28**, 094017 (2011), arXiv:1010.5781 [astro-ph.CO].
- [35] P. Amaro-Seoane, *Living Reviews in Relativity* **21**, 4 (2018), arXiv:1205.5240 [astro-ph.CO].
- [36] P. Amaro-Seoane, arXiv e-prints, arXiv:2011.03059 (2020), arXiv:2011.03059 [gr-qc].
- [37] S. Babak, J. Gair, A. Sesana, E. Barausse, C. F. Sopuerta, C. P. L. Berry, E. Berti, P. Amaro-Seoane, A. Petiteau, and A. Klein, *Phys. Rev. D* **95**, 103012 (2017), arXiv:1703.09722 [gr-qc].
- [38] J. R. Gair, S. Babak, A. Sesana, P. Amaro-Seoane, E. Barausse, C. P. Berry, E. Berti, and C. Sopuerta, *J. Phys. Conf. Ser.* **840**, 012021 (2017), arXiv:1704.00009 [astro-ph.GA].
- [39] H.-M. Fan, Y.-M. Hu, E. Barausse, A. Sesana, J.-d. Zhang, X. Zhang, T.-G. Zi, and J. Mei, *Phys. Rev. D* **102**, 063016

- (2020), arXiv:2005.08212 [astro-ph.HE].
- [40] Y. Levin, arXiv e-prints , astro-ph/0307084 (2003), arXiv:astro-ph/0307084 [astro-ph].
- [41] Y. Levin, *MNRAS* **374**, 515 (2007), arXiv:astro-ph/0603583 [astro-ph].
- [42] G. Sigl, J. Schnittman, and A. Buonanno, *Phys. Rev. D* **75**, 024034 (2007), arXiv:astro-ph/0610680 [astro-ph].
- [43] H. Tagawa, Z. Haiman, and B. Kocsis, *Astrophys. J.* **898**, 25 (2020), arXiv:1912.08218 [astro-ph.GA].
- [44] Z. Pan and H. Yang, *Phys. Rev. D* **103**, 103018 (2021), arXiv:2101.09146 [astro-ph.HE].
- [45] Z. Pan, Z. Lyu, and H. Yang, *Phys. Rev. D* **104**, 063007 (2021), arXiv:2104.01208 [astro-ph.HE].
- [46] M. C. Miller, M. Freitag, D. P. Hamilton, and V. M. Lauburg, *Astrophys.J.Lett.* **631**, L117 (2005), arXiv:astro-ph/0507133 [astro-ph].
- [47] X. Chen and W.-B. Han, *Communications Physics* **1**, 53 (2018), arXiv:1801.05780 [astro-ph.HE].
- [48] Y. Y. Wang, F. Y. Wang, Y. C. Zou, and Z. G. Dai, *Astrophys.J.Lett.* **886**, L22 (2019), arXiv:1911.04117 [astro-ph.HE].
- [49] Y. Raveh and H. B. Perets, *MNRAS* **501**, 5012 (2021), arXiv:2011.13952 [astro-ph.GA].
- [50] S. Naoz, S. C. Rose, E. Michaely, D. Melchor, E. Ramirez-Ruiz, B. Mockler, and J. D. Schnittman, *Astrophys.J.Lett.* **927**, L18 (2022), arXiv:2202.12303 [astro-ph.HE].
- [51] L. Zwick, P. R. Capelo, E. Bortolas, L. Mayer, and P. Amaro-Seoane, *MNRAS* **495**, 2321 (2020), arXiv:1911.06024 [astro-ph.GA].
- [52] L. Zwick, P. R. Capelo, E. Bortolas, V. Vázquez-Aceves, L. Mayer, and P. Amaro-Seoane, *MNRAS* **506**, 1007 (2021), arXiv:2102.00015 [astro-ph.GA].
- [53] V. Vázquez-Aceves, L. Zwick, E. Bortolas, P. R. Capelo, P. Amaro-Seoane, L. Mayer, and X. Chen, arXiv e-prints , arXiv:2108.00135 (2021), arXiv:2108.00135 [astro-ph.GA].
