TAILS OF COMPOSITE RANDOM MATRIX DIAGONALS: THE CASE OF THE WISHART INVERSE*

Aris L. Moustakas

Physics Department, National Capodistrian University of Athens
157 84 Zografou, Athens, Greece

(Received April 11, 2011)

We analytically compute the large-deviation probability of a diagonal matrix element of two cases of random matrices, namely $\beta = \left[\boldsymbol{H}^\dagger \boldsymbol{H}\right]^{-1}_{11}$ and $\gamma = \left[\boldsymbol{I}_N + \rho \boldsymbol{H}^\dagger \boldsymbol{H}\right]^{-1}_{11}$, where $\boldsymbol{H}$ is a $M \times N$ complex Gaussian matrix with independent entries and $M \geq N$. These diagonal entries are related to the "signal to interference and noise ratio" (SINR) in multi-antenna communications. They depend not only on the eigenvalues but also on the corresponding eigenfunction weights, which we are able to evaluate on average constrained on the value of the SINR. We also show that beyond a lower and upper critical value of β, γ, the maximum and minimum eigenvalues, respectively, detach from the bulk. Responsible for this detachment is the fact that the corresponding eigenvalue weight becomes macroscopic (*i.e.* $O(1)$), and hence exerts a strong repulsion to the eigenvalue.

DOI:10.5506/APhysPolB.42.1105
PACS numbers: 84.40.Ua

1. Introduction

Random matrix theory has recently seen a flurry of applications in communications. The random matrix under study may be the matrix of random channel amplitudes between transmitting and receiving multi-antenna arrays, [1, 2] or an array of pseudo-random code vectors used in a multi-user communications setting in order to scramble the signals from other users [3, 4].

One metric to characterize the performance of the communications is the mutual information, which gives the ultimate number of bits per channel use that can be transmitted without error. The ergodic mean and the fluctuations of this quantity has been analyzed under a wide range of assumptions

* Presented at the XXIII Marian Smoluchowski Symposium on Statistical Physics, "Random Matrices, Statistical Physics and Information Theory", Kraków, Poland, September 26–30, 2010.

regarding the channel statistics [2, 5, 6]. For slowly time-varying channels a better metric for the performance is the so-called outage capacity which provides an achievable transmission rate given a probability that this rate cannot be supported from the underlying fading channel [7]. As a result, a number of works showed that for large antenna numbers the fading statistics become Gaussian [8, 9]. More recently, the tails of the distribution were also calculated using the Coulomb Gas approach [10].

To obtain the above full advantages from multiple antennas, it is necessary to have an optimal receiver structure, which however is quite complex to implement in real systems. Instead, low complexity, albeit suboptimal, linear receivers offer as a practical alternative.

Such receivers include the so-called MMSE (minimum mean square error) and the zero-forcing (ZF) receivers. The information throughput performance depends on the ability of the linear receiver structure to mitigate interference. One very useful method to quantify the performance is through the asymptotic analysis of the signal to interference and noise ratio (SINR) for the MMSE receiver in the limit of large antenna numbers using tools from random matrix theory.

As in the case of the mutual information, when the channel is slowly varying, it is important to evaluate the full probability distribution of the SINR to obtain the probability of outage for a given target SINR value. This is important when the number of antennas is not too large, in which case the fluctuations play an important role.

In a seminal work [11] the authors proved the asymptotic normality of the SINR for the MMSE and ZF receivers when all transmitters have equal power. More recently, [12, 13] showed the normality of the MMSE SINR. Unfortunately and in contrast to the total mutual information, the Gaussian approximation for the SINR behaves badly unless the number of channels is quite large. As a result, inspired by the fact that the SINR for the equal power MIMO ZF receiver has a Gamma distribution[11, 14], several works were devoted in approximating the SINR statistics with other distributions, notably the Gamma and generalized Gamma probability densities [15, 16, 17, 18], by matching the first three moments. Nevertheless, this methodology, although perhaps providing good agreement under certain conditions, is *ad hoc* and does not offer any intuition on the SINR statistics. Finally, it should be pointed out that the exact distribution of the SINR has been calculated recently in terms of ratios of determinants [19]. Nevertheless, such an analysis is quite tedious and does not provide any intuition about the result.

In this paper, we take a different approach. Instead of trying to prove Gaussian behavior close enaugh to the peak of the distribution of SINR, we apply the Coulomb Gas methodology, which allows us to calculate the

distribution of the SINR arbitrarily far from its most probable, ergodic value. The Coulomb Gas model, which was introduced originally by [20] and more recently has seen numerous applications [21, 22, 23, 24, 25, 10], treats each eigenvalue of a random matrix as a point charge in the presence of an external potential while repelling the others. To apply this model, we rely on the fact that the SINR can be written in terms of a diagonal matrix of a random matrix and hence as a sum over the eigenvalues of the matrix. Nevertheless, since this sum depends not only on the eigenvalues but also the weights of the corresponding eigenfunctions on the matrix element, we need to generalize the Coulomb Gas approach to take into account the effects of the fluctuating weights. It should also be mentioned that single matrix element distributions of related quantities have been evaluated elsewhere [26], but in a different context and without exemplifying the interaction between eigenvalues and eigenfunctions.

Outline: In the next section we present the channel model and introduce the concepts of the SINR for the ZF and MMSE receiver. In Section 3 we present our analytical results, while in Section 4 we demonstrate their validity numerically and we conclude in Section 5.

2. Problem statement

In this section we define the channel model. We consider a wireless communications system with an N antenna transmitter array and an M antenna receiver array. It is typically assumed that $M \geq N$ and the ratio is defined as $\alpha = M/N \geq 1$. The M-dimensional received signal vector $\boldsymbol{y}$ can be written as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{z}\,, \tag{1}$$

where the vector $\boldsymbol{x}$ represents the transmitted signal with identically distributed elements and variance $\mathbb{E}[\boldsymbol{x}\,\boldsymbol{x}^{\dagger}] = \rho \boldsymbol{I}_N$. $\boldsymbol{z}$ is the noise vector, with independent complex Gaussian elements $\sim \mathcal{CN}(0,1)$. The channel matrix $\boldsymbol{H}$ is assumed to have independent elements $\sim \mathcal{CN}(0, 1/N)$.

