Draft version June 20, 2012
Preprint typeset using LaTeX style emulateapj v. 5/2/11

QUALIFIER EXAM SOLUTIONS

Chenchong Zhu
(Dated: June 20, 2012)

Contents

1. Cosmology (Early Universe, CMB, Large-Scale Structure) 7
1.1. A Very Brief Primer on Cosmology 7
1.1.1. The FLRW Universe 7
1.1.2. The Fluid and Acceleration Equations 7
1.1.3. Equations of State 8
1.1.4. History of Expansion 8
1.1.5. Distance and Size Measurements 8
1.2. Question 1 9
1.2.1. Couldn't photons have decoupled from baryons before recombination? 10
1.2.2. What is the last scattering surface? 11
1.3. Question 2 11
1.4. Question 3 12
1.4.1. How do baryon and photon density perturbations grow? 13
1.4.2. How does an individual density perturbation grow? 14
1.4.3. What is violent relaxation? 14
1.4.4. What are top-down and bottom-up growth? 15
1.4.5. How can the power spectrum be observed? 15
1.4.6. How can the power spectrum constrain cosmological parameters? 15
1.4.7. How can we determine the dark matter mass function from perturbation analysis? 15
1.5. Question 4 16
1.5.1. What is Olbers's Paradox? 16
1.5.2. Are there Big Bang-less cosmologies? 16
1.6. Question 5 16
1.7. Question 6 17
1.7.1. How can we possibly see galaxies that are moving away from us at superluminal speeds? 18
1.7.2. Why can't we explain the Hubble flow through the physical motion of galaxies through space? 19
1.7.3. Can galaxies with recession velocities $v > c$ slow down until $v < c$? 19
1.8. Question 7 19
1.8.1. How does nucleosythesis scale with cosmological parameters? 20
1.8.2. How do we determine primordial densities if D is easily destroyed in stars? 20
1.8.3. Why is there more matter than antimatter? 20
1.8.4. What are WIMPs? 20
1.9. Question 8 20
1.9.1. Describe systematic errors. 21
1.9.2. Describe alternate explanations to the SNe luminosity distance data, and why they can be ruled out? 22
1.9.3. Can SNe II be used as standard candles? 23
1.10. Question 9 23
1.10.1. What is the fate of the universe, given some set of Ωs? 23
1.10.2. How do we determine, observationally, the age of the universe? 24
1.10.3. Is Λ caused by vacuum energy? 25
1.11. Question 10 25
1.11.1. What are the consequences of these numbers on the nature of the universe? 26
1.11.2. How do we determine Ω_r from the CMB? 26
1.11.3. How are other values empirically determined? 26
1.11.4. What are the six numbers that need to be specified to uniquely identify a ΛCDM universe? 27
1.11.5. Why is h often included in cosmological variables? 27
1.12. Question 11 27
1.12.1. What are the possible fates the universe? 28
1.13. Question 12 29

cczhu@astro.utoronto.ca

1.13.1. How does the CMB power spectrum support the inflation picture? 31
1.13.2. Derive the horizon size at recombination. 31
1.13.3. Why is the CMB a perfect blackbody? 31
1.13.4. How is the CMB measured? 32
1.13.5. Why did people use to think CMB anisotropy would be much larger than it is currently known to be? 32
1.13.6. What is the use of CMB polarization? 32
1.14. Question 13 32
1.14.1. Why is BAO often used in conjunction with CMB? 35
1.14.2. What is the BAO equivalent of higher-l CMB peaks? 35
1.15. Question 14 35
1.15.1. How is weak lensing measured? 36
1.15.2. Can strong lensing be used to determine cosmological parameters? 36
1.16. Question 15 36
1.16.1. What caused inflation? 37
1.16.2. How does inflation affect the large scale structure of the universe? 37
1.16.3. Is inflation the only way to explain the three observations above? 37
1.17. Question 16 37
1.17.1. Is the anthropic principle a scientifically or logically valid argument? 38
1.18. Question 17 38
1.18.1. Describe galaxy surveys. 39
1.18.2. What about three or higher point correlation functions? 39
1.19. Question 18 40
1.20. Question 19 40
1.20.1. What about He reionization? 41

2. Extragalactic Astronomy (Galaxies and Galaxy Evolution, Phenomenology) 43
2.1. Question 1 43
2.1.1. What does the Hubble sequence miss? 44
2.1.2. What is the Tully-Fisher relationship? 45
2.1.3. What is the fundamental plane? 45
2.1.4. What other criteria could be used for galaxy classification? 45
2.2. Question 2 45
2.2.1. Why can't we simply use the inner 10 kpc data derived from 21-cm line emission and a model of the halo to determine the mass of the halo? 46
2.2.2. How is the total mass of an elliptical galaxy determined? 46
2.3. Question 3 47
2.3.1. Is He and metals ionized? Is any component of the IGM neutral? 47
2.3.2. What are the properties of these neutral H clouds? Can they form galaxies? 48
2.4. Question 4 48
2.5. Question 5 50
2.5.1. How are SMBHs formed? 52
2.5.2. What correlations are there between the properties of the SMBH and the host galaxy? 52
2.6. Question 6 52
2.7. Question 7 53
2.7.1. What determines the emission spectrum of an AGN? 54
2.7.2. Are there backgrounds at other wavelengths? 54
2.8. Question 8 54
2.8.1. What are cold flows? 55
2.8.2. What is feedback? 55
2.8.3. What is downsizing? 56
2.9. Question 9 56
2.9.1. What are cooling flows? 57
2.10. Question 10 57
2.11. Question 11 59
2.11.1. Describe population synthesis. 60
2.11.2. What do the spectra of real galaxies look like? 60
2.12. Question 12 60
2.12.1. What have we learned about high-z galaxies from LBGs? 62
2.12.2. Are there other "breaks" that could be used to detect galaxies? 62
2.12.3. Can we find Lyman-break galaxies at low redshifts? 62
2.13. Question 13 62
2.13.1. What is the cause of the various emission features of AGN? 64
2.14. Question 14 64

2.14.1. Are there non-EM backgrounds? 66
2.15. Question 15 66
2.15.1. Where are most AGN located? 69
2.15.2. What evidence is there that AGN are powered by supermassive black holes? 69
2.16. Question 16 69
2.16.1. How are high-z cluster found? 71
2.17. Question 17 71
2.18. Question 18 73
2.18.1. What are SZ surveys conducted with? 74
2.18.2. What is the kinematic SZ effect? 74

3. Galactic Astronomy (Includes Star Formation/ISM) 75
3.1. Question 1 75
3.1.1. What is the IMF useful for? 75
3.1.2. Is there a universal IMF? 75
3.1.3. Why are the lower and upper limits of the IMF poorly understood compared to that of the middle (several $M_\odot$ stars)? What constraints are there? 76
3.1.4. What's the difference between a field and stellar cluster IMF? 77
3.1.5. How do you determine an a present-day mass function (PDMF) from an IMF? 77
3.2. Question 2 77
3.2.1. What Orbits are Allowed in an Axisymmetric Potential? 78
3.2.2. How did each population of stars gain their particular orbit? 78
3.2.3. Orbits in Elliptical Galaxies 79
3.2.4. What are the different populations in the galaxy, and what are their ages and metallicities? 80
3.2.5. What is the spheroid composed of (globular clusters?)? 80
3.3. Question 3 80
3.3.1. What errors are in your analysis? 80
3.3.2. Can you give some real statistics for SNe Ia? 81
3.4. Question 4 81
3.4.1. What stars are collisional? 82
3.4.2. Gas is collisional. Why? 82
3.5. Question 5/6 82
3.5.1. How does H_2 form? 82
3.5.2. Why is H_2 necessary for star formation? 83
3.5.3. How do Population III stars form? 83
3.6. Question 7 84
3.6.1. What observational differences are there between GCs and dSphs? 84
3.6.2. What is a Galaxy? 84
3.7. Question 8 85
3.7.1. How does velocity dispersion increase over time? 85
3.7.2. Why is the mean [Fe/H] not $-\infty$ at the birth of the MW? 86
3.8. Question 9 87
3.9. Question 10 88
3.9.1. Given a cross-section, how would you calculate the amount of extinction? 90
3.9.2. What percentage of the light is absorbed (rather than scattered) by the light? 90
3.9.3. Why is dust scattering polarized? 90
3.9.4. What is the grain distribution of the ISM? 91
3.10. Question 11 91
3.10.1. Under what conditions does the above formulation of dynamical friction hold? 92
3.10.2. Under what conditions does it not? 93
3.11. Question 12 93
3.11.1. How about for different galaxies? 94
3.12. Question 13 95
3.12.1. State the assumptions in this problem. 96
3.13. Question 14 96
3.13.1. What complications are there? 98
3.13.2. What triggers the collapse? 98
3.13.3. What causes these overdensities? 100
3.13.4. Why do D and Li fusion occur before H fusion? Do they change this cycle? 100
3.13.5. How does fragmentation work? 100
3.14. Question 15 100
3.14.1. How do halo properties scale with the properties of luminous matter in late-type galaxies? 101
3.14.2. How do we observationally determine the rotational profiles of galaxies? 101
3.14.3. What is the difference between late and early type rotation curves? 101

3.15. Question 16 102
3.15.1. Molecular Gas 103
3.15.2. Cold Neutral Medium 103
3.15.3. Warm Neutral Medium 103
3.15.4. Warm Ionized Medium 103
3.15.5. Hot Ionized Medium (Coronal Gas) 104
3.15.6. Why Do the Phases of the ISM Have Characteristic Values At All? 104
3.16. Question 17 105
3.16.1. The Galactic Disk 105
3.16.2. The Galactic Bulge 106
3.16.3. Open Star Clusters 107
3.16.4. Globular Clusters 107
3.16.5. (Ordinary) Elliptical Galaxies 107
3.16.6. What do these properties tell us about the formation and evolution of these objects? 108
3.17. Question 18 108
3.17.1. Details? 109
3.17.2. How is density determined? 110
3.17.3. How is abundance determined? 112
3.17.4. Can you use the same techniques to determine the properties of other ISM regions? 112

4. Stars and Planets (Includes Compact Objects) 114
4.1. Question 1 114
4.1.1. Describe protostellar evolution 118
4.1.2. What does this look like on a ρ-T diagram 118
4.2. Question 2 118
4.2.1. What other factors could change the mass-radius relation of an object? 119
4.3. Question 3 120
4.3.1. How does radiative transfer occur inside a star? 121
4.4. Question 4 122
4.4.1. What are the different classes of variable stars? 124
4.4.2. Why do different stars have different pulsation frequencies? Describe the period-luminosity relation. 124
4.4.3. What kinds of pulsations are there? 124
4.5. Question 5 125
4.5.1. What is the Virial Theorem? 126
4.5.2. What is the instability criterion for stars? 126
4.5.3. What about the convective turnover time? 126
4.6. Question 6 127
4.6.1. Describe the dynamical evolution of a supernova. 129
4.6.2. What are the energy scales and peak luminosities of these supernovae? Describe their light curves and spectra. 130
4.6.3. What is the association between supernovae and gamma-ray bursts? 131
4.6.4. What are the nucleosynthetic processes involved? 131
4.6.5. What is the spatial distribution of supernovae? 132
4.7. Question 7 132
4.7.1. How are asteroids related to Solar System formation? 133
4.7.2. How does rotation change things? 133
4.8. Question 8 133
4.8.1. Derive the Ledoux Criterion. 134
4.8.2. Describe convection. 135
4.8.3. How does convective energy transport change radiative energy transport? 135
4.8.4. What are the main sources of opacity in the stellar atmosphere? 135
4.8.5. Why is convection ineffective near the photosphere? 135
4.8.6. What is the Hayashi Line? 135
4.9. Question 9 136
4.9.1. Compare the interiors of the inner planets. 138
4.9.2. Why are there tectonic plates on Earth, but not other planets? 138
4.9.3. Why has Mars lost its atmosphere while Venus has not? 138
4.9.4. Why does Venus have a much thicker atmosphere than Earth? 139
4.10. Question 10 139
4.10.1. What is the Eddington accretion rate? 140
4.10.2. Under what conditions does the Eddington luminosity not apply? 140
4.11. Question 11 140
4.11.1. What assumptions have you made, and how good are they? 140

4.11.2. Why are the centres of stars so hot? 140
4.12. Question 12 141
4.12.1. Why do we use the radiative diffusion equation rather than a convection term? 142
4.12.2. Where does this approximation fail? 142
4.12.3. What about extreme masses on the main sequence? 142
4.12.4. What is the scaling relation when the pressure support of the star itself is radiation-dominated? 143
4.13. Question 13 143
4.13.1. Does more detailed modelling give better results? 143
4.14. Question 14 144
4.14.1. Estimate the Chandrasekhar Mass for a WD. Estimate the equivalent for an NS. 145
4.14.2. Why is proton degeneracy unimportant in a WD? 145
4.14.3. What is the structure of a neutron star? 146
4.14.4. How can rotational support change the maximum mass? 146
4.14.5. Could you heat a WD or NS to provide thermal support? 146
4.15. Question 15 147
4.15.1. What are the primary sources of line broadening? 150
4.15.2. What is the curve of growth? 151
4.15.3. How is a P Cygni feature produced? 151
4.15.4. What are the main differences between supergiant and main sequence spectra, assuming the same spectral classification? 151
4.16. Question 16 152
4.16.1. What methods are there to detect planets? 153
4.16.2. What are common issues when using radial velocities and transits? 154
4.16.3. How do Hot Jupiters form? 154
4.17. Question 17 154
4.17.1. How are YSOs classified spectroscopically? 156
4.17.2. How do YSOs evolve on the HR diagram? 157
4.17.3. What classes of PMS stars are there? 158
4.17.4. How is this picture changed for massive star formation? 158
4.18. Question 18 158
4.18.1. Why does the disk have to be flared? 159
4.19. Question 19 160
4.19.1. Can you determine an order of magnitude estimate for Io's heating? 161
4.20. Question 20 161
4.20.1. What would change if the object were rotating? 163
4.20.2. Charged Compact Objects 163
4.20.3. What would happen if you fell into a black hole? 164
4.21. Question 21 164
4.21.1. The CNO Cycle 165
4.21.2. What about fusion of heavier nuclei? 166
4.21.3. Why is quantum tunnelling important to nuclear fusion in stars?0 167
4.21.4. What is the Gamow peak? 167

5. Math and General Physics (Includes Radiation Processes, Relativity, Statistics) 169
5.1. Question 1 169
5.1.1. What happens if lensing is not done by a spherically symmetric object on a point source? 170
5.1.2. What is lensing used for? 170
5.2. Question 2 170
5.2.1. Complications? 172
5.2.2. What is aperture synthesis? 173
5.2.3. What is interferometry used for? 173
5.2.4. Name some interferometers. 173
5.3. Question 3 173
5.3.1. What if the lens or mirror is not perfect? 175
5.3.2. How does adaptive optics work? Does it truly create diffraction-limited images? 176
5.4. Question 4 176
5.4.1. How are supermassive black holes (SMBHs) formed? 176
5.5. Question 5 176
5.5.1. Can the LHC produce microscopic black holes? 177
5.5.2. What if there were more dimensions than 4 to the universe? 177
5.5.3. Can you motivate why temperature is inversely propotional to mass for black hole thermodynamics? 177
5.6. Question 6 178
5.6.1. What is cyclotron radiation? 178

5.6.2. What are common sources of synchrotron radiation? 179
5.6.3. What do the spectra of other major radiation sources look like? 179
5.7. Question 7 179
5.7.1. Derive Einstein's coefficients. 180
5.7.2. How do atoms end up in states where they can only transition back out through a forbidden line? 180
5.8. Question 8 180
5.8.1. Why are polytropes useful? 181
5.8.2. What numerical methods are used to for calculating polytropes? 181
5.8.3. How is the Chandrasekhar Mass estimated using polytropes? 181
5.9. Question 9 182
5.9.1. Why do we believe neutrinos have masses? 182
5.10. Question 10 183
5.10.1. Why does the star expand homologously? 184
5.10.2. What if you increased fusion tremendously inside a star? Would that not destroy it? 184
5.11. Question 11 184
5.12. Question 12 184
5.12.1. How can you tell if something is degenerate? 185
5.12.2. Why are white dwarfs composed of an electron gas (rather than, say, a solid)? Are WDs pure Fermi degenerate gases? 185
5.12.3. Sketch the derivation of the Fermi pressure. 185
5.13. Question 13 186
5.13.1. How will your results change if accretion was spherical? 187
5.13.2. What is Bondi accretion? 187
5.14. Question 14 187
5.14.1. What is the significance of only being able to find this coordinate system at P? 188
5.14.2. What is the strong equivalence principle? 188
5.15. Question 15 188
5.15.1. Which of these devices are photon counters? 190
5.16. Question 16 190
5.17. Question 17 190
5.17.1. What if the sky were non-negligible? What if dark current and readout noise were non-negligible? 191
5.17.2. What if you were doing this in the radio? 191
5.18. Question 18 191
5.18.1. What are Stokes' parameters? 193
5.18.2. Why is polarization useful in optics? 193
5.18.3. Give some examples of linear and circular polarization in nature. 193
5.19. Question 19 193
5.19.1. Derive the Étendue conservation rule. 194
5.20. Question 20 194
5.21. Question 21 196
5.21.1. Derive the propagation of uncertainty formula, Eqn. 295. 197
5.21.2. What observational and instrumental simplifications have you made to answer this question? 197
5.22. Question 22 197
5.23. Question 23 199
5.23.1. What is hypothesis testing? 199
5.23.2. In this problem, what could observationally go wrong? 200

1. COSMOLOGY (EARLY UNIVERSE, CMB, LARGE-SCALE STRUCTURE)

1.1. *A Very Brief Primer on Cosmology*

Just like in stars far too many question depend on the equations of stellar structure, in cosmology too many questions depend on the basic underpinnings of the FLRW universe. We will summarize the results below. This information comes from Ch. 4 and 5 of Ryden (2003).

1.1.1. *The FLRW Universe*

In accordance with the cosmological principle (that there be a set of observers that see the universe as homogeneous and isotropic), the spatial extent of the universe must have uniform curvature (unless we move to truly non-trivial geometries). This restricts our metric to be of a form known as the Robertson-Walker metric

$$ds^2 = cdt^2 - a(t)^2 \left(\frac{dx^2}{1 - \kappa x^2/R^2} + x^2 d\Omega^2 \right) \tag{1}$$

where $\kappa = -1,\, 0\, 1$ and R scales κ. Another way of writing this metric (and making it perhaps more palatable) is

$$ds^2 = cdt^2 - a(t)^2 \left(dr^2 + S_\kappa^2(r) d\Omega^2 \right) \tag{2}$$

where

$$S_\kappa = \begin{cases} R \sin(r/R) & \text{if } \kappa = 1 \\ r & \text{if } \kappa = 0 \\ R \sinh(r/R) & \text{if } \kappa = -1 \end{cases}, \tag{3}$$

Writing the metric in this form shows that if κ is non-zero, angular lengths are either decreased (for positive curvature) or increased (for negative). Just as importantly, this metric indicates that, like in Minkowski space, time is orthogonal to position (meaning we can foliate the spacetime such that each hypersurface slice can be associated with a single time t) and radial distances are independent of curvature S_κ.

The solution to the RW metric is known as the Friedmann-Lemaître equation, and describes how the scale factor $a(t)$ changes with time:

$$\left(\frac{\dot{a}}{a} \right)^2 = H^2 = \frac{8\pi G}{3} \rho - \frac{\kappa c^2}{R^2} \frac{1}{a^2} + \frac{\Lambda c^2}{3} \tag{4}$$

where ρ is the matter-energy density of the universe, $\kappa c^2/R^2$ the curvature term, λ the cosmological constant and a the scale factor in the RW metric. This expression can also be derived (but becomes unscrutable because the terms make little sense) by representing the universe by a homologously expanding sphere (i.e. an explosion at $t = 0$), and considering the dynamics of a test mass within this universe. If $r = r_0 a$ (noting this automatically produces $v = r_0 \dot{a}$, so $v/r = \frac{\dot{a}}{a} = H$, reproducing Hubble's law), we can integrate $\frac{d^2 r}{dt^2} = \frac{4\pi G r \rho}{3}$[1] to get Eqn. 4 (with Λ subsumed into a constant of integration). Even more easily, we can do the same with energy balance, $K + U = E$.

In a Λ-free universe, if a value of H^2 is given, ρ and κ/R^2 are linked, and there is a critical density

$$\rho_c = \frac{3H^2}{8\pi G}, \tag{5}$$

for which the universe is flat. We define $\Omega \equiv \rho/\rho_c$. We can then rewrite the FL solution as $1 - \Omega = \frac{-\kappa c^2}{R^2 a^2 H^2}$. Note that the right side cannot change sign! This means that if at any time $\rho > \rho_c$, the universe will forever be closed; if $\rho < \rho_c$, the universe will forever be open, and if $\rho = \rho_c$, the equality will hold for all time and the universe will be flat.

1.1.2. *The Fluid and Acceleration Equations*

From the first law of thermodynamics and an assumption of adiabaticity, $dE + PdV = 0$. $\dot{V}/V = 3\dot{a}a$, and $dE/dt = \rho c^2 dV/dt + c^2 V d\rho/dt$. This gives us

$$\dot{\rho} + 3\frac{\dot{a}}{a}(\rho + P/c^2). \tag{6}$$

This can be combined with the FL equation to get

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3} \left(\rho + 3\frac{P}{c^2} \right) + \frac{\Lambda c^2}{3} \tag{7}$$

[1] Recall that $2\dot{a}\ddot{a} = \frac{d}{dt}\dot{a}^2$.

1.1.3. *Equations of State*

Cosmologists generally use equations of state that look like

$$P = \omega c^2 \rho. \tag{8}$$

The ideal gas law, for example, has an ω that is dependent on temperature, and therefore time. "Dust", which is pressure-free, has $\omega = 0$ - stars exert little enough pressure to be considered a dust. A relativistic equation of state always has $P = \frac{1}{3}\rho c^2$, including photons. Dark energy has $\omega = -1$. In substances with positive ω, $\sqrt{\frac{P}{\rho}} = c_s \leq c$, which restricts $w \leq 1$.

Combining Eqn. 6 with $P = \omega c^2 \rho$ gives us $\rho = \rho_0 a^{-3(1+\omega)}$. From this we determine that matter density $\rho = \rho_0 a^{-3}$ and radiation density $\rho_r = \rho_{r,0} a^{-4}$. We can compare the densities of any component to the critical density to obtain Ω. For example, $\rho/\rho_c = \rho/(3H^2/8\pi G) = \frac{8\pi G}{3}\rho/H^2$. We then note that $\frac{8\pi G}{3}\rho = \frac{8\pi G}{3}\rho\frac{\rho_0}{\rho_0} = H_0^2\Omega_{m,0}a^{-3}$ - conversions such as this will be useful in the following section. Note that $\rho_\Lambda = \frac{\Lambda c^2}{8\pi G}$, giving $\Omega_\Lambda = \frac{\Lambda c^2}{3H^2}$.

Taking ratios of Ωs gives us the energy component that dominates (ex. radiation to matter is $\Omega_r/\Omega_m = \rho_{r,0}/\rho_{m,0}\frac{1}{a} \approx \frac{1}{3600a}$ if $a_0 = 1$, indicating there was a time when radiation dominanted the energy of the universe.

A related question is how the radiation field temperature scales with time. Assuming adiabaticity, $dQ = dE + PdV$ (the work is being done by the radiation field on the universe). Since $P = \frac{1}{3}U = \frac{1}{3}aT^4$ we obtain $\frac{1}{T}\frac{dT}{dt} = -\frac{1}{3V}dVdT$, which implies (through integration and the fact that V scales like a^{-3}) that $T \propto a^{-1}$.

1.1.4. *History of Expansion*

Let us consider several possibilities:

- An empty flat universe is static. An empty, open universe goes like $a = ct/R$. An empty, closed universe is impossible.
- A flat universe with a single component would have $\dot{a}^2 = \frac{8\pi G}{3}\rho_0 a^{-(1+3\omega)}$. This gives $a \propto t^{2/(3+3\omega)}$.
- A Λ-dominated universe would have $\dot{a}^2 = \frac{\Lambda c^2}{3}a^2$, which gives $a \propto \exp(\sqrt{\Lambda c^2/3}t)$. We note we could have snuck Λ into the energy density of the universe if we set $\omega = -1$ and $\rho_0 = c^2/8\pi G$.

We may now consider a universe with radiation, stars, and a cosmological constant. Since $\kappa/R^2 = \frac{H_0^2}{a^2}(\Omega_0 - 1)$ (so that R may be written as $\frac{c}{H_0}\sqrt{|\Omega_\kappa|}$), we can actually write the FL equation as $H^2 = \frac{8\pi G}{3}\rho - \frac{H_0^2}{a^2}(\Omega_0 - 1)$, where ρ includes radiation, matter and dark energy, and if we divided by H_0^2, we get

$$\frac{H^2}{H_0^2} = \frac{\Omega_{r,0}}{a^4} + \frac{\Omega_{m,0}}{a^3} + \Omega_{\Lambda,0} + \frac{1-\Omega_0}{a^2} \tag{9}$$

where $\Omega_0 = \Omega_{r,0} + \Omega_{m,0} + \Omega_{\Lambda,0}$. Note how the curvature still responds to the total matter-energy density in the universe, but the expansion history may now be altered by Λ. Assuming that radiation is negligible, Fig. 16 describes the possible fates and curvatures of the universe.

Using measured values of the Ωs, we find the universe to have the expansion history given in Fig. 2.

1.1.5. *Distance and Size Measurements*

The redshift z is given by

$$1 + z = \frac{\lambda_0}{\lambda_e} = \frac{a_0}{a_e} \tag{10}$$

where subscript e stands for "emitted".

Taylor expanding the current a, we obtain $a(t) \approx a(t_0) + \dot{a}|_{t=t_0}(t-t_0) + \frac{1}{2}\ddot{a}|_{t=t_0}(t-t_0)^2$. Dividing both sides by a_0 (which is equal to 1) we get $1 + H_0(t-t_0) - \frac{1}{2}q_0H_0^2(t-t_0)^2$. $q_0 = -\ddot{a}/aH^2|_{t=t_0}$ is known as the deceleration parameter, and helps constrain the makeup of the universe, since $q_0 = \frac{1}{2}\sum_\omega \Omega_\omega(1+3\omega)$.

The comoving distance (interpretable as how distant the object would be today) to an object whose light we are seeing is given by

$$d_c(t_0) = c\int_{t_e}^{t_0}\frac{dt}{a}, \tag{11}$$

which can be converted into $d_c = c\int_0^z \frac{dz}{H}$. Since radial distances are not curvature-dependent, the proper (physical) distance is simply given by (Davis & Lineweaver 2004)

$$d_p(t) = a(t)d_c \tag{12}$$

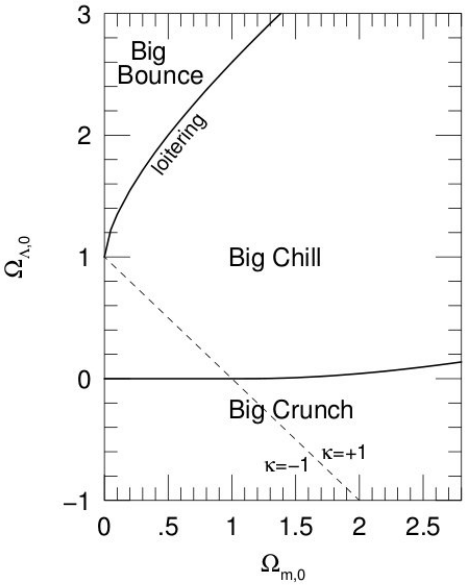

FIG. 1.— Fate of the universe as a function of Ω_m and Ω_Λ. From Ryden (2003), her Fig. 6.3.

