Optimal exposure for DSLR and how to find it.

Preface

During astrophotography sessions many of you probably wandered if the settings on your camera were correct? Or, may be, it could be changed in subtle way and the result could be drastically better? Probably not, but these doubts remained in the back of your mind and kept bothering you at night. Well, then this article is for you, since I would like to at least try to clarify this murky water.

First of all, I hope that you've familiarized yourself with the basics such as triangle exposure, what is RAW image, and how most cameras sensors arranged and how they process images. I also strongly encourage you to read rather short but quite useful Jim Solomon's Astrophotography Cookbook. He briefly covers there such crucial topics such as planning, image acquisition and calibration, stacking, color calibration etc.

Having read all of this we can finally move to the main topic of this article. I'll base all my reasoning on the quite old but still useful article "The Noise about Noise" written by Blair MacDonald. Even though it is a serious peer-reviewed scientific paper it is not as heavy-reading as most of them and I highly advise you to read it as well.

Theory

Before jumping straight into astrophotography, I suggest you first conduct a thought experiment. Suppose we have an object with a perfectly uniform surface, such as a gray card that professional photographers like to use to set the white balance. If we shoot it from a sufficiently large distance, then all the smallest inhomogeneities on it will average out and we should get a uniformly gray background: that is, all pixels should have exactly the same brightness, for example, 128 units. In a real picture, if we move eyedropper tool over it in Photoshop, we will see that the average brightness is the same, but there are also deviations from it. We will call these deviations from the mean value noise.

In real life photons from an object arrive randomly, like raindrops. We can say how many of them, on average, fall on a certain area, but at what specific points in time they land on it, we cannot say. If N of them arrive in a unit time interval, then in the next moment of time or to another similar site, the same number of them will arrive with a deviation $\pm$ the root of N. This fluctuation is called "Photon noise".

It would seem that the larger N, the greater the spread around the mean, and the greater the noise. On the one hand, it is. On the other hand, an absolute error is rarely informative. Assume that the error in distance is one thousand kilometers. But if this is the distance to the Andromeda galaxy, then the value will be very accurate. Therefore, it is much more convenient to express it in relation to the measured value, and here, with the growth of the signal, the error decreases, exactly by the root of N times. Let's call this quotient the signal-to-noise ratio:

$$snr = \frac{N}{\sqrt{N}} \qquad 1$$

So, we got that for an ideal camera, the more photons we've collected before the effects of overexposure take place, the less noise we observe in the image. This will be our photon component. What else do we have? Well, real camera instead of the ideal one has to operate in nonideal conditions and it has to somehow save the information into file the further processing. The photons that hit the photosensitive element increase (or decrease, it doesn't matter to us now) the charge of the pixel. Suppose a pixel has accumulated some amount of charge. Then we can apply this signal to the analog-to-digital converter (ADC) and write the value given to us. On the other hand, we can first multiply our signal by a certain constant, and only after that digitize it. If the ADC itself was ideal and didn't add its noise, there would be no difference whether we digitized the signal before or after multiplication. However, it not ideal. Therefore, if we say the level of this noise is around 6 electrons, and we have accumulated only 12 of them in a pixel, then we will introduce quite a noticeable noise during the digitization process. However, if we first "increase the sensitivity" of our matrix, say, by a factor of 4, then the same 6 noise electrons against the value of 48 useful ones will stand out much less. We can say that in the first case, the signal-to-noise ratio (snr) was 2, and we increased it to 8.

It would seem, why then not increase the sensitivity to the maximum possible value? Firstly, any photographer who is used to shooting in RAW knows such a thing as "overexposure" — when the dynamic range of the camera is not enough to display the brightest detail of the scene being shot without information loss. Simply because the pixel capacity is limited (some old DSLRs even suffered from the blooming effect). If our pixel is capable of holding, for example, 2^{15} photoelectrons, and on the resulting frame the corresponding area has already become completely white, then by multiplying this number by 4, we will effectively be quartering the depth of the potential well for photoelectrons, because now already 2^{15-2} photoelectrons will correspond to white color, and all greater values — even more so. And secondly, we still don't obtain fundamentally new information from such a multiplication — only slightly reduce the contribution of ADC noise.