The basic communications problem at the receiver is to deduce $\boldsymbol{x}$ from $\boldsymbol{y}$, given the knowledge of the channel $\boldsymbol{H}$. Even though the optimal receiver structure leads to the maximum throughput per channel use there are several suboptimal receivers, which are popular because they are linear in their implementation. The most common ones are the so-called minimum-mean-square-error (MMSE) and the zero-forcing (ZF) receivers. In both cases the vector $\boldsymbol{y}$ is multiplied by a matrix $\boldsymbol{P}$ in an effort to mitigate the noise and interference.

2.1. MMSE receiver

In this case the matrix $\boldsymbol{P}_{\rm mmse}$ is

$$\boldsymbol{P}_{\rm mmse} = \left[\boldsymbol{I}_M + \rho \boldsymbol{H}\boldsymbol{H}^\dagger\right]^{-1} \boldsymbol{H}^\dagger \,. \tag{2}$$

This matrix has the property of minimizing the average square error of the signal in the presence of the noise $\boldsymbol{z}$. The output signal can be then expressed as

$$\hat{\boldsymbol{x}} = \boldsymbol{P}_{\rm mmse}\, \boldsymbol{y} \,. \tag{3}$$

The resulting signal-to-interference-and-noise ratio (SINR) for each signal stream x_i for $i = 1, \ldots, N$ is given by [3]

$$\gamma_i \;=\; \frac{1}{\left[\left(\boldsymbol{I}_N + \rho \boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1}\right]_{ii}} - 1 \,. \tag{4}$$

It will turn out to be convenient to parameterize this quantity by $z_i = \gamma_k/\rho$ and re-write the above equation as

$$\begin{aligned} \frac{1}{1+\rho z_i} \;&=\; \left[\left(\boldsymbol{I}_N + \rho \boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1}\right]_{ii} \\ &=\; \sum_{j=1}^{N} \frac{|u_{ji}|^2}{1+\rho x_j} \,. \end{aligned} \tag{5}$$

In the second line we have expressed the ith diagonal element of the matrix in the r.h.s. of (5) in terms of the eigenvalues x_j and the matrix elements of the unitary matrix $\boldsymbol{U}$, which diagonalizes $\boldsymbol{H}^\dagger\boldsymbol{H}$. Since the elements of $\boldsymbol{H}$ are $\sim \mathcal{CN}(0, 1/N)$, $\boldsymbol{U}$ is Haar-unitary matrix. As a result, the quantities $|u_{ji}|^2$ for fixed i and $j = 1, \ldots, N$ are uniformly distributed in $(0, 1)$ with the constraint

$$\sum_{j=1}^{N} |u_{ji}|^2 = 1 \,. \tag{6}$$

2.2. ZF receiver

Similarly the SINR of the zero-forcing (ZF) receiver can be obtained. In this case the projector matrix $\boldsymbol{P}_{zf}$ is simply the pseudo-inverse of the matrix $\boldsymbol{H}$ ($\boldsymbol{H}^+$), which of course exists with probability one only when $M \geq N$. As a result, the output vector is

$$\hat{\boldsymbol{x}} = \boldsymbol{P}_{zf}\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{H}^+\boldsymbol{z} \,. \tag{7}$$

We see that multiplication with $\boldsymbol{H}^{+}$ on $\boldsymbol{y}$ kills all self-interference of signals, since all terms involving signals x_q, with $q \neq i$ are forced to zero (hence the name "zero-forcing"). Of course this comes at the cost of increasing the noise. The corresponding SINR β_i can be written as

$$\begin{aligned} \frac{\rho}{\beta_i} &= \frac{1}{z_i} = \left[\left(\boldsymbol{H}^{\dagger}\boldsymbol{H}\right)^{-1}\right]_{ii} \\ &= \sum_{j=1}^{N} \frac{|u_{ji}|^2}{x_j} \end{aligned} \tag{8}$$

with the quantities u_{ji} and x_j defined as above.

It is worth pointing out that the SINR of both MMSE and ZF cases above may be written as a sum over a function of eigenvalues weighted by the corresponding eigenvector weight

$$\sum_{j=1}^{N} |u_{ji}|^2 \, s(x_j) \, . \tag{9}$$

Also, it is important to mention that in the limit of large ρ, z_{mmse} coincides with z_{zf}, *i.e.* $z_{zf} = \lim_{\rho\to\infty} z_{\mathrm{mmse}}$ and thus $\beta(\rho) = \rho \lim_{\rho'\to\infty} \gamma(\rho')/\rho'$. As a result, we will focus on the distribution of the MMSE SINR first, from which we will be able to derive all results for the ZF SINR by taking the appropriate limit.

3. Technical analysis

In this section we will go through the basic steps of the calculation of the probability distribution function (PDF) of the normalized SINR z_i, omitting the index i when necessary. Keeping in mind the statistics of u_{ji} and their constraint (6) we write the PDF of z as

$$\mathbb{P}(z) = \frac{1}{N|s'(z)|} \, \mathbb{E}_{\boldsymbol{x},\boldsymbol{t}} \left[\delta\left(N s(z) - \sum_{j=1}^{N} s(x_j) t_j \right) \right] , \tag{10}$$

where the expectation is over the vector $\boldsymbol{x}$ of the $\boldsymbol{H}^{\dagger}\boldsymbol{H}$ eigenvalues and the random vector $\boldsymbol{t}$ with distribution identical to the quantities $N|u_{ji}|^2$, for fixed i and $j = 1, \ldots, N$. We have used the compact notation $s(x)$ to indicate both ZF and MMSE cases above. For simplicity we will omit the dependence of $s(z)$ on z as well as the overall proportionality factor unless

explicitly mentioned. We may explicitly integrate over $\boldsymbol{t}$ by first expressing the above δ-function as well as the constraint

$$\sum_{j=1}^{N} t_j = N \tag{11}$$

as Fourier integrals. As a result we obtain

$$\begin{aligned}\mathbb{P}(z) \;&\propto\; \int dk \int d\lambda\, \mathbb{E}_{\boldsymbol{x}} \left[e^{N(ks+\lambda)} \prod_{j=1}^{N} \int_0^N dt_j e^{-(\lambda+ks(x_j))t_j} \right] \\ &\propto\; \int dk \int d\lambda\, \mathbb{E}_{\boldsymbol{x}} \left[e^{N(ks+\lambda)} \prod_{j=1}^{N} \left[\frac{1-e^{-N(ks(x_j)+\lambda)}}{(ks(x_j)+\lambda)} \right] \right] . \end{aligned} \tag{12}$$

