Proving the short-wavelength approximation in Pulsar Timing Array gravitational-wave background searches

Chiara M F Mingarelli and Angelo B Mingarelli

1 Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave, New York, NY, USA, 10010
2 School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6
3 Author to whom any correspondence should be addressed.

E-mail: cmingarelli@flatironinstitute.org and angelo@math.carleton.ca

Keywords: gravitational-wave background, Hellings and Downs curve, pulsar timing array, supermassive black holes, pulsars, neutron stars

Abstract

A low-frequency gravitational-wave background (GWB) from the cosmic merger history of supermassive black holes is expected to be detected in the next few years by pulsar timing arrays. A GWB induces distinctive correlations in the pulsar residuals—the expected arrival time of the pulse less its actual arrival time. Simplifying assumptions are made in order to write an analytic expression for this correlation function, called the Hellings and Downs curve for an isotropic GWB, which depends on the angular separation of the pulsar pairs, the gravitational-wave frequency considered, and the distance to the pulsars. This is called the short-wavelength approximation, which we prove here rigorously and analytically for the first time.

1. Introduction

Gravitational waves (GWs) are ripples in the fabric of space-time, originating from some of the most violent events in the Universe, including the mergers of supermassive black holes. High frequency GWs from the merger of stellar-mass black holes were first detected by the Laser Interferometer Gravitational-wave Observatory (LIGO) in September 2015 [1], hailing the dawn of gravitational-wave astronomy. However, LIGO can only detect high frequency GWs, in the 100–1000Hz range. Similarly to electromagnetic radiation, different GW detectors are needed to probe different GW frequencies. Currently there are plans to launch a space-based GW detector in 2034—the Laser Interferometer Space Antenna (LISA) [2]—which will probe the millihertz GW frequency regime, thought to be populated primarily by merging supermassive black holes (SMBHs) in the 105–106 M_☉ range. At the very low-frequency end of the GW spectrum, one expects to find nanohertz GWs from very massive inspiraling SMBHs, in the 109–1010 M_☉ range. These can be detected by timing millisecond pulsars, called a Pulsar Timing Array (PTA) [3–6]. Millisecond pulsars are excellent clocks, and delays or advances in their arrival times—inducing a timing residual—could signal the presence of GWs. PTA experiments are very active, and have been taking data for over a decade [7–9]. With a PTA, one can detect not only GWs from inspiraling SMBH binaries (SMBHBs), see e.g. [10, 11], but the GW background (GWB) from the cosmic merger history of SMBHBs [12–14]. This GWB is expected to be detected in the next few years [15, 16], with the details depending on the underlying astrophysics of the SMBH mergers. More details on PTAs can be found in recent review articles, e.g. [17–20], and an outline GW astrophysics covering nanohertz to kilohertz frequencies can be found in [21].

Indeed, a rigorous exploration and examination of the tools which will be used to make the first detection of a GWB is crucial. An isotropic GWB will induce characteristic correlations in the pulsar timing residuals. By cross-correlating these residuals, one expects to see a characteristic correlation called the Hellings and Downs curve [3]. Deviations from an isotropic GWB can be induced by nearby and/or particularly loud SMBHBs, inducing anisotropy in the GWB. Anisotropic GWBs will induce different correlations patterns, and have been explored by [22–26].

© 2018 The Author(s). Published by IOP Publishing Ltd
Here we prove analytically, and for the first time, that the Hellings and Downs curve can be extracted from the cross-correlated pulsar residuals, without making assumptions that the pulsars are all at the same distance $L$ from the Earth. Part of this proof is a consequence of the application of the Riemann-Lebesgue lemma and the Lebesgue Dominated Convergence Theorem — well-known in the mathematics community, but less well-known in the field of GWs. We emphasize that no previous work has been able to do this analytically, though computer-aided integration has been used to verify one’s intuition numerically.

2. The characteristic strain

The International PTA (IPTA) published combined data on 49 millisecond pulsars in their first data release [14]. These millisecond pulsars are the most stable natural astrophysical clocks known [27], and are regularly monitored by 8 radio telescopes: 5 in Europe [8], 2 in North America [7] and one in Australia [9]. PTAs take advantage of the precise arrival times of millisecond pulsars to enable GW detection.

The GWB is described in terms of its characteristic strain, $h_c(f)$, with amplitude $A$ at a reference frequency of 1/yr (e.g. [28]):

$$h_c(f) = A \left( \frac{f}{\text{yr}^{-1}} \right)^{-2/3}. \quad (2.1)$$

The current upper limits on $A$ are difficult to compare, since it was recently discovered that errors in planetary masses and positions (called the solar system ephemeris model) can directly affect the limit on $A$ [12, 29], and in some cases mimic a GWB signal.

While the current upper limit on $A$ from NANOGrav can take this into account, and limit $A < 1.35 \times 10^{-15}$, other PTAs have not yet published updates to their limits. Projections the characteristic strain accessible with future IPTA and Square Kilometer Array (SKA) [30–32] detectors are shown in figure 1.