- [54] A. Galametz, D. Stern, P. R. M. Eisenhardt, M. Brodwin, M. J. I. Brown, A. Dey, A. H. Gonzalez, B. T. Jannuzi, L. A. Moustakas, and S. A. Stanford, *Astrophys. J.* **694**, 1309 (2009), arXiv:0901.1109 [astro-ph.CO].
- [55] M. Macuga, P. Martini, E. D. Miller, M. Brodwin, M. Hayashi, T. Kodama, Y. Koyama, R. A. Overzier, R. Shimakawa, K.-i. Tadaki, and I. Tanaka, *Astrophys. J.* **874**, 54 (2019), arXiv:1805.06569 [astro-ph.GA].
- [56] P. Goldreich and S. Tremaine, *Astrophys. J.* **233**, 857 (1979).
- [57] P. Goldreich and S. Tremaine, *Astrophys. J.* **241**, 425 (1980).
- [58] H. Tanaka, T. Takeuchi, and W. R. Ward, *Astrophys. J.* **565**, 1257 (2002).
- [59] H. Tanaka and W. R. Ward, *Astrophys. J.* **602**, 388 (2004).
- [60] L. Arzamasskiy, Z. Zhu, and J. M. Stone, *MNRAS* **475**, 3201 (2018), arXiv:1710.11128 [astro-ph.EP].
- [61] Z. Zhu, *MNRAS* **483**, 4221 (2019), arXiv:1812.01262 [astro-ph.EP].
- [62] B. Kocsis, N. Yunes, and A. Loeb, *Phys. Rev. D* **84**, 024032 (2011), arXiv:1104.2322 [astro-ph.GA].
- [63] N. Yunes, B. Kocsis, A. Loeb, and Z. Haiman, *Phys. Rev. Lett.* **107**, 171103 (2011), arXiv:1103.4609 [astro-ph.CO].
- [64] C. Cutler, D. Kennefick, and E. Poisson, *Phys. Rev. D* **50**, 3816 (1994).
- [65] H. Cohn and R. M. Kulsrud, *Astrophys. J.* **226**, 1087 (1978).
- [66] H. Cohn, *Astrophys. J.* **234**, 1036 (1979).
- [67] D. F. Chernoff and M. D. Weinberg, *Astrophys. J.* **351**, 121 (1990).
- [68] S. L. Shapiro and A. B. Marchant, *Astrophys. J.* **225**, 603 (1978).
- [69] J. Binney and S. Tremaine, *Galactic dynamics* (1987).
- [70] P. C. Peters, *Phys. Rev.* **136**, B1224 (1964).
- [71] S. Tremaine, D. O. Richstone, Y.-I. Byun, A. Dressler, S. M. Faber, C. Grillmair, J. Kormendy, and T. R. Lauer, *Astronomical Journal* **107**, 634 (1994), arXiv:astro-ph/9309044 [astro-ph].
- [72] W. Dehnen, *MNRAS* **265**, 250 (1993).
- [73] R. Perna, Y.-H. Wang, W. M. Farr, N. Leigh, and M. Cantiello, *Astrophys.J.Lett.* **878**, L1 (2019), arXiv:1901.03345 [astro-ph.HE].
- [74] S. Tremaine, K. Gebhardt, R. Bender, G. Bower, A. Dressler, S. M. Faber, A. V. Filippenko, R. Green, C. Grillmair, L. C. Ho, J. Kormendy, T. R. Lauer, J. Magorrian, J. Pinkney, and D. Richstone, *Astrophys. J.* **574**, 740 (2002), arXiv:astro-ph/0203468 [astro-ph].