Note that that although the integral over k and λ is along the imaginary line, the saddle-point will lie on the real axis and hence we omit the imaginary i for simplicity. The expectation over $\boldsymbol{x}$ is performed with the eigenvalue distribution $P(\boldsymbol{x})$ given by

$$P(\boldsymbol{x}) \;\propto\; \Delta(\boldsymbol{x})^2 \prod_{j=1}^{N} x_j^{M-N} e^{-Nx_j} \equiv e^{-N^2 F(\boldsymbol{x})} \tag{13}$$

with $\Delta(\boldsymbol{x}) = \prod_{i>j}(x_i - x_j)$ the Vandermonde determinant and the second equation being the definition of $F(\boldsymbol{x})$. When N is large, the eigenvalues of $\boldsymbol{H}^\dagger \boldsymbol{H}$ form a tight density which can be represented as a density $p(x)$ corresponding to the quantity

$$p(x) = \frac{1}{N} \sum_j \delta(x - x_j) \,. \tag{14}$$

As a result we may rewrite (12) as

$$\mathbb{P}(z) \;\propto\; \int dk\, d\lambda \int Dp e^{-N^2 F[p]} e^{-N E_0[p]} \,, \tag{15}$$

where $\int Dp$ represents a path integral over non-negative, normalized $p(x)$, $F[p]$ is the energy functional associated with the probability distribution of eigenvalues (13) [27, 22, 10] and $E_0[p]$ is the functional obtained from the exponent in (10)

$$F[p] = \int_a^b dx p(x) \left(x - (\alpha - 1)\ln(x) - \int_a^b dx' p(x') \ln|x - x'| \right) \quad (16)$$

$$E_0[p] = -(ks + \lambda) - \int_a^b dx p(x) \ln \left[\frac{1 - e^{-N(ks(x)+\lambda)}}{ks(x) + \lambda} \right] . \quad (17)$$

It is crucial to point out that the functional $F[p]$ in (13) is multiplied by N^2 while $E_0[p]$ is only multiplied by N. Hence the fluctuations of $F[p]$ will be far smaller than those of $E_0[p]$. As a result, to leading order in N we may first find the optimal distribution that minimizes $F[p]$. This distribution is the celebrated Marcenko–Pastur distribution given by [22, 27, 10]

$$p_0(x) = \frac{\sqrt{(x-a)(b-x)}}{2\pi x} , \quad (18)$$

where the limits of the support are

$$a, b = \left(\sqrt{\alpha} \pm 1\right)^2 . \quad (19)$$

Subsequently, using this $p_0(x)$ we may find the optimal values of k and λ that minimize $E_0[p_0]$. This two-tiered approach works, as mentioned before, because, for large N, the eigenvalue distribution has much smaller fluctuations compared to the fluctuations of the unitary matrix elements $|u_{ji}|^2$ and k, λ.

We next analyze the above equations in two separate regimes, depending on whether the quantity $\lambda + ks(x)$ is positive. When it is, the exponential factor inside the logarithm of (17) is negligible, and we may therefore omit it. This corresponds to the situation when all weights of the eigenvalues are of similar size, *i.e.* $|u_{ji}|^2 = O(1/N)$, or equivalently $t(x_j) = O(1)$. The analysis of this region will be discussed next in Section 3.1. When $\lambda + ks(x) < 0$, we need to take the exponential explicitly into account, which we will do in Section 3.2. In this case, as we shall see, the average weight of one eigenvalue becomes macroscopic, *i.e.* $t(x) = O(N)$.

3.1. Region with $\lambda + ks(x) > 0$

In this case, the exponential inside the logarithm of (17) is exponentially small in N and therefore may be neglected. As mentioned above, all typical values of t_j are of order of unity, *i.e.* $t_j = O(1)$, with their sum fixed to N (11). In fact, the integral over $p(x)$ in (17) represents the entropy of the

random variables $\boldsymbol{t}$ for given $p(x)$ and the mean value of the $t_j = N|u_{ji}|^2$ is equal to

$$t(x_j) = N\,\mathbb{E}[|u_{ji}|^2] = \frac{1}{\lambda + ks(x_j)}\,. \tag{20}$$

The saddle-point equations for λ and k are obtained by differentiating $E_0[p_0]$ with respect to λ and k, respectively, and setting the derivative to zero

$$\int_a^b dx \frac{p_0(x)}{\lambda + ks(x)} \;=\; 1\,, \tag{21}$$

$$\int_a^b dx \frac{p_0(x)s(x)}{\lambda + ks(x)} \;=\; s(z)\,. \tag{22}$$

By identifying $1/(ks(x_j) + \lambda)$ as the average value of t_j, we immediately see that the first equation is nothing else but the normalization condition (11). Similarly, the second equation simply states that $\sum_j s(x_j)t_j = Ns$, *i.e.* imposes the δ-function constraint in (10). We note that combining the two equations we get the identity

$$\lambda + ks(z) = 1\,. \tag{23}$$

This, together with *e.g.* (22) will provide us with the optimal values of k, λ to plug into (17) and thus evaluate the leading term in the exponent of the probability distribution of s (respectively z) by evaluating it at the saddle-point.