The observed residuals due to the presence of a GWB with characteristic strain $h_c(f)$ is described by the cross-power spectral density of pulsar 1 and pulsar 2 by

![Figure 1](https://example.com/figure1.png)

**Figure 1.** The spectrum of gravitational radiation from low-frequency to high-frequency. At very low frequencies pulsar timing arrays can detect both the GWB from supermassive black hole binaries, in the $10^8$–$10^{10}$ $M_\odot$ range, as well as radiation from individual binary sources which are sufficiently strong. For the IPTA sensitivity we assume 20 pulsars with 100 ns timing precision with a 15 year dataset, and for the SKA we assume 100 pulsars timed for 20 years with 30 ns timing precision. Both estimates assume 14-day observation cadence. The PTA spans the size of the Galaxy, and is therefore a ‘galaxy-based’ GW detector. LISA is a space-based GW detector scheduled to launch in 2034 [2]. aLIGO a ground-based high-frequency GW detector, and is currently the only detector to directly detect GWs from compact binary coalescences, which currently include binary black holes and binary neutron star mergers [1, 33]. Note that these GW detectors are all complementary, and that LIGO cannot, for example, detect GWs from supermassive black hole binaries, just as PTAs cannot detect high-frequency GWs from merging stellar-mass black holes. A review of current and future GW detectors across the spectrum is available in [21].
$S_{1,2}(f) = \frac{\Gamma_{1,2}(f_{1,2}, \beta_{1,2}, \zeta) h_\zeta^2(f)}{12\pi^2 f^3}, \tag{2.2}$

see e.g. [34], where $\Gamma_{1,2}$ is the so-called overlap reduction function, which describes the GWB-induced correlation signature in the pulsar residuals. This is a function of the frequency of the GWB, the distance to the pulsars $L_{1,2}$, and the angular separation of the pulsars, $\zeta$. PTA geometry is explored in detail in figure 2. For an isotropic GWB, this is called the Hellings and Downs curve [5], and for anisotropic GWBs see [22–25].

3. The Hellings and Downs curve

In analogy with [22, 24], we present an overview of how one arrives to the Hellings and Downs curve. A source of GWs in direction $-\hat{\Omega}$, see figure 2, generates a metric perturbation $h_{ij}(t, \hat{\Omega})$, which we describe as a plane wave:

$$h_{ij}(t, \hat{\Omega}) = \sum_A \int_{-\infty}^{\infty} df \int_{S^2} d\hat{\Omega} h_A(f, \hat{\Omega}) e^{i2\pi f(t - \hat{\Omega} \cdot \hat{n})} e_A^i(\hat{\Omega}).$$

This can be decomposed over two polarization tensors $e_A^i(\hat{\Omega})$, and two independent polarization amplitudes $h_A(t, \hat{\Omega})$ [35, 36]:

$$h_{ij}(t, \hat{\Omega}) = e_A^i(\hat{\Omega}) h_A(t, \hat{\Omega}) + e_A^j(\hat{\Omega}) h_A(t, \hat{\Omega}). \tag{3.2}$$

We note that General Relativity predicts only two independent polarizations, plus $+$, and cross $\times$, while other theories predict additional polarizations, such as breathing modes [37–39]. He restrict ourselves to the well-known tensor polarizations, $A = +, \times$.

The $e_A^i(\hat{\Omega})$ polarization tensors are uniquely defined by specifying $\hat{m}$ and $\hat{n}$—the GW principal axes, illustrated in figure 2:

$$e_A^i(\hat{\Omega}) = \hat{n}_i \hat{n}_j - \delta_{ij} \hat{n}_j, \quad e_A^j(\hat{\Omega}) = \hat{m}_i \hat{m}_j + \delta_{ij} \hat{m}_j. \tag{3.3}$$

For a stationary, Gaussian, and unpolarized GWB, the polarization amplitudes satisfy (see e.g. [40]):

$$\langle h_A^*(f, \hat{\Omega}) h_B(f', \hat{\Omega}') \rangle = \delta^{AB}(\hat{\Omega}, \hat{\Omega}') \delta_{AA'} \delta(f - f') H(f). \tag{3.4}$$

The metric perturbation will change the proper distance between the Earth and the pulsars, inducing an advance or delay in the pulsar pulse’s arrival time at the Earth. Consider for example a millisecond pulsar with frequency $\nu_0$ whose location in the sky is described by $\hat{\Omega}$, at a distance $L$ from the Earth. The metric perturbation affects the frequency of the radio pulses, $\nu$, received at the radio telescope. This frequency shift is given by

$$\Delta \nu(t) = \frac{\nu(t) - \nu_0}{\nu_0} = \frac{1}{2} \hat{\Omega} \cdot \hat{p} \Delta H(t, \hat{\Omega}), \tag{3.5}$$

where $\hat{p}$ is the pulsar’s proper motion. 

The advance of periastron, $\dot{\Omega}$, is a direct consequence of the GWB-induced frequency shift:

$$\dot{\Omega} = \frac{\Delta \nu(t)}{\nu_0} \frac{\hat{p} \cdot \hat{p}}{2} \Delta H(t, \hat{\Omega}). \tag{3.6}$$

Figure 2. Pulsar 1 is on the $z$-axis at a distance $L_1$ from the origin, and Pulsar 2 is in the $x-z$ plane at a distance $L_2$ from the origin making an angle $\zeta$ with Pulsar 1. Here, $\hat{\Omega}$ is the direction of GW propagation, with principal axes $\hat{m}$ and $\hat{n}$, such that $\hat{m} \times \hat{n} = \hat{\Omega}$. The angles $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi]$ are the polar and azimuthal angles, respectively.
where

\[ \Delta h_{ij}(t, \hat{\Omega}) \equiv h_{ij}(t_e, \hat{\Omega}) - h_{ij}(t_p, \hat{\Omega}) \]  \hspace{1cm} (3.6)

is the difference between the GW-induced metric perturbation at the Earth \( h_{ij}(t_e, \hat{\Omega}) \) and at the pulsar \( h_{ij}(t_p, \hat{\Omega}) \), with coordinates \( t_e \), \( \hat{x}_e \), and \( t_p \), \( \hat{x}_p \):

\[ t_p = t_e - L, \quad \hat{x}_p = L \hat{p}, \quad \hat{x}_e = 0. \]  \hspace{1cm} (3.7)

The indices ‘e’ and ‘p’ refer to the Earth and the pulsar, however, it is standard to write \( t_e = t \), see e.g. \([22, 24, 41, 42]\).

