**Astronomy 112: The Physics of Stars**

*Class 19 Notes: The Stellar Life Cycle*

In this final class we'll begin to put stars in the larger astrophysical context. Stars are central players in what might be termed "galactic ecology": the constant cycle of matter and energy that occurs in a galaxy, or in the universe. They are the main repositories of matter in galaxies (though not in the universe as a whole), and because they are the main sources of energy in the universe (at least today). For this reason, our understanding of stars is at the center of our understanding of all astrophysical processes.

## I. Stellar Populations

Our first step toward putting stars in a larger context will be to examine populations of stars, and examine their collective behavior.

### A. Mass Functions

We have seen that stars' masses are the most important factor in determining their evolution, so the first thing we would like to know about a stellar population is the masses of the stars that comprise it. Such a description is generally written in the form of a number of stars per unit mass. A function of this sort is called a mass function. Formally, we define the mass function $\Phi(M)$ such that $\Phi(M)\,dM$ is the number of stars with masses between $M$ and $M + dM$.

With this definition, the total number of stars with masses between $M_1$ and $M_2$ is

$$N(M_1, M_2) = \int_{M_1}^{M_2} \Phi(M)\,dM.$$

Equivalently, we can take the derivative of both sides:

$$\frac{dN}{dM} = \Phi$$

Thus the function $\Phi$ is the derivative of the number of stars with respect to mass, i.e. the number of stars $dN$ within some mass interval $dM$.

Often instead of the number of stars in some mass interval, we want to know the *mass* of the stars. In other words, we might be interested in knowing the total mass of stars between $M_1$ and $M_2$, rather than the number of such stars. To determine this, we simply integrate $\Phi$ times the mass per star. Thus the total mass of stars with masses between $M_1$ and $M_2$ is

$$M_*(M_1, M_2) = \int_{M_1}^{M_2} M\Phi(M)\,dM$$

or equivalently

$$\frac{dM_*}{dM} = M\Phi(M) \equiv \xi(M).$$

Unfortunately the terminology is somewhat confusing, because $\xi(M)$ is also often called the mass function, even though it differs by a factor of $M$ from $\Phi(M)$. You will also often see $\xi(M)$ written using a change of variables:

$$\xi(M) = M\Phi(M) = M\frac{dN}{dM} = M\frac{d\ln M}{dM}\frac{dN}{d\ln M} = \frac{dN}{d\ln M}.$$

Thus $\xi$ gives the number of star per logarithm in mass, rather than per number in mass. This has an easy physical interpretation. Suppose that $\Phi(M)$ were constant. This would mean that there are as many stars from $1-2\ M_\odot$ as there are from $2-3\ M_\odot$ as there are from $3-4\ M_\odot$, etc. Instead suppose that $\xi(M)$ were constant. This would mean that there are equal numbers of stars in intervals that cover an equal range in logarithm, so there would be the same number from $0.1-1\ M_\odot$, from $1-10\ M_\odot$, from $10-100\ M_\odot$, etc.

Often we're more concerned with the distribution of stellar masses than we are with the total number or mass of stars. That is because the distribution tends to be invariant. If we examine two clusters of different sizes, then $dN/dM$ will be different for them simply because they have different numbers of stars. However, they may have the same fraction of their stars in a given mass range. For this reason, it is generally common to normalize $\Phi$ or $\xi$ so that the integral is unity, i.e. to choose a pre-factor for $\Phi$ or $\xi$ such that

$$\int_0^\infty \Phi(M)\, dM = 1$$

and similarly for $\xi(M)$. If a mass is normalized in this way, then $\Phi(M)\, dM$ and $\xi(M)\, dM$ give the fraction of stars (fraction by number for $\Phi$ and fraction by mass for $\xi$) with masses between $M$ and $M + dM$.

B. The IMF

One can construct mass functions for any stellar population. However, the most useful sort of mass function is the one for stars that have just formed, since that determines the subsequent evolution of the population. This is known as the initial mass function, or IMF for short. The IMF is not the same as the mass function at later times, because stars lose mass over their lives, and some go supernova and disappear completely. Thus the IMF is distinct from the present-day mass function, or PDMF.

Observationally, one can attempt to determine the IMF in two ways. The most straightforward way is to look at star clusters so young that no stars have yet lost a significant amount of mass, and none have yet gone supernova. Since the lifetime of a massive star is only $3-4$ Myr, such clusters must be younger than this. Young clusters of this sort are rare, so we don't have a lot of examples where we can do this. To make matters worse, such young clusters also tend to be still partially enshrouded by the dust and gas out of which they formed, making it difficult to determine stars' masses accurately.