We start by making the following convenient change of variables from k to c through

$$k = -\frac{\lambda}{s(c)}\,, \tag{24}$$

which for the MMSE case becomes $k = -\lambda(1 + \rho c)$. Once this variable is determined, the other, *e.g.* λ can be obtained from it through (23)

$$\lambda = \frac{s(c)}{s(c) - s(z)}\,. \tag{25}$$

Plugging (24) into (22) we get, after some re-arrangements,

$$\int_a^b dx\, \frac{p_0(x)}{x - c} \;=\; \frac{1}{z - c}\,, \tag{26}$$

where we have also used the fact that $s(z) = 1/(1+\rho z)$. It is interesting to point out that this equation is independent of ρ and holds also for the ZF receiver. Also it represents a balance of forces for a (yet fictitious) charge located at c: from one side we have the repulsion of the Coulomb sea, while from the other there is another (fictitious) charge located at the position dictated by the normalized SINR z. Before proceeding to integrate the l.h.s. we note that, in order to get a convergent answer, c has to take values outside the support of $p(x)$, *i.e.* $c \notin (a,b)$. We then have

$$z = \frac{\mathrm{sgn}(1+\alpha-c)\sqrt{(b-c)(a-c)} + c + \alpha - 1}{2}. \tag{27}$$

We see that the region of z for which the above equation has solutions is $|z-\alpha| \leq \sqrt{\alpha}$. This corresponds to values of c in the regions $-\infty < c < a$ (for $\alpha - \sqrt{\alpha} < z < \alpha$) and $b < c < +\infty$ (for $\alpha < z < \alpha + \sqrt{\alpha}$). Solving the above equation for c gives us

$$c(z) = z\left(1 + \frac{1}{z-\alpha}\right). \tag{28}$$

The values of $c(z)$ for which the solutions above break down are $c = a$ ($z(a) = \alpha - \sqrt{\alpha}$) and $c = b$ ($z(b) = \alpha + \sqrt{\alpha}$).

We may now calculate the exponent of the PDF for $|z-\alpha| \leq \sqrt{\alpha}$. To do so, we simply need to plug in the above values of k and λ (obtained directly from c) into (17) and calculate the corresponding integrals. As discussed before [22, 27, 10] the value of $F[p_0]$ does not depend on z and is therefore a constant (we have also omitted the dependence of E_0 on p_0)

$$\begin{aligned} E_0(z) &= -\ln\left[\frac{z-c}{\rho^{-1}+z}\right] + \frac{c+\rho^{-1}}{2} \\ &+ \frac{1}{2}\left(\mathrm{sgn}(1+\alpha-c)\sqrt{(b-c)(a-c)} - \sqrt{(\rho^{-1}+a)(\rho^{-1}+b)}\right) \\ &+ (\alpha+1)\ln\left[\frac{\sqrt{|b-c|}+\sqrt{|a-c|}}{\sqrt{\rho^{-1}+b}+\sqrt{\rho^{-1}+a}}\right] \\ &- (\alpha-1)\ln\left[\frac{\sqrt{a|b-c|}+\sqrt{b|a-c|}}{\sqrt{a(\rho^{-1}+b)}+\sqrt{b(\rho^{-1}+a)}}\right]. \end{aligned} \tag{29}$$

Plugging in the dependence of $c(z)$ we obtain the following simplified formula:

$$
\begin{aligned}
E_0(z) \;=\; & z - \alpha \ln z + \ln(\rho^{-1}+z) + \frac{\rho^{-1}+1-\alpha}{2} \\
& - \frac{\sqrt{(\rho^{-1}+a)(\rho^{-1}+b)}}{2} + \ln 4 + \frac{(\alpha+1)\ln\alpha}{2} \\
& - (\alpha+1)\ln\left[\sqrt{\rho^{-1}+a}+\sqrt{\rho^{-1}+b}\right] \\
& + (\alpha-1)\ln\left[\sqrt{b(\rho^{-1}+a)}+\sqrt{a(\rho^{-1}+b)}\right] . \qquad (30)
\end{aligned}
$$

This exponent has two interesting properties. First, it is a maximum at the ergodic value of z, which corresponds to the ergodic average of the SINR. This corresponds to $k=0$, and, following (23) also $\lambda=1$. As a result, (22) equates $s(z)$ to its ergodic average over the Marcenko–Pastur distribution, $\mathbb{E}_{p_0}[s(x)]$. As we shall see, this corresponds to the peak of the Gaussian distribution of the SINR distribution. The MMSE SINR is then given by

$$
\gamma_{\rm erg} \;=\; \rho z_{\rm erg,mmse} = \frac{\sqrt{(1-(\alpha-1)\rho)^2+4\alpha\rho}+\rho(\alpha-1)-1}{2} . \qquad (31)
$$

This can be seen by directly maximizing (30) over z. Indeed, expanding (30) close to $z_{\rm erg}$ we find

$$
E_0(z) \;\approx\; \frac{(z-z_{\rm erg})^2}{2v_{\rm erg}} , \qquad (32)
$$

where $v_{\rm erg,mmse}$ is the variance of the MMSE SINR given by [28]

$$
v_{\rm erg,mmse} \;=\; \frac{(\alpha-1)\sqrt{(1-(\alpha-1)\rho)^2+4\alpha\rho}+\rho(\alpha-1)+\alpha+1}{2\sqrt{(1-(\alpha-1)\rho)^2+4\alpha\rho}} . \qquad (33)
$$

The corresponding values of $\beta_{\rm erg}$ and $v_{\rm erg}$ for the case of the ZF receiver can be obtained by taking the limit $\rho\to\infty$ but keeping z fixed, and then multiplying with ρ to get the SINR $\beta_{\rm erg}$

$$
\beta_{\rm erg} \;=\; \rho z_{{\rm erg},zf} = \rho(\alpha-1) , \qquad (34)
$$

$$
v_{{\rm erg},zf} \;=\; \alpha-1 . \qquad (35)
$$

We see that when $\alpha=1$ the Gaussian approximation breaks down [11, 28].