We can now write (3.6), using (3.1) and (3.2) as

\[ \Delta h_{ij}(t, \hat{\Omega}) = \sum_A \int_{-\infty}^{\infty} df \int_S d\Omega A^A(f, \hat{\Omega}) h_{ij}(f, \hat{\Omega}) e^{i2\pi \hat{f}L} \times \left[ 1 - e^{-i2\pi \hat{f}(1 + \hat{\Omega} \cdot \hat{p})} \right]. \]  \hspace{1cm} (3.8)

The fractional frequency shift, \( z(t) \), produced by a stochastic GWB is simply given by integrating equation (3.5) over all directions. Using (3.1) and (3.8), we obtain:

\[ z(t) = \int d\Omega z(t, \hat{\Omega}) = \sum_A \int_{-\infty}^{\infty} df \int_S d\Omega F^A(f, \hat{\Omega}) h_{ij}(f, \hat{\Omega}) e^{i2\pi \hat{f}L} \times \left[ 1 - e^{-i2\pi \hat{f}(1 + \hat{\Omega} \cdot \hat{p})} \right], \]  \hspace{1cm} (3.9)

where \( F^A(\hat{\Omega}) \) are the antenna beam patterns for each polarization \( A \), which we write as

\[ F^A(\hat{\Omega}) = \left[ \frac{1}{2} \frac{\hat{\eta} \cdot \hat{p}}{1 + \hat{\Omega} \cdot \hat{p}} \ e^{i\hat{A}_0(\hat{\Omega})} \right]. \]  \hspace{1cm} (3.10)

Searches for the GWB rely on looking for correlations induced by GWs in the timing residuals of pulsar pairs. Indeed, the observed quantity in PTA experiments is the timing residual \( r(t) \), which is simply the integral of equation (3.9) in time:

\[ r(t) = \int \! dt' z(t'). \]  \hspace{1cm} (3.11)

The expected value of the correlation between a residual from pulsar 1 at time \( t_j \) with that from a different pulsar, say pulsar 2 at time \( t_k \), depends on terms of the form:

\[ \langle r^s(t_j) r^u(t_k) \rangle = \left\{ \int_{t_j}^{t_k} dt' \int_{t_k}^{t_j} dt'' z_i^s(t') z_j^u(t'') \right\}, \]  \hspace{1cm} (3.12)

\[ \langle r^s(t_j) r^u(t_k) \rangle = \left\{ \int_{t_j}^{t_k} dt' \int_{t_k}^{t_j} dt'' \int_{-\infty}^{\infty} df e^{-i2\pi \hat{f}(t' - t'')} H(f) \Gamma(f L_1, f L_2, \zeta) \right\}, \]  \hspace{1cm} (3.13)

where \( H(f) \) contains the information of the spectrum of radiation. In analogy with \([22, 23, 36]\), we define the quantity above that depends on the angular separation of the pulsars, \( \zeta \), their distances from the Earth, \( L_1, L_2 \), and the GW frequency \( f \), as the overlap reduction function

\[ \Gamma(f L_1, f L_2, \zeta) \equiv \int d\hat{\Omega} \kappa(f, \hat{\Omega}) \left[ \sum_A F^A_1(\hat{\Omega}) F^A_2(\hat{\Omega}) \right], \]  \hspace{1cm} (3.14)

where

\[ \kappa(f L_1, L_2, \zeta) \equiv \left[ 1 - e^{-i2\pi \hat{\Omega} \cdot \hat{p}_1} \right] \left[ 1 - e^{-i2\pi \hat{\Omega} \cdot \hat{p}_2} \right]. \]  \hspace{1cm} (3.15)

In order to write a closed-form, analytic solution to (3.15), we choose a reference frame where one pulsar is placed along the \( z \)-axis and the other in the \( x-z \) plane as seen in figure 2. Specifically, we write

\[ \hat{p}_1 = (0, 0, 1), \]  \hspace{1cm} (3.16a)

\[ \hat{p}_2 = (\sin \zeta, 0, \cos \zeta), \]  \hspace{1cm} (3.16b)

\[ \hat{\Omega} = (\sin \theta \cos \phi, \sin \theta \sin \phi, \cos \theta), \]  \hspace{1cm} (3.16c)

\[ \hat{m}_1 = (\sin \phi, -\cos \phi, 0), \]  \hspace{1cm} (3.16d)

\[ \hat{m}_2 = (\cos \theta \cos \phi, \cos \theta \sin \phi, -\sin \theta). \]  \hspace{1cm} (3.16e)
We remind the reader that \( \hat{\rho}_1 \) and \( \hat{\rho}_2 \) are the unit vectors pointing to pulsars 1 and 2, respectively, \( \hat{\Omega} \) is the direction of GW propagation and \( \hat{m} \) and \( \hat{n} \) are the GW principal axes, see figure 2. Note that in this reference frame \( F_0^x = 0 \) by (3.11), making it a convenient choice.

For an isotropic GWB one is free to choose whichever coordinate system is most convenient, as was done here. However, one must be more careful when considering reference frames which are used to describe pulsar locations in an anisotropic GWB, as was done by [22].

### 4. Main results

We choose the coordinate system defined in equation (3.17), and apply it to equation (3.15). The result is equations (4.1) and (4.2).

**Claim.** Let \( L_1, L_2, f \), be real positive constants. Then, for each \( \zeta \in (0, \pi] \), as \( fL_1 \to \infty \) and \( fL_2 \to \infty \), we have

\[
\int_0^\pi d\theta \int_0^{2\pi} d\phi K_2(\zeta, \phi, \theta)[1 - e^{2\pi i fL_1 (1 + \cos \theta)}] \\
\times [1 - e^{-2\pi i fL_2 \cos \phi + \sin \theta \sin \zeta \cos \phi}] \quad \to \quad \int_0^\pi d\theta \int_0^{2\pi} d\phi K_2(\zeta, \phi, \theta) \quad (4.1)
\]

except when \( \zeta = 0 \) and \( L_1 = L_2 \), a case covered in [24]. Here,

\[
K_2(\zeta, \phi, \theta) = \frac{\sin \theta (1 - \cos \theta)}{1 + \cos \theta \cos \zeta + \sin \theta \sin \zeta \cos \phi} \quad (4.2)
\]

Note that the above integrals are now written in terms of the coordinate system constructed in equation (3.17), and illustrated in figure 1, which was applied to (3.15) and (3.16).