A less simple method is to survey field stars that are not part of clusters and measure the PDMF, and then try to extrapolate back to an IMF based on an understanding of mass loss and stellar lifetimes as a function of mass. This is tricky because we only understand those things at a rough level. The great advantage of the method is that it gives us an absolutely immense number of stars to use, and thus provides great statistical power. This is important in determining the IMF for stars that are rare, and thus are unlikely to be present in the few clusters where we can use the first method.

[Slides 1 and 2 – IMFs from clusters and field stars]

Regardless of which method is used, observations tend to converge on the same result for the IMF of stars larger than $\sim 1\ M_\odot$. For these stars, $\Phi(M) \propto M^{-2.35}$, or equivalently $\xi(M) \propto M^{-1.35}$. This value of $-2.35/-1.35$ for the exponent is known as the Salpeter slope, after Edwin Salpeter, who first obtained the result. This result means that massive stars are rare both by number and by mass, since $\Phi$ and $\xi$ are strongly declining functions of $M$.

At lower masses the IMF appears to flatten out, reaching a peak somewhere between $0.1\ M_\odot$ and $1\ M_\odot$ before declining again below $0.1\ M_\odot$. Some people claim there is a rise again at lower masses, but such claims are still highly controversial – very low mass objects are extremely difficult to find due to their low luminosities, and it is not easy to infer their masses. These two factors make these observations quite uncertain.

I should point out that all of these results are empirical. We don't have a good theory for why the IMF looks like it does. It seems to be very constant in the local universe, but we can't rule out the possibility that it might have been different in the distant past, or that it might depend on the environment where the star formation takes place, and change in environments that are simply not found in our own galaxy. A number of people are working on the problem.

For convenience we sometimes just ignore the flattening below $1\ M_\odot$, and assume that the Salpeter slope holds over a range from $M_{\rm min} = 0.1 M_\odot$ to $M_{\rm max} = 120 M_\odot$. A mass function of this sort is known as a Salpeter IMF. The normalization constant in this case, obtained by requiring that the integral be 1, is

$$\begin{aligned} 1 &= \int_{M_{\rm min}}^{M_{\rm max}} \Phi(M)\, dM = \int_{M_{\rm min}}^{M_{\rm max}} A M^{-2.35}\, dM = \frac{A}{-1.35}\left(M_{\rm max}^{-1.35} - M_{\rm min}^{-1.35}\right) \\ A &= \frac{1.35}{M_{\rm min}^{-1.35} - M_{\rm max}^{-1.35}} = 0.060, \end{aligned}$$

where we are working in units of solar masses. Similarly, for the IMF in terms of mass,

$$\begin{aligned} 1 &= \int_{M_{\rm min}}^{M_{\rm max}} \xi(M)\, dM = \int_{M_{\rm min}}^{M_{\rm max}} B M^{-1.35}\, dM = \frac{B}{-0.35}\left(M_{\rm max}^{-0.35} - M_{\rm min}^{-0.35}\right) \\ B &= \frac{0.35}{M_{\rm min}^{-0.35} - M_{\rm max}^{-0.35}} = 0.17. \end{aligned}$$

Thus if we take

$$\Phi(M) = 0.060 M^{-2.35} \text{ and } \xi(M) = 0.17 M^{-1.35},$$

for $M = 0.1-120\ M_\odot$, these functions give us a reasonable estimate of the fraction of stars by number and by mass in a given mass range.

As an example, suppose we wanted to compute what fraction of stars (by number) are more massive than the Sun in a newborn population. The answer is

$$f_N(> M_\odot) = \int_1^{120} 0.060 M^{-2.35} = \frac{0.060}{1.35}\left(1^{-1.35} - 120^{-1.35}\right) = 0.045.$$

Similarly, the fraction of the mass in stars above $M_\odot$ in mass is

$$f_M(> M_\odot) = \int_1^{120} 0.17 M^{-1.35} = \frac{0.17}{0.35}\left(1^{-0.35} - 120^{-0.35}\right) = 0.40.$$

Thus only 4.5% of stars are more massive than the Sun, but these stars contain roughly 40% of all the mass in newborn stars.

C. Star Clusters

As we've discussed several times, stars a born in clusters that are relatively coeval, i.e. all the stars in them are born at the same time plus or minus a few Myr. This means that for many purposes we can approximate the stars in a star cluster as all having been born in a single burst. Everything that happens subsequently is due simply to aging of the stellar population. Star clusters therefore constitute the simplest example of what happens as stellar populations age.