Keeping only the dependence on z in (30), we see that, to leading exponential order,

$$
\begin{aligned}
\mathbb{P}_{\rm mmse}(z) \;&\propto\; \frac{z^M}{(\rho^{-1}+z)^N} e^{-Nz} , \\
\mathbb{P}_{\rm mmse}(\gamma) \;&\propto\; \frac{\gamma^M}{(1+\gamma)^N} e^{-N\gamma/\rho} . \qquad (36)
\end{aligned}
$$

This formula is remarkable for two reasons. First, it is surprising that this simple formula peaks at the ergodic value of z_{erg} in (31). Second, it settles a year-old conjecture, that the distribution of MMSE SINR should be (approximately) a Gamma distribution. Several papers in the literature [15, 17, 16] tried to fit the distribution to the Gamma distribution by fitting their moments. We see that this asymptotic form illustrates that although simple in form, it is not a Gamma distribution.

By letting $\rho \to \infty$ we can recover the distribution of z_{zf}

$$\begin{aligned} \mathbb{P}_{zf}(z) &\propto z^{M-N} e^{-Nz}, \\ \mathbb{P}_{zf}(\beta) &\propto \beta^{M-N} e^{-N\beta/\rho}. \end{aligned} \tag{37}$$

It turns out that this result is in fact exact [11, 14].

Finally, we can also evaluate the average weight of each eigenvalue $x \in [a, b]$ constrained on the value of z. Using (20), (25) and (28) we obtain

$$\mathbb{E}[t(x|z)] = \frac{1}{\lambda + ks(x)} = \frac{s(z) - s(c)}{s(x) - s(c)} \tag{38}$$

for $|z - \alpha| < \sqrt{\alpha}$, which is valid for both ZF and MMSE. This result, can also be obtained by noting that the distribution of $t(x|z)$ is exponential $\sim \exp[-(\lambda + ks(x))t]$, as seen in (12).

3.2. Region with $\lambda + ks(x) \leq 0$

Before moving on, it is worth pointing out that the above behavior is bound to break down at *some* value of z. Indeed, assuming $\lambda + ks(x) > 0$ and using (21), (22) and the fact that $s(x)$ is a decreasing function of x, we get the following inequality

$$\begin{aligned} s(z) &= \int_a^b dx \frac{p(x)s(x)}{\lambda + ks(x)} \leq s(a), \\ s(z) &= \int_a^b dx \frac{p(x)s(x)}{\lambda + ks(x)} \geq s(b). \end{aligned} \tag{39}$$

Therefore, for $z < a$ and $z > b$, the assumption $\lambda + ks(x) > 0$ and the resulting equations (21) and (22) have to break down.

To see how, we need to analyze the situation outside the region $|z - \alpha| < \sqrt{\alpha}$. We thus need to consider the situation when for some eigenvalue(s) the exponent in (17) becomes positive, *i.e.* when $\lambda + ks(x) < 0$. For the

section of the support of $p(x)$ where this occurs, the exponent in (17) will be positive, so we will need to include an additional term in $E_0[p]$, namely

$$NE_0[p] \approx -(ks(z)+\lambda)N + N\int_a^b dx p(x)\ln|ks(x)+\lambda| + N^2\int_{\mathcal{R}} dx p(x)(ks(x)+\lambda)\,, \tag{40}$$

where $\mathcal{R}$ is the region of the support of $p(x)$ (possibly including only a finite number of eigenvalues) with $\lambda + ks(x) < 0$. This extra term is an additional potential of strength $N^2 ks(x)$ exerting a force on the charge density $p(x)$. Since it is $O(N^2)$ we can no longer assume it is small and we have to take it into account explicitly together with $N^2 F[p]$ in the determination of the optimal $p(x)$. First, we need to estimate whether the number of eigenvalues affected is finite or scales with N. To answer this we start by observing that for these eigenvalues the corresponding typical value of t_j becomes of $O(N)$. Due tot the constraint $\sum_j t_j = N$ (11), there can be at most a finite number of such eigenvalues with corresponding $t_j = O(N)$. We will initially assume that it is only one such eigenvalue and later on show that this is consistent. As a result of these considerations we need to treat this eigenvalue separately from the others. Therefore, we separate the continuous part of the eigenvalue density and express is as

$$q(x) = \frac{1}{N-1}\sum_{j=1}^{N-1}\delta(x-x_j) \tag{41}$$

and denote the position of the N eigenvalue (which may be the largest or smallest depending on whether we are analyzing the case $z > \alpha + \sqrt{\alpha}$ or $z < \alpha - \sqrt{\alpha}$, respectively) by y. The exponent in (15) can be expressed as

$$-(N-1)^2 F[q] - NE_+[q]\,, \tag{42}$$

where $F[q]$ is the same energy functional as in (15). As a result, the optimal $q(x)$ is still the Marcenko–Pastur distribution (18). Thus $E_+[p_0]$ is given by

$$E_+[p,y] = y - (\alpha-1)\ln y - 2\int_a^b dx\, p_0(x)\ln|x-y| - (ks(z)+\lambda) + \int_a^b dx\, p_0(x)\ln(ks(x)+\lambda) - \int_a^b dx p_0(x)\ln\left[1-e^{-N(ks(x)+\lambda)}\right] + \frac{1}{N}\left(\ln|\lambda+ks(y)| - \ln\left[e^{-N(ks(y)+\lambda)}-1\right]\right)\,. \tag{43}$$

This need of explicitly splitting one eigenvalue from the bulk and treating it in a special way has appeared also in the context of bipartite entanglement [25, 29]. Since only $ks(x) + \lambda \geq 0$ for $x \neq y$ the last term in the second line above will only contribute subleading terms and therefore may be neglected. We now need to find the saddle-point jointly for y, λ and k.

$$1 = \int_a^b dx \frac{p_0(x)}{\lambda + ks(x)} + \frac{1}{N(\lambda + ks(y))} + \frac{e^{-N(\lambda+ks(y))}}{e^{-N(\lambda+ks(y))} - 1}, \tag{44}$$

$$s(z) = \int_a^b dx \frac{p_0(x)s(x)}{\lambda + ks(x)} + \frac{s(y)}{N(\lambda + ks(y))} + \frac{s(y)e^{-N(\lambda+ks(y))}}{e^{-N(\lambda+ks(y))} - 1}, \tag{45}$$

$$1 = 2\int_a^b dx \frac{p_0(x)}{y-x} + \frac{\alpha-1}{y} + \frac{k\rho s(y)^2}{N(\lambda+ks(y))} + \frac{k\rho s(y)^2 e^{-N(\lambda+ks(y))}}{e^{-N(\lambda+ks(y))} - 1}. \tag{46}$$