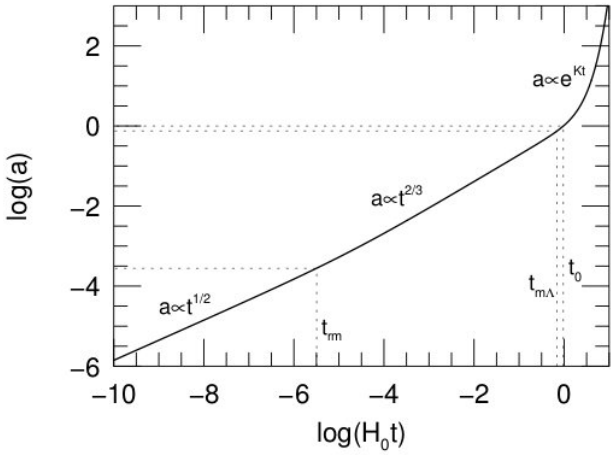

FIG. 2.— Fate of the universe, using measured values of Ω_i. From Ryden (2003), her Fig. 6.5.

where $a_0 = 1$ is assumed, and t represents time since the beginning of the universe. The luminosity distance is defined as

$$d_L = \sqrt{\frac{L}{4\pi F}} = S_\kappa(r)(1+z). \tag{13}$$

The second expression is due to two factors - first, the expansion of space drops energy with $1+z$, and increases the thickness of any photon shell dr by $1+z$ as well. The area covered by the wave of radiation is $4\pi S_\kappa^2(r)$ ($S_\kappa = r$ for a flat universe), where r should be interpreted as the comoving distance d_c. The angular diameter distance $d_A = \frac{l}{d\theta}$ (l is the *physical* length of an object at the time the light being observed was emitted) is given by the fact that $ds = a(t_e)S_\kappa(r)d\theta$. If the length l is known, then $ds = l$ and we obtain

$$d_A = \frac{S_\kappa}{1+z}. \tag{14}$$

Note that the angular diameter distance is related to the luminosity distance by $d_A = d_L/(1+z)^2$. For $z \to 0$, all these distances are equal to cz/H_0, but at large distances they begin to differ significantly.

1.2. *Question 1*

QUESTION: What is recombination? At what temperature did it occur? How does this relate to the ionization potential of Hydrogen?

Most of this information comes from Ryden (2003), filtered through Emberson (2012). Subscript 0 will represent present-day values.

Recombination is when the universe cooled to the point at which protons combined with electrons to form hydrogen atoms. The resulting massive decrease in opacity caused the universe to become optically thin, and the photon field of the universe decoupled from its matter counterpart[2]. Recombination does not refer to a single event or process: the epoch of recombination is the time at which the baryonic component of the universe went from being ionized to being neutral (numerically, one might define it as the instant in time at which the number density of ions is equal to the number density of neutral atoms). The epoch of photon decoupling is the time at which the rate at which photons scatter from electrons becomes smaller than the Hubble parameter (at the time). When photons decouple, they cease to interact with the electrons, and the universe becomes transparent. Third, the epoch of last scattering is the time at which a typical CMB photon underwent its last scattering from an electron. The three processes are related because hydrogen opacity is the driver of all three.

For simplification, let us assume the universe is made completely of hydrogen atoms. ^{4}He has a higher first ionization energy significantly higher than that of H, and therefore helium recombination would have occured at an earlier time.

The degree of H ionization is determined by the Saha equation, which can be derived (especially for H, where it is easy) from the grand partition function Ξ (Carroll & Ostlie 2006),

$$\frac{N_{i+1}}{N_i} = \frac{2Z_{i+1}}{n_e Z_i}\left(\frac{m_e k_B T}{2\pi\hbar^2}\right)^{3/2} e^{-\chi_i/k_B T} \tag{15}$$

where i indicates the degree of ionization, χ_i is the ionization energy from degree i to degree $i+1$. Suppose we ignore excited internal energy states (note that Ryden does this, but does not make it explicit); then $Z_{H^+} = Z_p = g_p = 2$ and $Z_H = g_H = 4$, for all possible spin states of the nucleus and electron. This gives us (multiplying the lefthand side of Eqn. 15 by V/V)

$$\frac{n_H}{n_p} = n_e\left(\frac{m_e k_B T}{2\pi\hbar^2}\right)^{-3/2} e^{\chi/k_B T} \tag{16}$$

where $\chi = 13.6$ eV and $n_e = n_p$. Using the fact that the number of photons is $0.243\left(\frac{k_B T}{\hbar c}\right)^3$ and $n_p = n_\gamma \eta$, with $\eta \approx 5.5 \times 10^{-10}$ (this does not change by much throughout the era of recombination), we may eliminate n_e and solve for when the left side is equal to one, i.e. when X, the ionization fraction, is equal to 1/2. This gives us $T = 3740$ K, $z = 1370$ (0.24 Myr for a matter dominated and flat universe). Past this temperature, photons became too cold to ionize H. The evolution of X with redshift is given in Fig. 3.

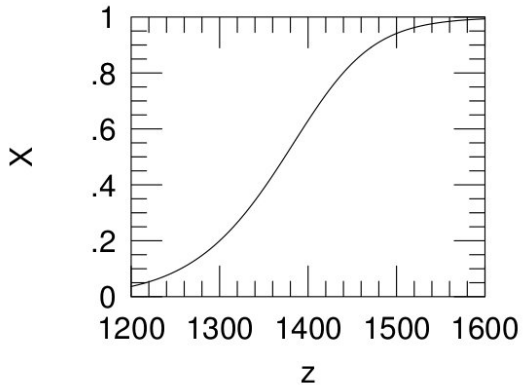

FIG. 3.— Change in ionized fraction X as a function of redshift. From Ryden (2003), her Fig. 9.4.

1.2.1. *Couldn't photons have decoupled from baryons before recombination?*

Photons are coupled to baryons through photoionization and recombination, though in the very hot universe the dominant interaction would have been Thomson scattering off free electrons, with a rate given by $\Gamma = n_e \sigma_e c$. Values can be calculated using $n_e = \frac{0.22\text{m}^3}{a^3}$ (if $a = 1$ today) and $\sigma_e = 6.65 \times 10^{-29}$. Photons decouple (gradually) from baryons when Γ exceeds H, equivalent to saying the mean free path λ exceeds c/H. The critical point $\Gamma = H$ occured at $z \approx 42$, $T \approx 120$ K, long past recombination.

[2] Since atoms were always ionized before this, "recombination" is almost a misnomer!

If we perform the same calculation during the era of recombination, setting n_e using the analysis above and obtaining H from $H(t) = H_0\Omega_m \frac{a_0^3}{a^3} = H_0\Omega_m \frac{1}{(1+z)^3}$, we obtain $z \approx 1100$ and $T \approx 3000$ (exact answers are difficult without modelling, since during the final stages of recombination the system was no longer in LTE).

1.2.2. *What is the last scattering surface?*

The last scattering surface is the $\tau = 1$ surface for photons originally trapped in the optically thick early universe. The age $t_0 - t$ of this surface can be found using

$$\tau = 1 = \int_t^{t_0} \Gamma(t) dt \tag{17}$$

In practice this is difficult, and so we again estimate that $z \approx 1100$ for last scattering.

1.3. *Question 2*

QUESTION: The universe is said to be "flat", or, close to flat. What are the properties of a flat universe and what evidence do we have for it?

This information comes mainly from Emberson (2012), with supplement from Carroll & Ostlie (2006).

As noted in Sec. **?**, the FLRW universe may only have three types of curvature. When $\kappa = 1$, the universe is positively curved, since $R\sin(r/R) < r$ (i.e. the actual size of the object would be smaller than its physical size, consistent with the fact that two straight lines intersecting on a circle will eventually meet again) and when $\kappa = -1$ the universe is negatively curved, since $R\sinh(r/R) > r$.

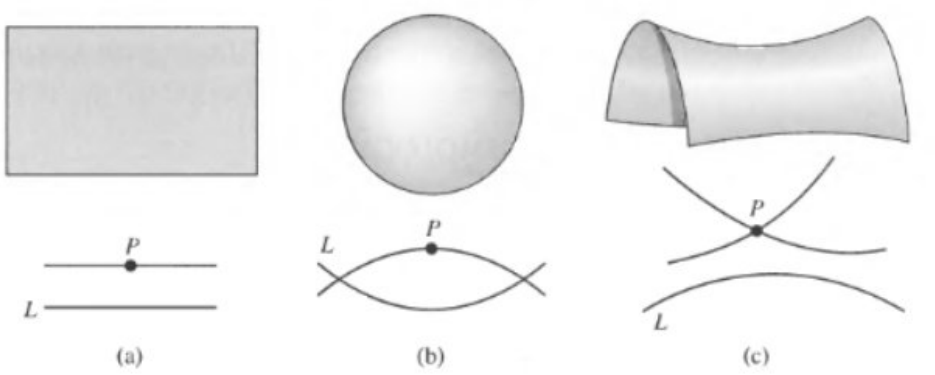

FIG. 4.— Schematic of two dimensional analogues to possible curvature in an FLRW universe. The behaviour of lines parallel at a point P within each space is also drawn. A Euclidian plane has no curvature, a sphere has positive curvature and a saddle has negative curvature. From Carroll & Ostlie (2006), their Fig. 29.15.

Fig. 4 shows the two primary geometric features of a flat, closed and open universe. In open and closed universes, parallel lines tend to diverge (open) or converge (closed), while for a flat universe two parallel lines remain parallel indefinitely. Open and flat universes are infinite, while a closed universe may have a finite extent, since it "curves back" on itself. In $\Lambda = 0$ universes, the geometry of the universe is intimately related to the matter-energy density of the universe.

Measurement of the curvature of the universe is difficult. In a $\Lambda = 0$ universe it actually is greatly simplified, since curvature and expansion history are linked, and the age of the universe, combined with H_0, can be used to determine the curvature (or H_0 and q_0). In a universe with a non-zero cosmological constant, however, the age of the universe is decoupled from the curvature. Instead, we use a standard ruler: the first peak of the CMB power spectrum. This peak, due to the length of the sound horizon at decoupling, is

$$r_s(z_{rec}) = c \int_{z_{rec}}^{\infty} \frac{c_s}{H(z)} dz \tag{18}$$

where $c_s = (3(1 + 3\rho_{bary}/\rho_{ph}))^{-1/2}$ (Vardanyan et al. 2009). Detailed measurements of higher order peaks and their spacings in the CMB allow us to constrain both $H(z)$ and c_s, and obtain a preferred length scale (Eisenstein 2005). This is our standard ruler, and if we measure its current angular size θ, we can use Eqn. 14 alongside Eqn. 3 to determine

$$\frac{\theta}{1+z} = \frac{l}{S_\kappa} \tag{19}$$

Note that l is known, but S_κ depends on the co-moving distance between us and the CMB. This requires some knowledge of the subsequent expansion history of the universe, or else there is a degeneracy between Ω_m, Ω_Λ and Ω_κ (Komatsu et al. 2009). An additional constraint, such as a measurement of H_0, or the series of luminosity distance measurements using high-z SNe, allows us to constrain Ω_κ (Komatsu et al. 2009). See Fig. 5.

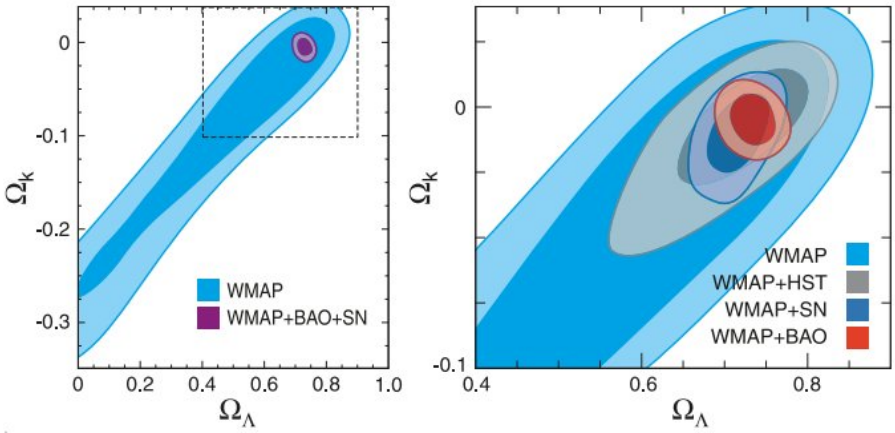

FIG. 5.— Joint two-dimensional marginalized constraint on the dark energy density Ω_Λ, and the spatial curvature parameter, Ω_κ. The contours show the 68% and 95% confidence levels. Additional data is needed to constrain Ω_κ: HST means H_0 from Hubble measurements, SN means luminosity distances from high-z SN, and BAO means baryon acoustic oscillation measurements from galaxy surveys. From Komatsu et al. (2009), their Fig. 6.

1.4. *Question 3*

QUESTION: Outline the development of the Cold Dark Matter spectrum of density fluctuations from the early universe to the current epoch.

Most of this information is from Schneider (2006), Ch. 7.3 - 7.5.

The growth of a single perturbation (described as one of the follow-up questions) in a matter-dominated universe can be described in the following way. We define the relative density contrast $\delta(\vec{r},t) = (\rho(\vec{r},t) - \bar{\rho})/\bar{\rho}$; from this $\delta(\vec{r},t) \leq -1$. At $z \sim 1000$ $|\delta(\vec{r},t)| << 1$. The mean density of the universe $\bar{\rho}(t) = (1+z^3)\bar{\rho}_0 = \bar{\rho}_0/a(t)^3$ from Hubble flow. Like in the classic Newtonian stability argument of an infinite static volume of equally space stars, any overdense region will experience runaway collapse (and any underdense region will become more and more underdense). In the linear perturbative regime, the early stages of this collapse simply make it so that the the expansion of the universe is delayed, so $\delta(\vec{r},t)$ increases. As it turns out, $\delta(\vec{r},t)$ can be written as $D_+(t)\delta_0(\vec{x})$ in the linear growth regime. $D_+(t)$ is normalized to be unity today, and $\delta_0(\vec{x})$ is the linearly-extrapolated (i.e. no non-linear evolution taken into account) density field today.

The two-point correlation function $\xi(r)$ (Sec. 1.18) describes the over-probability of, given a galaxy at $r = 0$, there will be another galaxy at r (or x, here). It describes the clustering of galaxies, and is key to understanding the large-scale structure of the universe. We define the matter power spectrum (often shortened to just "the power spectrum") as

$$P(k) = \int_{-\infty}^{-\infty} \xi(r) exp(-ikr) r^2 dr \tag{20}$$

Instead of describing the spatial distribution of clustering, the power spectrum decomposes clustering into characteristic lengths $L \approx 2\pi/k$, and describes to what degree each characteristic contributes to the total overprobability.

Since the two-point correlation function depends on the square of density, if we switch to co-moving coordinates and stay in the linear regime,

$$\xi(x,t) = D_+^2(t)\xi_0(x,t_0). \tag{21}$$

Likewise,

$$P(k,t) = D_+^2 P(k,t_0) \equiv D_+^2 P_0(k), \tag{22}$$

i.e. everything simply scales with time. This the evolution of the power spectrum is reasonably easily described.

The initial power spectrum $P_0(k)$ was generated by the quantum fluctuations of inflation. It can be argued (pg. 285 of Schneider (2006)) that the primordial power spectrum should be $P(k) = Ak^{n_s}$, where A is a normalization factor

that can only be determined empirically. $P(k)$ when $n_s = 1$ is known as the Harrison-Zel'dovich spectrum, which is most commonly used.

An additional correction term needs to be inserted is the transfer function to account for evolution in the radiation-dominated universe, where our previous analysis does not apply. We thus introduce the transfer function $T(k)$, such that $P_0(k) = Ak^{n_s}T(k)^2$. $T(k)$ is dependent on whether or not the universe consists mainly of cold or hot ($k_BT << mc^2$, where T is the temperature at matter-radiation equality) dark matter. If hot dark matter dominate sthe universe, they freely stream out of minor overdensities, leading to a suppression of small-scale perturbations. Since our universe is filled with cold dark matter, this need not be taken into account (and indeed gives results inconsistent with observations). $T(k)$ also accounts for the fact that $a(t) \propto t^{1/2}$ rather than $t^{2/3}$ during radiation domination, and that physical interactions can only take place on scales smaller than $r_{H,c}(t)$ (the co-moving particle horizon) - on scales larger than this GR perturbative theory must be applied.

Growth of a perturbation of length scale L is independent of growths at other length scales. The growth of a density qualitatively goes like this:

1. In the early universe, a perturbation of comoving length L has yet to enter the horizon. According to relativistic perturbation theory, the perturbation grows $\propto a^2$ in a radiation-dominated universe, and $\propto a$ in a matter-dominated universe.

2. At redshift z_e, when $r_{H,c}(z_e) = L$, the perturbation length scale becomes smaller than the horizon. If the universe is still radiation-dominated, the Mészáros effect prevents effective perturbation growth, and the overdensity stalls (Mészáros 2005).

3. Once the universe becomes matter dominated ($z < z_{\rm eq}$), the perturbation continues to grow $\propto a$.

There is therefore a preferred length scale $L_0 = r_{H,c}(z_{\rm eq}) \approx 12(\Omega_m h^2)^{-1}$ Mpc. The transfer function then has two limiting cases: $T(k) \approx 1$ for $k << 1/L_0$, and $T(k) \approx (kL_0)^{-2}$ for $k >> 1/L_0$. This generates a turnover in $P_0(k)$ where $k = 1/L_0$. Note that due to the dependence on the sound horizon on $\Omega_m h^2$ we often define the shape parameter $\Gamma = \Omega_m h$.

One last modification must be made to this picture: at a certain point growth becomes non-linear, and our analysis must be modified.

Fig. 6 shows the schematic growth of a perturbation, as well as both the primordial Harrison-Zel'dovich spectrum and the modern-day power spectrum for a series of different cosmological parameters.

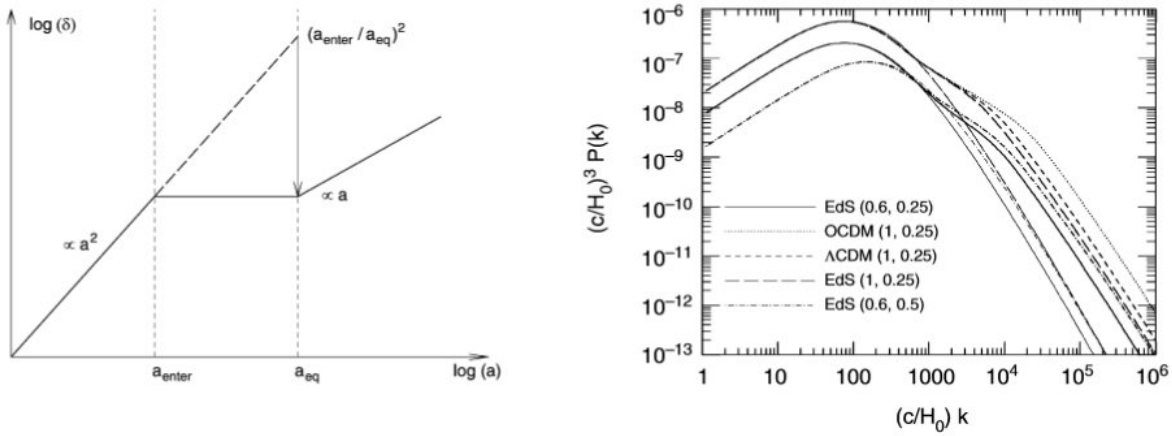

FIG. 6.— Left: evolution of a density perturbation. The $(a_{\rm enter}/a_{\rm eq})^2$ line indicates the degree of suppression during radiation domination. Right: the current power spectrum of density fluctuations for CDM models. The various curves have different cosmological models (EdS, $\Omega_m = 1$, $\Omega_\Lambda = 0$, OCDM, $\Omega_m = 0.3$, $\Omega_\Lambda = 0$, ΛCDM, $\Omega_m = 0.3$, $\Omega_\Lambda = 0.7$). Values in paretheses specify (σ_8, Γ). The thin curves correspond to power spectra linearly extrapolated, while the thick curves include non-linear corrections. From Schneider (2006), his Figs. 7.5 and 7.6.

1.4.1. *How do baryon and photon density perturbations grow?*

This information is from Schneider (2006), pg. 288 - 289.

Baryon and photon density perturbations grew alongside dark matter perturbations until z_e, at which point baryon acoustic oscillations began, stimying any growth until recombination, $z_r < z_{\rm eq}$. Following this, the photon overdensities escaped while the baryon overdensities began to track the dark matter overdensities.

1.4.2. *How does an individual density perturbation grow?*

This is described in much greater detail in Schneider (2006), Ch. 7.2.

If we assume a pressure-free ideal fluid, we can write the Euler's and continuity equations in comoving coordinates and linearize them to obtain $\frac{\partial^2\delta}{\partial t^2} + \frac{2\dot{a}}{a}\frac{\partial\delta}{\partial t} = 4\pi G\bar{\rho}\delta$. This means we can separate $\delta(\vec{x},t)$ into $D(t)\delta_0(\vec{x})$; i.e. at all comoving points $\vec{x}$ the overdensity rises in the exact same manner over time. Our equation of motion then becomes

$$\ddot{D} + \frac{2\dot{a}}{a}\dot{D} - 4\pi G\bar{\rho}(t)D = 0. \tag{23}$$

There are two solutions to this equation, and we call the increasing one the growth factor, $D_+(a)$. In general $D_+(a)$ is normalized so that $D_+(1) = 1$, so that $\delta_0(\vec{x})$ is the density distribution we would have today if no non-linear effects take hold. We can show that the general increasing solution to this equation, when we switch from time to a, is

$$D_+(a) \propto \frac{H(a)}{H_0}\int_0^a \frac{da'}{(\Omega_m/a' + \Omega_\Lambda a'^2 + \Omega_k)^{3/2}} \tag{24}$$

For an Einstein-de Sitter universe, we can show, through an ansatz that $D \propto t^q$ and Eqn. 23 that $D_+(t) = (t/t_0)^{2/3}$. During matter domination, overdensities grew with the scale length.

Eventually $D(t)\delta(\vec{x})$ approaches 1, and the linear approximation fails. Growth increases dramatically (Fig. 7).

In the case of a uniform homogeneous sphere with density $\rho = \bar{\rho}(1+\delta)$, where δ is the average density perturbation in the sphere, and we have switched back to proper distances rather than comoving. The total mass within the sphere is $M \approx \frac{4\pi}{3}R_{\rm com}^3\rho_0(1+\delta_i)$, where $R_{\rm com} = a(t_i)R$ is the initial comoving sphere radius (R is the physical radius of the sphere), and $\rho_0 = \bar{\rho}/a^3$ is the present average density of the universe. We may then model the mass and radius of the sphere as a miniature universe governed by the Friedmann equations. If the initial density of the system is greater than critical, the sphere collapses. Because of the time-reversibility of the Friedmann equations, if we know the time $t_{\rm max}$ where $R_{\rm com}$ is maximum, we know the time $t_{\rm coll}$ when the universe collapses back into a singularity.

This collapse is unphysical. In reality violent relaxation will occur - density perturbations in the infalling cloud will create knots due to local gravitational collapse; these knots then scatter particles, which create more perturbations, creating more knots. This creates an effective translation of gravitational potential energy to kinetic (thermal) energy, within one dynamical time. It turns out a virialized cloud has density

$$\langle\rho\rangle = (1+\delta_{\rm vir})\bar{\rho}(t_{\rm coll}), \tag{25}$$

where $1+\delta_{\rm vir} \approx 178\Omega_m^{-0.6}$.

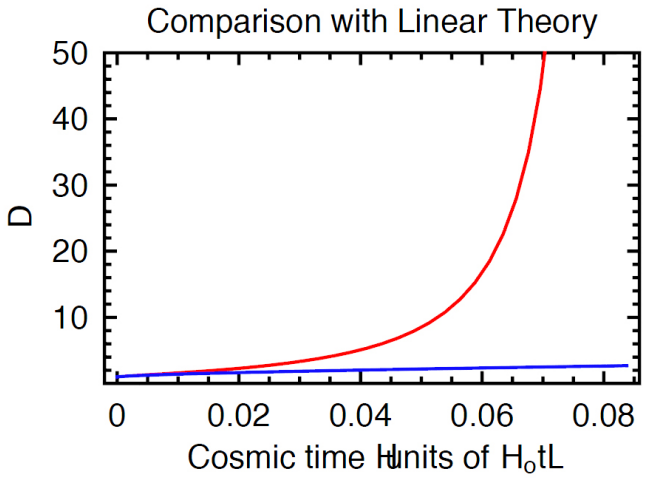

FIG. 7.— Growth of a density fluctuation taking into account non-linear evolution, versus the equivalent linear evolution. The singularity that eventually forms in the non-linear case is not physical, as the dust approximation eventually fails and virialization occurs. In baryonic material dissipative cooling also occurs. From Abraham (2011b).

1.4.3. *What is violent relaxation?*

This information is from Schneider (2006), pg. 235 and 290.

Violent relaxation is a process that very quickly establishes a virial equilibrium in the course of a gravitational collapse of a mass concentration. The reason for it are the small-scale density inhomogeneities within the collapsing matter distribution which generate, via Poisson's equation, corresponding fluctuations in the gravitational field. These then scatter the infalling particles and, by this, the density inhomogeneities are further amplified.

This virialization occurs on a dynamical time, and once virialization is complete, the average density of the perturbation becomes, as noted earlier, $\langle\rho\rangle \approx 178\Omega_m^{-0.6}\bar{\rho}(t_{\rm collapse})$

1.4.4. *What are top-down and bottom-up growth?*

This information is from Schneider (2006), pg. 286.

In a universe dominated by hot dark matter, all small perturbations cease to exist, and therefore the largest structures in the universe must for first, with galaxies fragmenting during the formation of larger structures. This top-down growth is incompatible with the fact that galaxies appear to have already collapsed, while superclusters are still in the linear overdensity regime. In a universe dominated by cold dark matter, small overdensities collapse first, and this bottom-up growth is consistent with observations.

1.4.5. *How can the power spectrum be observed?*

The matter power spectrum, if one assumes that baryons track dark matter, can be determined observationally recovering the two-point correlation function from galaxy surveys (Sec. 1.18).

1.4.6. *How can the power spectrum constrain cosmological parameters?*

This information is from Schneider (2006), Ch. 8.1.

