

Shotnoise in the context of heterodyne detection and direct detection:

One of the more remarkable things about interferometric detection of a signal lies with the fact that it allows us to achieve shot noise detection sensitivity.

Shot noise detection limit can be easily understood in the context of direct detection. Consider a source with power P and optical frequency f . Suppose we detect the signal using a photodetector with detection efficiency ε (the ratio of photon to signal count conversion in the detector) (perfect efficiency = 1) and collect for a time period of τ .

In this situation, we can expect to detect $\left(\frac{\varepsilon P \tau}{hf}\right)$ photons. The uncertainty in the detected photon count is $\sqrt{\left(\frac{\varepsilon P \tau}{hf}\right)}$. The signal to noise ratio (SNR) is given then by:

$$SNR = \left(\frac{\varepsilon P \tau}{hf}\right) / \sqrt{\left(\frac{\varepsilon P \tau}{hf}\right)} = \sqrt{\left(\frac{\varepsilon P \tau}{hf}\right)}. \quad (1)$$

Notice that the SNR is dependent on the detection efficiency; we do better when the detection efficiency is as close to the maximum possible.

SNR: What IS the convention?

If you have been working in the field of optics, you will very quickly come to realize that the definition of SNR can be very confusing.

In the normal engineering convention,

$$SNR = \frac{\text{signal amplitude}}{\text{noise amplitude}}$$

$$SNR_{dB} = 20 \log_{10}(SNR)$$

In most optics related publications, (for example for OCT applications)

$$SNR_{OCT} = \left(\frac{\text{signal amplitude}}{\text{noise amplitude}}\right)^2$$

$$SNR_{OCT,dB} = 10 \log_{10}(SNR_{OCT})$$

At the end of the day, if we are discussing SNR in dB's, both conventions gets you to the same answer. But, if you are simply discussing SNR, the results are actually off by a square factor.

I am going to stick with the optics convention for the rest of this essay and subsequent essays. Just keep in mind the difference!

In the spirit of sticking with optics convention, The above equation have to be rewritten as:

$$SNR = \left(\frac{\epsilon P \tau}{hf} \right) / \sqrt{\left(\frac{\epsilon P \tau}{hf} \right)^2} = \left(\frac{\epsilon P \tau}{hf} \right). \quad (1)$$

While we are on the subject of SNR, let's explore one more phenomenon (which we shall later see, has some connections to our original query) – bright field and dark field detection. Suppose you have a weak light source with power of P_{weak} . You want to be able to detect this source in the presence of a much stronger source (power of P_{strong}).

Assuming that $P_{weak} \ll P_{strong}$, we can see that the presence of the first light source will be detectable only if the associated photon counts for it for the total detection period of τ exceeds the uncertainty in the overall detected photon count which includes both photons from the weak and strong source:

$$\left(\frac{\epsilon P_{weak} \tau}{hf} \right) \geq \sqrt{\left(\frac{\epsilon (P_{weak} + P_{strong}) \tau}{hf} \right)^2} \approx \sqrt{\left(\frac{\epsilon P_{strong} \tau}{hf} \right)^2} \quad (2)$$

We can associate a SNR quality to the detection process. In this situation, the signal is the detected weak light source photon count. The noise is the uncertainty in the overall detected photon count. The achieved SNR is:

$$SNR_{direct, bright\ field} = \left(\frac{\epsilon P_{weak} \tau}{hf} \right) / \sqrt{\left(\frac{\epsilon P_{strong} \tau}{hf} \right)^2} = \left(\frac{\epsilon \tau}{hf} \right) \frac{P_{weak}^2}{P_{strong}} \quad (3)$$

Naturally, the above equation makes a lot of sense. The longer you detect, the easier it is to distinguish the first light source. The brighter the second light source, the harder it is to detect the first light source. This situation is typically referred to as the bright field detection mode.

In dark field detection mode, we are simply trying to detect a weak light source (the first source from the above example) in the absence of other light sources. In this case, the job becomes easier; the only shot noise fluctuation comes from the light source itself (P_{weak}).

The presence of the first light source is detectable as long as the photon counts for it for a given period of τ significantly exceeds the uncertainty in the photon count:

$$\left(\frac{\varepsilon P_{weak} \tau}{hf} \right) \geq \sqrt{\left(\frac{\varepsilon P_{weak} \tau}{hf} \right)} \quad (4)$$

The achieved SNR is:

$$SNR_{direct, dark\ field} = \left(\left(\frac{\varepsilon P_{weak} \tau}{hf} \right) / \sqrt{\left(\frac{\varepsilon P_{weak} \tau}{hf} \right)} \right)^2 = \left(\frac{\varepsilon \tau P_{weak}}{hf} \right). \quad (5)$$

We can easily see that dark field detection mode will yield greater sensitivity than bright field detection mode.

Let's turn back to our original inquiry.