From the first equation we conclude that the values of $\lambda + ks(y) = O(1/N)$. Otherwise, if $\lambda + ks(y) = O(1)$, the integral in the r.h.s. of (44) would have to vanish, which is inconsistent with the fact that $\lambda + ks(x) > 0$. Thus, setting $\lambda + ks(y) = -w/N$ for w still unknown, we find that to leading order,

$$k = \frac{1}{s(z) - s(y)} \tag{47}$$

and w is a solution of the equation

$$\int_a^b dx \, \frac{p_0(x)(s(z) - s(y))}{s(x) - s(y)} = \frac{1}{w} - \frac{1}{e^w - 1}. \tag{48}$$

Putting this together, we finally get the equation for y:

$$\int_a^b \frac{p(x)dx}{y - x} = 1 - \frac{\alpha - 1}{y} - \frac{1}{y - z}. \tag{49}$$

This last equation is both surprising and intuitive. It is firstly surprising that the value of y does not depend on the specific form of the SINR function $s(z)$ but only on z itself, the normalized SINR. Second, it tells us that the position of this eigenvalue is determined by external forces exerted on the other eigenvalues and, in addition, it feels the repulsion of a unit charge located in the position z. Also it is interesting to note that it feels only *half* the repulsion from the Marcenko–Pastur continuous eigenvalue density. The

other half has been *screened* away due to the interaction of this eigenvalue with the matrix elements of the diagonalizing unitary matrix (through λ and k). It should be pointed out that (49) is valid for both cases $z < \alpha - \sqrt{\alpha}$ and $z > \alpha + \sqrt{\alpha}$, with $y \leq a$ and $y > b$, respectively.

In the above analysis we have assumed there is only one eigenvalue that detaches from the bulk. Let us assume there were $r > 1$ such eigenvalues. In that case, they would all have to stick together satisfying $\lambda + ks(y_j) = -w_j/N$. Otherwise, if say only one y_1 satisfied this relation, all others with $y_j - y_1 = O(1)$, for $j = 2, \ldots, r$, would necessarily have $\lambda + ks(y_j) > 0$, which would not be sufficient to provide the "kick" to get out of the bulk. However, on the other hand, if these r eigenvalues are within $O(1/N)$ from each other, their repulsion $1/(y_i - y_j)$ will be large ($O(N)$) and hence would dominate (46). As a result, only one eigenvalue can be detached from the bulk.

We may now integrate the l.h.s. of (49) to get

$$\frac{\sqrt{(a-y)(b-y)} + y - \alpha + 1}{2y} = 1 - \frac{\alpha - 1}{y} - \frac{1}{y - z}\,. \tag{50}$$

Solving for y gives

$$y \;=\; z\left(1 + \frac{1}{z - \alpha}\right), \qquad |z - \alpha| > \sqrt{\alpha} \tag{51}$$

which is identical with (28), although obtained through a completely different method, and with z here taking different values. One way to jointly interpret c and y is that they correspond to the location of a "state", which when located outside the continuum of nearby states forms a bound state (y), while when it enters the continuum, it becomes a "resonance".

We may now plug in the above results into (43) to obtain the exponent of the PDF in this region of z. The final result is

$$\begin{aligned} E_+(z) \;=\; & -\ln\left[\frac{|z-y|}{\rho^{-1}+z}\right] + \frac{y+\rho^{-1}}{2} - (\alpha-1)\ln y \\ & + \frac{1}{2}\left(\mathrm{sgn}(y-\alpha)\sqrt{(y-b)(y-a)} - \sqrt{(\rho^{-1}+a)(\rho^{-1}+b)}\right) \\ & - (\alpha+1)\left(\ln\left[\sqrt{|y-b|}+\sqrt{|y-a|}\right] + \ln\left[\sqrt{\rho^{-1}+b}+\sqrt{\rho^{-1}+a}\right]\right) \\ & + (\alpha-1)\left(\ln\left[\sqrt{a|b-c|}+\sqrt{b|a-c|}\right]\right. \\ & \left. +\; \ln\left[\sqrt{a(\rho^{-1}+b)}+\sqrt{b(\rho^{-1}+a)}\right]\right)\,. \end{aligned} \tag{52}$$

Inserting (51) into (52) we find that, up to a constant, $E_+(z)$ is identical to $E_0(z)$ in (30), thus extending the validity of the latter for all values of $z > 0$.

Finally, we evaluate the average eigenvalue weights constrained on the value of SINR (or equivalently to z), which are the analogues of (38). Using (47) we find the identical expression as (38)

$$\mathbb{E}[t(x|z)] = \frac{1}{\lambda + ks(x)} = \frac{s(z) - s(y)}{s(x) - s(y)} \,. \tag{53}$$

In addition, we may calculate the mean weight of the detached eigenvalue y. As with the other weights, its distribution is exponential $\sim \exp[-t(\lambda + ks(y))]$, (12). We thus find that in this case

$$\begin{aligned} \mathbb{E}[|u_{Nk}|^2] &= \frac{\mathbb{E}[t_y(z)]}{N} = 1 - \frac{1}{w} + \frac{1}{e^w - 1} \\ &= 1 - \frac{z((1+\rho\alpha)(z-\alpha) + \rho\alpha)}{(1+\rho z)(z-\alpha)^2(z-\alpha+1)} \end{aligned} \tag{54}$$

for $|z - \alpha| > \sqrt{\alpha}$, where $w = -N(\lambda + ks(y))$ appears in (48). We see that in the limit $|z - \alpha| = \sqrt{\alpha}$, $\mathbb{E}[t_y(z)] = 0$ (*i.e.* $= O(1/N)$). We plot