We have already seen that lifetimes of stars decrease monotonically with mass, so it is convenient to introduce for a given cluster the turn-off mass, $M_t$, defined as the mass of star that is just now leaving the main sequence. In a given cluster no stars with masses above $M_t$ remain on the main sequence, while those with lower masses are all on the main sequence. As clusters age, $M_t$ decreases, since lower and lower mass stars evolve off the main sequence.

We can use the turn-off mass plus the IMF to figure out how the population of stars in the cluster changes with time. As a simple example, we can estimate what fraction of the stars in a cluster by number will still be on the main sequence. The total fraction of stars by number is simply

$$f_N = \int_{M_{\rm min}}^{M_t} \Phi(M)\, dM = \frac{0.060}{1.35}\left(M_{\rm min}^{-1.35} - M_t^{-1.35}\right).$$

This expression does not fall below 0.5 unless $M_t < 0.168$, which never happens, since the universe is not old enough for stars with masses so low to have left the main sequence. Thus the majority of the stars in a cluster are always on the main sequence, even for the oldest clusters. This is because the most common stars are those with low masses, which have not yet had time to leave the main sequence.

As a somewhat more complex problem, we can try to estimate what fraction of the stellar mass in the cluster will remain when the turnoff mass is $M_t$. Stars that go supernova will eject most of their mass at speeds well above the escape speed from a cluster, so the supernova ejecta will simply escape, reducing the mass of the cluster. Most neutron stars probably escape as well, because the supernovae are not perfectly symmetric, and tend to give the neutron stars kicks that are well above the escape velocity. Similarly, the mass ejected from stars that are turning into white dwarfs will also escape the cluster due to its high temperature, although the white dwarfs themselves will remain.

Putting all this together, the mass in the cluster comes in two parts: main sequence stars and remnant white dwarfs. The fraction of the original mass contributed by main sequence stars is given by a calculation just like the one we just did for number, except using $\xi(M)$ instead of $\Phi(M)$ :

$$f_{M,MS} = \int_{M_{\min}}^{M_t} \xi(M)\, dM = \frac{0.17}{0.35} \left( M_{\min}^{-0.35} - M_t^{-0.35} \right).$$

For white dwarfs, we will make the simple approximation that they all have masses of 0.6 $M_\odot$ regardless of their initial mass, so that the fraction of the original star's mass left in the white dwarf is $0.6/M$, where $M$ is in $M_\odot$. We also approximate that all stars with initial masses below $M_{NS} = 8M_\odot$ form white dwarfs, while more massive ones go supernova and leave nothing behind in the cluster. Thus the fraction of the original mass in the form of leftover white dwarfs is

$$\begin{aligned} f_{M,WD} &= \int_{M_t}^{M_{NS}} \frac{0.6}{M} \xi(M)\, dM = 0.6 \times 0.17 \int_{M_t}^{M_{NS}} M^{-2.35}\, dM \\ &= \frac{0.6 \times 0.17}{1.35} \left( M_t^{-1.35} - M_{NS}^{-1.35} \right). \end{aligned}$$

Adding these two up, we obtain an expression for the fraction of the original cluster mass that remains:

$$\begin{aligned} f_M &= \frac{0.17}{0.35} \left( M_{\min}^{-0.35} - M_t^{-0.35} \right) + \frac{0.6 \times 0.17}{1.35} \left( M_t^{-1.35} - M_{NS}^{-1.35} \right) \\ &= 1.09 - 0.49 M_t^{-0.35} + 0.076 M_t^{-1.35} \end{aligned}$$

Obviously this expression is valid only when $0.6M_\odot < M_t < M_{NS}$. If $M_t > M_{NS}$, then only main sequence stars are left, and we only get their contribution:

$$f_M = 1.09 - 0.49 M_t^{-0.35}.$$

The slide shows this function.

[Slide 3 – fraction of mass remaining in a cluster]

Thus clusters lose about 35% of their original mass once the turnoff mass declines to 0.6 $M_\odot$. They lose more than 20% from supernovae during their first $\sim$ 10 Myr of life, when the turnoff mass declines to around 10 $M_\odot$. This mass loss process can be important in disrupting star clusters. They are held together by gravity, and as mass is ejected they become less tightly bound. Some of them dissolve entirely as a result of mass loss.