Until now, one was only able to show this result by picking some values of pulsar distance 1, \( L_1 \), and pulsar distance 2, \( L_2 \) and solve (4.1) numerically assuming some GW frequency \( f \). In the literature, e.g. [22, 24], the authors invoke the reader’s physical intuition to support the numerical result—that if the exponents in (4.1) are large, \( fL \gg 1 \), these oscillatory pieces rapidly converge to zero. This is often referred to as the ‘short wavelength approximation’, and has been used without proof, which we will now provide.

#### 4.1. Proof of claim

To prove this result, we estimate each of the four integrals (4.4)–(4.8) below which make up (4.1) separately. We apply the Lebesgue Dominated Convergence theorem (see appendix), Fubini’s Theorem, and the two-dimensional Divergence theorem to get the required limiting value (4.1). A key result used in the proofs which follow is a variant of the *Riemann-Lebesgue Lemma* in harmonic analysis (see [43], p. 277 and [44], p.2): let \( a, b \) be finite (though this is not necessary). Then for a Lebesgue integrable function \( f \) (a comprehensive definition and examples of this are given in appendix) over \([a, b]\),

\[
\int_a^b dt \ e^{itf(t)} \to 0 \quad \text{as} \quad x \to \infty. \quad (4.3)
\]

The aforementioned Dominated Convergence Theorem, equation (A.1), basically gives us conditions under which we can interchange the operation of taking the limit of an integral with the integral of the limit.

First, we show that \( K_2(\zeta, \phi, \theta) \) can be made continuous—and so absolutely integrable over its domain of definition—for all values of \( \theta \in [0, \pi], \zeta \in (0, \pi] \), and \( \phi \in [0, 2\pi] \).

We use the identity

\[
1 + \cos \theta \cos \zeta + \sin \theta \sin \zeta \cos \phi = 1 + \cos(\theta + \zeta) + \sin \theta \sin \zeta (1 + \cos \phi),
\]

to show that the denominator of (4.2), i.e.,

\[
1 + \cos \theta \cos \zeta + \sin \theta \sin \zeta \cos \phi \gg 0
\]

for all \( \theta \in [0, \pi], \zeta \in (0, \pi] \), and \( \phi \in [0, 2\pi] \). It follows that the only singularities of \( K_2 \) must occur when the denominator vanishes, and this occurs precisely when \( 1 + \cos(\theta + \zeta) = 0 \) and \( \sin \theta \sin \zeta (1 + \cos \phi) = 0 \), since both these quantities are necessarily non-negative. This, in turn, implies that for given \( \zeta, \theta = \pi - \zeta \) and \( \phi = \pi \) or \( \zeta = 0, \phi \) any, or \( \zeta = \pi, \phi \) any. Each of these cases is handled by limiting arguments.
For example, we note that
\[
\lim_{\zeta \to 0^+} K_2(\zeta, \phi, \pi - \zeta) = 0, \quad \lim_{\zeta \to \pi} K_2(\zeta, \phi, \pi - \zeta) = 0,
\]
and
\[
\lim_{\phi \to \pi} K_2(\zeta, \phi, \pi - \zeta) = \frac{2\sin^3\zeta}{1 - \cos \zeta}.
\]
The previous equation gives a zero limit as \(\zeta \to 0^+\) and is otherwise finite. It follows from this that \(K_2\) can be defined to be a continuous function for any given value of \(\zeta \in (0, \pi]\) and all values of \(\theta \in [0, \pi]\) and \(\phi \in [0, 2\pi]\). Thus \(K_2\) is Lebesgue integrable over the region \([0, 2\pi] \times [0, \pi]\). Next,
\[
\int_0^\pi d\theta \int_0^{2\pi} d\phi \ K_2(\zeta, \phi, \theta) \left[1 - e^{2\pi i 1/(1 + \cos \theta)}\right]
\times \left[1 - e^{-2\pi i 1/(1 + \cos \theta) \cos \phi \sin \theta \cos \phi}\right] =
\]
(4.4)
\[
- \int_0^\pi d\theta \int_0^{2\pi} d\phi K_2(\zeta, \phi, \theta) e^{2\pi i 1/(1 + \cos \theta)}
\]
(4.5)
\[
- \int_0^\pi d\theta \int_0^{2\pi} d\phi K_2(\zeta, \phi, \theta) e^{-2\pi i 1/(1 + \cos \theta) \cos \phi \sin \theta \cos \phi} \ d\phi \ d\theta
\]
(4.6)
\[
+ \int_0^\pi d\theta \int_0^{2\pi} d\phi K_2(\zeta, \phi, \theta) e^{2\pi i 1/(1 + \cos \theta) - 1/(1 + \cos \theta) \cos \phi \sin \theta \cos \phi} \ d\phi \ d\theta
\]
(4.7)
\[
+ \int_0^\pi d\theta \int_0^{2\pi} d\phi K_2(\zeta, \phi, \theta)
\]
(4.8)
the last of which is identical to the required integral, (4.16). Note that each of the previous four integrals is necessarily finite since the region of integration is finite and \(K_2\) is absolutely integrable over it.

Write \(\lambda = \mathcal{L}_1\), \(\mu = \mathcal{L}_2\). Now we treat each of the previous three integrals (4.5)–(4.7) separately, and fix \(\zeta \in (0, \pi]\).