Suppose we have a weak light source of power P_{weak} that we would like to detect. We next assume that we have another light source of power $P_{reference}$ that is coherently in phase with the first light source. (By coherently in phase, I simply mean a light source that can interfere with the original light source.)

If we combine the two light sources in a way that achieves optimal interference, the resulting combined power can be expressed as:

$$P_{combined} = P_{reference} + P_{weak} + 2\sqrt{P_{weak} P_{reference}} \cos(k\Delta x) \quad (6)$$

The last term of the above expression can be extracted in isolation. One approach will be to modulate it by varying the optical path length mismatch, Δx , of the two light beams and then extract only the AC component. The shot noise in this situation is given by

$$\sqrt{\left(\frac{\varepsilon(P_{weak} + P_{reference})\tau}{hf} \right)} \quad (\text{basically derived from the combined DC of the first two terms}).$$

We can detect the presence of the weak light source by using the AC component as the signal. So the criterion for being able to detect the first light source becomes:

$$\frac{\varepsilon 2\sqrt{P_{weak} P_{reference}} \tau}{hf} \geq \sqrt{\left(\frac{\varepsilon(P_{weak} + P_{reference})\tau}{hf} \right)} \approx \sqrt{\left(\frac{\varepsilon P_{reference} \tau}{hf} \right)} \quad (7)$$

The final term is arrived at by arranging the reference light source to be very strong. The resulting SNR can be expressed as:

$$SNR_{\text{interference, dark field}} = \left(\frac{\varepsilon 2 \sqrt{P_{\text{weak}} P_{\text{reference}}} \tau}{hf} / \sqrt{\left(\frac{\varepsilon P_{\text{reference}} \tau}{hf} \right)} \right)^2 = \left(\frac{2\varepsilon \tau P_{\text{weak}}}{hf} \right)$$

We can see that the achieved SNR is actually equivalent to the one obtained for dark field direct detection (within a factor 2). The shot noise of the detection scheme is given by

$$\sqrt{\left(\frac{\varepsilon (P_{\text{weak}} + P_{\text{reference}}) \tau}{hf} \right)},$$

which is similar to the bright field situation and yet the SNR is completely independent of the second light source's power (when it is much stronger than the weak light source).

This apparent paradox can be resolved by focusing our attention on the difference between the interferometric detection case and the bright field case.

In the bright field case, any increase in the reference light source power will increase the shot noise term. This necessarily degrades our SNR. In the interferometric detection case, any increase in the reference light source power will correspondingly increase our useful signal - the detected modulated signal (by virtue of the cross-term $\frac{\varepsilon 2 \sqrt{P_{\text{weak}} P_{\text{reference}}} \tau}{hf}$).

So, as the shot noise increases, our signal will also increase. The proportionality of the increases are both square root in nature, so they effectively cancel each other out in the SNR ratio.

The realization of shot noise limited detection with interferometric based methods brings this class of signal detection method to par with dark field detection approach. And in fact, it can perform dramatically better in the following aspects:

1. The detector efficiency – as we can see in the above analysis, we do progressively better when the detection process approaches perfect detection. Given that a typical photodiode has the same quantum efficiency as a PMT tube, it stands to argue that interferometric detection approach is more robust in that it allows for the use of a photodiode which is more sturdy and far more tolerant of bright illumination conditions than a PMT tube.
2. The detector dark noise – A sometimes-used argument against heterodyne detection highlights the fact that photodiode intrinsically have a much higher dark noise (noise that is present in a detector even when no light is incident upon it) than a PMT. This necessarily pulls up the noise term for interferometric detection and, therefore, interferometric method for detection is inferior to direct detection with a PMT tube (which has very low dark counts).

This argument is formally correct. However, there's a trick that can be played with interferometric detection to makes it irrelevant. Namely, if we make the reference light

source sufficiently intense that its shot noise dominates over the dark noise. By doing so, we can obviate the role of dark noise in our consideration.

3. Background rejection. Interferometric detection can completely eliminate the contribution of other light sources that isn't in coherence with the light sources involved in the detection process. In other words, as long as the other light sources are generated via some other approaches, they will not show up as contributions in our final modulated signal. The same cannot be said of direct detection approaches. In those cases, unless those contributions can be filtered out by wavelength, polarization or some other ways, they will be detected and counted as noise in the system.

Given these advantages, it stands to question why interferometric based detection is not used in far more situations. The answer lies with the fact that the method can only be applied in situations where it is possible to have a reference light source that is in coherence with the targeted light source. Practically, this implies that the two light sources are derived from one and the same source. Such a condition is achievable in OCT for example, because the reflected light from the sample is in coherence with the portion of the light siphoned from the same light source to create the reference beam.

Fluorescence emission from a sample will not be in coherence with fluorescence emission from another sample (even if both samples are identical), as such there is no possibility of using interferometric based detection in that context.