A final game we can play with the IMF and cluster is to ask how their luminosities evolve over time. The initial luminosity of the cluster is simply given by

$$L_0 = N_* \int_{M_{\rm min}}^{M_{\rm max}} L(M)\Phi(M)\, dM,$$

where $N_*$ is the total number of stars in the cluster and $L(M)$ is the luminosity of a star of mass $M$. In other words, we find the total luminosity simply by integrating the luminosity per star as a function of mass against the fraction of stars per unit mass, all multiplied by the total number of stars.

Later on, when some stars have turned off the main sequence, the luminosity is given by a similar expression, but with $M_{\rm max}$ replaced by $M_t$:

$$L_1 = N_* \int_{M_{\rm min}}^{M_t} L(M)\Phi(M)\, dM.$$

This implicitly assumes that white dwarfs contribute negligible luminosity, which is a pretty good approximation.

The fraction of the original luminosity that remains is therefore given by

$$f_L = \frac{L_1}{L_0} = \frac{\int_{M_{\rm min}}^{M_t} L(M)\Phi(M)\, dM}{\int_{M_{\rm min}}^{M_{\rm max}} L(M)\Phi(M)\, dM}$$

To evaluate this, we will again make a very simple approximation: we will take $L(M) = L_\odot(M/M_\odot)^3$ for all stars. Obviously this breaks down at both the low and high mass ends, but it is good enough to give us a rough picture of what happens to clusters' luminosities as they age. Inserting this value for $L(M)$ into our expression for $f_L$ and canceling constants that appear in both the numerator and denominator, we have

$$f_L = \frac{\int_{M_{\rm min}}^{M_t} M^{0.65}\, dM}{\int_{M_{\rm min}}^{M_{\rm max}} M^{0.65}\, dM} = \frac{M_t^{1.65} - M_{\rm min}^{1.65}}{M_{\rm max}^{1.65} - M_{\rm min}^{1.65}}.$$

Since $M_t \gg M_{\rm min}$ even in the oldest clusters (which have $M_t \sim 0.7 M_\odot$), we can drop the $M_{\rm min}$ terms, and we have

$$f_L \approx \left(\frac{M_t}{M_{\rm max}}\right)^{1.65}.$$

Thus the luminosity decreases as something like the turnoff mass to the 1.7 power. Recall that the lifetime of a 12 $M_\odot$ star is only a few tens of Myr, so this means that $M_t/M_{\rm max} \sim 0.1$ in a cluster that is about 20-30 Myr old, which in turn means that its luminosity has decreased by a factor of about $10^{1.7} = 50$! Thus clusters fade out very quickly. The young ones are extremely bright, but they lose much of their brightness in their first few tens of Myr. Thereafter they dim greatly.

It should be noted that this calculation does ignore one significant effect, which is important for very old clusters: the luminosity of red giants. Although there aren't many such stars present in any given cluster at a given time due to the short times that stars spend as red giants, they are so bright that they can dominate the total luminosity once the massive stars have faded from view. In very old stellar populations, with $M_t \lesssim M_\odot$, the total luminosity is generally dominated by the red giants.

Although we have done this for the total luminosity, we could just as easily have done it for luminosity in some specific color, say blue light, but taking into account how the surface temperature varies with mass as well. Since the surface temperatures of massive stars are higher than those of low mass stars, they emit more in the blue, and thus the blue luminosity fades even more quickly than the total luminosity. The red luminosity falls more slowly. Of course if one really wants to get precise answers, the way to do it is with numerical models of the stars' mass-dependent luminosity and surface temperature, not with rough analytic fits.

D. Stellar Population Synthesis

Figuring out how clusters fade in time, and how the fraction of the mass in white dwarfs and main sequence stars varies with time, is just the simplest example of a more general idea called stellar population synthesis. For star clusters, we assumed all the stars formed in a single burst with a given IMF, and then we computed how the stellar population would look at later times. Obviously we could easily generalize this to the case of a cluster that, for some reason, had two distinct bursts of star formation at different times. The total stellar population would just have properties given by the sum of the two bursts.

From there, however, it is clear that we can generalize even further and consider an arbitrary star formation history, i.e. we consider an object within which the star formation rate is a specified function $\dot{M}(t)$, where $t$ is a negative number representing the time before the present. At every time $t$, the problem of figuring out how that stellar population looks today is exactly the same as the calculation we just went through for star clusters, and one can then simply add up, or integrate, over all times $t$ to figure out the present-day appearance of the stellar population. This is the basic idea of stellar population synthesis.