The turnover of the power spectrum is determined by the wavenumber corresponding to the sound horizon at matter-radiation equality. This allows us to determine the shape parameter $\Omega_m h$, which can be combined with measurements of H_0 to obtain Ω_m. Detailed modelling of the power spectrum shows that the transfer function depends on Ω_b as well as Ω_m. As a result, this modelling can also derive the baryon to total matter ratio.

One important use of the dark matter power spectrum is to determine the shape and frequency of the baryon acoustic oscillations (Sec. 1.14).

1.4.7. *How can we determine the dark matter mass function from perturbation analysis?*

This information is from Schneider (2006), pg. 291 - 292.

As noted previously, a spherical region with an average density δ greater than some critical density will collapse. We can therefore back-calculate $\delta(\vec{x}, t)$ from the power spectrum[3], smooth it out over some comoving radius R to determine the average density, and determine using the critical density (given the redshift and cosmological model), the normalized number density of relaxed dark matter halos. Since the power spectrum has a normalization factor that must be determined empirically, this normalized number density can then be scaled to the true number density using observations. The result is the Press-Schechter function $n(M, z)$ which describes the number density of halos of mass $> M$ at redshift z. See Fig. 8.

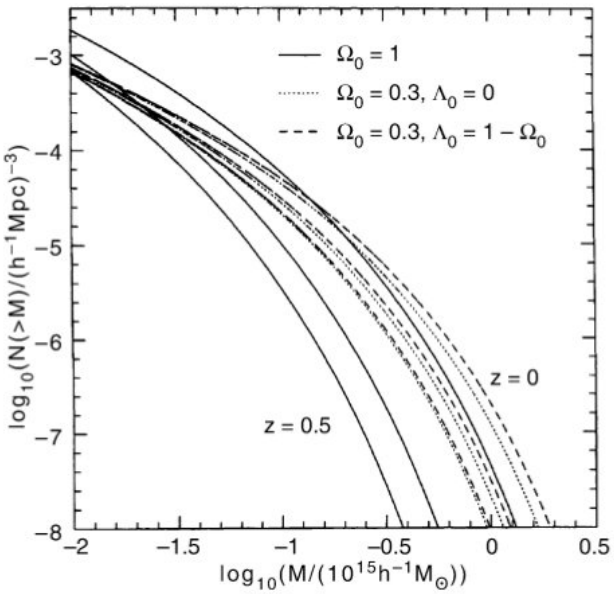

FIG. 8.— Number density of dark matter halos with mass $> M$ (i.e. a reverse cumulative model), computed from the PressSchechter model. The comoving number density is shown for three different redshifts and for three different cosmological models. The normalization of the density fluctuation field has been chosen such that the number density of halos with $M > 10^{14}h^1$ $M_\odot$ at z = 0 in all models agrees with the local number density of galaxy clusters. From Schneider (2006), his Fig. 7.7.

[3] This only works if $P(k)$ alone is sufficient to describe δ, but because this distribution is Gaussian (for complicated reasons), $P(k)$ completely constrains δ.

Since there is only one normalization, a survey of galaxy cluster total masses at different redshifts (using hot gas and a mass-to-light conversion, gravitational microlensing, etc) can be used to determine cosmological parameters. This is because the minimum overdensity for collapse $\delta_{\rm min}$ is dependent on both the growth rate of overdensities and the expansion of the universe. Increasing Ω_m, for example, decreases $n(M,z)/n(M,0)$, since massive halo growth is more extreme the higher Ω_m is. A large Ω_Λ dampens massive halo growth.

1.5. *Question 4*

QUESTION: State and explain three key pieces of evidence for the Big Bang theory of the origin of the Universe.

This information is cribbed from Emberson (2012).

The Big Bang theory is the theory that the universe started off in an extremely hot, dense state, which then rapidly expanded, cooled, and became more tenuous over time. The Big Bang theory requires that at some point in the past a). the universe was born, b). the universe was extremely hot and c). objects were much closer together. The three key pieces of evidence are:

1. **Hubble's Law**: galaxies isotropically recede from our position with the relationship

$$\vec{v} = H_0 \vec{r} \tag{26}$$

known as Hubble's Law. As it turns out, moving into the frame of another galaxy ($\vec{r}' = \vec{r} - \vec{k}$, $\vec{v}' = \vec{v} - H_0\vec{k} = H_0(\vec{r} - \vec{k}) = H_0\vec{r}'$) does not change any observations. At larger distances, Hubble's Law breaks down (see Sec. 1.7), but the rate of expansion only increases with distance. Because of this isotropic radial motion outward, we can back-calculate a time when all the galaxies ought to be together at one point. This time is $t_0 = r/v = 1/H_0 \approx 14$ Gyr, the Hubble Time. This gives an age to the universe, and indicates that in the distant past everything was closer together.

2. **The Cosmic Microwave Background**: the cosmic microwave background (CMB) is a near perfect isotropic blackbody with a (current) $T_0 \approx 2.73$ K. For a blackbody, $\lambda_{\rm peak} = 0.0029\,{\rm mK}/T$, $U = aT^4$ and $n = \beta T^3$, which gives us $n \approx 400$ cm^{-3}, $\epsilon \approx 0.25$ eV cm^{-3}, and $\lambda \approx 2$ mm. In Big Bang cosmology, this microwave background is the redshifted ($T \propto a^{-1}$) vestige of the surface of last scattering, when $T \approx 3000$ K and the universe became neutral enough for photons to travel unimpeded. This is evidence that the universe used to be hot.

3. **Big Bang Nucleosynthesis**: in the Big Bang theory, the lightest elements were created out of subatomic particles when the temperature dropped enough that the average photon was significantly below the binding energy of light elements. A detailed calculation of nucleosynthetic rates of H, D, He and Li during the first few minutes of the universe is consistent with the current abundances of light elements in the universe. See Sec. 1.8.

Additionally, no object has been found to be older than the currently accepted age of the universe, 13.7 Gyr. As we look back in time, we notice that the average galaxy in the universe looked considerably different - this evolution is consistent with ΛCDM cosmology, which has small, dense cores of dark matter forming due to gravitational instability, and then merging to form larger cores.

1.5.1. *What is Olbers's Paradox?*

Olbers's paradox is the apparent contradiction one has when an infinitely old, infinitely large universe with a fixed stellar density is assumed. In such a universe every single line of sight would eventually reach a star's photosphere. Since a typical photospheric temperature is ~ 5000 K and surface brightness is independent of distance, we would expect the entire sky to be at ~ 5000 K, roasting the Earth. Setting a finite age to the universe is one solution to the paradox; another would be that stars only formed in the last several billion years, and light from more distant stars have yet to reach us.

1.5.2. *Are there Big Bang-less cosmologies?*

It is impossible to generate a matter dominated universe for which there is no Big Bang. It is possible, however, for a Λ-dominated universe to be infinitely old, since an exponential (see Sec 1.1.4) never goes to zero. This is consistent with the steady state theory (Sec. 1.6).

1.6. *Question 5*

QUESTION: Define and describe the "tired light hypothesis" and the "steady state universe" as alternatives to the Big Bang. How have they been disproved observationally?

This information is cribbed from Emberson (2012).

The tired light hypothesis was an attempt to reconcile the steady state theory of the universe with redshift. It posited that as light travels across cosmological distances, it loses energy at a rate

$$\frac{d(h\nu)}{dl} = -H_0 h\nu, \tag{27}$$

i.e. galaxies are not receding from us, but rather light itself "grows tired". The CMB, and to a lesser extent the nature of galaxies at high redshift, are both strong evidence against the static universe the tired light hypothesis is in support of.

A more direct argument against this hypothesis is that the flux received by an observer from a source of luminosity L should, in a tired light steady state universe, be $f = \frac{L}{4\pi d_c^2(1+z)}$, while in an expanding (curved) universe it is $f = \frac{L}{4\pi S_\kappa^2(1+z)^2}$; the second $1+z$ is due to the stretching out of a shell of radiation with thickness dr. This can be tested using standardizable candles, such as Cepheids and SNe Ia. Another direct argument is that in a tired light universe the angular size of an object should scale with the square of the co-moving distance, $d\Omega = dA/r^2$. In an expanding universe, however, this value should be $d\Omega = dA(1+z)^2/S_\kappa^2$ (which also means that surface brightness $dF/d\Omega \propto (1+z)^{-4}$ instead of remaining constant as in a static universe). Observations of elliptical galaxies at $z \approx 0.85$ have show that the angular size of objects $d\Omega$ is consistent with an expanding universe.

The steady state universe assumes the "perfect" cosmological principle, that the universe is spatially homogeneous and isotropic, as well as homogeneous in time. It does assume that the universe is expanding, though to keep the universe temporally homogeneous this would mean $\dot{a}/a = H_0$, i.e. the universe expands exponentially. To maintain the same density over time, matter is thought to be created *ex nihilo* throughout the universe, and the exponential expansion of the universe would allow distant radiation and stellar material to be removed from the observable universe.

A number of observational evidence speak against this model of the universe as well. Galaxies evolve with redshift - $z > 0.3$ galaxies in rich clusters tend to be bluer than their counterparts at low redshift, indicating that the universe has had a variable star formation rate throughout its history. Even more series is the CMB. It is conceivable that the CMB is the result of either the radiation field, or the emission of stars downscattered to microwaves by intergalactic dust. The first is difficult to believe because the sum contribution of emission from the creation field at all redshifts would have to precisely equal a blackbody. The second suffers from the same problem, as well as the issue that, because the universe is isotropic, this dust would have to exist locally. The densities required to produce a CMB would also require $\tau = 1$ by $z = 2$, meaning that we would not be able to see radio galaxies with $z > 2$. This is not the case, further invalidating the steady state universe.

1.7. *Question 6*

QUESTION: Sketch a graph of recession speed vs. distance for galaxies out to and beyond the Hubble distance.

This answer is cribbed from Emberson (2012), with a helpful dallop of Davis & Lineweaver (2004).

The expansion of the universe imparts a wavelength shift on the spectra of galaxies, denoted $z = \Delta\lambda/\lambda$. In the low-$z$ limit the relationship between z and proper distance is $c(1+z) = Hd$. This shift was originally attributed to a "recession velocity" (i.e. the galaxies are moving away from us at a velocity v) and called Hubble's Law:

$$v = H_0 d \tag{28}$$

From the RW metric, at any given moment in time an object has proper distance (from us) $d_p(t) = a(t)d_c$, giving us

$$\frac{d(d_p)}{dt} = \dot{a}d_c = H(t)d_p(t). \tag{29}$$

Therefore Hubble's Law is *universal*, in that given some time t the universe undergoes homologous expansion as given by $H(t)$. If we sketched a graph of recession speed vs. comoving distance (i.e. proper distance today), it would be a straight line!

We, however, cannot measure the co-moving distance, so a more reasonable relationship to draw would be the relationship between recessional velocity and redshift z. We can calculate this by noting that $v(t) = H(t)d_p(t) = \frac{\dot{a}}{a}ad_c = \dot{a}c\int_{t_{\rm em}}^{t_0}\frac{1}{a}dt'$. t is separate from both the emission and reception times of the photon, and this is because there is no unique recessional velocity (it changes over time!) that we could point to. Converting this expression to an integration over redshift ($da = -1/(1+z)^2 dz = -dz/a^2$) and assuming $a_0 = 1$, we obtain

$$v(t,z) = \dot{a}\int_0^z \frac{c}{H}dz' \tag{30}$$

and if we assume $t = t_0$ (we want the *current*) recessional velocity, then $v(t_0, z) = \frac{1}{H_0}\int_0^z \frac{dz'}{H}$. This is plotted against z in Fig. 9.

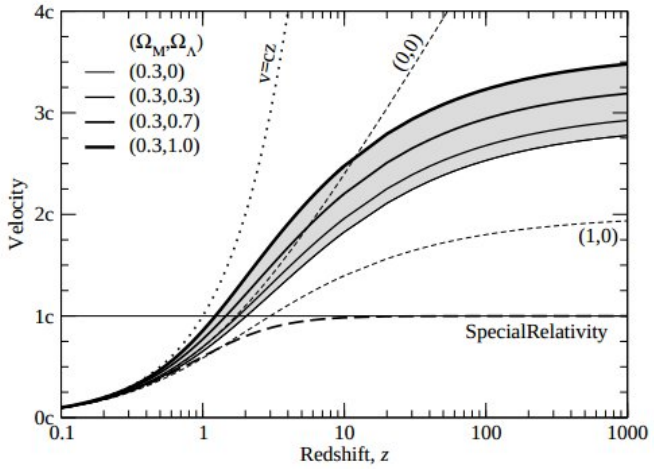

FIG. 9.— A plot of the recession velocity vs. redshift for several different FLRW cosmologies. How peculiar velocity would scale with redshift is also plotted. From Davis & Lineweaver (2004), their Fig. 2.

1.7.1. *How can we possibly see galaxies that are moving away from us at superluminal speeds?*

That nothing may move faster than the speed of light applies only to movement through space. Because it is space itself that expands, there is no restriction on maximum "speed" (more specifically $d_c da/dt$). Locally, galaxies (and light!) move in accordance with special relativity, but globally galaxies (and light!) can move away from us at superluminal speeds.

That being said, we can still see light being emitted by these objects from the distant past, so long as H is not constant! Suppose at $t_{\rm em}$ a photon is emitted by a galaxy at a critical distance $d_{p,\rm crit} = c/H$. The photon, while moving toward us in comoving space (and slowing down in comoving space) will initially have a fixed distance from us d_p, and will remain so as long as H is constant. If H decreases, however, which is the case for any universe accelerating less than exponentially, $d_{p,\rm crit}$ will increase, and d_p will be within the critical distance. The photon can then begin its journey toward us. In Fig. 10, notice how the light cone extends into the region where $d = c/H$.

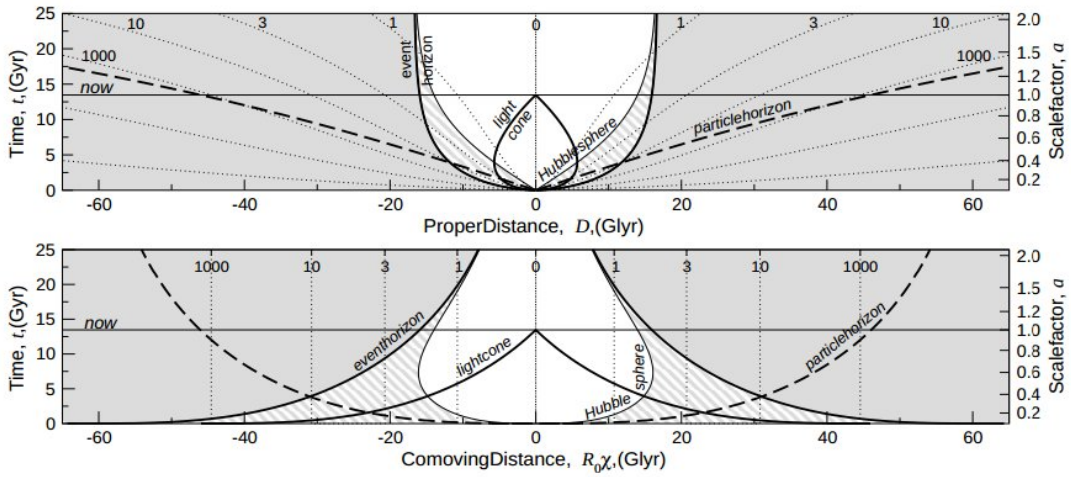

FIG. 10.— A space-time diagram of ΛCDM concordance cosmology. Above, time is plotted against proper distance, and below, time is plotted against comoving distance. "Light cone" represents the cone within which light could have reached us by now, and "event horizon" represents the furthest distance from which we will ever receive information on events at time t. The "particle horizon" represents the furthest distance objects that have ever been in in causal contact with us has gone. The "Hubble sphere" is the distance at which $d = c/H$, and the hatched region between the Hubble sphere and the event horizon represents events travelling superluminally away from us that we will one day be able to see. From Davis & Lineweaver (2004), their Fig. 1.

1.7.2. *Why can't we explain the Hubble flow through the physical motion of galaxies through space?*

If space were not expanding, but galaxies are moving away from us isotropically, then

$$v = c\frac{(1+z)^2 - 1}{(1+z)^2 + 1}. \tag{31}$$

If we assume $v = H_0 d_c$ applies to find the co-moving distance (in SR we have no way of accommodating further redshifting after photon emission, so we assume the galaxy still has the same velocity today, and follows the Hubble flow), we can use Eqn. 13 to determine the luminosity distance. We also use Eqns. 11 and 13 to determine the luminosity distance in GR. We compare this to the calculated luminosity distance using SNe Ia (any standardizable candle allows one to properly calculate the luminosity distance). The result is plotted in Fig. 11, and shows a clear bias against the special relativistic model. The reason why in the figure SR does even worse than Newtonian is simply because as $v \to c$, $d_c \to c/H_0$, resulting in a linear relationship between luminosity distance and redshift. This is not the case in either Newtonian or GR.

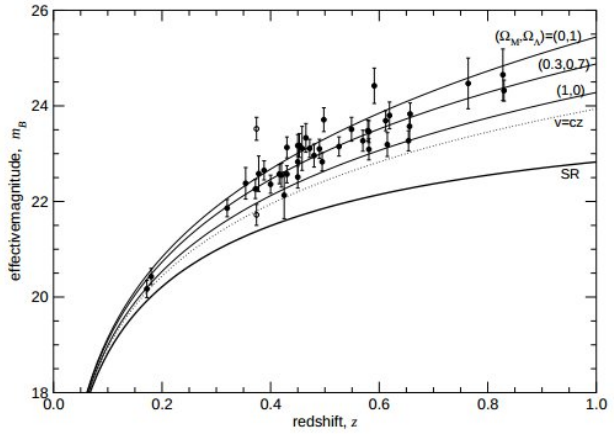

FIG. 11.— A plot of the magnitude-redshift relation, with a comparison between SR, Newtonian ($v = cz$) and several ΛCDM universes. Magnitude is calculated from luminosity distance. From Davis & Lineweaver (2004), their Fig. 5.

1.7.3. *Can galaxies with recession velocities $v > c$ slow down until $v < c$?*

Certainly! $v = H(t)d_p(t) = \dot{a}(t)d_c$, and therefore $v/c = \dot{a}(t)d_c/c$. For a matter-dominated universe $\dot{a}(t) \propto t^{-1/3}$ and therefore over time objects "slow down" (we cannot observe this, of course; light from these objects has yet to reach us!). This can be seen by the Hubble sphere expanding in Fig. 10.

1.8. *Question 7*

QUESTION: What happened in the first 3 minutes after the Big Bang? Why is only He (and tiny traces of Li) synthesized in the Big Bang?

A whole bunch of things happened in the first few minutes after the Big Bang, including inflation, CP symmetry breaking, neutrino decoupling. These features are summarized in Sec. 1.12. This question speaks mainly, however, of Big Bang nucleosynthesis (BBN).

The energy scale of BBN is set by the binding energy of nuclei - deuterium binding is about 10^5 times greater than the ionization energy of a hydrogen atom, and as a result BBN occured when $T \approx 4 \times 10^8$ K. The universe grew too cold to maintain such temperatures when it was only several minutes old.

The basic building blocks of matter are protons and neutrons. A free neutron has 1.29 MeV more energy than a proton, and 0.78 MeV more than a proton and electron. $n \to p + e^- + \bar{\nu}_e$, then, is energetically (and entropically) highly favourable, and the half-life of a neutron is about 890 seconds.

At age $t = 0.1$ s, $T \approx 3 \times 10^{10}$ K, and the mean energy per photon was about $E \approx 10$ MeV, high enough to easily begin pair production. Neutrons and protons will be at equilibrium with each other via $n + \nu_e \leftrightharpoons p + e^-$ and $n + e^+ \leftrightharpoons p + \bar{\nu}_e$, and given LTE, their densities will be given by the Maxwell-Boltzmann equation,

$$n = g\left(\frac{mkT}{2\pi\hbar^2}\right)^{3/2} \exp\left(-\frac{E}{k_B T}\right), \tag{32}$$

where the energy scale we consider is the rest mass of a proton vs. a neutron. The relative balance of neutrons and protons, then, is given by

$$\frac{n_n}{n_p} = \left(\frac{m_n}{m_p}\right)^{3/2} \exp\left(-\frac{(m_n - m_p)c^2}{k_B T}\right) \approx \exp\left(-\frac{Q}{k_B T}\right), \tag{33}$$

where Q = 1.29 MeV, which corresponds to $\sim 1.5{\times}10^{10}$ K. This shows a high preference for protons at low temperatures. In truth, however, $n+\nu_e \leftrightharpoons p+e^-$ is a weak reaction and the cross-sectional dependence of a weak reaction is $\sigma_w \propto T^2$. Since in a radiation-dominated universe $T \propto t^{-1/2}$, $\omega_w \propto t^{-1}$, and the neutron density is greater than $\propto t^{-3/2}$ (from $\rho \propto a^{-3}$ and the fact that neutron numbers are decreasing with temperature). As a result, Γ_w falls dramatically. When $\Gamma \approx H$, the neutrinos decouple from the neutrons and protons. This occurs (empirically) at about 0.8 MeV, or $T_{\rm freeze} = 9 \times 10^9$ K. Using Eqn. 33, we obtain 1 neutron for 5 protons.

The lack of neutrons prevented BBN from fusing to nickel. Proton-proton fusion is difficult due to Coulombic repulsion, and in the Sun the pp-chain has a timescale of several Gyr. This means that in the several minutes when the temperature of the universe was sufficiently high for nuclear fusion to occur, $p + n \leftrightharpoons D + \gamma$ dominated (neutron-neutron fusion has a very small cross-section). if every neutron binded to a proton, and the only nucleosynthetic product was ^{4}He, the fraction of ^{4}He/H would be (2 neutrons + 2 protons)/(6 free protons) = 1/3.

This fusion happened in several stages. The time of deuterium fusion (when $n_D/n_n = 1$) occured at $T \approx 7.6 \times 10^8$ K, or $t \approx 200$ s - this can be derived from the Saha equation ($g_D = 3$). Deterium can then be fused into tritium (^{3}H, half-life 18 years) or ^{3}He, and from there quickly fused into ^{4}He. ^{4}He is very tightly bound (hence α-decay), and there are no stable nuclei with atomic weight 5 or 8. Small amounts of ^{6}Li and ^{7}Li can be made via $^4\mathrm{He} + D \rightleftharpoons^6 \mathrm{Li} + \gamma$ and $^4\mathrm{He} +^3 \mathrm{H} \rightleftharpoons^7 \mathrm{Li} + \gamma$, which are fairly slow reactions. By the time the temperature has dropped to $T \approx 4 \times 10^8$ K at $t = 10$ min, BBN is over, and neutrons are locked up in ^{4}He and a small amount of Li.

1.8.1. *How does nucleosythesis scale with cosmological parameters?*

Nucleosynthesis depends critically on η, the baryon-to-photon ratio. A high ratio increases the temperature at which deuterium synthesis occurs, and hence gives an earlier start to BBN, allowing a greater conversion of D to ^{4}He. ^{7}Li is produced both by fusing ^{4}He and ^{3}He (decreases with increased baryon fraction) and by electron capture of ^{7}Be (increases). Fig. 12 shows the nucleosynthetic products of BBN as a function of baryon density.

1.8.2. *How do we determine primordial densities if D is easily destroyed in stars?*

One way is to look at Ly-α transitions in the neutral, high-redshift ISM. The greater mass of the D nucleus shifts slightly downward (i.e. more negative energy) the energy levels of the electron, creating a slightly shorter Ly-α transition.

1.8.3. *Why is there more matter than antimatter?*

When the temperature of the universe was greater than 150 MeV, quarks would roam free, and photons could pair-produce quarks. The various flavours of quarks were in LTE with each other, and very nearly equal. CP violation, however, produced a $\sim 10^{-9}$ bias in favour of quarks, and when the temperature cooled enough that quark pair production was no longer favourable, the quarks and antiquarks annihilated, producing an enormous photon to baryon ratio, and leaving only quarks. A similar situation occured for leptons.

1.8.4. *What are WIMPs?*

WIMPs (weakly interacting massive particles) are dark matter particle candidates that, due to their small cross-sections, would have stopped interacting with baryons at about the same time as neutrino decoupling. If WIMPs have masses < 1 MeV, they would be ultrarelativistic today and would have the same number density as neutrinos. This gives $\Omega_{\rm WIMP}h^2 \approx \frac{m_{\rm WIMP}}{91.5\,\rm eV}$ (this also applies to neutrinos). The mass of an individual WIMP, then must be < 100 eV. If instead the WIMP is massive, then it is not relativistic, and, then $\Omega_{\rm WIMP}h^2 \approx \left(\frac{m_{\rm WIMP}}{1\,\rm TeV}\right)^2$.

1.9. *Question 8*

QUESTION: Explain how Supernovae (SNe of Type Ia in particular) are used in the measurements of cosmological parameters.

This is adopted from my own qual notes.

Suppose a very bright standard candle exists throughout the history of the universe; since the luminosity of the candle is known, we would be able to use it to measure the luminosity distance (Eqn. 13, using $R = \frac{c}{H_0}\sqrt{|\Omega_\kappa|}$ and sinn to represent sin, sinh, etc.):

$$d_L = (1+z)\frac{c}{H_0}\frac{1}{\sqrt{|\Omega_\kappa|}}\mathrm{sinn}\left(\sqrt{|\Omega_\kappa|}H_0\int_0^z \frac{dz}{H}\right) \tag{34}$$

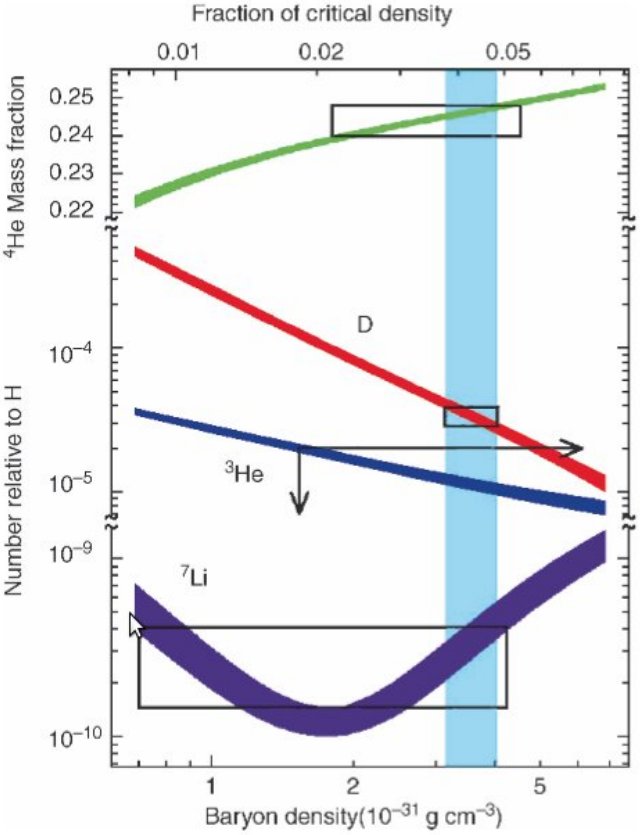

FIG. 12.— BBN predictions of the primordial abundances of light elements as a function of todays baryon density. If a photon density is known, then the x-axis scales with η. Rectangles indicate measured abundance values. Ω_b can be obtained by determining the baryon fraction consistent with measured abundances. From Schneider (2006), his Fig. 4.14.