And if it is summer and the matrix is warm (they noticeably heat up during operation) then inside the photodiode, which accumulates light, processes of spontaneous generation of electron-hole pairs are possible, even when there is no light hitting the sensor. This is the so-called dark current. Astrophotographers are well acquainted with it and curse it with all the bad words, because in the morning, when the process of shooting the main object is already completed, ~~they~~ we are forced to make calibration frames with the same duration as the light-frames, just to evaluate this spurious signal, which is even different in each pixel.

Since all these measurement errors — photon noise, dark current, readout noise — operate absolutely independently of each other, all their contributions to the errors are summed by the Pythagorean theorem, as a root from the sum of the squares. As a result, we get the formula for snr in the form:

$$snr = \frac{n_l t}{\sqrt{n_l t + n_d t + RN^2}} \qquad 2$$

where n_l is the frequency of occurrence of photoelectrons from the observed object, n_d is the frequency of occurrence of photoelectrons as a result of thermal processes without matrix illumination, RN is the readout noise in electrons, and t is the exposure time. For the case of shooting faint objects at the light-polluted sky, the photon noise from the sky must also be added to the denominator and under the root: $n_s t$. That's why shooting in the cities is so futile: photon noise from the sky clogs our potential well with photoelectrons, but these are not the ~~droids that you are looking for~~ the photons we need.

Estimating of the contributions for each type of noise

After we have decided what exactly we called noise, we need to evaluate the contribution of each of them to the final result.

I will start with the dark current, because I want to leave it out from further consideration, at least for now, since the task is already overgrown with numerous factors. First, thermal/dark noise is called so because its magnitude depends on temperature. So, the colder our sensor the better.

Let's go back to the other two noises: photon noise from the object + the sky and the read noise. Unfortunately, we cannot turn off the light pollution — we can only drive as far as possible. Obviously, the longer the shutter speed we set, the more useful signal we will collect. And this is true as long as the object we shoot does not go into the overexposure region.

It is interesting, however, to find out how strong the useful signal must be so that the last term, which is responsible for the unremovable readout noise, ceases to worry us? It can be seen that in the formula for snr all our values are in the denominator under the root. Let's express the photon noise in terms of read noise and plot how much of the total noise the read noise contributes:

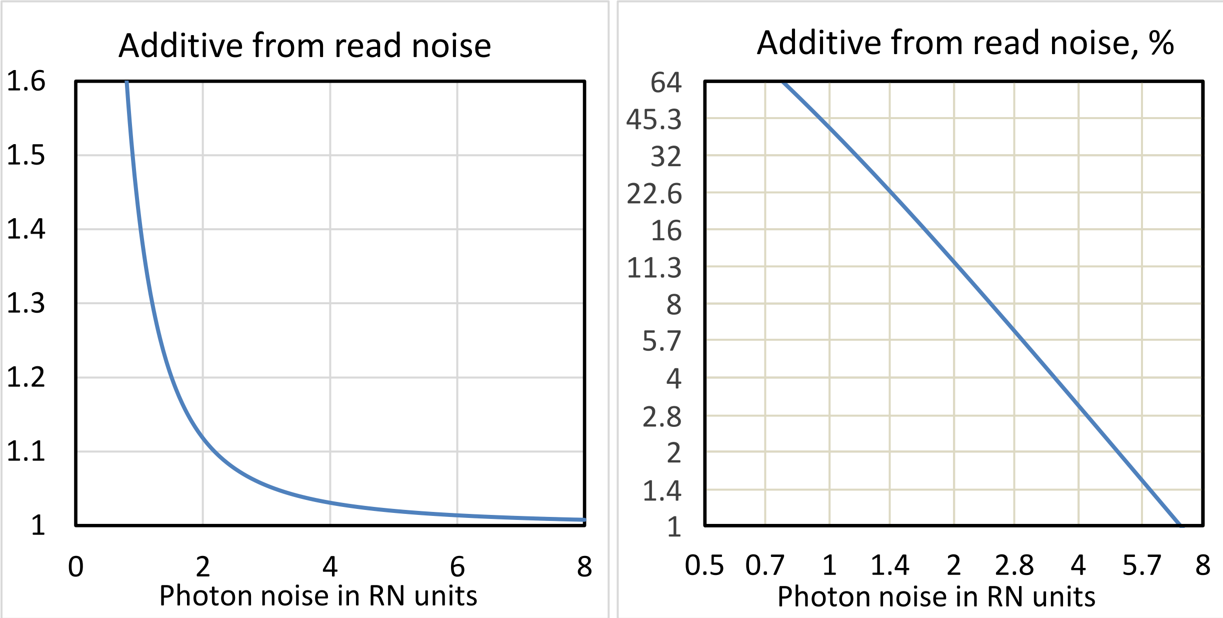

Figure 1 The contribution of read noise to the total noise pot. On the right is a double log percentage plot.

The graphs above were obtained as follows: we take the read noise as one and calculate the total error sum, and then divide by the case when the read noise is taken to be zero:

$$R = \frac{\sqrt{x^2 + 1}}{\sqrt{x^2 + 0}} = \sqrt{1 + \frac{1}{x^2}} \qquad 3$$

It can be seen that when the photon noise is twice as large as the readout noise, the contribution of the latter decreases to 11%. Further reduction is more and more difficult to obtain, which can be seen both from the graphs and from the table:

Table 1

Contribution from read noise, %	Required photon noise level
11	2.07
5	3.12
4	3.50
3	4.05
2	4.98
1	7.05

And since the photon noise grows as the root of the number of photoelectrons, this increases the required shutter speed quadratically! Therefore, I suggest not to be greedy and limit yourself to such a background level at which the readout noise will add no more than 3% into the final image, i.e. $(\Delta n_l t)^2 = RN$ or $n_l t = 16RN^2$. Of course, you can go further and try to squeeze the background to such values that the addition to the final noise will be 1%, but this will require $49/16 = 3$ times longer shutter speed. And the dynamic range, as we remember, is not infinite, and bright objects such as stars still continue to accumulate light like the background, so we risk getting overexposed blobs (each doubling of the shutter speed takes us at least one EV stop on bright objects).

But before that, I said the following phrase:

> *So, we got that for an ideal camera, the more photons we've collected before the effects of overexposure take place, the less noise we observe in the image.*

t seems to be in conflict with the fact that we, even having the opportunity to put a slow shutter speed, are trying to limit it for some reason! But in fact, by choosing the optimal sub-exposure length, we will use the wonderful fact that the sum of the Poisson distributions is also Poisson. Moreover, with an average value lying on the sum of the original distributions and with a variance equal to the sum of the variances. This is what allows us, without any complex mathematical transformations, to simply sum our images to obtain a "cleaner" result. When adding n independent frames that have the same snr, the resulting frame will have a square root of n times snr. So, the full formula for snr will be written as:

$$snr = \sqrt{n}\frac{n_l t}{\sqrt{n_l t + n_d t + RN^2}} \qquad 4$$

Which is logical. If we had a matrix with an unlimited well, a mount with perfect guidance and a sky without its sunrises/sunsets/clouds/haze, we could shoot everything in one long shot. On the other hand, if not for the noise of the ADC, we could cut our total observational time T into pieces of arbitrary duration: be it one hour or one millisecond. All the photoelectrons that we would have caught would go into the final sum, and the only benefit would be the convenience of shooting: the requirements for tracking accuracy are lower, the clouds that have flown in can simply be thrown out of the final sum, etc. But this nasty readout noise ruins everything. It is necessary to accumulate a certain minimum package of information in order to get it across the border without paying too much.

With which I congratulate us. We've reached the end of theoretical part. Now it is time to verify if these conclusions hold true.

Experiment

Now that the theory has been sorted out a little, let's start testing. Firstly, how all these theoretical researches are compared with the reality given to us in hardware sensations. And secondly, let's try to evaluate how well the rule of three percent contribution from read noise works. Would this turn out to be underestimation? I will be experimenting with my trusty old Canon EOS 450Dα, which, according to my measurements, have native ISO = 200. "Native" here means that for each additional electron there will be an additional change in Analog-to-Digital Unit (ADU). I will shoot using it so not to bother taking into account the influence of gain. The readout noise at this ISO is RN=10.6 be it photoelectrons or ADUs, they are equivalent here.