4.2. Equation (4.5) tends to zero
Here we show that the first of the three equations with the exponential pulsar terms, (4.5), tends to zero. Using the above notation, and using the fact that \(K_2\) is absolutely integrable over \(\mathcal{R}\), Fubini’s theorem on the interchange of iterated integrals yields the equality,
\[
I_1(\lambda) := \int_0^\pi d\theta \int_0^{2\pi} d\phi \ K_2(\zeta, \phi, \theta) e^{2\pi i 1/(1 + \cos \theta)}
\]
(4.9)
\[
\quad = \int_0^{2\pi} d\phi \left\{ \int_0^\pi d\theta \ K_2(\zeta, \phi, \theta) e^{2\pi i 1/(1 + \cos \theta)} \right\},
\]
(4.10)
which, after the change of variable \(u = 1 + \cos \theta\), gives us,
\[
I_1(\lambda) = \int_0^{2\pi} d\phi \left\{ \int_0^\pi du \ K_2^2(\zeta, \phi, u) e^{2\pi i 1/u} \right\},
\]
(4.11)
where \(K_2^2(\zeta, \phi, u) = K_2(\zeta, \phi, \theta) / \sin \phi\) in the new variables is still absolutely integrable. Next, since \(K_2^2\) is absolutely integrable over its domain, the ordinary two-dimensional version of the Riemann-Lebesgue Lemma, equation (4.3), implies that
\[
\lim_{\lambda \to \infty} \int_0^\pi du \ K_2^2(\zeta, \phi, u) e^{2\pi i 1/u} = 0.
\]
Since the previous integral is itself \(\mathcal{O}(||K_2||_\infty)\), the Lebesgue Dominated Convergence theorem (A.1) can be used to interchange the order of the limit and the integral. We find:
\[
\lim_{\lambda \to \infty} I_1(\lambda) = \int_0^{2\pi} d\phi \left\{ \lim_{\lambda \to \infty} \int_0^\pi du \ K_2^2(\zeta, \phi, u) e^{2\pi i 1/u} \right\} = 0.
\]
Thus, (4.5) tends to zero as \(\lambda \to \infty\).

4.3. Equation (4.6) tends to zero
**Preamble.** In order to extend the previous idea to more general exponents, we apply integration by parts to double integrals via the Divergence Theorem. In order to prove either (4.6) or (4.7) it suffices that we obtain the decay estimates \(\mathcal{O}(1/\mu)\) or \(\mathcal{O}(1/\lambda)\) as \(\mu \to \infty\) or \(\lambda \to \infty\). What follows is the general idea which we then apply to the various cases. We need to estimate limits of the form
\[ I(\omega) = \int_D d\theta \, d\phi \, f(\theta, \phi)e^{i\omega I(\theta, \phi)} = \int_D dA \, fe^{i\omega g}, \]  
(4.9)

as \( \omega \to \infty \). (Note that we used Fubini’s theorem to justify the interchange of the order of integration in the iterated integral \((4.6)\). Here \( D \) along with its boundary (or perimeter), \( C \), are completely contained in \( R \) and are chosen so that \( \nabla g(\theta, \phi) = 0 \) on and inside \( D \cup C \) (which necessarily has no points in common with \( R \)). By construction, the gradient of \( g \), \( \nabla g \) does not vanish on \( D \cup C \) and therefore the quantity

\[ u = \frac{\nabla g}{|\nabla g|^2} f \]  
(4.10)

is well-defined on \( D \cup C \).

We need to estimate the integral in \((4.9)\) for large \( \omega \). First, observe that (suppressing the variables for clarity of exposition)

\[ \nabla \cdot (u e^{i\omega g}) = (\nabla e^{i\omega g}) \cdot u + e^{i\omega g} (\nabla \cdot u), \]

where we assume, in addition, that \( f \) is sufficiently smooth so that \( \nabla \cdot u \) is defined. Since \( \nabla g \cdot u = f \) we have \( \nabla (e^{i\omega g}) \cdot u = i\omega fe^{i\omega g} \), which when inserted into the previous display and integrated over \( D \) yields,

\[ \int_D dA \, \nabla \cdot (u e^{i\omega g}) = \int_D dA (\nabla \cdot u)e^{i\omega g} + i\omega \int_D dA f e^{i\omega g} \]

\[ = \int_D dA (\nabla \cdot u)e^{i\omega g} + i\omega I(\omega). \]

An application of the divergence theorem to the integral on the left gives us,

\[ \int_D dA \, \nabla \cdot (u e^{i\omega g}) = \int_C d\sigma (u \cdot n) e^{i\omega g}, \]

where \( n \) is the unit normal to \( C \), itself oriented in the positive direction, and \( \sigma \) is arc length. Combining the two previous displays we get,

\[ I(\omega) = -\frac{i}{\omega} \int_C d\sigma (u \cdot n)e^{i\omega g} + \frac{i}{\omega} \int_D dA (\nabla \cdot u)e^{i\omega g}. \]  
(4.11)

Once we know that both integrands are absolutely integrable over \( C \) and \( D \) respectively, we get \( I(\omega) = O(1/\omega) \) or \( I(\omega) \to 0 \) as \( \omega \to \infty \), over \( D \cup C \). The results \((4.6)\) or \((4.7)\) are obtained by a careful limiting analysis of the case where \( D \cup C \) approaches \( R \) which then gives us the desired decay estimate.