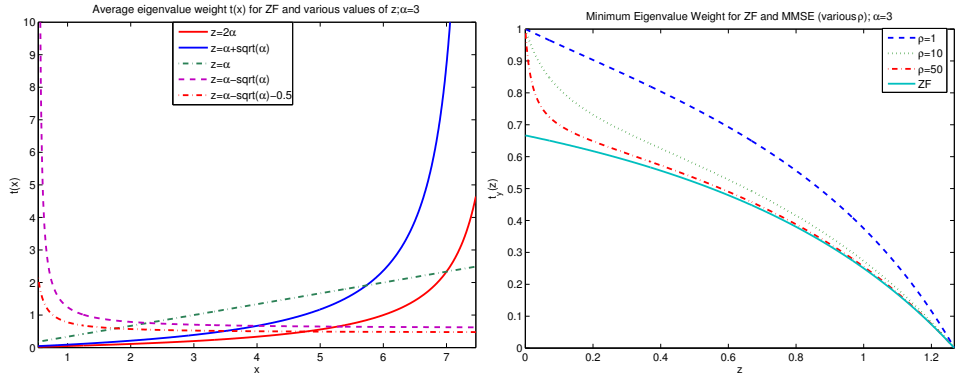

Fig. 1. Left: Weight for ZF. Average weight of the eigenfunction j on the i element, $\mathbb{E}[N|u_{ji}|^2]$ for the ZF case, as a function of the corresponding eigenvalue x_j, constrained on different values of z. The five curves include to the lower critical ($z = \alpha - \sqrt{\alpha}$), upper critical ($z = \alpha + \sqrt{\alpha}$) and an intermediate ($z = \alpha$) value of z. The remaining two curves have z below the lower critical value $z = \alpha - \sqrt{\alpha} - 0.5$ and above the upper critical value $z = 2\alpha$. In the first two we clearly see the divergence at the lower and higher edges of the spectrum, corresponding to the fact that beyond these values the weight of the edges becomes macroscopic $O(N)$. Right: Weight of $x_{\min}$. Average weight of the minimum eigenvalue for $0 < z < \alpha - \sqrt{\alpha}$. Note that now we plot the *macroscopic* occupation of the eigenvalue, *i.e.* $\mathbb{E}[|u_{\min,k}|^2]$. We plot the case of MMSE for various ρ as well as the case of ZF.

4. Numerical simulations

To test the applicability of this approach, we have performed Monte Carlo simulations and have compared our large deviations Coulomb Gas (CG) approach with Monte Carlo (MC) simulations and the Gaussian approximation. In Fig. 2 we plot the normalized probability density of the SINR for small $\rho = 1$, while in Fig. 3 we plot it for larger $\rho = 10$. In both cases we see good agreement.

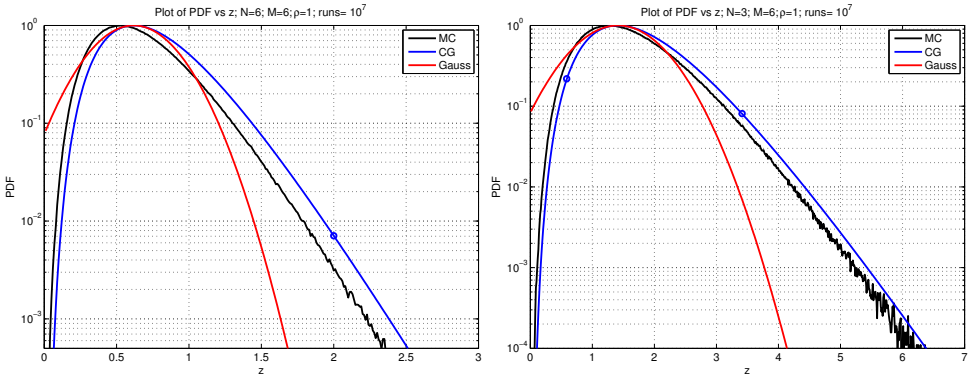

Fig. 2. Left: $\alpha = 1$. Right: $\alpha = 2$. PDF of MMSE SINR for $M = 6$, $\rho = 1$. The agreement of the Coulomb Gas (CG) curve with Monte Carlo (MC) simulations is good, even for such small matrices, especially compared to the Gaussian approximation. Denoted with circles are the values of z at which the "inner" and "outer" solutions match and we see no discontinuity in the numerics.

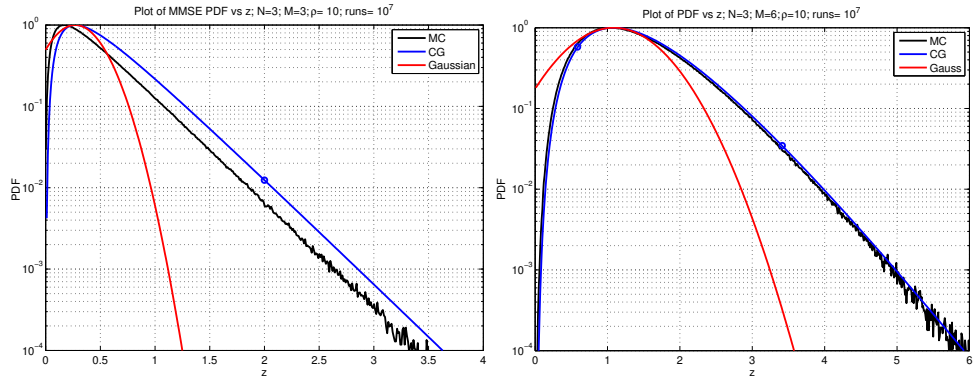

Fig. 3. The same as in the previous figures but for $\rho = 10$. Here the agreement is much better.