So, given:

- Monitor filled with white light.
- A camera with a telephoto on a tripod and sharpness set to infinity.

What we are about to do:

- Shoot the monitor, varying the shutter speed in steps of one EV, starting from the maximum possible shutter speed, when there is still no overexposure, and take 5-6 steps down with an interval of 1 EV.
- Split the CFA image into its components, as in the analysis of the gain, and look at the statistics.
- Visualize the data in two ways: convert the CFA into linear RGB images, set a neutral white balance and, by successive multiplication by factors of two, bring them to the same brightness and see what happens. And then we will process these same files the way we would process a standard photo: with a non-linear histogram stretching, increased color saturation and additional contrasting.

The previews after the first stage look like this:

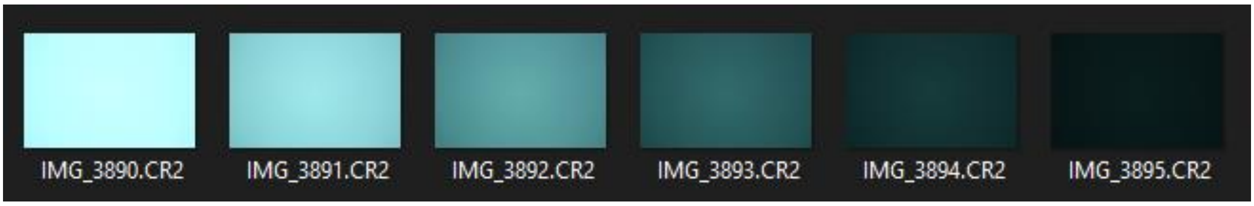

Figure 2 Raw images straight from the camera.

As we've expected. The next two steps give us the following data

Table 2

№	t, s	ADU	σ	snr	snr_{norm}
1	1/10	9408.6	97.3	86.17	1
2	1/20	5201.1	67.8	61.61	0.71
3	1/40	3107.4	46.9	44.42	0.52
4	1/80	2057.8	33	31.33	0.361
5	1/160	1540.8	24.2	21.36	0.25
6	1/320	1283.8	18.9	13.75	0.16

Everything is outrageously according to the theory: a decrease in the number of photoelectrons by half, increases the relative error by the root of two times. And snr, respectively, decreases at each step by the same amount (approximately root of the two times). From the value of the normalized snr, you can even find how many such short frames we would need if we wanted to get snr as on the longest frame:

$$n = \frac{1}{snr_{norm}^2(i)} \qquad 5$$

Second power is due to character of snr growth as a square root of n. So:

Table 3

№	1	2	3	4	5	6
t, s	1/10	1/20	1/40	1/80	1/160	1/320
n	1	1.95 → 2	3.76 → 4	7.57 → 8	16.3 → 16	39.3 → 39

As you can see, on the last frame, the readout noise has already got large enough to notice it, because. it will be required not 32 frames, but 22.8% more. If we recalculate what photon noise should be in order for the readout noise to contribute 23% to our images, it turns out that it should be 1.4RN or 14.84 ADU. If we calculate the total noise on the frame as:

$$noise_{tot} = \sqrt{14.84^2 + 10.6^2} = 18.2 \qquad 6$$

Could someone remind me the value of sigma on the last image? 18.9? Bull's eye! This means that the "optimality" of exposure in terms of readout noise has been proven with us as a pleasant bonus. And now let's see how visually noticeable the increase in noise by a few percent:

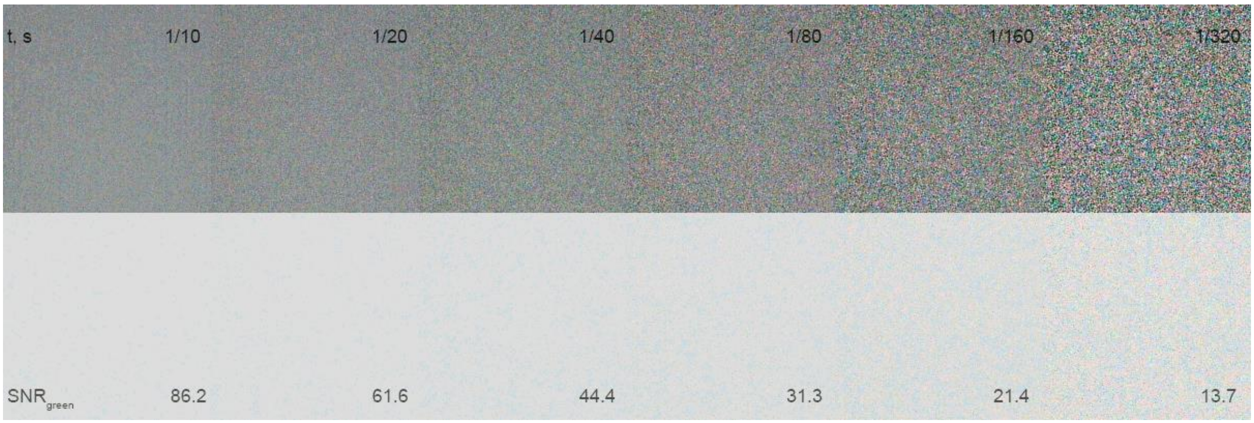

Figure 3 Dependence of snr over exposure.

The difference at the edges is simply colossal, but there we have 32 times difference in the amount of light hitting the sensor, and *snr* has changed by more than 5.6 times. And if the eye still catches the difference on neighboring panels, then if we divide it by another eight, and I remind you, the difference between our neighbors is 41%, and we are trying to understand if we will see a five percent addition, it becomes quite obvious that this is impossible. After all, even in a linear, extremely stretched representation, which is reflected in the upper half of the picture, the difference, although visible, looks insignificant even at 40%. And if we look at the lower half, which shows how this area of the histogram will look like after standard processing, then we will once again only make sure that extending the exposure longer than the optimal one, even twice, does not give anything, but the dynamic range cuts significantly.

Intermediate conclusions

- The main limiting factor when shooting dim objects is readout noise. It would be possible to shoot with short exposures and then add a bunch of images, but alas, this noise cannot be eliminated, and therefore a compromise must be found.
- Such a mathematically reasonable compromise is a shutter speed at which the readout noise is three to four times less than the photon noise from the sky, whether it is light-polluted or not. A further increase in exposure leads to diminishing returns. In the sense that more and more effort is required: the requirements for the mount, the quality of the sky, the dynamic range of the pixel are increasing, and the return is almost exponentially decreasing.
- The theory is brilliantly confirmed by experiment.

It remains only to consider the ISO factor in the problem.

Factoring the ISO

In order to evaluate the effect of gain/amplification/ISO on the image, we can no longer do without real data. It is probably possible to do this, but the resulting ratios will be extremely general and, in the end, will still require real numbers. As a demonstration, I will again take my workhorse, a modified Canon 450Dα. The fact that his IR-UV filter is replaced by a Baader analogue for better transmission of the hydrogen Hα-line will not affect the calculations and conclusions in any way, because we won't even shoot a single light, because we have already seen that photon noise can be remarkably described through theory.

To begin with, we need to make some preliminary measurements, namely:

- Determine not only the native ISO, at which gain=1, but also the rest of the coefficients corresponding to other available ISOs.
- Take several shots at minimum shutter speed with closed lens cap in order to measure the readout noise at different sensitivities.

In my case, this will require determining the gain at 5 different ISOs from 100 to 1600 and another 5 offset frames (this is what frames are called in astrophoto where there is nothing else besides readout noise).