The power of the technique is that we can run it in reverse. We can take an observed stellar population and try to figure out what formation history would yield something that looks like that. For example, if we see a high luminosity and a blue color for a given mass of stars, we can infer that the stars must have formed quite recently. In contrast a low luminosity and reddish color imply an older stellar population.

Given a bunch of stars in an HR diagram this technique can get quite sophisticated. An example is a recent paper by Williams et al. that used this method to determine the star formation history in different parts of a nearby galaxy. They divided the galaxy into annuli, and in each annulus they made an HR diagram using stars as observed by the Hubble Space Telescope. They then tried to find a star formation history that would reproduce the observed HR diagram.

[Slides 4-6 – star formation history in NGC 2976 inferred by Williams et al. using stellar population synthesis]

This is one example of stellar population synthesis. One can also do this in more distant galaxies where individual stars cannot be resolved by adding up the spectral features for stars at different ages. Whenever you hear a galaxy or a cluster described as consisting of young or old stars, it's a good bet that a technique like this was used to reach that conclusion.

II. Stars and the ISM

Stars are only one component of the baryonic (normal matter) mass in a galaxy. In between them is a sea of gas known as the interstellar medium, which we discussed very briefly a few weeks ago in the context of star formation. The ISM is very diffuse: its mean density is $\sim 1$ atom per cm$^3$. However, there are a lot of cubic cm in interstellar space, and, as a result, the mass of the ISM is considerable. In the Milky Way, the ISM has a mass equal to about 10% of the total stellar mass. There is a continuous cycling of matter between stars and the ISM, and it is this cycling that will be the final topic in the class.

A. Star Formation

One side of the cycle is the conversion of interstellar gas into stars, a process that we discussed briefly in the context of star formation a few weeks ago. Star formation is a major problem in astrophysics, and we don't have a full theory for what controls the rate at which gas turns itself into stars. Thus this discussion is mainly going to focus on empirical results and problems, with little hints at possible solutions.

Star formation appears to take place only in gas that is in molecular form – i.e. $H_2$ rather than atomic hydrogen. This is apparent just from looking at images of galaxies, and quantitative comparison confirms it. This is likely to be for the reason we outlined when we discussed star formation last time. Only molecular gas is cold enough to enable collapse to stars. Other types of gas are warm enough that their pressure prevents collapse.

[Slides 7-8 – star formation plus HI maps, and correlation between SFR, atomic, and molecular gas]

Within molecular gas, however, the star formation rate is significantly lower than one might expect, as can be illustrated using the example of the Milky Way. Our galaxy contains about $10^9$ $M_\odot$ in molecular clouds. Recall that we calculated the free-fall time, the time gas requires to collapse when it is not supported, to be

$$t_{\mathrm{ff}} = \sqrt{\frac{3\pi}{32G\rho}},$$

where $\rho$ is the gas density. Typical densities in giant molecular clouds are $n \sim 100$ cm$^{-3}$ (much more than the average $n \sim 1$ cm$^{-3}$), so $\rho \sim n m_{\mathrm{H}} \sim 2 \times 10^{-22}$ g cm$^{-3}$. Plugging this in, we have $t_{\mathrm{ff}} \sim 5$ Myr.

Thus if these clouds were collapsing to form stars in free-fall, the star formation rate would be

$$\dot{M} \sim \frac{10^9\ M_\odot}{5\ \mathrm{Myr}} \sim 200\ M_\odot\ \mathrm{yr}^{-1}.$$

The problem is that the observed star formation rate is about 100 times smaller than this. There is a similar discrepancy in other galaxies. Therefore something must be inhibiting the collapse of molecular clouds into stars.

No one is entirely sure what the origin of the discrepancy is, and that's another big problem in astrophysics. One possibility is that the winds and radiation from young, newly-formed massive stars disrupts the molecular clouds out of which they form. This process limits the star formation rate. For example, the ionizing radiation from stars with high surface temperatures can heat up cold molecular gas to $\sim 10^4$ K, causing it to expand rapidly and disrupt the cloud of which it is part. Disruptions like this are observed, and often make for spectacular images.

B. Gas Return and Chemical Enrichment

The cycle of material isn't simply one-way. Star formation converts interstellar medium gas into stars, but stars also return gas to the interstellar medium, via mechanisms we have already described: ejection of the envelopes of red giants, and formation of supernovae. The calculation of what fraction of the mass is returned is just the inverse of the one we already performed to see what fraction of the mass would remain in stellar form in a cluster.