5. Conclusion

In this paper we have used a large deviation approach to calculate the probability density of the "signal to interference and noise ratio" (SINR) for multi-antenna arrays for two popular receiving algorithms, namely the MMSE and the ZF algorithms. The approach is formally valid for large N antenna numbers, but is not restricted to the behavior close to the peak of the distribution, which has been shown to be asymptotically Gaussian when the number of antennas is very large. Instead we calculate the probability of the SINR being arbitrarily away from its ergodic peak. Surprisingly, the leading term of the exponent of the distribution is very simple, and the distribution is neither Gamma, nor Beta and certainly not Gaussian. In the ZF case, we recover the known chi-square result. We also test the MMSE results numerically and find good agreement even for relatively small antenna arrays. From a technical point of view, since the SINR of the two algorithms are related to the diagonal matrix elements of the matrices $\left[\boldsymbol{I}_N + \rho \boldsymbol{H}^\dagger \boldsymbol{H}\right]^{-1}$ and $\left(\boldsymbol{H}^\dagger \boldsymbol{H}\right)^{-1}$, the task is to find the distribution of a single diagonal matrix element. The methodology we applied is based on the so-called Coulomb Gas model, in which each eigenvalue can be seen as a point charge interacting with an external potential and repelling each other. In this particular case however, the eigenvalues interact not only with each other but also with the weights of their corresponding eigenfunctions in the particular matrix element. As a by-product of our analysis we are able to calculate the average weight of each eigenvalue in the particular matrix element, constrained on the value of the SINR or the matrix element. The interaction between the eigenfunction weights and the corresponding eigenvalues can be quite strong and as a result, below and above critical values of z the lowest and largest eigenvalues detach from the bulk. Nevertheless, it seems that there is no discontinuity involved in this detachment, at least to leading order. In hindsight, this is not surprising. A given diagonal matrix element depends on a number of $O(N)$ random variables of the matrix, which has $O(N^2)$ random variables. In our approach we have expressed this diagonal matrix in terms of the eigenvalues, which depend on the whole matrix. Somehow, we expect that the interaction with the eigenvalue weights will "wash" out this dependence from the full matrix.

The author is grateful for the hospitality of the Jagiellonian University in Kraków, Poland where the ideas of this work were formulated. He would also like to acknowledge insightful discussions with S.N. Majumdar that took place during the 23rd Marian Smoluchowski Symposium on "Random Matrices, Statistical Physics and Information Theory", Kraków, Poland, Sept. 26–30, 2010. Also, it is a pleasure to acknowledge useful discussions with G. Caire and P. Kazakopoulos during the initial stages of the work.

REFERENCES

[1] G.J. Foschini, M.J. Gans, *Wireless Personal Communications* **6**, 311 (1998).
[2] I.E. Telatar, *Eur. Trans. Telecomm.* **10**, 585 (1999).
[3] S. Verdú, *Multiuser Detection*, Cambridge University Press, Cambridge 2003.
[4] S. Verdú, S. Shamai, *IEEE Trans. Inform. Theory* **45**, 622 (1999).
[5] A.L. Moustakas *et al.*, *Science* **287**, 287 (2000) `http://xxx.lanl.gov/abs/cond-mat/0009097`
[6] R.R. Müller, *IEEE Trans. Inform. Theory* **48**, 2495 (2002).
[7] L.H. Ozarow, S. Shamai, A.D. Wyner, *IEEE Trans. Veh. Technol.* **43**, 359 (1994).
[8] A.L. Moustakas, S.H. Simon, A.M. Sengupta, *IEEE Trans. Inform. Theory* **49**, 2545 (2003).
[9] W. Hachem *et al.*, *IEEE Trans. Inform. Theory* **54**, 3987 (2008).
[10] P. Kazakopoulos, P. Mertikopoulos, A.L. Moustakas, G. Caire, *IEEE Trans. Inform. Theory* **57**, 1984 (2011).
[11] D N. Tse, O. Zeitouni, *IEEE Trans. Inform. Theory* **46**, 171 (2000).
[12] Y.-C. Liang, G. Pan, Z.D. Bai, *IEEE Trans. Inform. Theory* **53**, 4173 (2007).
[13] A. Kammoun, M. Kharouf, W. Hachem, J. Najim, *IEEE Trans. Inform. Theory* **55**, 5048 (2009).
[14] D. Gore, R.W. Heath, A. Paulraj, in Proc. of the 2002 IEEE Intern. Symposium on Information Theory, Lausanne, Switzerland, June 2002, p. 159.
[15] P. Li, D. Paul, R. Narasimhan, J. Cioffi, *IEEE Trans. Inform. Theory* **52**, 271 (2006).
[16] A. Kammoun, M. Kharouf, W. Hachem, J. Najim, *IEEE Trans. Inform. Theory* **55**, 4386 (2009).
[17] A.G. Armada, L. Hong, A. Lozano, in Proc. of the 2009 IEEE Intern. Conference on Communications, Piscataway NJ, USA, IEEE Press, p. 3836.
[18] H. Li, A.G. Armada, *IEEE Trans. Veh. Technol.* **60**, 313 (2011).
[19] M. Kiessling, J. Speidel, in Proc. of the IEEE Vehicular Technology Conference, 2003, vol. 3, p. 1738.
[20] F. Dyson, *J. Math. Phys.* **3**, 140 (1962).
[21] P. Vivo, S.N. Majumdar, O. Bohigas, *Phys. Rev. Lett.* **101**, 216809 (2008).
[22] D.S. Dean, S.N. Majumdar, *Phys. Rev.* **E77**, 041108 (2008).
[23] S.N. Majumdar, *Random Matrices, the Ulam Problem, Directed Polymers and Growth Models, and Sequence Matching*, Les Houches, Eds. M. Mézard and J.P. Bouchaud, Elsevier, 2006.
[24] C. Nadal, S.N. Majumdar, *Phys. Rev.* **E79**, 061117 (2009).
[25] C. Nadal, S.N. Majumdar, M. Vergassola, *Phys. Rev. Lett.* **104**, 110501 (2010).
[26] D.V. Savin, H.J. Sommers, Y.V. Fyodorov, *JETP Lett.* **82**, 544 (2005).
[27] P. Vivo, S.N. Majumdar, O. Bohigas, *J. Phys.* **A40**, 4317 (2007).
[28] K.R. Kumar, G. Caire, A.L. Moustakas, *IEEE Trans. Inform. Theory* **55**, 4398 (2009).
[29] C. Nadal, S.N. Majumdar, M. Vergassola, `arXiv:1006.4091v1`.