Using the fact that Iris allows users to write their own sequences of actions in “.pgm” files I get the following information:

Table 4

ISO	gain	RN(ADU)	RN(e^-)
100	0.472	10.3	21.836
200	0.943	10.6	11.236
400	1.857	12	6.462
800	3.679	15.2	4.131
1600	7.358	22	2.990

In theory, it would be possible to carry out all calculations not in electrons, but in ADU units, but we would still have to remember that these units, firstly, correspond to different dispersions at different ISOs, and these ADUs accumulate at different rates during exposure. Easy to get lost. So, it's more convenient to work with electrons: the ISO value does not affect their accumulation in pixels. However, measurements clearly show that with increasing gain, the readout noise in the electronic equivalent clearly decreases. And this is understandable, if the ADC introduces a certain constant noise into the signal, then the more we amplify the signal before digitizing, the less this parasitic contribution will be.

As we've shown previously, it is reasonable to choose as the optimal shutter speed one when the readout noise is four times less than the photon noise, then its influence in snr is limited to 3 percent. Accordingly, for each ISO, we get a different level of required photon noise in the image:

Table 5

ISO	RN (e$^-$)	$phot_{noize}$ = 4RN	signal	t_{rel} (ISO=800)
100	21.84	87.34	7629	27.94
200	11.24	44.94	2020	7.4
400	6.46	25.85	668	2.45
800	4.13	16.53	273	1
1600	2.99	11.96	143	0.52

The last column shows the total time assuming that at ISO 800 I need one second of exposure. Thus, it can be seen that even though the signal multiplication before digitization does not affect the actual number of photoelectrons in a pixel, the effect of readout noise is so significant that setting the ISO to only four times higher than the native one, makes it possible to reduce the required shutter speed without losing weak signal in the shadows, at least on the Canon 450D, not by 4, but as much as seven and a half times! However, it is also obvious that such conclusions cannot be thoughtlessly transferred to other cameras. In each case, a series of preparatory measurements is required to determine the gain and readout noise of the camera you plan to shoot with. Or at worst, you can use cheat sheets in the form of this resource. According to my canon cameras, the results are pretty similar, although there are slight differences that fit within ten percent in terms of readout noise and gain. Fortunately, it is enough to do all these calculations once and never return to them again. If only to just to refresh memory from time to time. It is also convenient to transfer the values in electrons back to ADU after calculations, because it is with them, we will deal in processing programs. For example, in my case it would be:

Table 6

ISO	100	200	400	800	1600
ADU_{min}	3599	1906	1241	1005	1052

The only real trouble that comes when we raise the ISO is the reduction in dynamic range. Of course, it is possible to restore burned-out stars using the HDR technique, when several frames with different shutter speeds are stitched into one. But since it is quite complicated procedure one would like to avoid this. But at the same time, we want to understand what share of Dynamic Range we are losing. Here I am not sure of my approach, but I did the following.

1. Convert electrons to ADU.
2. For the upper limit of the recorded signal, I take a value that, even after adding three standard deviations, does not fly into oversaturation (overexposure).
3. I take into account the lowest bits to which the camera adds the bias current, so even for 14 bits we have $(16384 - offset)$ the number of samples available.
4. I take into account the broadening of the Poisson distribution, which occurs when $gain > 1$ is applied.
5. For the minimum-significant recorded reading I take one that is the root of two times less than the readout noise, i.e. such that its $snr \sim 0.7$.

So, for the upper limit I have:

$$ADU_{max} + 3 \cdot gain\sqrt{ADU_{max}} = 16384 - offset$$

Solution of this, essentially a quadratic equation, gives an upper bound. For the lower, as I said, multiplying RN, expressed in ADU, by 0.7, we get the lower bound. There is only one caveat. We filled the bottom bits of all pixels to the ADU values calculated in the table 5. Which effectively further reduces the available dynamic range.

Having calculated ADU_{max}, and subtracting from it the values from the previous table, where we collected values for the ADU to achieve a given photon noise value, we can already divide them by the minimum count, take the base 2 logarithm and get the range in photographic stops.

As a result, we have such a table.

Table 7

ISO	100	200	400	800	1600
EV	7.77	7.87	7.54	6.82	5.61

It can be seen that, yes, with increasing gain, the well is reduced and the dynamic range drops, however, this reduction is still slower than one EV stop per one ISO stop and using a double or even four times higher gain turns out to be justified, especially considering that this reduces the required exposure by a factor of almost 2.5 and 7.5, respectively.