This allows us to measure the expansion history of the universe.

As a (particularly appropriate) series of examples, let us consider the following. If the universe were empty, $\Omega_\kappa = -1$ and $H = H_0(1+z)$, giving us $d_L = (1+z)\frac{c}{H_0}\sinh\left(\frac{1}{H_0^2}\int_0^z \frac{dz}{1+z}\right)$. Assuming the universe is flat (Sec. 1.3), the luminosity distance reduces to

$$d_L = (1+z)d_c = (1+z)c\int_0^z \frac{dz}{H} \tag{35}$$

If $\Omega_\Lambda = 0$, then$\Omega_m = 1$, which gives $H = 2/3t = H_0(1+z)^{3/2}$. A critical matter-dominated universe has on average smaller d_L than an empty universe. If $\Omega_\Lambda = 1$, then $H = H_0$ and while the comoving distance is much smaller than the curvature radius, d_L will be larger than either the empty or the critical matter-dominated universe.

SNe Ia are standardizable candles: they have a maximum luminosity spread of $\sigma \approx 0.4$ mag in B-band, which is empirically correlated with the rise and decay times of the light curve. This "Phillips Relation" allows for the determination of the true peak absolute magnitude of any SN Ia. (It should be noted that colour is almost just as important as duration; see below.) Since redshift can easily be determined from SNe Ia spectra (Si II P Cygni profile, intermediate element absorption lines, etc.), and SNe Ia are visible out to enormous distances, a plot of luminosity distance vs. redshift, or luminosity distance modulus vs. redshift covering a large portion of the universe's history is feasible.

Fig. 13 shows the result, from Riess et al. (2004). The results strongly rule out the empty and critical universes discussed earlier, in favour of a universe with a non-zero Ω_Λ. These results are in agreement with the ΛCDM concordance cosmology values $\Omega_m = 0.27$ and $\Omega_\Lambda = 0.73$.

1.9.1. *Describe systematic errors.*

The two common methods used to determine the luminosity distance moduli to SNe are the stretch and multi-light curve method (MLCS). The stretch method fits a dataset of SNe Ia to a "stretch parameter" α (which adjusts the light curve based on the fall-off timescale equivalent to δm_{15}) and a luminosity/colour slope β. The MLCS method uses a low-z set of SNe to determine the Phillips and colour relations, and then uses these values for high-z SNe. MLCS also treats intrinsic SNe reddening with extinction from the galaxy independently. Both features can add systematics into the analysis. The two methods, however, give remarkably similar results.

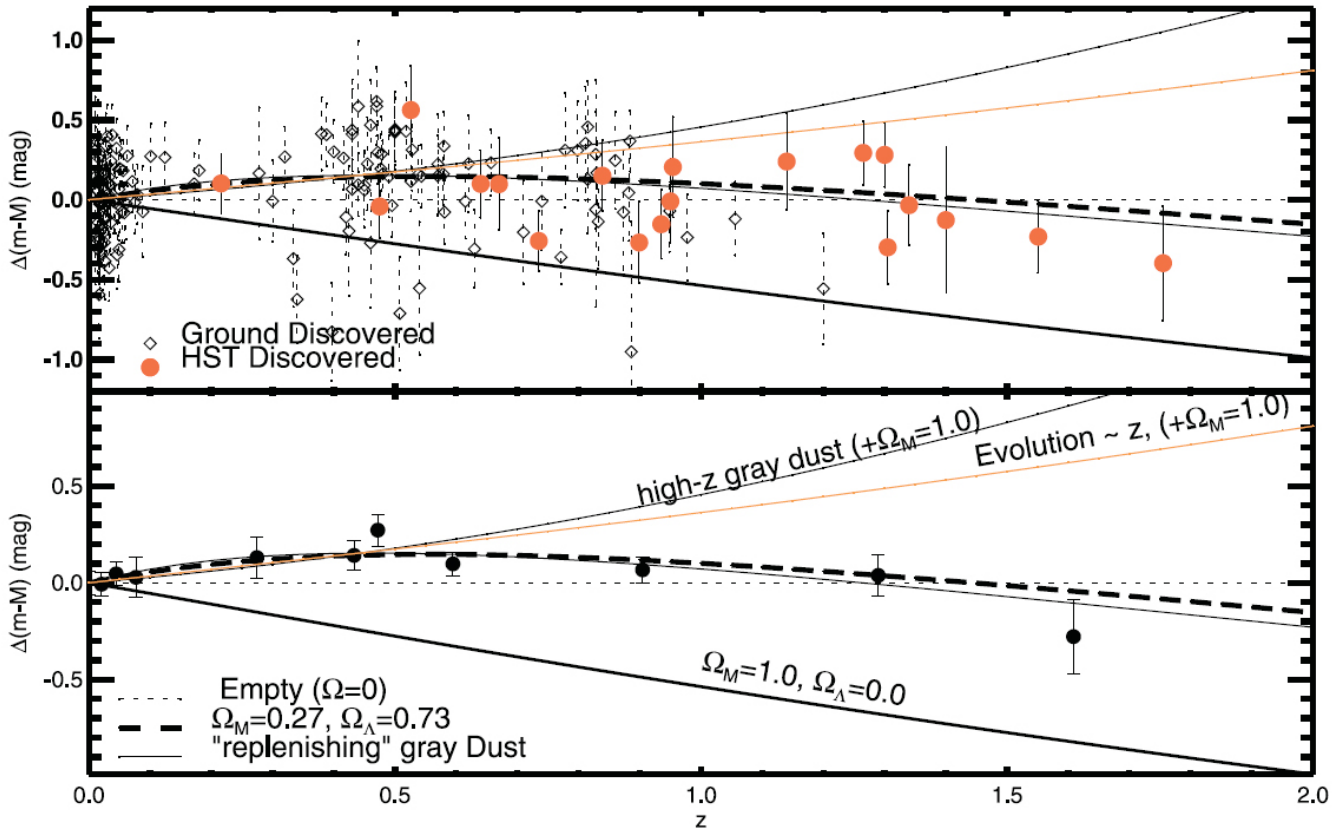

FIG. 13.— Above: the difference in luminosity distance modulus (positive means more distant, since $\Delta(m-M) = 5\log(d_L/10\,\mathrm{pc})$) between observed values from SNe Ia and theoretical values in an empty universe, as a function of redshift. Various cosmological scenarios are also depicted. Below: the same data, except averaged in redshift bins for a cleaner fit. A best fit $\Omega_m = 0.27$, $\Omega_\Lambda = 0.73$ is also drawn. Note that at lower z the observed deviation is positive, indicating a greater luminosity distance (equivalent to reduced flux) than expected for an empty universe. Making Ω_m non-zero only increases $E(z)$ which decreases the luminosity distance, so this increase must be accounted for by a non-zero Ω_Λ. At large z the curve approaches the same slope as the $\Omega_m = 1$ line, indicating matter dominance in the early universe. The "grey dust" lines are for a uniform extinction medium (dust that provides dimming without reddening) - the "replenishing grey dust" assumes a constant dust density even though the universe is expanding. The orange line assumes SNe Ia become dimmer by a percentage monotonically increasing with z (i.e. early universe SNe Ia are much dimmer than modern-day SNe Ia). From Riess et al. (2004), their Fig. 7.

There are a large number of systematic errors. They include:

1. Calibration: k-corrections require good knowledge of both the SED of a typical SN Ia as well as the filter system used.
2. UV spread: SNe Ia have a higher spread in the UV than in optical or infrared (in the IR SNe Ia actually have an uncalibrated σ of 0.15!), which becomes problematic when redshift pulls the U band into the B band.
3. Reddening: intrinsic reddening of the SNe themselves and reddening due to dust should be handled separately, but in practice they are hard to deconvolve. Intrinsically fainter SNe Ia are redder than normal Ias, an effect almost as important as the luminosity/falloff time relationship, but the same effect occurs with dust extinction.
4. Galactic evolution: fainter SNe Ia tend to be embedded in older stellar populations. This translates to a $\sim 12\%$ brightness increase for $z = 1$ SNe Ia due to increased star formation. While this is not a problem if the stretch factor corrects for light curves, it turns out that current fitting methods do not bring SNe Ia in high and low-mass galaxies to the same absolute magnitude (the difference is ~ 0.08 mag).
5. Since the progenitors of SNe Ia are unknown, it is not known if the physical nature of SNe Ia changes with redshift.

1.9.2. *Describe alternate explanations to the SNe luminosity distance data, and why they can be ruled out?*

Alternate scenarios for why the luminosity distance appears to increase for mid-z values include a 20% drop in SNe flux due to a "grey dust" (uniform opacity dust). This does not work, since at higher redshift the obscuration would be more significant, and this trend is not seen in the dataset. The idea that SNe Ia have fundamentally changed from the young universe to today is also difficult to support: the measured luminosity distance approaches what we would expect for a critical matter dominated universe at high redshift (a consequence of Λ becoming prominent only recently). It is not obvious how a change in the nature of SNe Ia could produce the same trend.

1.9.3. *Can SNe II be used as standard candles?*

Yes; in particular the recombination bump of SNe II-P can be used as a standardizable candle (Kasen & Woosley 2009). In these SNe, there is a tight relationship between luminosity and expansion velocity (as measured from ∼5000 ÅFe II absorption lines), explained by the simple behavior of hydrogen recombination in the supernova envelope (Kasen & Woosley 2009). There is sensitivity to progenitor metallicity and mass that could lead to systematic errors, and overall SNe II are dimmer than SNe Ia, however (Kasen & Woosley 2009).

1.10. *Question 9*

QUESTION: Rank the relative ages of the following universes, given an identical current-day Hubble constant for all of them: an accelerating universe, an open universe, a flat universe.

Most of this information comes from Emberson (2012) and Ryden (2003).

In general, the age of the universe can be determined (assuming a is normalized such that $a_0 = 1$) by solving the Friedman-Lemâtre equation, Eqn. 4. Equivalently, we note that $H(z(t)) = \dot{a}/a$, which can be rewritten as

$$dt = \frac{da}{aH} \tag{36}$$

There are, of course, time dependencies on both sides, but since a and z are simply functions of time (if a and z are not positive-definite, then we can parameterize both t and a or z as a function of some parameterization θ) and we can treat z as an independent variable to solve for t by rewriting this equation with Eqn. 9 to obtain:

$$t = \frac{1}{H_0}\int_0^\infty \frac{1}{1+z}\frac{1}{(\Omega_{r,0}(1+z)^4 + \Omega_{m,0}(1+z)^3 + (1-\Omega_0)(1+z)^2 + \Omega_{\Lambda,0})^{1/2}}dz. \tag{37}$$

Needless to say, for non-trivial cosmologies this requires some kind of numerical simulation. (For true order-of-magnitude enthusiasts, however, the right side is ~ 1, giving us the Hubble time.) Let us consider a few more trivial cosmologies.

As noted in Sec. 1.1.4, in a Λ-dominated universe, $H = H_0$, and as a result $a = Ce^{H_0 t}$. Such a universe is eternally old, since a never goes to zero. In an empty, $\Lambda = 0$ universe (which naturally would be open), $H^2 = \frac{c^2}{R^2}\frac{1}{a^2}$, meaning $a = \frac{c}{R}t$, and $H = 1/t$. The age of the universe is then $1/H_0$. In a matter-dominated critical universe, $\kappa = 0$, which results in $a = H_0^{1/3}\left(\frac{3}{2}t\right)$, which gives $H_0 = 2/3t$, and therefore the age is $2/3H_0$. For a flat universe in general, we can assume the energy density goes like $a^{-(1+3w)}$, and we obtain $t = \frac{2}{3H_0(1+w)}$ - a critical radiation-dominated universe has an age $1/2H_0$. The shorter time is due to the fact that in all matter dominated, non-empty universes $H(t)$ decreases over time, and for the same H_0 today $H(t)$ must have been much larger in the past, leading to a shorter amount of time needed to expand the universe to its current size. Open universes will have ages that lie somewhere between $1/H_0$ and $2/3H_0$, while supercritical universes, which cannot easily be calculated (see below), will have ages shorter than $2/3H_0$ because the slowdown of $H(t)$ becomes even more extreme than for a critical universe. See Fig. 15.

Our answer then, is that a flat universe is younger than an open universe, which is younger than an accelerating universe, if we assume the open and flat universes are matter dominated, and the accelerating universe is Λ dominated. If H_0 is fixed, then, in general, if Ω_κ is kept fixed while $Omega_\Lambda$ is increased, the age of the universe increases. If Ω_Λ is kept fixed while increasing Ω_κ, the age decreases. (Ω_m must vary to compensate for fixing one Ω while moving the other.)

From concordance cosmology, the universe is 13.7 Gyr.

1.10.1. *What is the fate of the universe, given some set of Ωs?*

In a matter-dominated universe, where $\Omega_\kappa = 1 - \Omega_m$, we may rewrite the FL equation as

$$\frac{H^2}{H_0^2} = \frac{\Omega_m}{a^3} + \frac{1-\Omega_m}{a^2}. \tag{38}$$

Without solving for anything, we can easily see that if $\Omega_m > 1$ there will be a maximum size to the universe when $H^2 = 0$. This problem actually can be solved analytically (pg. 106 of Ryden) to yield $a(\theta) = \frac{1}{2}\frac{\Omega_m}{\Omega_m - 1}(1 - \cos(theta))$ and $t(\theta) = \frac{1}{2H_0}\frac{\Omega_m}{(\Omega_m-1)^{3/2}}(\theta - \sin\theta)$. The universe, therefore, begins and ends in finite time. Similarly, an analytical solution also exists for $\Omega_m < 1$, though from our previous discussion it is obvious that this universe expands forever. In matter-dominated universes, therefore, matter determines fate.

When Λ is added to the mix, we solve for

$$\frac{H^2}{H_0^2} = \frac{\Omega_m}{a^3} + \frac{1-\Omega_m-\Omega_\Lambda}{a^2} + \Omega_\Lambda. \tag{39}$$

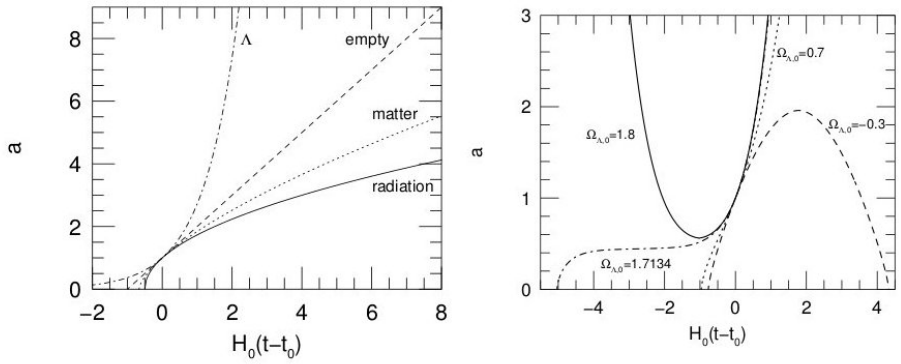

FIG. 14.— Left: a plot of scale factor a as a function of time for a Λ-dominated, empty $\Lambda = 0$, critical matter-dominated and critical radiation-dominated universe. Curves have been scaled so that they all correspond to H_0 today. The point at which $a = 0$ is the age of each universe. Right: the scale factor a as a function of time for universes with $\Omega_m = 0.3$ and varying Ω_Λ. From Ryden (2003), her Figs. 5.2 and 6.4.

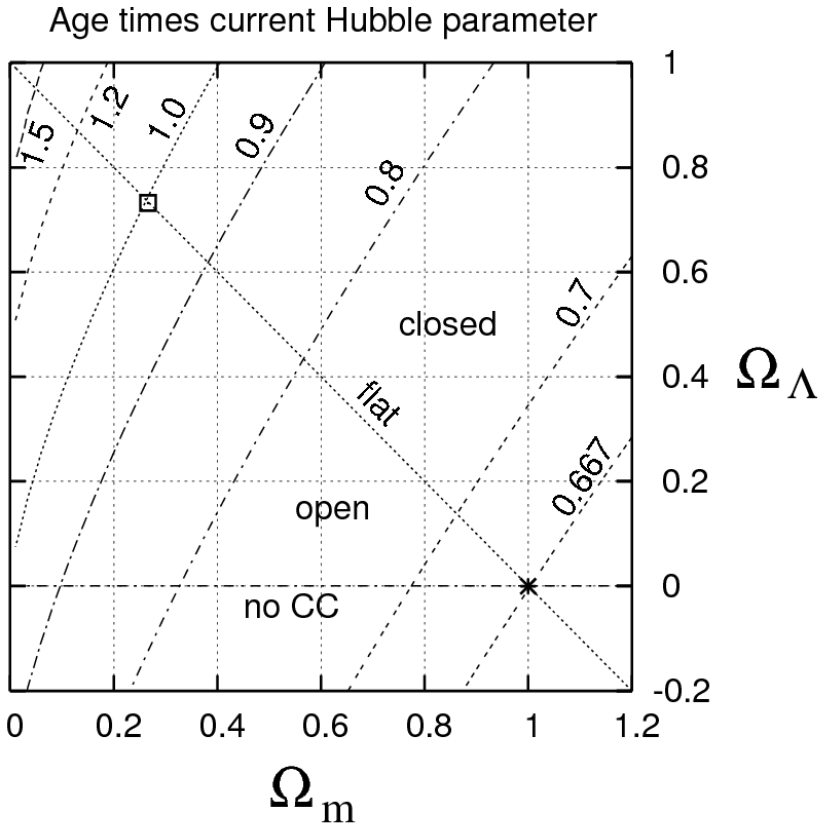

FIG. 15.— Age of the universe, scaled to the concordance cosmology value of 13.7 Gyr.

It is then possible for a closed universe to expand forever. This is because Λ has a constant negative energy density, and the negative pressure caused by the cosmological constant only increases with time. If Ω_Λ is taken to extreme values, it may be impossible for a to drop below a certain value (as H^2/H_0^2 becomes negative) - such a universe must start and end with $a \to \infty$, the "Big Bounce". On the precipice of creating a Big Bounce, a universe can loiter at a fixed a (as curvature attempts to "fight" expansion from Ω_Λ) for long periods of time. Fig. 16 summarizes this discussion.

Concordance ΛCDM suggests the universe will continue to expand forever, approaching exponential expansion.

1.10.2. *How do we determine, observationally, the age of the universe?*

Lower limits on the age of the universe can be determined by looking at the ages of its contents. Globular cluster main sequence turnoff, for example, can be used to date GCs. These generally give $\sim 10 - 13$ Gyr. ^{9}Be, an isotope

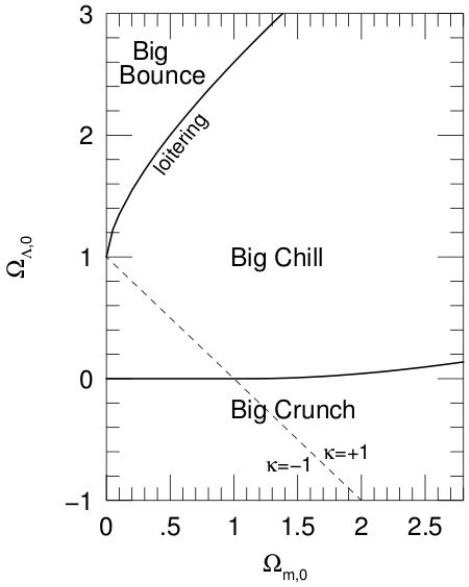

FIG. 16.— A parameter space study of possible fates of the universe given an Ω_m and an Ω_Λ. Curvature is assumed to be $1 - \Omega_m - \Omega_\Lambda$. From Ryden (2003), her Fig. 6.3.

not created by the Big Bang (or fused in stars before main sequence turnoff), is produced through the spallation of heavy elements due to galactic cosmic rays. Over time, the abundance of ^{9}Be increases, and therefore it serves as a clock. Observations of ^{9}Be content in GC MS turnoff stars suggest the MW is 13.6 ± 0.8 Gyr.

1.10.3. *Is* Λ *caused by vacuum energy?*

This information comes from Shaw & Barrow (2011).

The contribution of vacuum energy to the energy-momentum tensor is $-\rho_{\rm vac} g^{\mu\nu}$; i.e. vacuum energy density is fixed (so total vacuum energy scales with volume), and therefore $w = -1$, and Eqn. 8 gives a negative pressure. Formally, the value of $\rho_{\rm vac}$ is actually infinite, but if quantum field theory is only valid up to some energy scale E, then $\rho_{\rm vac} \propto E^4$. We know that QFT could be valid up to supersymmetry breaking (1000 GeV), the electroweak scale (100 GeV) or the Planck scale (10^{18} GeV). The vacuum energy density is therefore anywhere between 10^{12} to 10^{72} GeV^4. Actual measurements of ρ_Λ give 10^{-48} GeV^4, meaning that the energy density of ρ_Λ is approximately 60 to 120 orders of magnitude smaller than the vacuum energy density.

This problem gets worse; see Sec. 1.17.

WTF IS A LENGTH SCALE

1.11. *Question 10*

QUESTION: What are the currently accepted relative fractions of the various components of the matter-energy density of the universe? (i.e., what are the values of the various $Omega_{\rm i}$'s)

The relative fractions (with respect to the critical density $3H_0^2/8\pi G$) of Ω_b (baryon density), Ω_c (dark matter density), Ω_Λ (dark energy density), Ω_r(radiation density), Ω_ν (neutrino density) and Ω_κ (curvature) are:

$$\begin{aligned}
\Omega_b &= 0.0458 \pm 0.0016 \\
\Omega_c &= 0.229 \pm 0.015 \\
\Omega_m &= 0.275 \pm 0.015 \\
\Omega_\Lambda &= 0.725 \pm 0.016 \\
\Omega_r &= 8.5 \times 10^{-5} \\
\Omega_\nu &< 0.0032 \\
\Omega &= 1.000 \pm 0.022 \\
\Omega_\kappa &= 0.000 \pm 0.022
\end{aligned}$$

These values, except for Ω_r, come from Komatsu et al. (2011) and Jarosik et al. (2011) (for Ω_ν), or are calculated from them (in the case of Ω_m, Ω and Ω_κ). In both papers, WMAP 7-year data, alongside BAO and H_0, were used. Ω_r was

calculated from $\Omega_m a_{rm}$, where a_{rm}. Additional parameters, such as SNe Ia measurements of luminosity distance, can help further constrain values such as Ω_κ.

These values are for the relative fractions of various components *today*, and so should all have subscript 0 attached to them (this has been ommitted for clarity). To determine what these values were at any time t, we evolve them in accordance with their equation of state (i.e. $a^{-(1+3w)}$) and divide by the critical density at the time $3H^2/8\pi G$. This is covered in depth in Sec. 1.1.3.

1.11.1. *What are the consequences of these numbers on the nature of the universe?*

The fact that Ω_0 is 1 within error (i.e. $\Omega_\kappa = 0$) indicates that the universe is flat, and therefore adheres to standard Euclidean geometry, and is likely spatially infinite. The current Ω_m and Ω_Λ indicate that the universe was once matter dominated and in the future will be much more Λ-dominated ($\frac{\Omega_m}{\Omega_\Lambda} = \Omega_{m,0}\Omega_{\Lambda,0}\frac{1}{a^3}$). This means that the expansion of the universe was once $\propto t^{2/3}$ and in the far future will approach $\propto e^{Ht}$. From a similar argument, we can find that in the very distant past (since $\Omega_m/\Omega_r = 1/a$) the universe was dominated by radiation. See Fig. 2 for a schematic.

Using the concordance model, we can determine that the co-moving distance to the furthest galaxy within our particle horizon is 16 Gpc (46 Gly). Due to the Hubble sphere receding, this horizon is larger than one would expect if galaxies receding with $v_{\rm pec} > c$ could never be seen (see Sec. 1.7)

1.11.2. *How do we determine Ω_r from the CMB?*

This information comes from Dodelson (2003), pg. 40, but integrating the Planck function is better explained in Hickson (2010), pg. 52.

The current photon energy density from the CMB can be calculated from the energy density of a blackbody:

$$U = \frac{4\pi}{c}\int_0^\infty I_\nu d\nu \tag{40}$$

where I_ν is the Planck Function, $\frac{2h\nu^3}{c^2(e^{h\nu/k_BT}-1)}$. This integration gives $U = \frac{8\pi^5 k_B^4}{15h^3c^3}T^4 = aT^4$. If we input the COBE FIRAS CMB temperature of 2.725 ± 0.002 K and divide by the current critical density of the universe $3H_0/8\pi G$, we obtain $\Omega_r \approx 4.7 \times 10^5$ today.

The present-day galaxy luminosity is 2×10^8 $L_\odot/\mathrm{Mpc}^3$. From this the energy density of radiation from starlight can be estimated, and it turns out to be about $\sim 3\%$ of the CMB radiation. A more detailed estimate gives $\sim 10\%$. A miniscule amount of Ω_b has been turned into Ω_r.

1.11.3. *How are other values empirically determined?*

There are multiple techniques to determine the various Ω_i values empirically. These techniques are complementary, and generally a method of fitting is used to constrain a universe with χ (usually 6) numbers of unknown variables.

Techniques include (Dodelson 2003, Ch. 2.4):

1. There are four established ways of measuring Ω_b. The simplest way is to observe the luminosity of galaxies, and the X-ray luminosity of the hot gas between clusters (which house the majority of baryons in the universe), and convert to a mass. Another is to look at the column density of H from absorption in quasar spectra. The third is to use CMB anisotropy (see Sec. 1.13). Lastly, the matter density (or matter/radiation density ratio, since Ω_r is easily measured) helps determine the outcome of BBN.

2. The most obvious way of determining Ω_m is to use observations sensitive to Ω_b/Ω_m and then use the apparent value of Ω_b to determine Ω_m. For example, the temperature of virialized IGM in a cluster is sensitive to the total mass of the cluster, and can be determined using X-ray observations or the Sunayev-Zel'dovich effect (Sec. 2.18). Counting large clusters and matching the number to the expected number of large halos from N-body simulations also gives an estimate (Schneider 2006, pg. 314). Ω_m can be also be determined through baryon acoustic oscillations (Sec. 1.14) and the cosmic peculiar velocity field (peculiar velocities of galaxies, discernable by redshift surveys, are related to the rate at which overdensities grow, which depends on Ω_m).

3. The cosmic neutrino background remains unobserved. We know that Ω_ν acts much like a radiation field (since neutrinos are relativistic), but they decoupled from baryons just before BBN, when temperatures were still high enough to generate electron-positron pairs. At the end of pair production, the radiation field inherited some of the annihilation energy, while the neutrinos did not. Nevertheless, a detailed calculation (Dodelson 2003, pg. 44 - 46) can be used to show $\Omega_\nu \approx 0.68\Omega_r$ (number from Ryden). The component is also constrained when other Ω values, and the curvature of the universe, are known, hence why WMAP results, combined with HST H_0 and BAO, can place upper limits on it.