Formally, we define the return fraction $\zeta$ for a stellar population as the fraction of mass that is ejected. Making the same assumptions as before, the fraction that is ejected by stars that go supernova is taken to be 1 (which is not a bad approximation, since such stars eject at least 80% of their mass). The mass that remains in white dwarfs is taken to be 0.6 $M_\odot$ independent of the mass of the progenitor. Finally, stars that are still on the main sequence return a negligible amount of their gas.

Thus the return fraction is

$$\zeta = \int_{M_t}^{M_{NS}} \frac{M - 0.6M_\odot}{M} \xi(M)\, dM + \int_{M_{NS}}^{M_{\max}} \xi(M)\, dM,$$

where the first term represents stars that make white dwarfs, which return a fraction $(M - 0.6M_\odot)/M$ of their mass to the ISM, and the second term represents stars that go supernova and return all their gas to the ISM.

Evaluating the integrals for the Salpeter mass function just gives one minus the stellar mass fraction we found earlier:

$$\zeta = 1 - f_M = 0.49 M_t^{-0.35} - 0.076 M_t^{-1.35} - 0.09,$$

where $M_t$ is in solar masses, as before. For $M_t = 0.7$, roughly the turnoff mass for the oldest stellar populations, this gives $\zeta = 0.34$. Thus old stellar populations eventually return roughly 1/3 of their gas to the ISM. The first 20% or so of this is returned via supernovae in a few tens of Myr. The rest comes out much more slowly via red giants, asymptotic giants, and planetary nebulae produced by lower mass stars that take a very long time to reach the main sequence. A majority of the mass is in stars that stay on the main sequence longer than the age of the universe.

The gas that is returned to the ISM can make new generations of stars. Perhaps more important, it carries with it the products of nuclear burning: metals. The universe was born composed almost entirely of hydrogen and helium, and all the heavier elements were made in stars and then ejected in supernovae or by mass-losing giant stars. This process gradually alters the chemical composition of the gas, enriching it with metals.

The legacy of that process is reflected in the present-day chemical composition of stars and galaxies. Recall that the metal fraction in the Sun is $Z = 0.02$. Other stars have different metal fractions, and there is a correlation between the mass in metals and the age of the star. This is a signature of the gradual enrichment of the ISM by stellar processes over the age of the universe.

[Slide 10 – age-metallicity relation for solar neighborhood stars from Carraro et al. 1998]

We also see a correlation between the mass of a galaxy and its metallicity. This is a signature of two effects. First, more massive galaxies have formed more stars and have turned a larger fraction of their mass into stars, thereby producing more metal processing. Second, massive galaxies have larger escape speeds, which makes it harder for supernova ejecta to escape from the galaxy.

[Slide 11 – mass-metallicity relation from Kewley & Ellison 2008]

Thus the metal content of stars and galaxies provides us with direct evidence for stellar mass return to the ISM.

C. Cosmological Infall

The cycling between gas and stars in a galaxy is part of the story, but it's not all of the story. The reason is that, if you try to balance the books between mass going into stars and mass returned to the ISM, things don't add up. Stars only return 1/3 of the mass that goes into them over the entire age of the universe, which means that most of the mass that goes into stars doesn't come back into gas.

If there were a large enough gas supply in galaxies to keep fueling star formation for the age of the universe, this wouldn't be a problem. Unfortunately, we don't have that much gas. The total mass of interstellar gas in the Milky Way inside the Sun's orbit is a few times $10^9\ M_\odot$, and we have already mentioned that the star formation rate is a few $M_\odot$ per year. This means that it should take roughly $10^9$ yr, or 1 Gyr, to use up all the available gas.

The problem is that the Milky Way and the universe are about 10 Gyr old, so there isn't enough gas to keep things going. There might have been more gas in the past, but then we face the uncomfortable proposition that we live at a special time, as do all other star-forming galaxies, when the gas supplies are just about to dry up. This seems to require an unreasonable amount of luck and coordination.

Instead, the preferred explanation is that the Milky Way isn't done growing. Our galaxy continues to acquire new material from intergalactic space. The majority of the baryonic mass in the universe must be out there in the gas between the galaxies – there simply isn't enough mass in the galaxies to account for the amount that we believe is there based on the models of the early universe.

This gas is called the intergalactic medium, and hunting for it, and its infall onto the Milky Way and similar galaxies, is a major project in astronomy right now. So far, no one has found it, although there are tantalizing hints that it may have been observed.