Let me remind you (and myself) once again: the fact that we reduce the exposure does not mean that we lose the opportunity to accumulate light from the object. If we have, say, 4 hours of observation time, then we use it anyway. Simply, for example, at ISO 800 we will make 2.4 times more frames than at ISO 400. And the total number of photons and noise in the final image will be the same. The only thing we will lose is 0.7 EV stops in dynamic range on the final sum. But if we are shooting a subject with a large spread in brightness, such as the Orion Nebula, for example, we will inevitably either underexposure the shadows or a burn the nebula's center. So, anyway, we will have to shoot 10-20 frames with an exposure 10 times shorter than the main series in order to restore burnt details. And if so, shall we worry about 0.7 stops of dynamic range, if instead of a heavy mount we can take a small tracker with us and blast a bunch of short exposures?.. The answer is obvious to me. The main thing is to calculate such a table once and make a cheat sheet, like the one I made for myself:

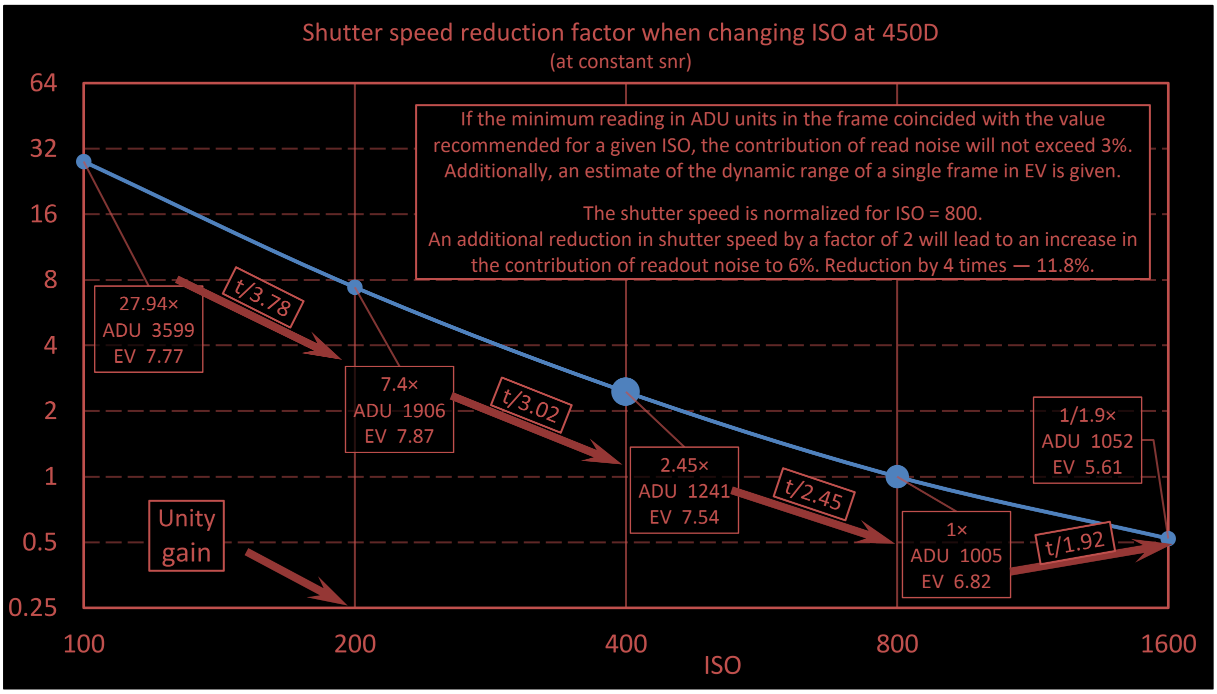

Figure 4 Astro- cheat-sheet.