4. Ω_Λ can be constrained by observing the expansion history of the universe. Luminosity distances from SNe Ia can rule out to high precision that $\Omega_\Lambda > 0$. A combination of results from very different epochs (ex. CMB and SNe Ia out to $z = 1$) is necessary to properly remove the $\Omega_\kappa/\Omega_\Lambda$ ambiguity (Sec. 1.3).

1.11.4. *What are the six numbers that need to be specified to uniquely identify a ΛCDM universe?*

The six numbers are the (current) baryon fraction, the dark matter fraction, the dark energy fraction (ΛCDM assumes that $w = -1$ for dark energy), the curvature fluctuation amplitude Δ^2_R, the scalar spectral index n_s and the reionization optical depth τ. From these values, all others can be derived.

1.11.5. *Why is h often included in cosmological variables?*

h is the dimensionless Hubble parameter, defined as $H_0/(100$ km/(s Mpc)). It is often included in measurements of cosmological parameters that depend on H_0. Any error in the measurement of the current Hubble parameter can then be removed from the variable in question by including h in the expression (ex. $\Omega_c h^2$ for dark matter density), and relegating errors on H_0 to h.

Current estimates give $h = 0.72$.

1.12. *Question 11*

QUESTION: Outline the history of the Universe. Include the following events: reionization, baryogenesis, formation of the Solar system, nucleosynthesis, star formation, galaxy formation, and recombination.

This short chronology of the universe comes from Wikipedia (2011d), checked with Emberson (2012).

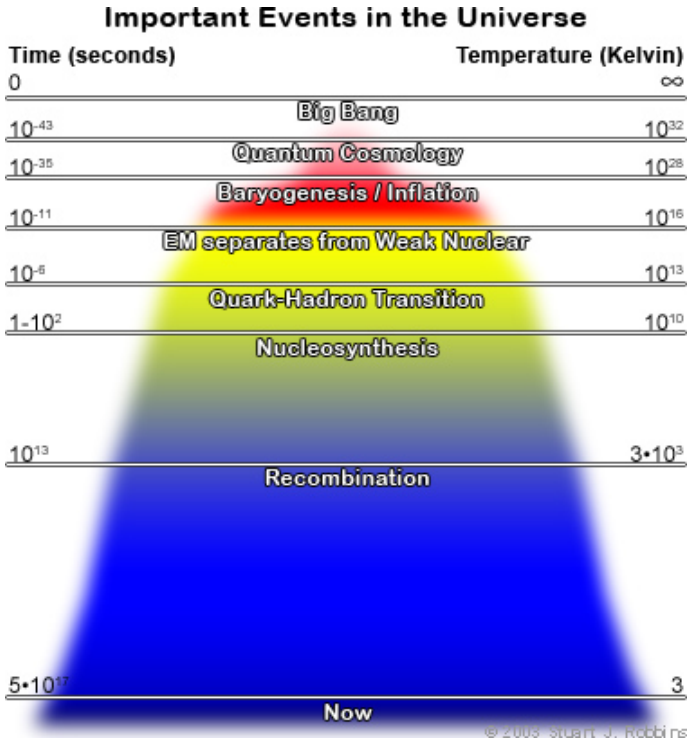

FIG. 17.— The history of the universe, with different cosmological eras, and their corresponding temperature and age ranges, labelled. From Robbins & McDonald (2006).

1. **Planck Epoch** (0 - 10^{-43} s?) - the Planck time is the time it takes light to travel one Planck length, or $t_P = l_P/c = \sqrt{\hbar G/c^5}$, and gives the length of time over which quantum gravity effects are significant. A physical theory unifying gravity and quantum mechanics is needed to describe this era. The four fundamental forces are merged.

2. **Grand Unification Epoch** (10^{-43} - 10^{-36} s?) - gravity separates from electronuclear force, and the universe can now be described by some flavour of grand unified theory (GUT). Separation of the strong nuclear force from the electroweak force at the end of the GUT Epoch produced magnetic monopoles in great quantity.

3. **Inflation** ($10^{-36} - 10^{-34}$ s?) - at 10^{-36} (?) seconds (and 10^{15} GeV) cosmological inflation (Sec. 1.16) is triggered. While the true microphysical cause of inflation is not known, Carroll & Ostlie (2006) describes inflation as a bubble of true vacuum (created via quantum fluctuation) surrounded by a false vacuum with a large energy

density (which can be described by a scalar "inflaton field"). For a cosmological constant-like vacuum energy, $w = -1$, and $P = -\rho c^2$; therefore the bubble of true vacuum expanded dramatically. Due to the high energy density of the false vacuum the universe likely cooled to far below 10^{15} GeV by the end of inflation.

Note that the exact timespan over which inflation occured, and how long after the birth of the universe it occured, is not yet well-understood. Times for the Planck and GUT Epochs are those that would have passed assuming no inflation had occurred.

4. **The Electroweak Epoch** (10^{-34} - 10^{-12} s) - the start of inflation coincided with the separation of the strong nuclear and electroweak forces. When inflation ended, the release of potential energy corresponding to the decay of the inflaton field reheated the universe and saturated it with a quark-gluon plasma. At the end of the Electroweak Epoch (10^{-12} seconds and 100 GeV) the weak nuclear force and electromagnetic force decoupled via the Higgs mechanism. Particles gain mass.

5. **Quark Epoch** (10^{-12} - 10^{-6} s) - massive quarks and leptons mingle with gluons in the quark gluon plasma. Temperatures are still too high for hadron formation.

6. **Hadron Epoch** (10^{-6} - 10^{0} s) - the universe cools to the point where quarks binding together into hadrons is energetically favourable. At 1 second (1 MeV) neutrinos decouple from other particles, creating a cosmic neutrino background analagous to the CMB. At the end of this epoch the universe becomes too cold to create hadrons/anti-hadron pairs, and all remaining pairs annihilate, leaving a small hadron excess due to baryogenesis. At some point before this the universe satisfied the Sakharov conditions for baryogenesis, resuting in the universe being dominated by matter over antimatter. This likely occured much earlier in the Electroweak Epoch. Neutron number becomes set at neutrino decoupling.

7. **Lepton Epoch** (10^{0} - 10^{1} s) - hadrons and anti-hadrons annihilate, leaving lepton and anti-leptons. After ~ 10 seconds the universe cools to below the point at which lepton/anti-lepton pairs are created, and the remaining pairs annihilate (except for the small lepton excess due to baryogenesis).

8. **Photon Epoch** (10^{1} - 3.8×10^{5} yr) - photons dominate the universe, interacting with charged particles. At ~ 3 minutes, nucleosythesis (creation of nuclei other than atomic hydrogen) begins, and lasts for ~ 17 minutes, after which the universe becomes too cold for fusion to take place. At 7×10^{4} yrs the matter and photon energy density equalize, and past this time the universe is matter-dominated. Recombination occurs at 3700 K, resulting in the decoupling between matter and photons and the creation of the cosmic optical background (later redshifted into the CMB) at 3000 K (3.8×10^{3} yr).

9. **Dark Epoch** (3.8×10^{5} yrs - 500 Myr) - the universe is largely dark, except for the cosmic background and 21-cm transition of neutral hydrogen.

10. **Reionization Epoch** (500 Myr - $\sim$ 1 Gyr) - the first stars and galaxies form around 500 Myr ($z \sim 10$) after the Big Bang, initiating the reionization of neutral hydrogen across the universe. Reionization begins in pockets, which grow and merge, until the entire universe is reionized at around 1 Gyr ($z \sim 6$). Once the first Pop III stars die, metals are injected into the universe which permit the formation of Pop II stars and planets.

11. **The Modern Universe** ($\sim$ 1 Gyr - 13.7 Gyr) - our own galaxy's thin disk likely formed about 8 Gyr ago. Our Solar System formed at about 4.5 Gyr ago.

1.12.1. *What are the possible fates the universe?*

These scenarios are obviously more speculative.

- **The Big Freeze** - ($> 10^{5}$ Gyr from today) the universe eventually runs out of new material to form stars, and the last stars die out. Black holes eventually evaporate. Some variants of grand unified theories predict eventual proton decay, in which case planets and black dwarfs will also eventually evaporate. The universe eventually turns into a cold bath of particles approaching thermal equilibrium. Under this scenario the universe may eventually approach "Heat Death" at 10^{141} Gyr.

- **The Big Crunch** - (> 100 Gyr from today) the universe recollapses. Unlikely to occur in the ΛCDM concordance model.

- **The Big Rip** - (> 20 Gyr from today) if the EOS of dark energy had $w < -1$, then the dark energy proper density (i.e. measured in kg/m^3) would increase over time, resulting in a runaway expansion.

- **Vacuum Metastability** - (?) if our current universe contains a scalar field that were to suddenly decay in some region of space through quantum tunnelling (much like what occured during inflation), the result would be an inflation-like event where that region would rapidly expand into the rest of space.

1.13. *Question 12*

QUESTION: Explain how measurements of the angular power spectrum of the cosmic microwave background are used in the determination of cosmological parameters.

This information comes from Ch. 30.2 of Carroll & Ostlie (2006), Ch. 8.6 of Schneider (2006), and of course Emberson (2012).

To first order, the CMB is a perfect blackbody peaked at 2.73 K, or 2 mm, a consequence of inflation (which generated nearly identical initial conditions) and the same evolution since (see below for why the blackbody is perfect). To higher order, this is not true.

Following inflation, the quantum (matter, not just baryon) density fluctuations were expanded to much larger than the particle horizon. As these fluctuations come into the particle horizon, they could react to each other, communicating their overdensities via acoustic waves (since photon propagation was stymied). The photons sloshed with the baryons, and therefore at the time of recombination there were characteristic overdensities of photons as well. Effects ("primary anisotropies") directly pertaining to inhomogeneities at the time of last scattering include:

- **The Sachs-Wolfe effect**: photons are redshifted due to passing out of an overdense dark matter region. They are also time dialated, and therefore do not cool as much as their surroundings. The combined effect, called the Sachs-Wolfe effect, is to make photons cooler in overdensities (compared to the mean) and hotter in the underdensities.
- **Peculiar velocities**: velocities of individual regions of the universe Doppler shift the photons trapped in them.
- **Enhanced baryon density**: in regions of dark matter overdensity baryons are also overdense. On scales larger than the horizon at recombination the distribution of baryons traces the distribution of dark matter. On smaller scales, the pressure of the baryon-photon fluid result in adiabatic heating when compressed.
- **Silk damping**: on small scales photons and baryons are not coupled as there is a finite mean free path of photons (set by time between Thompson scattering events). Temperature fluctuations is smeared out by photon diffusion below ~ 5'.

The first three effects are highly coupled to one another. On scales larger than the sound horizon at recombination the first two effects partly compensate for each other. On scales smaller, dark matter densities drive enhanced baryon densities, while pressure in the baryon-photon soup act as a restoring force, creating forced damped oscillations. These are known as baryon acoustic oscillations (BAO), and are qualitatively described below.

Secondary effects include Thompson (decreases anisotropy) and inverse Compton (Sunayev-Zel'dovich effect) scattering of CMB photons, gravitational lensing of CMB photons, and the "integrated Sachs-Wolfe effect". The last effect is due to photons falling into and out of a potential well. If the potential well did not change over time, the net effect is zero (possibly some lensing), but if the potential did change, the net effect is measureable.

The overall CMB anisotropy is written as $\delta T(\theta,\phi)/T = \frac{T(\theta,\phi)-T}{T}$ (T is the mean temperature):

$$\frac{\delta T(\theta,\phi)}{T} = \sum_{l=1}^{\infty}\sum_{m=-l}^{l} a_{l,m} Y_m^l(\theta,\phi), \tag{41}$$

where $Y_m^l(\theta,\phi)$ is a spherical harmonic. To remove the arbitrary choice of where $\phi = 0$, we average over all m values. We then define:

$$C_l = \frac{1}{2l+1}\sum_{m=-l}^{l} |a_{l,m}|^2. \tag{42}$$

Notice how C_l is positive even if $a_{l,m}$ is negative - this means temperature peaks and troughs all contribute. The angular power spectrum is then written out as $\frac{\Delta_T^2}{T} = l(l+1)C_l/2\pi$ or $\Delta_T = T\sqrt{l(l+1)C_l/2\pi}$, as Δ_T tells us the contribution per logarithmic interval in l to the total temperature fluctuation (Ryden 2003, pg. 200). The result of performing such an analysis using, say, WMAP data, gives us Fig. 18.

Each region of the power spectrum describes some aspect of cosmological evolution. Note that we can convert between l and effective angular size simply through $l \sim 180°/\theta$.

The region far to the left of the first peak represents scales so large as to not be in acoustic causal contact at recombination. The anisotropy at these scales therefore directly reflects the matter fluctuation spectrum $P(k)$. For a Harrison-Zel'dovich spectrum, we expect Δ_T^2 to approximately be constant for $l < 50$.

THERE IS A REGION THAT IS IN GRAVITATIONAL CONTACT BUT NOT ACOUSTIC CONTACT - WHAT HAPPENS TO THOSE?

The first peak is determined by the largest structure that could have been in acoustic causal contact at last scattering. The soundspeed of a relativistic gas, from $\sqrt{P/\rho}$ is $c_s = c/\sqrt{3}$, where c is the speed of light. We estimate the size of

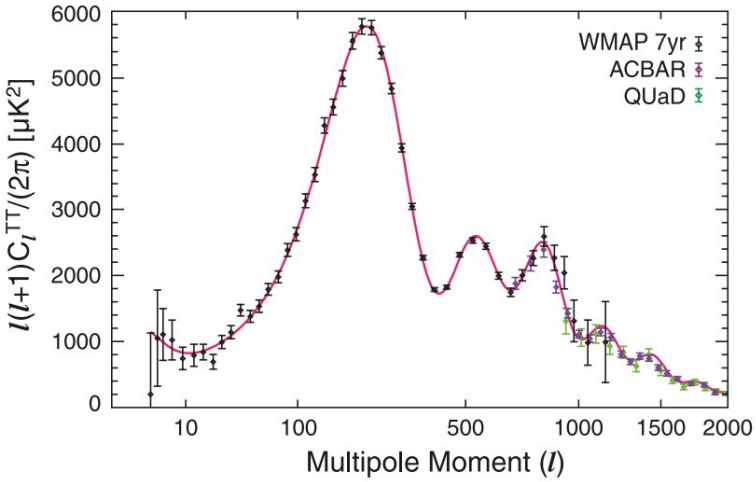

FIG. 18.— WMAP 7 year data CMB power spectrum, alongside data from ground-based ACBAR and QUaD. Note that this is the square of the CMB SED. The pink line is a best fit 6-parameter ΛCDM model. From Komatsu et al. (2011).

the horizon at about 1.8° (see below), which means the angular size of the first peak should be $\theta \sim 1.04°$, or $l \sim 175$. Since the system must react to the data it receives from the horizon, it is sensible that $l = 200$ is the true value (a more detailed calculation can show that $l = 200/\sqrt{\Omega_0}$.

We can model, *a la* pg. 1265 - 1266 of Carroll & Ostlie (2006) baryon-photon soup oscillations as a vertical piston with two regions of material. The upper and lower regions of the piston both start out with the same density, but because gravity pulls down, the lower region is compressed. The difference in densities between the two regions at maximum compression is much greater than the difference in densities between the two regions at maximum rarefactions, since the equilibrium position of the piston is one in which the upper region is less dense than the lower region.

Likewise, in the CMB, the initial distribution of matter over/underdensities was centred around the average density of the universe, not the equilibrium density configuration. As a result, a region that has just come into causal contact will first adiabatically contract. Since the baryon-photon fluid is essentially a trapped glob of photons, $P = \frac{1}{3}U/V$, meaning that radiation pressure will increase dramatically, acting as a restoring force, and the gas eventually expands again. This oscillatory motion changes the temperature of the fluid - hottest at maximum compression, coldest at maximum expansion (modulated by the Doppler effect, which also plays a role in shaping emission from these regions). Since the equilibrium position of the system is a (less extreme) compression, the compression will be stronger than the rarefaction. From all this, we conclude that the first peak in the CMB power spectrum is due to the maximum compression, on the largest possible scale at last scattering, of baryon-photon fluid. When photons decoupled from baryons at last scattering, the CMB retained the imprint of excess temperature at this scale.

The first trough is produced by an area somewhat smaller than the horizon. Its oscillation speed was faster than the first peak oscillation, and therefore it reached $\delta T = 0$ at time of last scattering.

The second peak is generated by rarefaction (recall that C_l included an absolute value!). As discussed earlier, rarefactions are weaker than compression because the equilibrium state is closer to compression, and so the first peak is greater in magnitude than the second peak, and the relative suppression of the second peak increases with Ω_b. This is because increasing Ω_b moves the equilibrium position to a greater compression, "loading down" the plasma (Hu & Dodelson 2002). **WAIT AT SOME POINT WE'D EXPECT BARYONS TO OUTNUMBER DARK MATTER, AND THIS WOULDN'T BE TRUE!** Indeed, all odd peaks are compressions, and all even peaks rarefactions, which accounts for, in our universe, the fact that odd peaks are slightly taller than even peaks.

The third peak, due to compression, is sensitive to the density of dark matter Ω_c. This is because when photons contribute substantially to the universe's overdensities, their dilution during maximum compression due to the expansion of the universe, helps to amplify acoustic oscillations (Hu & Dodelson 2002). Since the low l modes were launched when radiation dominated the most, they are the most amplified by this effect, explaining why the first peak of the CMB is taller than all those that follow it. When Ω_c is increased, the position of matter-radiation equality is shifted, lowering most of the peaks. Increasing Ω_c also increases baryon compression, however, which allows the odd peaks to retain some of their height. The relative height of the third to the second peak, then, helps pin down Ω_c.

For $l > 1000$, Silk damping becomes significant, and the peaks die off.

Fig. 20 shows detailed theoretical calculations of the power spectrum as functions of various cosmological parameters. Ω_0 is related to curvature, which serves to shift the peaks of the CMB (by changing the time of last scattering), and disturb the spectrum on very large scales (the latter is a function of the integrated Sachs-Wolfe effect, which strengthens with curvature). Changing Ω_Λ serves to distort very large scales (because expansion rate modifies the integrated Sachs-Wolfe effect) and shift the peaks of the CMB (because it changes, slightly, the time of last scattering). Increasing Ω_b increases the odd-numbered peaks, and decreases the strength of Silk damping (since it decreases the photon mean free path). Increasing Ω_m dramatically decreases peak amplitudes for reasons described earlier, and slightly increases the l on peaks (from changing the time of last scattering).

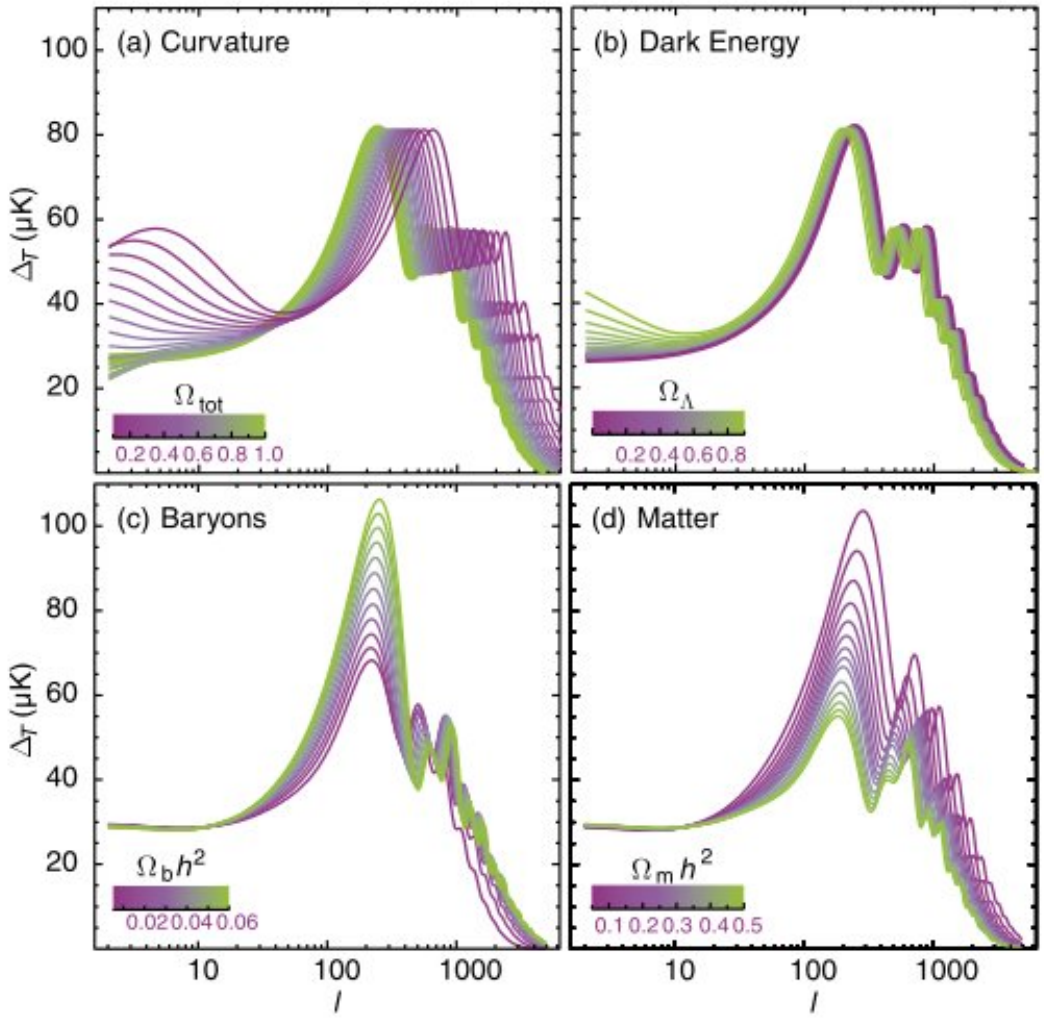

FIG. 19.— CMB power spectrum as a function of cosmological parameters Ω_0, Ω_Λ, Ω_b and Ω_m. From Schneider (2006), his Fig. 8.24.

1.13.1. *How does the CMB power spectrum support the inflation picture?*

In general, a collection of random density perturbations would not be in phase, and we would not see regular peaks in the CMB power spectrum. Because of inflation, the entire universe received the same initial perturbation structure (i.e. statistically homogeneous initial conditions), and oscillations began at a certain scale when the sound horizon reached that scale.

1.13.2. *Derive the horizon size at recombination.*

See pg 170 - 171 of Schneider (2006).

The proper distance of the particle horizon is $r_H = a(t)d_c$. At large z, if we assume matter domination, then $r_H \approx 2\frac{c}{H_0}\frac{1}{\sqrt{(1+z)^3\Omega_m}}$. If we assume radiation domination, then $r_H \approx \frac{c}{H_0\sqrt{\Omega_r}}\frac{1}{(1+z)^2}$. Note that in our universe, matter domination and flatness necessary gives $r_H \approx 2\frac{c}{H_0}\frac{1}{(1+z)^{3/2}}$, which, since $t = \frac{2}{3H_0}a^{3/2}$, gives $r_H = 3ct$.

The angular size of the horizon is given by r_H/d_A, and Eqn. 14 can be modified to read $d_a \approx \frac{c}{H_0}\frac{2}{\Omega_m z}$ for high redshift and matter domination. This gives us $\theta_H \approx \sqrt{\Omega_m}1.8°$, so the size of the horizon is about 1° across the present-day sky. This derivation is modified if Ω_Λ is involved - instead of Ω_m, $\Omega_m + \Omega_\Lambda$ is used - , which gives $\sim 1.8°$ for the horizon size.

1.13.3. *Why is the CMB a perfect blackbody?*

If we assume the universe is homogeneous, then there are no spatial temperature gradients. There is a temporal temperature gradient, which could produce absorption lines (Sec. 4.3), but conveniently $T \propto 1/a \propto 1 + z$ and $\lambda \propto 1 + z$ (and a redshifted blackbody looks like a blackbody), and therefore radiation produced during the entirety of recombination has the same blackbody spectrum.

1.13.4. *How is the CMB measured?*

The CMB was first observed by Arno Penzias and Robert Wilson at Bell Laboratories, using a horn-reflector radio antenna. Modern day observations are done using millimetre observatories situated in very dry regions (water absorption is strong at about 70 GHz). Examples include high altitude balloons such as Boomerang, ground-based observatories such as the South Pole Telescope, and spaceborne platforms such as COBE and WMAP.

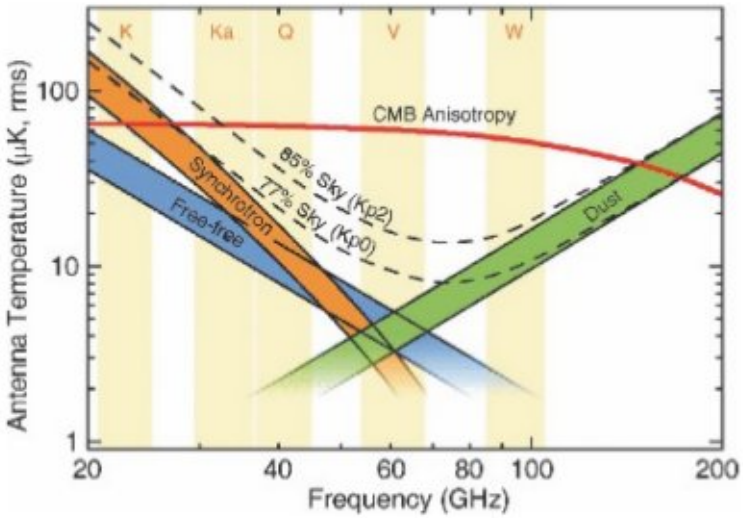

FIG. 20.— The CMB power spectrum (written in "antenna temperature", $\propto I_\nu \nu^{-2}$), with sources of noise superimposed. The five frequency bands of WMAP are also labelled. "Sky" is the summation of free-free, synchrotron and dust emission averaged over a certain fraction of the sky. From Schneider (2006), his Fig. 8.26.

When performing observations of the CMB, foreground sources from our Galaxy, namely synchrotron from relativistic electrons, bremsstrahlung from hot gas, and thermal emission from dust, (Fig. 20) must be removed. There are two ways to do this. The spectral index of synchrotron is -0.8 ($I_\nu \propto \nu^{-0.8}$, for bremsstrahlung it is 0, and for dust it is 3.5. For the CMB it is 2 (Rayleigh-Jeans tail). With this knowledge, a large number of spectral measurements could be made in the microwave, and the spectral indicies of foreground sources (highly different than CMB) could be fit for and eliminated. The other option is to use known tracers of each source of emission at other wavelengths where they dominate (radio for synchrotron, H-α for bremsstrahlung and FIR for dust), and use those measurements (and a model of emission) to remove them from the CMB map.