**Proof, Case 1.** \( \zeta \in (0, \pi) \). Set \( g(\theta, \phi) = -(1 + \cos \theta \cos \zeta + \sin \theta \sin \zeta \cos \phi), \omega = 2\pi \beta_2 \) in \((4.9)\). Then \( \nabla g(\theta, \phi) = (\sin \theta \sin \zeta \cos \phi, -\sin \theta \cos \zeta - \cos \theta \sin \zeta \cos \phi) \), so that \( \nabla g(\theta, \phi) = 0 \) if and only if \( \sin \theta \cos \phi = 0 \) and \( \sin \theta \cos \zeta - \cos \theta \sin \zeta \cos \phi = 0 \). For \( \zeta \in (0, \pi) \) this yields the eight (8) critical (or stationary) points

\[ (\theta, \phi) = \left(\frac{\pi}{2}, \frac{3\pi}{2}\right), \left(\frac{\pi}{2}, \frac{\pi}{2}\right), \left(\frac{\pi}{2}, \frac{3\pi}{2}\right), \]

\[ (\zeta, 0), (\pi - \zeta, 0), (\zeta, 2\pi), (\pi - \zeta, 2\pi), \]

(4.12)

all of which are located on the perimeter of \( R \). Since we want \( D \) to be critical-point-free, for given \( \varepsilon > 0 \), choose

\[ D = \{(\theta, \phi): \varepsilon < \theta < \pi - \varepsilon, \varepsilon < \phi < 2\pi - \varepsilon\}, \]

and its perimeter,

\[ C = \{(\theta, \phi): \phi = \varepsilon, \phi = 2\pi - \varepsilon, \theta = \varepsilon, \theta = \pi - \varepsilon\}. \]

Then, by construction, \( \nabla g = 0 \) in \( D \) as well as on its perimeter, \( C \). Defining \( u \) as in \((4.10)\) we then obtain \((4.11)\) for suitably smooth functions \( f, g \), i.e., \( I(\omega) \to 0 \) on \( D \cup C \), for every \( \varepsilon > 0 \). Now set \( f = K_1 \) and note that both integrals in \((4.11)\) are finite on their respective region of integration. Thus, for every \( \varepsilon > 0 \),

\[ \lim_{\beta_2 \to 0} \int_D d\theta \, d\phi K_2(\zeta, \phi, \theta) e^{i\omega g(\theta, \phi)} = 0, \]

i.e., so taking the limit as \( \varepsilon \) approaches zero, we must have

\[ \lim_{\varepsilon \to 0} \lim_{\beta_2 \to 0} \int_D d\theta \, d\phi K_2(\zeta, \phi, \theta) e^{i\omega g(\theta, \phi)} = 0. \]  
(4.13)
All that remains to be shown is that the interchange of the limits in the next expression is justified, i.e.,
\[
\lim_{\varepsilon \to 0} \lim_{\beta \to \infty} \int_{\mathcal{D}} d\theta \ d\phi \ K_2(\zeta, \phi, \theta) e^{i \omega g(\theta, \phi)} = \lim_{\beta \to \infty} \lim_{\varepsilon \to 0} \int_{\mathcal{D}} d\theta \ d\phi \ K_2(\zeta, \phi, \theta) e^{i \omega g(\theta, \phi)},
\]
\[(4.14)\]
as the right hand side of (4.14) is necessarily equal to (4.6) and so must vanish as well by (4.13), which is what we set out to prove. To this end, we note that, by continuity of the integrals,
\[
\lim_{\varepsilon \to 0} \int_{\mathcal{D}} d\theta \ d\phi \ K_2(\zeta, \phi, \theta) e^{i \omega g(\theta, \phi)} = \int_{\mathcal{R}} d\theta \ d\phi \ K_2(\zeta, \phi, \theta) e^{i \omega g(\theta, \phi)}
\]
\[(4.15)\]
and that, in fact, the convergence is uniform in \(\omega\), for \(\omega \in [0, \infty)\).

Indeed, observe that
\[
\left| \int_{\mathcal{R}} d\theta d\phi K_2(\zeta, \phi, \theta) e^{i \omega g(\theta, \phi)} - \int_{\mathcal{D}} d\theta d\phi K_2(\zeta, \phi, \theta) e^{i \omega g(\theta, \phi)} \right| \\
= \left| \int_{\mathcal{R} \setminus \mathcal{D}} d\theta d\phi K_2(\zeta, \phi, \theta) e^{i \omega g(\theta, \phi)} \right| \\
\leq \int_{\mathcal{R} \setminus \mathcal{D}} d\theta d\phi |K_2(\zeta, \phi, \theta)|
\]
and this last integral may be made arbitrarily small, independently of \(\omega\), if \(\varepsilon\) is sufficiently restricted. Hence the convergence in (4.15) is uniform in \(\omega\) (actually for any \(\omega \in \mathbb{R}\) but we only require this for \(\omega\) on the half axis, \([0, \infty)\)).

We are now in a position to apply a fundamental theorem on the interchange of such limits (see citepf, p. 395) to validate the equality in (4.14) and complete the proof in the case where \(\zeta \in (0, \pi)\).

**Case 2.** \(\zeta = \pi\). In this case \(g(\theta, \phi) = -1 - \cos \theta\) is independent of \(\phi\) and the resulting double integral can be handled in a similar way as (4.5), the only difference being the presence of a negative sign in the exponent. This, however, causes no difficulty with the argument in that section and so we omit the details.

**4.4. Equation (4.7) tends to zero**

We use the same basic technique as in the proof of (4.6). The proof of the limiting result for (4.7) can be obtained by reducing it to the case of (4.6) just proved. For example, (4.7) may be rewritten in the form
\[
\int_0^\pi d\theta \int_0^{2\pi} d\phi \ K_2(\zeta, \phi, \theta) e^{2\pi i \lambda (1 + \cos \theta)} \\
\times e^{-2\pi i \mu (1 + \cos \theta \cos \zeta + \sin \theta \sin \zeta \sin \phi)},
\]
where now it is \(K_2(\zeta, \phi, \theta)\) that is absolutely integrable over \(\mathcal{R}\), since \(K_2\) is and the exponential term has modulus equal to one. So, it follows from the methods above leading to (4.6) approaching zero as \(\mu \to \infty\) that (4.7) also tends to zero as \(\mu \to \infty\). Similarly, interchanging the \(\mu\) and \(\lambda\) terms in the preceding integral we obtain that (4.7) tends to zero as \(\lambda \to \infty\) as well.