The only thing that remains is the ability to somehow control this very minimal countdown in the RAW file. Moreover, it would be desirable to do this even without having a laptop at hand. Here, the ability to display a histogram on the camera screen will come to our aid. And as always, here too, it will not do without complications: this histogram is extremely non-linear. I took a whole series of measurements from photographs of the same monitor at different shutter speeds, then compiled a table of where the histogram peak was at each average histogram illumination and spent a lot of time trying to fit this behavior with some smooth and physically meaningful function. And it looks like I succeeded, because. then, when I took a camera with a different ADC bit rate: 12 instead of 14, I got similar histograms and my model predicted the position of the peaks quite well. In short: I determined the maximum reading in the RAW file, subtracted the offset value, applied a logarithmic histogram stretch, and then some kind of logistic curve was superimposed on top. As a result, I got the following reference table:

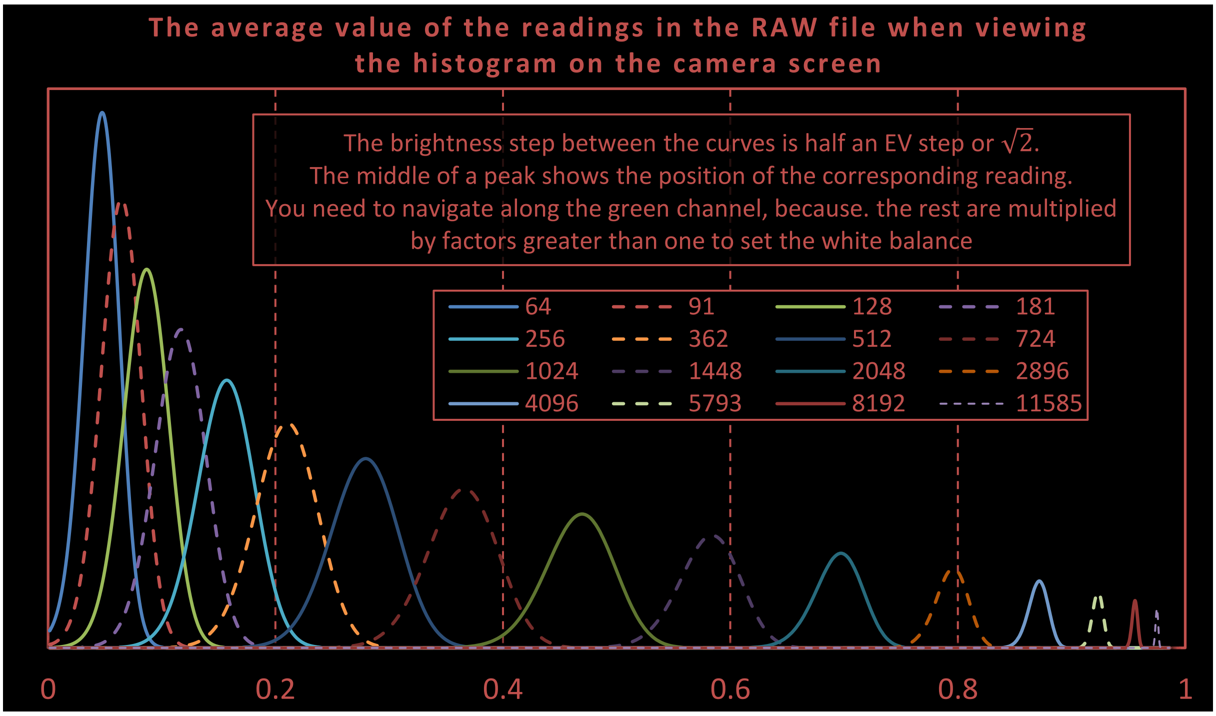

Figure 5 A diagram that allows us to understand where we should put the hump of histogram from the sky so that the exposure corresponds to the minimum optimal illumination.

If you have a camera that has 12-bit ADC instead of 14-bit (or your offset is different from 2048), then you can download my excel file and correct a couple of values there to get a similar infographic for your camera model. The maximum of the hump in a Canon histogram will almost certainly be cut off, but even looking at the edges, it is easy to see if we should drive it into first, second, or even third fifth of the total histogram area.

Well, in general, that's all I had to say on this non-trivial topic of choosing the optimal parameters for astrophotography. I hope that now the scope of your competence in astro- and ordinary photography has grown a little more, and you have become even more aware of the behavior of your camera.