1.13.5. *Why did people use to think CMB anisotropy would be much larger than it is currently known to be?*

In Sec. **??** we discuss the growth of matter overdensities, and noted that for an Einstein-de Sitter (flat, critical) universe the density $D \propto a$, i.e. density grows with redshift. The average supercluster has $\delta \sim 1$, meaning that their collapse histories are reasonably well-described by $D \propto a \propto 1/(1+z)$. This means that at $z \approx 1100$, $\delta \approx 10^{-3}$. CMB overdensities are of order 10^{-5}, which does not match. The solution is to invoke the fact that baryons, because they have yet to decouple with photons, would have not had as strong overdensities as dark matter, which did have overdensities of 10^{-3}.

1.13.6. *What is the use of CMB polarization?*

Polarization occurs due to Thompson scattering of CMB photons, and is a 1% effect. Since Thompson scattering cannot produce polarization, it is the CMB anisotropy scattering off the surface of last scattering that must have produced it. This polarization will eventually be modified by reionization, and therefore measuring CMB polarization gives information at the redshift that reionization occurred (10.5 according to WMAP).

Electron motion causes different patterns to appear in CMB polarization. E-modes are caused by bulk movement of electrons during baryon-photon fluid oscillation, while B-modes are caused by gravitational radiation compression a ring of electrons. To date, B modes have not been observed.

1.14. *Question 13*

QUESTION: Explain how measurements of bayron-acoustic oscillations can be used in the determination of cosmological parameters.

Most of this information comes from Bassett & Hlozek (2010).

Oscillations of the baryon-photon fluid occur, as noted in Sec. 1.13, because the fluid is attracted to matter-energy overdensities. Material streaming into these overdensities become adiabatically compressed, leading to their expansion. As this expansion is also adiabatic, the material cools as it expands, eventually losing radiation support. It then begins to re-collapse.

This picture describes a parcel of material being oscillated in the plasma. The *wave itself*, however, manifests as an outgoing overdensity of baryons, photons and neutrinos travelling at the sound speed $c_s = 0.577c$.

Consider an overdensity of photons, baryons and dark matter. The photon-baryon fluid has high overpressure and as a result photons and baryons (and neutrinos) launch a sound wave into the surrounding medium. The dark matter mainly stays in place, though is partly gravitationally affected by the outgoing wave (see Sec. 2.2 of Eisenstein et al. 2007 for details). The soundwave travels outward until recombination occurs, at which point the photons and baryons decouple. The photons become free to propagate outward, and become the first peak of the CMB, while the baryons, no longer affected by outward pressure, stall at the sound horizon at recombination, r_H. Over time, gravitational attraction to the central dark matter overdensity drives most baryons into the centre of the original overdensity. Since gravitational attraction works both ways, dark matter also clumps r_H, which then becomes permanently etched in the large scale structure of the universe.

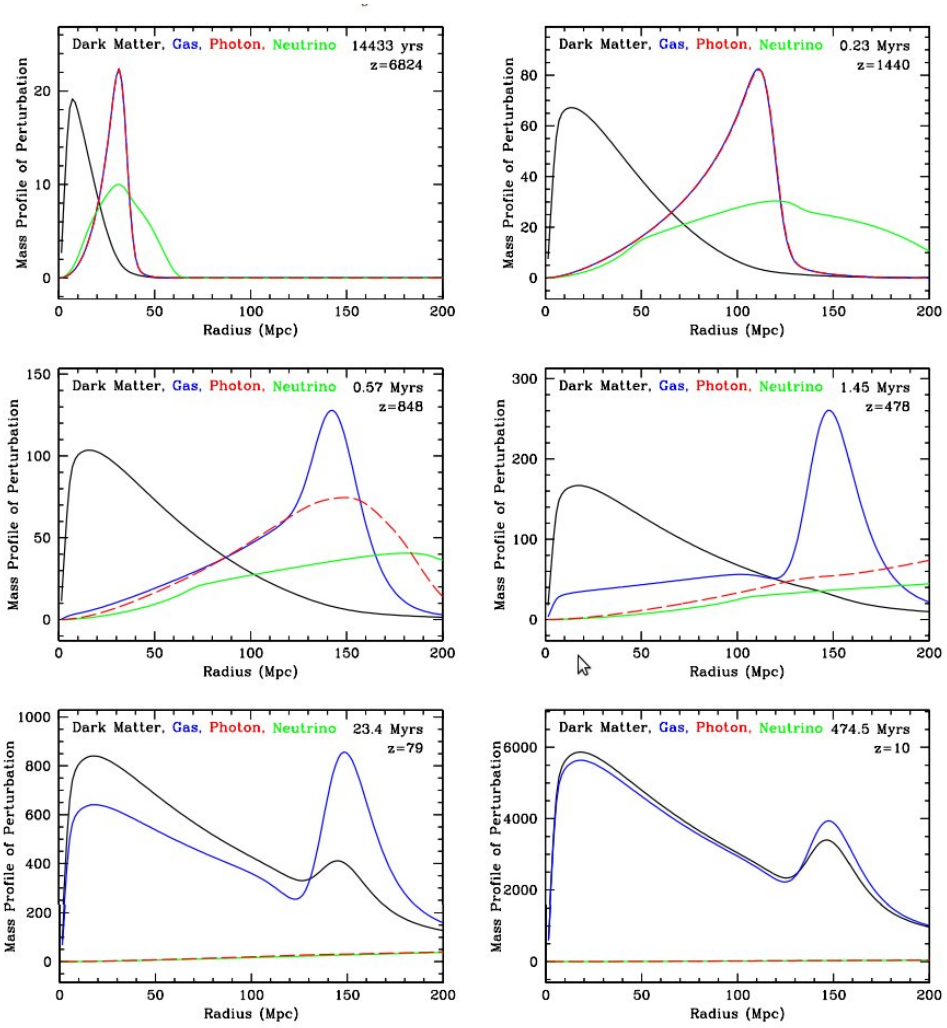

FIG. 21.— A series of snapshots of evolution of an overdensity of point-like initial overdensity which is present in baryons, dark matter, neutrinos and photons. On the y-axis is radial mass profile, and on the x-axis comoving radius. The overpressure of the baryon-photon fluid immediately drives an outgoing sound wave. After recombination the photons and neutrinos stream away, while the baryons, having lost outward pressure support, stall at a characteristic distance from the initial overdensity. Gravitationaly collapse now occurs, and most of the baryons stream back into the dark matter overdensity, but the characteristic bump at large distance is maintained. From Eisenstein et al. (2007), their Fig. 1.

There were many such perturbations in the photon-baryon fluid before recombination, and so we expect many such outgoing waves. Since at high redshift perturbations can adequately be described by linear theory, we can simply add all these waves up to form a large series of outgoing waves overlapping each other. Statistically, then, there would still

be a preferred length, r_H (comoving) at which galaxies prefer to be separated.

The preferred length scale can be measured using the two-point correlation function $\xi(r)$. The correlation function and the power spectrum $P(k)$ are related by the radial Fourier transform $P(k) = \int_{-\infty}^{-\infty} \xi(r) exp(-ikr) r^2 dr$. A δ-function in $\xi(r)$ due to the preferred length scale therefore results in a series of oscillations in $P(k)$. These are what are commonly called the baryon acoustic oscillations, or BAO (see Fig. 23).

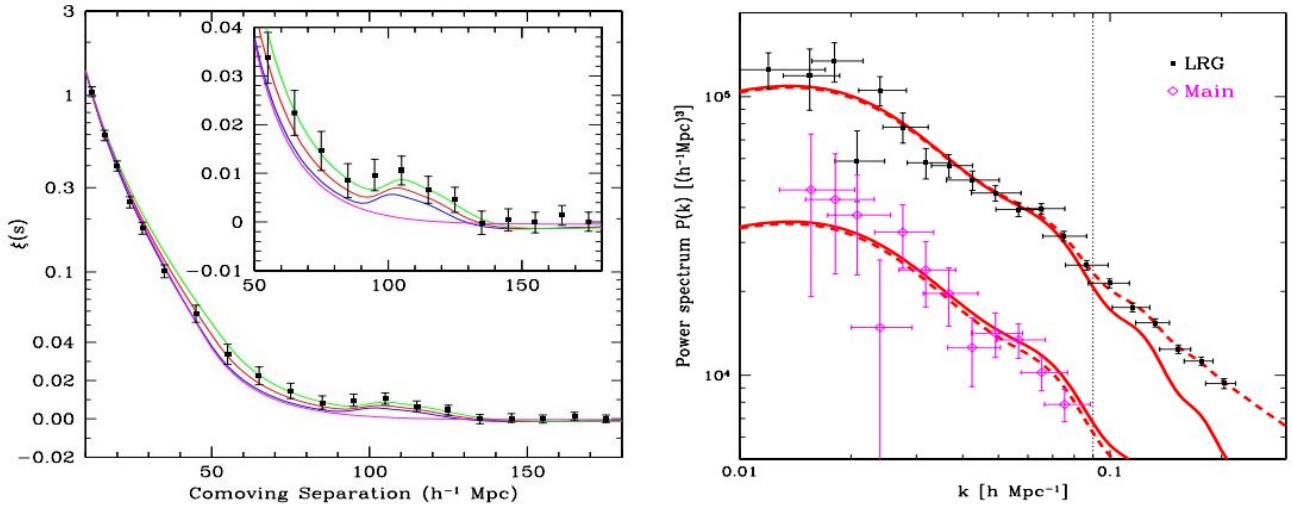

FIG. 22.— Left: the baryon acoustic peak in the two-point correlation spectrum ξ. Points are from the SDSS LRG sample, and solid lines represent models with $\Omega_m h^2 =$ (from top to bottom) 0.12, 0.13, 0.14 and 0 ($Omega_b h^2 = 0.024$ for all). Right: baryon acoustic oscillations in the SDSS power spectrum. Magneta points are the main SDSS sample, and black points are the LRG sample. Lines represent ΛCDM fits to the WMAP3 data, while the dashed lines indicate non-linear corrections. From Bassett & Hlozek (2010), their Figs. 1.1 and 1.2.

Suppose we had the results of galaxy clustering (in Fourier space using spherical coordinates) from a large galaxy survey. If we decompose the results into transverse (θ, ϕ) and radial (r) modes, we can inverse Fourier-transform these modes back into physical space to determine the preferred radial and transverse lengths galaxies cluster at (see Bassett & Hlozek 2010, pg. 12, for complications), represented by Δz and $\Delta\theta$, respectively. As seen in Fig. 23, we can measure

$$H(z) = \frac{c\Delta z}{s_\parallel} \tag{43}$$

and[4]

$$d_A(z) = \frac{s_\perp}{(1+z)\Delta\theta} \tag{44}$$

because we know the co-moving distance $s_\parallel = s_\perp$ from the r_H at last scattering and $z_{\rm recomb} \approx 1100$ from CMB measurements. We can then perform a simultaneous measurement of the Hubble parameter and the angular distance at a given redshift.

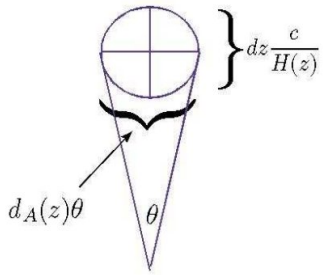

FIG. 23.— A schematic of what cosmological parameters can be measured from a spherical object of known radius. From Bassett & Hlozek (2010), their Fig. 1.6.

[4] Bassett & Hlozek (2010) write $s_\parallel(z)$ instead of $s_\parallel$. I am not sure why.

The Hubble parameter is given by Eqn. 9, and so BAO can constrain the (present-day) density ratios Ω_m, Ω_k, Ω_r and Ω_Λ. Having $H(z)$ is extremely important for determining the behaviour of dark energy over time (i.e. it is not obvious if the equation of state for dark energy does not change over time). Moreover, having both $d_A(z)$ and $H(z)$ allows a much greater constraint on possible cosmologies, as the two values are related to one another. For example, having d_A alone cannot constrain dark energy evolution due to a degeneracy between the dark energy EOS and Ω_k/a^2, but having both d_A and $H(z)$ immediately allows us to determine Ω_k. Theoretically, BAO can also constrain the growth of large-scale structure through changes in amplitude of the power spectrum.

1.14.1. *Why is BAO often used in conjunction with CMB?*

This answer comes verbatim from Emberson (2012).

The complementary probes of the CMB and galaxy clustering observations can be combined to break each others' parameter degeneracies in order to better constrain cosmological parameters. For instance, for a fixed primordial spectrum, increasing DM density shifts the matter power spectrum up to the right while shifting the CMB peaks down to the left. On the other hand, the addition of baryons boosts the amplitude of odd-numbered peaks in the CMB spectrum, but suppresses the power spectrum rightward of its peak as well as increasing its oscillation amplitude. Finally, increasing the abundance of hot dark matter (i.e. neutrinos) suppresses galaxy clustering on small scales while having essentially no effect on the CMB.

1.14.2. *What is the BAO equivalent of higher-l CMB peaks?*

I have no idea.

1.15. *Question 14*

QUESTION: Explain how weak lensing measurements can be used in the determination of cosmological parameters.

This information comes from Schneider (2006), Ch. 6.5 and 8.4.

Weak lensing is when a gravitational lens only moderately distorts the image of a background object. Typically, these background objects are at larger angles to the lens than strongly lensed objects. The distortion, or shear, is sufficiently small that it cannot be distinguished in a single image (since we do not know the true shape of any background object), but since the shear is over a large number of background objects, it can be detected statistically.

The distortion is known as a shear because it reflects the contribution of the tidal forces to the local gravitational field of the lens. The shear results from the derivative of the deflection angle (since the background object is tiny!), and the deflection angle is an integral over the surface density Σ of the lens,

$$\theta(\vec{\xi}) = \frac{4G}{c^2}\int \Sigma(\vec{\xi}')\frac{\vec{\xi}' - \vec{\xi}}{|\vec{\xi}' - \vec{\xi}|^2}d^2\xi \tag{45}$$

where we have projected the geometry of the system onto the sky (i.e. flattening along the radial axis) and are representing the 2D vector position of objects on the sky as $\vec{\xi}$. $\vec{\xi}'$ is the impact parameter vector of the light being lensed. measuring weak lensing, then, gives us a parameter-free method of determining the surface mass density of galaxy clusters, dark matter included. This method can also be used to search for clusters alongside the Sunayev-Zel'dovich effect.

According to Nick Tacik's qualifier notes, microlensing also features "convergence", which is a slight magnification of the background object.

There are a number of uses of weak lensing:

- **Cosmic shear** is microlensing due to the large-scale structure of the universe. This effect is extremely subtle (1% on angular scales of a few arcminutes). Mapping cosmic shear gives us statistical properties about the density inhomogeneities of the universe, much like galaxy surveys do. Indeed, we can determine the two-point correlation function of ellipticities, and relate this to the matter power spectrum $P(k)$. The matter power spectrum is directly related to cosmological parameters (Fig. 24). Microlensing is advantageous because no assumptions need to be made about whether or not dark matter and baryons track each other.

  The most significant result from cosmic shear has been Ω_m combined with the normalization σ_8 of the power spectrum. The two values are almost completely degenerate, and for an assumed $\Omega_m = 0.3$ we can obtain $\sigma_8 \approx 0.8$.

- One obvious cosmological usage of this is to determine the **mass to light ratio of galaxy clusters**, which places constraints on the baryon-dark matter ratio, if reasonable theoretical models for mass-to-light of baryonic objects can be created.

-

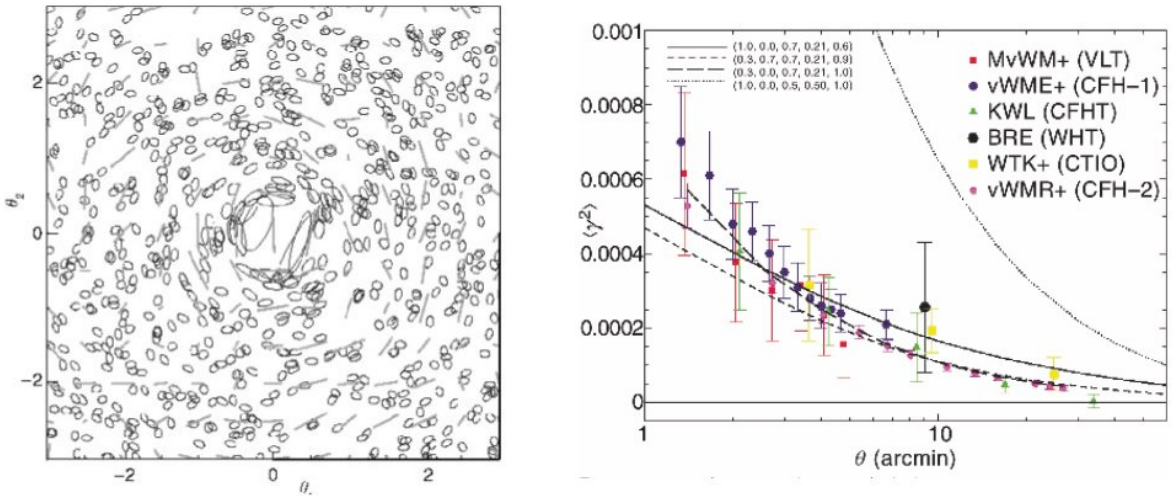

FIG. 24.— Left: computer simulation of gravitational microlensing of distant galaxies. Right: measurement of cosmic shear dispersion as a function of angular scale from multiple teams, overplotted on shear dispersion curves of universes with varying properties (labelled on the upper left; numbers mean Ω_m, Ω_Λ, h, shape parameter Γ and power spectrum normalization σ_8). From Schneider (2006), his Fig. 6.35 and 8.15.

1.15.1. *How is weak lensing measured?*

This information comes from Schneider (2006), Ch. 6.5.

To measure weak lensing we require a large number of well-resolved background objects, meaning we need a deep and wide image. Systematic observations of weak lensing have only become feasible in recent years due to the development of wide-field cameras, improvement of the dome seeing at many telescopes and development of dedicated analysis software.

Because measurement of cosmic shear requires high precision, one major source of error is actually insufficient knowledge of the redshift distribution of background galaxies needed for microlensing. High-redshift galaxy surveys are therefore needed to help reduce errors.

1.15.2. *Can strong lensing be used to determine cosmological parameters?*

Strong lensing tends to generate multiple images (culminating in an Einstein Ring if the source is directly behind the lens), and the photons from each image have different travel times (due to both having to move through a gravitational potential well and due to the geometric differences between different paths). The differences in travel time Δt can be measured because luminosity variations of the source are observed at different times in different images. Δt linearly scales with the current size of the universe, and therefore scales with H_0^{-1} due to the Hubble law. See Fig. 25.

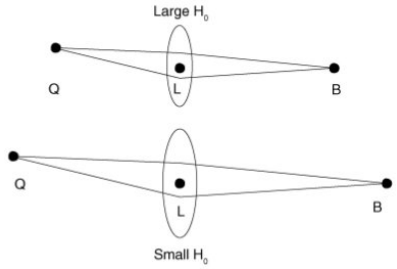

FIG. 25.— Schematic of how Δt can constrain H_0. Above is a large H_0 universe, while below is a small H_0 universe (larger because $v = H_0 d$), with all other observables kept constant. The Δt is larger for the small H_0 universe, and hence is $\propto H_0^{-1}$. From Schneider (2006), his Fig. 3.44.

1.16. *Question 15*

QUESTION: Describe cosmological inflation. List at least three important observations which it is intended to explain.

Most of this solution was from Ryden (2003).
Classical Big Bang theory (i.e. without inflation) has three significant problems

1. **The horizon problem**: the CMB we view is isotropic to one part in 10^5, and since we are just receiving these photons right now, the two "sides" of the universe could not have communicated with each other. In fact, the Hubble sphere at the time of last scattering has a diameter $2c/H(t_{\rm ls}) \approx 0.4$ Mpc, approximately 2 degrees on the sky. There should, therefore, have been no prior communciations between various patches of the sky before last scattering.

2. **The flatness problem**: we may rewrite Eqn. 4 as

$$1 - \Omega(t) = \frac{-\kappa c^2}{R^2 a^2 H^2} = -\frac{H_0^2(1-\Omega_0)}{H^2 a^2} \tag{46}$$

   From Sec. 1.11, we know that $|1 - \Omega_0| \lesssim 0.02$, but Eqn. 46 requires, then, that for earlier times the universe be even flatter (in fact, $1 - \Omega(t) \leq 10^{-60}$ during the Planck era). Classical Big Bang theory does not explain this fine-tuning.

3. **The monopole problem**: various GUT theories predict that as space cooled to below 10^{12} TeV, spontaneous symmetry breaking between the electroweak and strong forces occured. This phase trasition created various topological defects, including pointlike defects (magnetic monopoles), linelike defects (cosmic strings) and so on. These objects would have been highly non-relativistic, and therefore would have begun to dominate the evolution of the universe at around $t = 10^{-16}$ s. This did not occur, and we do not observe monopoles or other topological defects today.

The solution to all three problems is to invoke a short period of exponential a increase very early in the history of the universe. As an example, suppose $H^2 = H_i^2 = \frac{\Lambda c^2}{3}$ during inflation (before and after, the universe is radiation dominated). Then:

$$a(t) = \begin{cases} a_i(t/t_i)^{1/2} & \text{if } t < t_i \\ a_i \exp\left(H_i(t-t_i)\right) & \text{if } t_i < t < t_f \\ a_i \exp\left(N\right)(t/t_f)^{1/2} & \text{if } t > t_f \end{cases} \tag{47}$$

This gives $a(t_f)/a(t_i) = e^N$, where $N = H_i(t_f - t_i)$. Let us also assume $t_f - t_i = 100/H_i$ (i.e. 100 Hubble times); this is a reasonably short time during the GUT Epoch, since $1/H_i \sim 10^{-36}$ s. This means $N = 100$. $e^{100} = 10^{43}$, so length scales in the universe increased by 43 orders of magnitude. This easily allows the last scattering surface to have once been in causal contact[5]. We may perform a more detailed calculation by noting that the particle horizon distance is the proper (co-moving times a) distance travelled by a photon from $t = 0$ to $t = t_i$ or t_f (before or after inflation), and this shows that the horizon increased from 10^{-27} m to 1 pc. The flatness problem is easily addressed as well, since $\frac{1-\Omega(t_f)}{1-\Omega(t_i)} = \frac{\dot{a}(t_i)^2}{\dot{a}(t_f)^2}$. Using Eqn. 47, we see that $\frac{1-\Omega(t_f)}{1-\Omega(t_i)} = e^{-2N}$, meaning that the universe was much, much flatter after inflation, whatever its initial curvature. Lastly, since a increased by 43 orders of magnitude, volume increased by 129 orders of magnitude. As it turns out, we would expect one magnetic monopole for every 10^{61} Mpc^3.

1.16.1. *What caused inflation?*

There is no general consensus as to what caused inflation. The energy density of dark energy today is ~ 0.004 TeV/m^3, while $\rho_\Lambda = \frac{3H_i^2}{8\pi G}$ (since $H^2 = \frac{\Lambda c^2}{3}$ during inflation) gives a ridiculous 10^{105} TeV/m^3.

Suppose, then, there was a scalar field $\phi(\vec{r}, t)$ known as an "inflaton field", associated with a scalar potential $V(\phi)$. From the derivation in (Ryden 2003, pg. 247 - 248), if V changes very slowly, then it is in what is known as a "metastable false vacuum", and as a result $P_\phi \approx -V_\phi$, which, like a cosmological constant, would drive inflation. When ϕ finally changes sufficiently to minimize V, inflation ends. We note that inflation also significantly cools down the universe ($T \propto 1/a$); minimization of V may release enough energy to reheat the universe.

1.16.2. *How does inflation affect the large scale structure of the universe?*

Aside from flattening the universe and homogenizing it, inflation also carried Planck-scale energy fluctuations (10^{-35} m) to macrophysical scales (in our example, 10^8 m). This could be what seeded the initial spectrum of density fluctuations in the universe.

[5] One might wonder why all points were not in causal contact at the Big Bang in any universe. This is because $H = \infty$ when $t = 0$ for universes with a Big Bang.

1.16.3. *Is inflation the only way to explain the three observations above?*

This information is from Wikipedia (2012a).

The horizon problem can be solved by increasing the speed of light. The monopole problem exists primarily because the GUT is well-accepted - since no topological defects have ever been produced in experiment, it may be that they do not exist. Penrose recently found that, from a purely statistical standpoint, there are many more possible initial conditions in the universe which do not have inflation and produce a flat universe, than there are which do have inflation and produce a flat universe, suggesting that using inflation to solve the flatness problem introduces a far larger fine-tuning problem than leaving it be.

1.17. *Question 16*

QUESTION: Define and describe the 'fine tuning problem'. How do anthropic arguments attempt to resolve it?

Fine-tuning refers to circumstances when the parameters of a model must be adjusted very precisely in order to agree with observations (Emberson 2012).

One major cosmological fine tuning problem is the Ω_Λ fine-tuning problem. WMAP 7-year results give $\Omega = 0.725$; if we were to backtrack this to the Planck Era, we would get a remarkably tiny number. As a first order estimate, assume the universe is radiation dominated; then $\Omega_\Lambda = \Omega_{\Lambda,0}\frac{H_0^2}{H^2} = \Omega_{\Lambda,0}\frac{T_0^4}{T^4}$, and at the Planck scale $T_0/T = 1.88\times10^{-32}$ $\Omega_\Lambda \approx 10^{-127}$ (Dodelson 2003, pg. 392).

This is related to the problem elucidated in Sec. **??**, that the vacuum energy density is about 100 orders of magnitude higher the observed ρ_Λ. If QFT is correct, then vacuum energy exists and should greatly accelerate the universe, and the fact that it does not indicates that the true cosmological constant Λ_t is actually negative and cancels with the vacuum energy constant Λ_v (to *60 - 120 decimal places of accuracy*!) to produce the effective Λ_o ("observed Λ") we see (Shaw & Barrow 2011).

A related issue is the coincidence problem, which asks why the timescale $t_{\Lambda_o} \sim \Lambda_o^{-1/2}$ is approximately the age of the universe (or the nuclear timescale of stars), rather than much smaller (Shaw & Barrow 2011). In my view, this is a rewording of the problem, since saying that the Λ_o timescale is of order the nuclear timescale is equivalent to asking why the observed $\Omega_\Lambda = 0.725$ and not nearly 1 at present, which then requires that Ω_Λ be tiny during the Planck Epoch.

The fine-tuning of Λ_o to be nearly zero is problematic because we know of no known physical principle that constrains it (it is a fundamental constant of the universe) except for the vacuum energy density, which it is clearly not equal to. There are a number of possible solutions to this question (from Page (2011)):

1. $\Lambda_t - \Lambda_v = \Lambda_o$ simply by chance. This seems highly unsatisfactory: if we assume Λ_o is completely randomly distributed (or perhaps distributed with a peak near Λ_t) it seems highly unlikely Λ_o would be so close to zero. This claim is impossible to disprove, however.

2. $\Lambda_t - \Lambda_v = \Lambda_o$ (or is highly likely to be Λ_o) for a fundamental physical reason that we have yet to discover. For example, it may be required by the laws of physics that $\Lambda_t - \Lambda_v = 0$, and the nonzero Λ_o comes from a separate physical principle (Shaw & Barrow 2011). While there is no way to disprove this claim either, it becomes more unattractive.