- [75] K. Gültekin, D. O. Richstone, K. Gebhardt, T. R. Lauer, S. Tremaine, M. C. Aller, R. Bender, A. Dressler, S. M. Faber, A. V. Filippenko, R. Green, L. C. Ho, J. Kormendy, J. Magorrian, J. Pinkney, and C. Siopis, *Astrophys. J.* **698**, 198 (2009), arXiv:0903.4897 [astro-ph.GA].
- [76] T. Alexander and C. Hopman, *Astrophys. J.* **697**, 1861 (2009), arXiv:0808.3150 [astro-ph].
- [77] Z. Pan and H. Yang, *The Astrophysical Journal* **923**, 173 (2021).
- [78] A. S. Jermyn, A. J. Dittmann, M. Cantiello, and R. Perna, *Astrophys. J.* **914**, 105 (2021), arXiv:2102.13114 [astro-ph.HE].
- [79] In the previous work [44], we have approximated the effect of density waves as inclination-independent and have neglected the contribution from dynamical friction. As a result, the previous approximation is an over estimate for highly inclined orbiters.
- [80] E. Sirko and J. Goodman, *MNRAS* **341**, 501 (2003), arXiv:astro-ph/0209469 [astro-ph].
- [81] T. A. Thompson, E. Quataert, and N. Murray, *Astrophys. J.* **630**, 167 (2005), arXiv:astro-ph/0503027 [astro-ph].
- [82] S. A. Balbus and J. F. Hawley, *Astrophys. J.* **376**, 214 (1991).
- [83] S. A. Balbus and J. F. Hawley, *Reviews of Modern Physics* **70**, 1 (1998).
- [84] R. G. Martin, C. J. Nixon, J. E. Pringle, and M. Livio, *New Astronomy* **70**, 7 (2019), arXiv:1901.01580 [astro-ph.HE].
- [85] A. King and C. Nixon, *MNRAS* **453**, L46 (2015), arXiv:1507.05960 [astro-ph.HE].
- [86] K. Schawinski, M. Koss, S. Berney, and L. F. Sartori, *MNRAS* **451**, 2517 (2015), arXiv:1505.06733 [astro-ph.GA].
- [87] A. Soltan, *MNRAS* **200**, 115 (1982).
- [88] E. Barausse, *MNRAS* **423**, 2533 (2012), arXiv:1201.5888 [astro-ph.CO].
- [89] J. R. Gair, C. Tang, and M. Volonteri, *Phys. Rev. D* **81**, 104014 (2010), arXiv:1004.1921 [astro-ph.GA].
- [90] L. Barack and C. Cutler, *Phys. Rev. D* **69**, 082005 (2004), arXiv:gr-qc/0310125 [gr-qc].
- [91] A. J. Chua and J. R. Gair, *Class. Quant. Grav.* **32**, 232002 (2015), arXiv:1510.06245 [gr-qc].
- [92] A. J. Chua, C. J. Moore, and J. R. Gair, *Phys. Rev. D* **96**, 044005 (2017), arXiv:1705.04259 [gr-qc].
- [93] H.-K. Guo, J. Shu, and Y. Zhao, *Phys. Rev. D* **99**, 023001 (2019).
- [94] S. Barsanti, V. De Luca, A. Maselli, and P. Pani, arXiv e-prints , arXiv:2109.02170 (2021), arXiv:2109.02170 [gr-qc].
- [95] J. F. Navarro, C. S. Frenk, and S. D. M. White, *Astrophys. J.* **462**, 563 (1996), arXiv:astro-ph/9508025 [astro-ph].
- [96] H. Tagawa, S. S. Kimura, Z. Haiman, R. Perna, H. Tanaka, and I. Bartos, *Astrophys. J.* **927**, 41 (2022), arXiv:2112.01544 [astro-ph.HE].
- [97] L. Ferrarese, *Astrophys. J.* **578**, 90 (2002), arXiv:astro-ph/0203469 [astro-ph].

- [98] A. A. Dutton and A. V. Macciò, *MNRAS* **441**, 3359 (2014), [arXiv:1402.7073 \[astro-ph.CO\]](#).