**4.5. Final result: the Hellings and Downs curve**

We have shown that for an isotropic GWB, and for \(\beta \to \infty\), that the pulsar terms tend to zero. We can now write down the final form of the overlap reduction function: the ‘Hellings and Downs’ curve [3]:
\[
\int_0^\pi d\theta \int_0^{2\pi} d\phi \ K_2(\zeta, \phi, \theta) \\
= \frac{\sqrt{\pi}}{2} \left\{ 1 + \cos \frac{\zeta}{3} + 4(1 - \cos \zeta) \ln \left( \frac{\sin \zeta}{2} \right) \right\} (1 + \delta_{1,2}),
\]
\[(4.16)\]
for \(\zeta \in (0, \pi)\) by [5, 22, 41]. Several comments should be made about equation (4.16) regarding a choice of normalization for the Hellings and Downs curve, the failure of the short-wavelength approximation, and the subsequent approximation of the pulsar term by a delta function for the autocorrelation.

When evaluating the autocorrelation, one can easily see that the value of the overlap reduction function is \(4 \sqrt{\pi} / 3\). Indeed, it is a choice of normalization to set the autocorrelation equal to one when \(\zeta = 0\), which requires that equation (4.16) be multiplied by \(\sqrt{\pi} / 3\).

Next, one will note the \((1 + \delta_{1,2})\) term: this takes into account the failure of the short-wavelength approximation when evaluating the autocorrelation term. We approximate this as a delta function, however, in appendix C of [25], equation C4, it is shown that the exact solution is
\[
\Gamma_{\text{auto}}(f, L_1 = L_2) = \frac{1}{(2\pi)^2} \left\{ \frac{8\pi}{3} - \frac{1}{\pi(4L^2)}[1 - j_0(4\pi L)] \right\},
\]

where we adopt natural units (\(c = 1\)). Here \(j_0(x) = \sin x/x\) is a spherical Bessel function of the first kind. Since the oscillatory piece is suppressed by a factor of \(1/(4L^2)\), approximating the pulsar term by a multiplicative factor of 2 for the \(\zeta = 0\) autocorrelation case is justified.

Mingarelli and Sidery 2014 [24] also explored analytic expressions for the autocorrelation, and found that the \(\kappa(L_{1,2}, \Omega)\) term, equation (3.16), can be well approximated as \(2 - 2 \cos M\), where \(M = 2\pi fL(1 + \cos \theta)\). They showed numerically that the \(2 \cos M\) term contributed very little for large values of \(fL\).

We note that for small values of \(\zeta\) and \(fL\), there is an intermediate regime where one requires more than \((1 + \delta_{1,2})\) to approximate \(\kappa(L_{1,2}, \Omega)\). This case was explored in detail by Mingarelli and Sidery 2014 [24], who found that there are strong additional correlations from the pulsar term when the pulsars are separated by less than one gravitational wavelength. The authors also gave first and second order corrections for these cases.

5. Discussion and conclusion

We have shown analytically that the Hellings and Downs curve approaches the Earth–term only solution, even when the pulsars are arbitrarily distant from the Earth, and not themselves at the same distance \(L\) from the Earth. Of course, the case when \(fL_1 = fL_2 = fL\) is easily recovered, since in this case, \(\lambda = \mu\) and all terms (4.5)–(4.7) approach zero as the common value of this parameter approaches infinity.

The proofs indicate that the asymptotic estimate, equation (4.1), holds for sufficiently smooth kernels that are absolutely integrable over the region \(\mathcal{R}\), and not just kernels of the form \(K_\lambda(\zeta, \phi, \theta)\) as considered here.

The astrophysical interpretation of this result is that if one monitors any galactic millisecond pulsar, and cross-correlates it with a pulsar in e.g. the Large Magellanic Cloud, the Hellings and Downs curve would still be correct correlation function to use, under the assumption that the GWB is isotropic. Anisotropic GWBs can be handled similarly, but care is required when evaluating the new kernel.

To summarize, we have shown that for pulsars at distances \(L_1\) and \(L_2\) from the Earth, that the pulsar terms tend to zero as the \(fL_1 \to \infty\). The asymptotic estimate (4.1) is false if \(fL_1\) is fixed as there is no reason for the integral (4.6) to tend to zero as \(fL_1 \to \infty\), since it is independent of \(fL_1\). While this result is consistent with the previous intuition developed in the field of nanohertz GW astronomy, and indeed verified numerically for a few values of \(fL\) in [24], it has never before been proven analytically or generally for any \(fL_1\). This result is an important validation of a fundamental result in the field, and lends credibility and rigor to current GWB searches as we enter detection era in nanohertz GW astronomy.

Acknowledgments

The authors thank Yacine Ali-Haïmoud for useful comments and a thorough reading of this manuscript. Figure 1 was generated in part with the online tool ‘gwplotter’ [32]. The Flatiron Institute is supported by the Simons Foundation.

Appendix A. Overview of Lebesgue integration

Here we give a brief overview of the Lebesgue integral on the real line, \(\mathbb{R}\), see ([45]), Chapter 10. Readers familiar with this concept may proceed immediately to the Main Results in section 3 and their proofs.

One of the great advantages of the Lebesgue integral over the classical (Riemann) integral is in the handling of limiting processes such as limits of integrals of sequences of functions which, in the classical case, usually requires uniform convergence of the sequence in question—this is not the case for the Lebesgue integral.

Fundamental to this now-standard integral in mathematics is the notion of Lebesgue measure. One intuitive notion of measure assigns the value \(b - a\) for the length of an interval \([a, b]\) or, more generally, \((b - a)(d - c)\), for the area of a plane rectangle \(\mathcal{R} = [a, b] \times [c, d] = \{(x, y)\mid x \in [a, b] \text{ and } y \in [c, d]\}\) formed by the Cartesian product of the two intervals \([a, b]\) and \([c, d]\). The notion of measure allows one to extend this property (length, area, volume, etc) to Lebesgue measurable sets. Aside traditional examples such as rectangles, unions of disjoint intervals, etc one can consider the set \(Q\) of rational numbers on the real line, or its complement, the set of irrational numbers there. Both the latter two sets are Lebesgue measurable, the former has Lebesgue measure zero while the latter has positive Lebesgue measure.