3. There is a multiverse, and $\Lambda_t - \Lambda_v$ equal various values in different universes. We inhabit a universe where Λ_o is nearly zero because if it was even just a few orders of magnitude larger, atoms would disintegrate and life (that we know of) would not form.

4. The universe is fine-tuned so that life (that we know of) will form, and because of this Λ_o is nearly zero.

The cosmological constant fine-tuning problem is one of several fine-tuning problems, as apparently changing any one of the fundamental constants in the universe (those constants that are not constrained by any physical theory, ex. the relative strength of gravity to the other four forces) may lead to wildly different-looking universes. The four options above may apply to any one of them.

The last two options are variations of the anthropic principle, which is a philosophical consideration that observations of the physical universe must be compatible with the conscious life that observes it (Wikipedia 2012a). The fact, therefore, that we observe ourselves living in a fine-tuned universe or an unlikely member of the multiverse is because if the universe were different, we would not exist to live in it.

Indeed, Ω_κ was once among the ranks of fine-tuned universal constants. A small variation in the curvature would be greatly amplified (the flatness problem in Sec. 1.16) so that either the universe would quickly become empty, preventing the formation of large scale structure, or collapse in on itself. This issue was solved by option 2 - inflation was developed as a physically plausible mechanism to create a very flat universe.

1.17.1. *Is the anthropic principle a scientifically or logically valid argument?*

Page claims that one variant of the anthropic principle, that cosmological principles are fine-tuned to maximize the amount/chance of life coming into being, is in principle testable (at least theoretically through modelling different universes).

My issue with the anthropic principle is that it mistakes cause for effect. The fact that we are here to observe the universe requires that $\Lambda_t - \Lambda_v = \Lambda_o$, not the other way around. The reverse would be true only if the existence of the universe requires observers that think like us, which in itself requires a corroborating physical principle (ex. collapse of the "existence" wave function requires a specific "observer" operator that corresponds to physical "intelligence"). If this were the case, the anthropic principle would simply be a stepping stone to a physical principle that sets the coefficients of the universe, and not a final explanation.

1.18. *Question 17*

QUESTION: Define the two-point correlation function. How is it related to the power spectrum? How is the C_l spectrum of the CMB related to low redshift galaxy clustering?

Most of this information comes from Schneider (2006).

Suppose we chose a patch of space dV centred at a point in space $\vec{x}$; we wish to determine the probability that we will find a galaxy in dV. We cannot (due to chaos) actually describe the density structure of the universe except on a statistical level, which in this case means the probability of finding a galaxy in dV centred on $\vec{x}$, averaged over all possible universes with the same statistical properties (this washes out random inhomogeneities; see **?**, pg. 282). If the universe were statistically homogeneous, then this probability would be $P_1 = \bar{n}dV$ (i.e. without gravity, all inhomogeneities are random), where $\bar{n}$ is the average number density of galaxies in the universe. We now consider the probability that a galaxy will be found in dV centred around $\vec{x}$ *and* a galaxy will be found in dV centred around another point $\vec{y}$ (again, in the statistical sense). If galaxy scattering were completely uncorrelated, then we would obtain

$$P = \bar{n}^2 dV^2 \tag{48}$$

i.e. uncorrelated probabilities simply multiply. Since galaxies gravitationally cluster, however, this cannot be the case. We therefore modify the above equation into

$$P = \bar{n}^2 dV^2 (1 + \xi(\vec{x}, \vec{y})) \tag{49}$$

The term $\xi(\vec{x}, \vec{y})$ is known as the two-point correlation function. Correspondingly, the correlation function for matter density $\langle \rho(\vec{x})\rho(\vec{y})\rangle = \bar{\rho}(1 + \xi(\vec{x}, \vec{y}))$. If the universe obeyed the cosmological principle, then $\xi(\vec{x}, \vec{y}) = \xi(r)$, where $r = |\vec{x} - \vec{y}|$.

$\xi(r)$ is related to the power spectrum $P(k)$ by a spherical Fourier transform, as noted in Sec. 1.4:

$$P(k) = \int_{-\infty}^{-\infty} \xi(r) exp(-ikr) r^2 dr \tag{50}$$

(this is equivalent to $P(k) = 2\pi \int_{-\infty}^{-\infty} \xi(r) \frac{\sin(kr)}{kr} r^2 dr$). We recall that this is the method by which we can determine the matter power spectrum of the universe (if dark matter and baryons are coupled). Fig. **??** shows the relationship between the two functions.

We can now see the primary connection between the two-point correlation function and the CMB power spectrum, since there is an intimate connection between the matter power spectrum $P(k)$ and the CMB power spectrum Δ_T. Perhaps most importantly, the first peak of the CMB anisotropy spectrum corresponds to the sound horizon at last scattering. The co-moving preferred length of galaxy clustering as measured from the two-point correlation function, $\sim$140 Mpc, is a relic of the stalled baryon oscillation at the sound horizon at last scattering. In general, how the initial Harrison-Zel'dovich spectrum shifts over time to the current matter power spectrum depends on the expansion history of the universe and the fraction of baryons to dark matter. The expansion history depends on Ω_m, Ω_k and Ω_r, which all affect the CMB. Ω_b/Ω_c changes the amplitude of the BAO, as well as the height of the first peak of the CMB.

The CMB power spectrum has also been modified by secondary anisotropies, many of which are related to galaxy clustering. The integrated Sachs-Wolfe effect, for example, is due to the growth of dark matter halos, which also is what seeds the formation of galaxy clusters. At small angular scales, the image of the CMB is warped by microlensing from particularly concentrated halos. Hot gas in the centres of galaxy clusters inverse Compton-scatter CMB photons (the Sunayev-Zel'dovich effect), which changes the overall thermal structure of the CMB along the line-of-sight to these clusters.

1.18.1. *Describe galaxy surveys.*

This information is from Schneider (2006), Ch. 8.1.

With photometric sky surveys, the two-dimensional distribution of galaxies on the sphere can be mapped. To also determine the third spatial coordinate, it is necessary to measure the redshift of the galaxies using spectroscopy, deriving the distance from the Hubble law (or some model of $a(t)$). Actually performing such (spectroscopic!) surveys is daunting, and was not practical until the advent of CCDs and high-multiplexity spectrographs, which could take spectra of several thousand objects simultaneously. Modern surveys are defined by two parameters: the angular size of the sky covered, and the brightness cutoff of objects being observed.

Major galaxy surveys include the two-degree Field Galaxy Redshift Survey (2dFGRS, or 2dF for short), which covered approximately 1500 square degrees and objects with $B \lesssim 19.5$, and the Sloan Digital Sky Survey (SDSS), which covered a quarter of the sky, with a limiting surface brightness of about 23 magnitudes/arcsec^2 (SDSS 2012).

1.18.2. *What about three or higher point correlation functions?*

This information is from Schneider (2006), pg. 282.

Higher point correlation functions can be defined in the same manner that we have defined the two point correlation function. It can be shown that the statistical properties of a random field are fully specified by the set of all n-point correlations. Observationally, these functions are significantly harder to map, though.

1.19. *Question 18*

QUESTION: Consider a cosmological model including a positive cosmological constant. Show that, in such a model, the expansion factor eventually expands at an exponential rate. Sketch the time dependence of the expansion factor in the currently favoured cosmological model.

This question as been entirely answered in Secs. 1.1.4 and 1.10. Followups can be found there as well.

1.20. *Question 19*

QUESTION: Define and describe the epoch of reionization. What are the observational constraints on it?

This information is from a collection of sources (mostly Schneider (2006)); see my other document for details.

After recombination (3.8×10^5 yrs after the Big Bang) the vast majority of matter in the universe was neutral hydrogen. The epoch of reionization is the period in the universe's history over which the matter in the universe became ionized again. An understanding of reionization is important because of the role reionization plays in large-scale structure formation. The nature of reionization is directly linked to the nature of the reionizing sources, the first stars and active galactic nuclei in the universe (studies could shed light into everything from the early stages of metal enrichment in the uinverse and the clumpiness of the IGM, to the formation of the first supermassive black holes). Moreover, IGM ionization and temperature regulate galaxy formation and evolution.

It is currently believed that the first several generations of stars were the primary produces of photoionizing radiation during the epoch of reionization. The pre-population III star universe was a metal-free environment, meaning that the only methods of cooling involved H and He. All collapsing dark matter halos in the early universe had small masses, corresponding to low virial temperatures on the order of 10^3 K. Only H_2 emission cools baryons at $T_{\rm vir} \approx 10^3$ K efficiently; as a result, the baryon clouds formed $\gtrsim 100$ $M_\odot$ stars. When the baryon clouds corresponding to these small CDM halos collapse to form population III stars, the immediate regions surrounding these halos is ionized by the stars' radiation. Moreover, the radiation dissociates most of the H_2 in the universe (the dissociation energy is 11.3 eV, below the 13.6 eV minimum absorption energy of neutral ground-state H, so H_2-dissociating photons can travel without impedement) and prevent further star formation. After they die, population III stars seed the IGM with metals, which provide cooling for gas at much higher virial temperatures. Therefore baryons associated with larger halos can now collapse to form stars. The greater volume of ionizing photons from these stars begins ionizing more and more of the IGM. Eventually different expanding patches of ionized IGM merge, greatly accelerating reionization (since photos can now freely travel between bubbles). Eventually the last regions of the universe become fully ionized. This process can be seen in cartoon form in Fig. 26.

Because the recombination time for the IGM becomes longer than a Hubble time at z $\sim$ 8, very complicated reionization histories are allowed, and it is unclear exactly how longer reionization took. We know that reionization must occur over a time period of longer than $\Delta z > 0.15$, but it is very likely much longer.

Despite the fact that the universe is now fully ionized, the average density of HII in the universe is too low for the universe to again become opaque; assuming the universe somehow remained ionized *ad infinitum*, it would still have become transparent after 20 Myr.

Observational constraints on reionization include:

- **Ly-α observations** constrain the redshift at which recombination must have ended. Ly-α absorption lines (and, at high column densities, continuum absorption due to the Lyman limit at 912 Å) are created in quasar spectra by

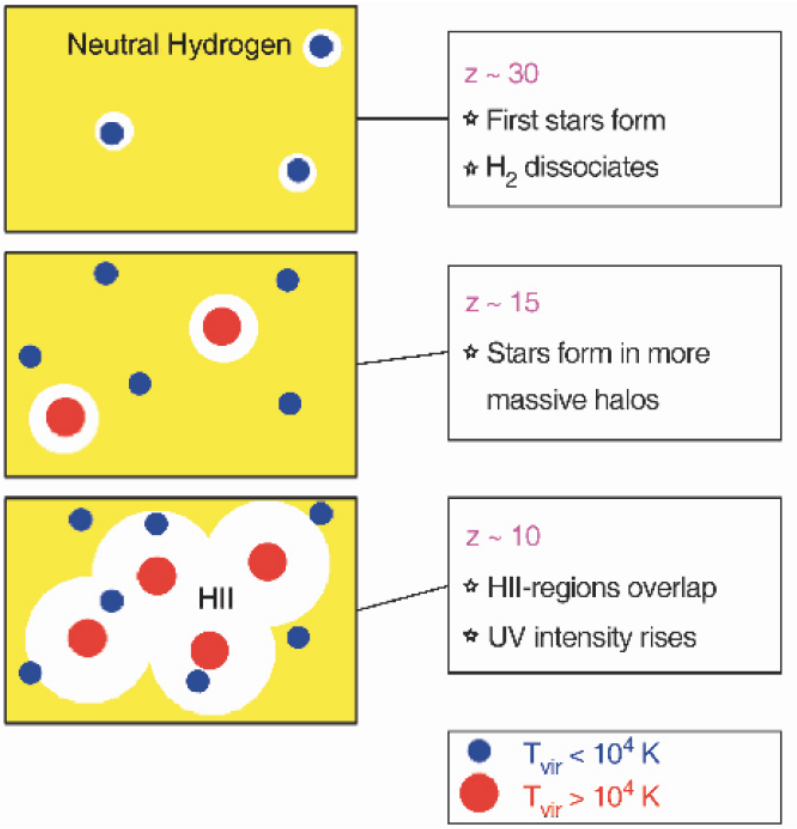

FIG. 26.— A cartoon of the two-step reionization process. At z $\gtrsim$ 12 population III stars form in low-mass halos, ionizing the nearby universe and dissociating H_2 across the universe. When these stars die, they seed the IGM with metals, allowing more metal-rich stars to form in much more massive halos. These metal-rich stars eject many more ionizing photons, and are primarily responsible for reionizing the entire universe. From Schneider (2006), his Fig. 9.30.

the H I floating in the line-of-sight between us and the quasar (neither free-free absorption nor scattering produces absorption lines). The n = 1 to n = 2 transition for a neutral H atom translates to a photon of $\lambda_{Ly\alpha} = 1216$ Å, meaning that for a QSO at $z = z_{\rm QSO}$, we could expect any QSO emission between $\lambda_{Ly\alpha}(1+z_{\rm QSO})^{-1}$ and $\lambda_{Ly\alpha}$ to potentially be absorbed by lines. If there were a large amount of H I at all redshifts from 0 to $z_{\rm QSO}$, the absorption lines would merger together into a trough - the fact that we do not see this at z $<\sim$ 5 (despite there being enough hydrogen density) indicates that the universe is currently ionized (the Gunn-Peterson test); otherwise the various pockets of overdense HI would create a "forest" of absorption lines (Schneider 2006). However, for spectra of QSR at z $\gtrsim$ 6, we see signs of strong Gunn-Peterson troughs, indicating that dense patches of H I still existed at z $\gtrsim$ 6. From this we can set the lower limit for the completion of reionization at z $\sim$ 6. Unfortunately, Ly-α emission cannot penetrate past regions with neutral fractions greater than about 10^{-3}, meaning that it cannot be used to probe the era of reionization itself. See Fig. 27.

- **The proximity effect** is when the QSO ionizes nearby H I clouds, reducing the Ly-α absorption at redshifts close to the emission redshift of the QSO spectrum. The size and shape of this damping can tell us the fraction of hydrogen that is neutral in the space around the QSO (giving us an indication of how far reionization has progressed).

- **CMB polarization** measurements can be used to constrain the start of reionization. H II region electrons have a tendency to Thomson scatter incoming CMB photons. Since this scattering is isotropic, Thomson scattering blends photons from different regions of the CMB together, washing out anisotropies. Thomson scattering also has a natural polarization axis. If in the electron rest frame the CMB were isotropic, no net polarization would be created, but since the CMB itself is anisotropic, CMB radiation can become polarized. The degree of CMB polarization is directly related to the Thomson scattering optical depth, which is related to when a significant amount of H II existed in the universe. WMAP three-year observations of CMB polarization suggest the depth is fairly large, indicating that reionization may have started quite early. Because it is measuring an integrated quantity (degree of polarization along the line of sight) WMAP results cannot easily constrain the period of time over which reionization occured - if it was instant, WMAP gives $z = 11.3$; if it occurred over an extended period of time from z $\sim$ 7 onward, reionization could have begun closer to $z = 15$.

- **21-cm line emission and absorption** from H I could prove an invaluable tool to studying reionization. In

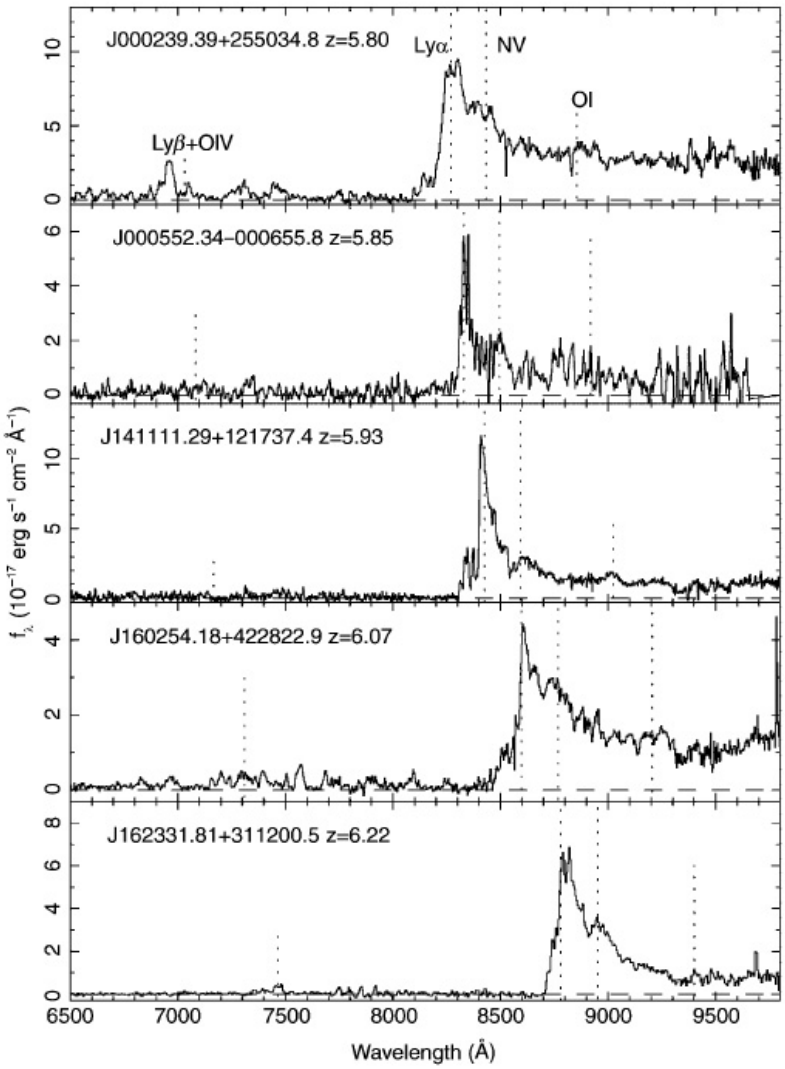

FIG. 27.— UV spectra of QSOs near z = 6. Each spectrum contains a prominent and broad Ly-α emission line (intrinsic to the quasar emission spectrum **(CHECK) THIS**), preceded by a Ly-α forest at lower wavelengths. Near z = 6 the forest is thick, and overall absorption along the line of sight is significant. Note in some cases there is slightly less absorption right near the Ly-α emission line - this is due to the proximity effect, where the QSO ionizes it surrounding IGM (this effect can be used to separate the ionizing flux of photons from the particular QSR and from the background). The forest flattens near z = 6, indicating the creation of significant Gunn-Peterson troughs past that redshift from passing the emission through thick HI regions. This indicates the universe became fully ionized around z = 6. From Schneider (2006), his Fig. 9.28.

the late ($z \lesssim 9$) reionization epoch, the 21-cm line intensity is simply proportional to the ionized fraction, and therefore it is straightforward to translate fluctuations in 21-cm emission to the progression of reionization. While QSO observations are restricted to $z \lesssim 6$ and CMB observations are integrated, 21-cm line emission and absorption could probe to much higher redshifts while maintaining fidelity in both time and space.

1.20.1. *What about He reionization?*

This information is from **?**.

HeII has an ionizing potential energy of 54.4 eV, and the soft emission from population III stars is unable to ionize it (in contrast, population III stars can easily ionize HeI, with an ionizing potential of 24.6 eV) (Furlanetto & Oh 2008). As a result He only fully ionizes when the quasar population is large enough for large numbers of hard quasar-emitted photons to percolate the universe, which occurs at z $\sim$ 3. The most significant observation evidence for reionization completion at z $\sim$ 3 comes from far-UV studies of the He Ly-α forest along lines of sight to bright quasars, which show a significant increase in HeII optical depth near z $\sim$ 3. Other, more controversial evidence also exists (such as a decrease in the HI recombination rate (which corresponds to a change in HI opacity), a change in the hard photon opacity in the universe, and a change in IGM temperature; observations of these effects are not as conclusive as the Ly-α studies). Despite the lack of theoretical studies done on He reionization, it should be noted that the universe at z = 3 is easier to understand and observe than z $\gtrsim$ 6, and therefore the physics of reionization can be constrained by further study of He reionization.

Due to time constraints **all these solutions are based off of Emberson (2012) and last year's qualifier notes**, including figures, unless specifically cited otherwise.

2.1. *Question 1*

QUESTION: Sketch out the Hubble sequence. What physical trends are captured by the classification system?

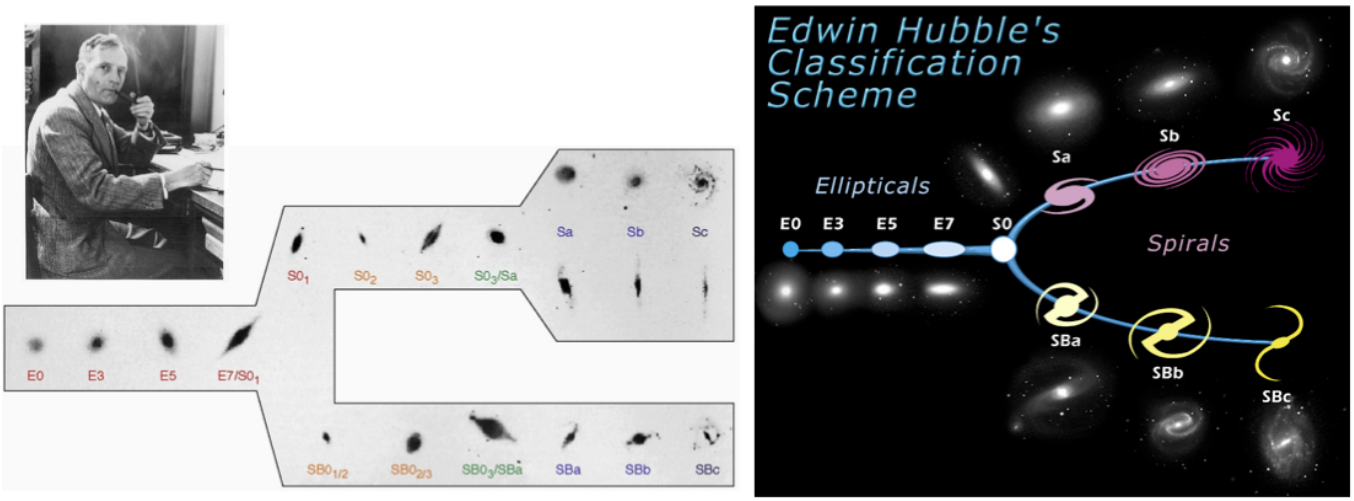

FIG. 28.— The Hubble sequence of galaxy classification. From Emberson (2012).

Hubble produced an empirical morphological classification scheme of galaxies; this is now known as the Hubble sequence (Fig. 28). According to the scheme, these main types of galaxies exists:

- **Elliptical Galaxies**: These have nearly elliptical isophotes, contours along which their surface brightness is constant. They are without any clearly defined structure. They are subdivided according to their ellipticity $\epsilon = 1 - b/a$, where a and b denote the semimajor and the semiminor axes, respectively. Ellipticals are found over a relatively broad range in ellipticity, $0 \leq \epsilon \lesssim 0.7$. The notation E$n$ is commonly used to classify the ellipticals with respect to ϵ, with $n = 10\epsilon$ (e.g., an E4 galaxy has an axis ratio of $b/a = 0.6$, and E0's have circular isophotes. In reality, elliptical galaxies are triaxial in nature, so ϵ is a function of viewing angle. Misaligned isophotes (i.e. isophotes not concentric with one another) is evidence of triaxiality. Ellipticals vary in size from 10 to 1000 kpc, have M_B from -8 to -25, and have masses between 10^7 and 10^{14} $M_\odot$. They are deficit in dust and gas, and do not form many new stars.
- **S0 (Lenticular) Galaxies**: These are a transition between ellipticals and spirals, subdivided into S0 and SB0, depending on whether or not they show a bar. They contain a bulge and a large enveloping region of relatively unstructured brightness which often appears like a disk without spiral arms. They may also have dust, which is parameterized by subscript numbers running from 1 to 3 ($S0_1$ has no dust, $S0_3$ has a lot of dust).
- **Spiral Galaxies**: These consist of a disk with spiral arm structure and a central bulge. They are divided into two subclasses: normal spirals (S's) and barred spirals (SB's). In each of these subclasses, a sequence is defined that is ordered according to the brightness ratio of bulge and disk, and that is denoted by a (a bulge-to-disk luminosity ratio of about 0.3), ab, b, bc, c (bulge-to-disk luminosity ratio of 0.05), cd, d. Spirals vary in size from 5 to 100 kpc, have M_B from -16 to -23, and have masses between 10^9 and 10^{12} $M_\odot$.
- **Irregular Galaxies**: These are galaxies with only weak (Irr I) or no (Irr II) regular structure. In particular, the sequence of spirals is often extended past Sd to the classes Sdm, Sm, Im, and Ir, and the sequence of barred spirals equivalently so (m stands for Magellanic; the Large Magellanic Cloud is of type SBm). Irregulars vary in size from 0.5 to 50 kpc, have M_B from -13 to -18, and have masses between 10^8 and 10^{10} $M_\odot$.

Ellipticals and S0 galaxies are often referred to as early-type galaxies, spirals as late-type galaxies. While Hubble himself believed this to be an evolutionary track (which could make sense if one assumed that spirals are collapsed ellipticals), we now known this is not at all the case.

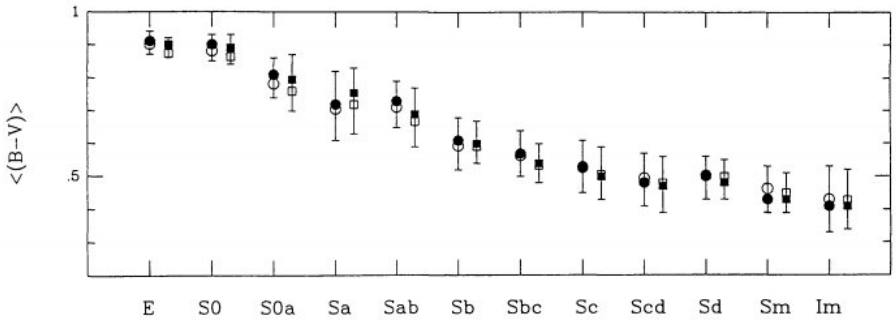

FIG. 29.— $B-V$ colour as a function of galaxy morphology. From Roberts & Haynes (1994), their Fig. 5.

If we look at the various differences in galaxy properties as a function of morphology, like Roberts & Haynes (1994) did, we would find the strongest trend to be colour (Fig. 29), and a trend past Sbc galaxies of later-type spirals generally being less massive, less bright, and slightly smaller. The colour trend broadly describes the average age of stars in the galaxies, which describes their star formation history. Since bulges tend to contain older, redder stars, while disks contain regions of active star formation, the fact that spiral galaxies are bluer than ellipticals is at least partly a function of bulge-to-disk luminosity ratio (it is also partly a function of metallicity, since metallicity makes stars redder). Bulges are also held in equilibrium via velocity dispersion σ rather than bulk rotational velocity v, and therefore later-type galaxies tend to be more v-supported rather than σ-supported. For spirals, later-type galaxies also tend to have less loosely wound spirals.