More precisely, a set \(E \subset \mathbb{R}\) is said to have Lebesgue measure zero if for any given \(\varepsilon > 0\) there is a sequence of intervals \([a_n, b_n]\), \(n = 1, 2, \ldots\) such that \(E\) is contained within the union of all these intervals where, in addition, \(\sum_{n=1}^{\infty} (b_n - a_n) < \varepsilon\). For example, the set of all rational numbers \(Q \subset \mathbb{R}\) has measure zero.
One of the connections between the Lebesgue theory and the ordinary Riemann theory of the integral is the following result: If a function $f$ is bounded on an interval $[a, b]$ then $f$ is Riemann integrable over $[a, b]$ if and only if the set of its discontinuities is a set of measure zero. By its very definition, the Lebesgue integrability of $f$ forces that the absolute value of the function, $|f|$, in question be Lebesgue integrable, which is not so in the case of the Riemann integral. Still, the advantage in using Lebesgue integrals is huge in that we can extend the class of functions and sets over which we are integrating and still get meaningful results.

Now we offer an extremely brief introduction to the Lebesgue integral, with a few key definitions and results. Briefly, with the Lebesgue integral we subdivide the range of a given function $f$, look for those parts where horizontal lines intersect the graph of $f$ and then drop rectangles onto the domain of $f$. (Recall that we subdivide the domain of $f$ in the case of the Riemann integral).

More generally, the measure of a set $E \subset \mathbb{R}$ is the greatest lower bound (see e.g. [45], p. 11) of the set of all numbers of the form $\sum_{n}^{\infty} \left( b_{n} - a_{n} \right)$ where the union of all the intervals $[a_{n}, b_{n}]$ contains $E$. A set $E$ is said to be Lebesgue measurable if it has a finite measure. A function $f$ is said to be measurable if the special set $F = \{x: f(x) > a\}$ is measurable for every real number $a$. These are precisely the functions that we can integrate in the Lebesgue sense, i.e., the Lebesgue integrable functions over $E$. First, the integral is defined for simple functions (i.e., those functions whose range is a finite set of points). Then, using the fact that for a given measurable function $f(x) \geq 0$ there is a monotonically increasing sequence of simple functions $s_{n}(x)$ that converges to $f(x)$ and whose integrals, $\int_{E} s_{n}(x)\,dx$, are bounded by a constant $C$ (that depends on $f$), we define the Lebesgue integral of $f$ over $E$ by

$$\int_{E} f(x)\,dx = \lim_{n \to \infty} \int_{E} s_{n}(x)\,dx,$$

Finally, when $f$ is a general measurable function (i.e., not necessarily non-negative) we define its integral using its decomposition into positive and negative parts, that is, we know that $f(x) = f^{+}(x) - f^{-}(x)$ where $f^{\pm}(x) = \max \pm f(x)$, 0). Thus, the Lebesgue integral of a general measurable function $f$ over $E$ is by definition,

$$\int_{E} f(x)\,dx = \int_{E} f^{+}(x)\,dx - \int_{E} f^{-}(x)\,dx,$$

It is then an easy matter to see that

$$\int_{E} |f(x)|\,dx = \int_{E} f^{+}(x)\,dx + \int_{E} f^{-}(x)\,dx.$$

The Lebesgue Dominated Convergence theorem states that if $f_{n}(x)$ is a sequence of measurable functions with $f_{n}(x) \rightarrow f(x)$ almost everywhere on $E$. (This means that $f_{n}(x) \rightarrow f(x)$ at all points $x$ except for a set of measure zero.) In addition, let $g \geq 0$ be Lebesgue integrable on $E$ with $|f_{n}(x)| \leq g(x)$ then

$$\lim_{n \to \infty} \int_{E} f_{n}(x)\,dx = \int_{E} f(x)\,dx. \quad (A.1)$$

This theorem is the main one being used in this paper in order to guarantee convergence of the various integrals. A similar theory and similar results hold in 2 or more dimensions.

The space $L^{1}[0, \pi]$ is by definition the space of all complex valued integrable functions $f$ such that

$$\int_{0}^{\pi} |f(x)|\,dx < \infty.$$ 

In addition, the space $L^{\infty}(\mathbb{R}^{+})$ is the space of all such functions $f$ for which there exists a constant, $C$, depending on $f$, such that $|f(x)| < C$ almost everywhere on $\mathbb{R}^{+}$.

**ORCID iDs**

Chiara M F Mingarelli © [https://orcid.org/0000-0002-4307-1322](https://orcid.org/0000-0002-4307-1322)
Angelo B Mingarelli © [https://orcid.org/0000-0003-3870-5965](https://orcid.org/0000-0003-3870-5965)

**References**

[1] Abbott B P et al 2016 *Phys. Rev. Lett.* 116 061102
[2] Amaro-Seoane P et al 2017 arXiv:1702.00786
[3] Szanin M V 1978 *Soviet Astronomy* 22 36
[4] Detweiler S 1979 *Astrophys. J.* 234 1100
[5] Hellings R W and Downs G S 1983 *Astrophys. J. Lett.* 265 L39
[6] Bertotti B, Carr B J and Rees M J 1983 *Mon. Not. R. Astron. Soc.* 203 945
[7] Arzoumanian Z et al 2018 *Astrophys. J.* 855 37
[8] Desvignes G et al 2016 *Mon. Not. R. Astron. Soc.* 458 3341
[9] Shannon R M et al 2015 *Science* 349 1322