2.1.1. *What does the Hubble sequence miss?*

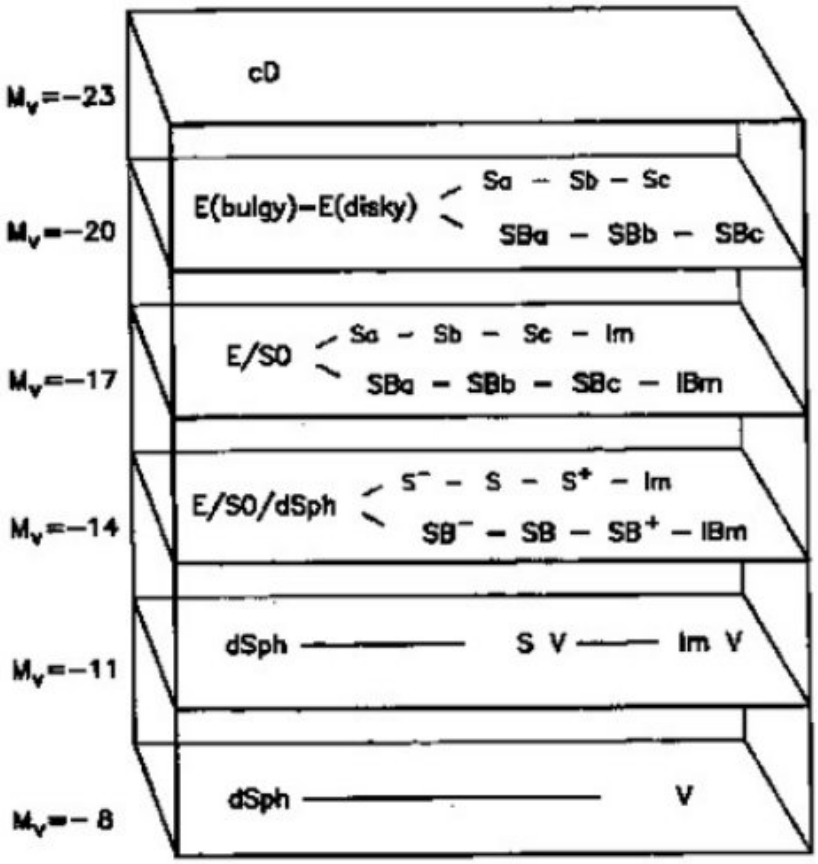

FIG. 30.— The extended Hubble sequence, plotted as a function of both morphology and luminosity. From Abraham (2011c).

One substantial feature missed in a classification system based only on morphology is luminosity (Fig. 30). As an example, elliptical galaxies would encompass

- **Normal ellipticals**: (with a range of M_B from -15 to -23, itself divided into giant ellipticals (gEs), intermediate luminosity ellipticals (Es) and compact ellipticals (cEs).

- **Dwarf ellipticals**: much smaller surface brightnesses and lower metallicity than cEs.
- **cD galaxies**: $M_B \sim 25$ and $R \lesssim 1$ Mpc galaxies only found near the centres of clusters. They appear to be ellipticals (with a de Vacouleurs profile) with an extended exponential envelope.
- **Blue compact dwarf galaxies** (BCDs): small ellipticals with a relatively blue population and significant amounts of gas.
- **Dwarf spheroidals** (dSphs): low luminosity and surface brightness ($M_B \gtrsim -8$).

There are direct correlations between luminosity and other properties of galaxies; see below for the Tully-Fisher relationship and the fundamental plane.

The Hubble sequence also misses out a large number of galaxy types. The classic sequence neglected irregulars, cDs, BCDs, dSphs, and other very small objects, which make up the vast majority of all galaxies (by number, not by mass). Active galactic nuclei (AGN), starburst galaxies (including luminous infrared galaxies, LIRGs, and ultra-luminous infrared galaxies, ULIRGs) are also neglected. Galactic evolution is also neglected: merging galaxies make up about 3% of all galaxies today, and are not accounted for by the Hubble sequence. At high redshift, however, they make up almost half of all galaxies (and the rest look more like irregulars than spirals and ellipticals).

2.1.2. *What is the Tully-Fisher relationship?*

The Tully-Fisher relation is a relationship between a late-type galaxy's luminosity and the maximum velocity of its rotation curve (roughly the the value of v in the flat region of the curve):

$$L \propto v_{\max}^{\alpha} \tag{51}$$

where $\alpha \approx 4$. This scaling can easily be derived by noting that $v_{\max}^2 \propto M/R$ and $L \propto M$ and assuming L/R^2 is roughly a constant. The Tully-Fisher relationship exists so long as the mass-to-light ratios and the surface brightnesses of late-types is roughly constant (the latter is related to Freeman's Law, an empirical statement that the central surface brightnesses of disks have a very low spread amongst sprials). This is of course not always true, and we can see shifts in the Tully-Fisher relationship from Sa to Sc due to changing M/L.

2.1.3. *What is the fundamental plane?*

The Faber-Jackson relation is a relationship between an early-type galaxy's luminosity and its central velocity dispersion σ_0:

$$L \propto \sigma_0^4. \tag{52}$$

This can also easily be derived, using the same assumptions as was used to derive the Tully-Fisher relationship except for replacing the centripetal motion equation with the virial theorem, $T \propto \sigma^2 \propto M/R$. As it turns out, if we actually plotted real early-type galaxies on an L vs. σ_0^4 plot, we would obtain a fairly large dispersion, and the situation is more complicated.

Empirically, $R_e \propto \langle I \rangle_e^{-0.83}$, where the averaging for I is done over the area of galaxy interior of R_e. If we use this additional piece of information, we can find that $L \propto R_e^2 \langle I \rangle_e^{-0.83}$, which gives us $\langle I \rangle_e \propto L^{-1.5}$. Luminosity is also somehow related to σ_0, as we noted earlier with the Faber-Jackson relation. We therefore have a relationship between R_e, $\langle I \rangle_e$ and σ_0, though if we plotted these relationships using real data, we would obtain large dispersions for each.

The true relationship between the three values, derived observationally, is known as the fundamental plane,

$$R_e \propto \sigma_0^{1.4} \langle I \rangle_e^{-0.85}, \tag{53}$$

and can be modelled if we assume that the mass-to-light ratio for early-type galaxies increases slightly with mass.

2.1.4. *What other criteria could be used for galaxy classification?*

Besides morphological criteria, colour indices, spectroscopic parameters (based on emission or absorption lines), the broadband spectral distribution (galaxies with/without radio and/or X-ray emission), as well as other features may also be used for galaxy classification.

2.2. *Question 2*

QUESTION: What is the total mass (in both dark matter and in stars) of the Milky Way galaxy? How does this compare to M31 and to the LMC? How is this mass determined?

From Binney & Tremaine (2008), pg. 18, the mass of the various components of the MW are: $2^{+3}{-}1.8 \times 10^{12}$ $M_\odot$ for the dark matter halo, $4.5 \pm 0.5 \times 10^{10}$ $M_\odot$ for the disk, and $4.5 \pm 1.5 \times 10^9$ $M_\odot$ for the bulge, with approximately

1/3 of the baryonic mass in gas. The dark matter dominates, by far, and its error is enough to wash out all other contributions to the galactic mass. The current consensus is that the Andromeda Galaxy (M31) has roughly the same, with M31 probably the slightly more massive of the two. The mass of the LMC is approximately $1 - 2 \times 10^{10}$ $M_\odot$, which includes its dark matter halo.

For massive nearby galaxies, as well as our own, the total mass is constrained within the first few tens of kpc using the Doppler-shifted 21-cm radio emission of the H I disk to determine rotational velocities. If we estimate the entire system, including the dark matter, as being axisymmetric out to the furthest extent of H I, then the interior mass $M(r)$ can be determined by determining the H I velocity at r

$$v(r)^2 = \frac{GM(r)}{r} \tag{54}$$

Past the edge of the H I disk, kinematics of particularly bright objects, such as planetary nebulae, globular clusters and satellite galaxies are used instead. Unfortunately, the uncertainties in such techniques are plagued by low sample sizes as well as the fact that there is seldom knowledge of the proper motions of the objects to complement their observed radial velocities and distances (because with low sample size higher order velocity moments cannot be determined). Because of the latter, assumptions must be made on the eccentricities of the satellite orbits thereby affecting the mass determination. (This is an incarnation of the "mass-anisotropy" degeneracy, where highly elliptical orbits will shift mass estimates.) As a sketch of how such an estimate is made, let us assume that, over the sampling of objects used, the total enclosed mass does not change significantly. Then, since $GM = v^2 r$, the mass can be estimated by

$$M = \frac{C}{G}\frac{1}{N}\sum_i^N v_{r,i}^2 r_i. \tag{55}$$

$v_{r,i}$ is the radial velocity of object i, $C \approx \frac{32}{\pi(3-2\langle e^2\rangle)}$ is a corrective term that accounts for the fact that $v_r^2 r/G$ samples the projected mass and not the true mass, and $\langle e^2 \rangle$ can be calculated if the population distribution of eccentricities is known. This expression converges to the true mass as $1/\sqrt{N}$. This method was used to constrain $M_{\rm MW} \approx 3 \times 10^{12}$ $M_\odot$ and $M_{\rm M31} \approx 1 \times 10^{12}$ $M_\odot$ using satellite galaxies.

To get the mass of the non-dark matter components of a galaxy, we can use photometry and spectroscopy. Photometry is used to determine the total luminosity of the galaxy, while spectroscopy is used to determine the stellar population, and therefore give a theoretical estimate for the mass-to-light ratio. A stars-to-gas ratio must also be determined.

These methods can also be used with smaller galaxies like the LMC, but as an irregular galaxy significantly affected by the MW's gravity the LMC does not have a very well-defined rotation curve. Another method of determining the mass of satellite galaxies is its tidal radius: since the LMC is gravitationally affected by the MW, past a certain distance from the "centre" of the LMC, stars and gas that originally belonged to the LMC would have been cannibalized by the MW. This distance can be determined by

$$r_t = R_{\rm LMC}\left(\frac{M_{\rm LMC}}{2M_{\rm MW}(R_{\rm LMC})}\right)^{1/3} \tag{56}$$

where $R_{\rm LMC}$ is the distance to the LMC from the MW, about 50 kpc, and $M_{\rm MW}(R_{\rm LMC})$ is the mass enclosed within $R_{\rm LMC}$. The tidal radius, estimated from the distribution of stars associated with the LMC, is ~ 10 kpc, and from the above equation we can determine $M_{\rm LMC} \approx 0.01 M_{\rm MW}$. Note that we have assumed sphericity of both galaxy potential wells, which is likely untrue.

2.2.1. *Why can't we simply use the inner 10 kpc data derived from 21-cm line emission and a model of the halo to determine the mass of the halo?*

While an NFW halo[6] or isothermal sphere is often assumed as the distribution of dark matter, as we have no real understanding of the nature of dark matter it would be foolish to assume that we do not require observations at large distance to more fully constrain the halo. Moreover, in lenticular and elliptical galaxies, the inner 10 kpc is often dominated by stellar mass, making it difficult to obtain good estimates of the dark matter halo properties.

2.2.2. *How is the total mass of an elliptical galaxy determined?*

This answer is mainly from my own work with Anne-Marie.

The 21-cm line is often non-existent in early-type galaxies, and satellite galaxy/PNe kinematics have the same low S/N issues as they do when used on the MW or Andromeda. As hinted at by Eqn. 55, the eccentricity and orbital orientations of a population of objects can change the enclosed mass estimate substantially. To obtain this information, higher-order terms in the line-of-sight velocity distribution (v_r and σ_r being the first two moments, skew and kurtosis being the third and fourth) are required. These cannot easily be obtained through satellite/PNe kinematics (due to

[6] Note that the NFW halo profile gives infinite mass if integrated out to infinity. A cutoff (usually the radius at which density drops to 200 times the critical density of the universe) is generally imposed to obtain a mass.

low S/N), but they can be using stellar kinematics. Early-type galaxies have their matter dominated by stars out to at least 1 R_e, and have low surface brightness outside 1 R_e. To probe the dark matter halo, therefore, requires either extensive long-slit spectroscopy or the use of an integral-field spectrograph (and tons of binning). For details, see Weijmans et al. (2009).

2.3. *Question 3*

QUESTION: How do we know that the intergalactic medium is ionized?

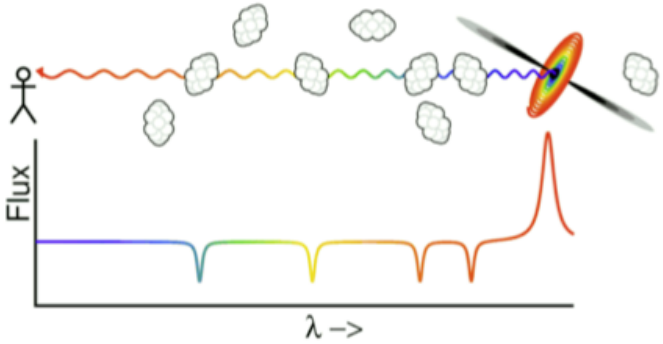

FIG. 31.— A cartoon of how pockets of neutral hydrogen absorb Ly-α from a quasar. From Emberson (2012).

This question is intimiately related to Sec. 1.20 and Sec. 2.14, and indeed are partially answered in those questions. We know the universe's hydrogen is ionized for two primary reasons:

- **Lack of Ly-α absorption**: the spectra of high-redshift quasars always display a large number of narrow absorption lines superimposed on the quasar's continuous spectrum; these lines are in addition to any broad absorption lines that are associated with the quasar itself. These narrow lines are formed when the light form a quasar passes through clouds of material in the line of sight between us and the quasar (Fig. 31). The reason why a forest of lines appears is that as the light propagates the entire spectrum, along with any absorption lines formed, redshift. Each absorption line corresponds to the redshifted Ly-α wavelength of an intervening clouds.

  Past $z = 6$, we see the development of a Gunn-Peterson trough, indicating that all light above about 1216 Åis being absorbed by intervening clouds. This indicates that a fraction of the IGM is neutral past $z = 6$. Because the Ly-α cross-section is enormous, an extremely low column density ($n_{HI} \gtrsim 10^{-13}$ cm^{-3}[7] is already sufficient to generate a Gunn-Peterson trough. We do not see this for nearby quasars, indicating that almost all hydrogen in the universe is ionized.

- **Temperature of the IGM**: the temperature of the IGM today is between 10^5 and 10^7 K. This requires that H I be almost completely ionized.

We can also perform an order of magnitude theoretical estimate, *a la* Jeffrey Fung. There are 10^{11} galaxies in the observable universe, with on average 10^9 stars in them, and let us assume 0.1% of these are O and B stars. Their luminosity is about 10^3 $L_\odot$ and their emission peaks in the UV, so we can assume all their 2×10^{47} photons generated per second ionize hydrogen. This is a total ionizing photon flux of 2×10^{63} s^{-1}. There are 10^{80} H-atoms in the universe, so it takes 1.6 Gyr to completely ionize the universe.

2.3.1. *Is He and metals ionized? Is any component of the IGM neutral?*

Absorption of quasar emission by He II and partly and totally ionized metals, primarily C and Mg, together with Si, Fe, Al, N, and O. The mix of elements is similar to that found in the interstellar medium of the MW, indicating that the material has been processed through stars and enriched in heavy elements. These lines are though to be formed in the extended halos or disks of galaxies found along the line of sight to the quasar.

From a theoretical standpoint, it takes $2.176 \times 10^{-18}\, Z^2$ J to completely ionize an element with atomic number Z. Setting $k_B T = 2.176 \times 10^{-18}\, Z^2$, and $T = 10^7$ K, we obtain $Z \approx 8$. This suggests that heavy elements such as Fe or U are not completely ionized.

[7] This can be determined using $\tau = 4 \times 10^{10} h^{-1} \frac{n_{HI}(z)}{(1+z)\sqrt{1+\Omega_m z}}$, where n_{HI} is in units of cm^{-3}. We set $\tau = 1$.

2.3.2. *What are the properties of these neutral H clouds? Can they form galaxies?*

We deduce the size of the intergalactic clouds by comparing the Ly-α forest in the spectra of pairs of lensed quasars. Many of the absorption lines are seen in both spectra, but some are not. This indicates that the clouds are, on average, about the size of the lensing galaxy. From the total calculated column density of hydrogen (ionized plus neutral), the mass of atypical cloud is somewhere around 10^7 $M_\odot$. At the temperate estimated for a typical cloud ($T \sim 3 \times 10^4$ K, its self-gravity would be too weak to keep it from dispersing. It may be held together by the pressure of less dense (but hotter) external IGM or by the presence of DM within the cloud.

The clouds are placed into three categories. The low column density Ly-α forest absorbers ($\Sigma_{HI} < 10^{16}$ cm^{-2}) are associated with the diffuse IGM. These systems probe low-density, highly ionized gas and are thought to trace the dark matter distribution throughout the IGM as well as contain the bulk of the baryons at high redshift and a significant amount of the baryons even today. At the other end, the high column density damped Ly-α absorbers (DLAs, $\Sigma_{HI} > 10^{20}$ cm^{-2}) appear associated with the main bodies of galaxies. These high-density, predominantly neutral systems serve as neutral gas reservoirs for high redshift star formation. The intermediate column density systems, known as Lyman Limit Systems, mark the transition from the optically thin Ly-α forest to the optically thick absorbers found in and around the extended regions of galaxies. Typically these absorbers are easy to identify in QSO spectra due to the characteristic attenuation of QSO flux by the Lyman limit at ~ 912 Å in the rest frame. In addition, they are optically thick enough to be harbouring neutral hydrogen cores.

2.4. *Question 4*

QUESTION: Describe as many steps of the distance ladder and the involved techniques as you can. What are the rough distances to the Magellanic Clouds, Andromeda, and the Virgo Cluster?

This is mostly from my own notes.

The cosmic distance ladder consists of a large number of means to determine (luminosity, for the most part) distances to objects. They are listed below, and extragalactic ones in Fig. 32. The differing techniques are generally calibrated to one another (to provide consistency), and therefore it is highly advantageous for differing rungs of the ladders to overlap.

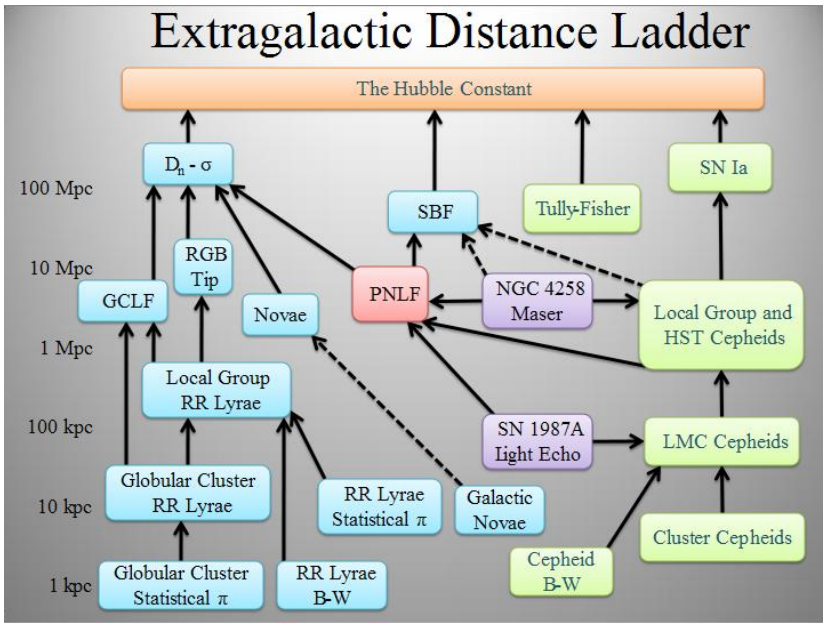

FIG. 32.— Plot of various extragalactic distance-determining techniques, with their maximum effective luminosity distance measures located to the left. The various colours represent applicability to different classes of object (but can effectively be ignored). Solid black lines indicate good calibration between two steps on the ladder; dashed lines indicate poor calibration. "GCLF" is the globular cluster luminosity function, "PNLF" is planetary nebula luminosity function, and "SBF" is surface brightness function. From Wikipedia (2011a).

Solely Galactic distance determining techniques:

- **Geometric parallax** is the measurement of the position of a distant objects from two different observation points separated by some physical distance called a "baseline". The baseline divided by the angle by which the objects shifts when moving between the two points of observation gives a distance estimate. The technique is usually only applicable to nearby stars (the Gaia satellite will change this). Commonly the baseline used is the

diameter of the Earth's orbit around the Sun, but the proper motion of the Sun can be used to create longer baselines.

- **The moving cluster method** uses the proper motion of cluster members, as well as cluster member radial velocities, to determine distance to the cluster.

- **Dynamical parallax** uses measurements of the period (or a related orbital value) and angular semi-major axis of binary stars to determine their distance. By plugging in the period, angular semi-major axis and a mass-luminosity relation into Kepler's Third Law, one obtains an equation solvable for the distance to the binary.

- **Main sequence fitting*** of stars in a cluster onto an HR diagram, and comparison of that fit with a similar fit to the Hyades cluster, can be used to determine open cluster distances up to 7 kpc.

- **Spectroscopic parallax** is the determination of the luminosity of a star from a combination of spectral class and line-broadening (the 2-dimensional Morgan-Keenan luminosity class). Combined with the apparent magnitude, this can be used to determine a distance modulus to the star. Technically spectroscopic parallax is useful up to 10 Mpc, but in practice it is only used up to 10^5 pc.

- **Expansion velocities of SNe ejecta** proper motion measurements, combined with measurements of radial velocity from Doppler shifts and the assumption that expansion is spherically symmetric can be combined to obtain distances.

Extragalactic distance determining techniques (some can also be used in our own Galaxy):

- **The Wilson-Bappu effect** (up to 0.1 Mpc) is a spectroscopic parallax effect where the width of a small emission line housed in the K absorption line of calcium is correlated to absolute magnitude for G, K and M stars. Calibrated to geometric parallax, the Wilson-Bappu effect is mostly used for distances up to the LMC.

- **Bright red supergiants** (up to 7 Mpc) which appear to have the same absolute V and bolometric magnitude. Requires distingishing individual stars, giving it the same range as spectroscopic parallax.

- **Tip of the RGB** (up to 7 Mpc) is a similar method that uses the tip of the red giant branch star luminosity function. Stars travelling up the RGB will eventually experience an He flash and transition off the RGB to the zero-age horizontal branch. An He flash occurs when the He core of an RGB reaches ~ 0.5 $M_\odot$, and the luminosity prior is dependent on properties of the H-burning shell, which in turn is dependent on the He core; this means that the most luminous a red giant can get is an almost constant value. A distance, then, can be estimated from the brightest RGB stars in a galaxy (since they will be very near the RGB tip).

- **Novae** (up to 20 Mpc) have, like their SNe Ia cousins, a luminosity vs. light curve decline time relationship that allows them to be calibrated into standard candles using $M_V^{max} = -9.96 - 2.31 log_{10}\dot{m}$, where $\dot{m}$ is the average rate of decline over the first 2 mag in magnitudes per day. This relationship comes from the fact that more massive WDs tend to accrete smaller amounts of matter before a nuclear runaway occurs (i.e. producing less bright an explosion), and this thinner layer is ejected more easily.

- **Variables** (up to 29 Mpc) have a correlation between pulsation and luminosity[8], allowing determination of a distance modulus (assuming extinction is also known). Classical Cephids have been seen up to 29 Mpc away, while RR Lyrae and W Virginis stars are generally fainter and therefore can only be used for closer distances.

- **Globular clusters*** (up to 50 Mpc) follow an empirical luminosity function. Because the function has a turnover, fitting a sampling of globular clusters around a distant galaxy with the luminosity function can be used to derive a relative distance.

- **Planetary nebulae*** (up to 50 Mpc) also appear to follow a luminosity function[9] with a rapid cutoff at around -4.6 mag. Fitting, and comparing to a fit with a known distance, provides a distance measure; alternatively, finding the brightest PN in a galaxy and assuming they reside near the cutoff also gives a distance. As the cutoff method uses brighter PN, it can be to larger distances (the fitting method only goes up to 20 Mpc).

- **The surface brightness method*** (up to 50 Mpc) is based on the fact that the number of bright stars (which contribute the majority of the brightness) per area element in a galaxy fluctuates by Poisson noise. Since a galaxy further away will subtend a smaller area on the sky, there will be more bright stars per angular area, and therefore less fluctuation in the surface brightness (since the fluctuation goes like $\sqrt{N}/N$).

[8] $M_V = -3.53 log_{10}(P_d) - 2.13 + 2.13(B-V)$.
[9] Here the measured luminosity is L_λ at $\lambda = 5007$ Å.

- **The Tully-Fisher relation*** ($>$ 100 Mpc) is a relation between a spiral galaxy's luminosity and its maximum rotational velocity described in Sec. 2.1. The analagous relation for ellipticals is the Faber-Jackson relation, and is much noisier and more difficult to use as a standard ruler. Physically, the Tully-Fisher relation means that the mass-to-light ratio and mean surface brightness of spirals is fairly constant. In fact, due to the changing fraction of baryons in gas instead of stars for lower mass spirals, the mass-to-light ratio does change - adding a correction term to the Tully-Fisher relation results in a much tighter relationship.

- **The $D - \sigma$ relation*** ($>$ 100 Mpc) is a relation between an elliptical galaxy's angular diameter D out to a constant surface brightness (20.75 B-mag per arcsec) and its velocity dispersion. Since surface brightness is independent of distance, D is inversely proportional to the distance to the elliptical galaxy. Physically, this relation is a natural outcome of the fundamental plane and the (fairly good) assumption that all ellipticals are self-similar.

- **Masers** ($>$ 1000 Mpc) are amplified microwave emissions coming from regions of interstellar media where populations (due to very low densities) can become inverted. Cosmic masers exist around some galactic nuclear regions with line luminosities up to 10^4 $L_\odot$; these maser sources orbit the supermassive black hole, and observations of both the radial and proper motions of these sources can be made. If we assume circular orbits, these two values can be combined to give a distance. Source radial velocities can also be combined with an independent measure of the black hole mass to determine distances.

- **Supernovae Ia** ($>$ 1000 Mpc) have (or at least a sub-class of them have) a strong relation between light curve decay time and peak luminosity, the Phillips relation (Sec. 1.9). While most optical and infrared observations can be used, this relation is has the smallest spread in the infrared. Other supernovae have similar relationships that could be used to determine luminosity (ex. SNe IIP plateau longevity is correlated with maximum luminosity), though these relationships are generally not as well defined and/or studied, and SNe Ia are brighter in the optical and IR and other SNe.

- **Brightest galaxies in clusters*** ($>$ 1000 Mpc) fits a luminosity function to the galaxy cluster, and determines a distance modulus from the fit. It assumes that the luminosity function of galaxies remains fixed over time.

- **Cosmological redshift** ($>$ 1000 Mpc) requires finding characteristic spectral features (lines, bumps, etc.). Redshift can be made to correspond to a physical distance through Hubble's Law ($v = H_0 d$) at low redshifts, and to d_c (assuming a cosmological model) at large redshift. Use of this measure must take into account peculiar velocities of objects.

All techniques denoted with a * are secondary distance indicators, which only give distance scaling, and therefore require a calibration galaxy with a distance known by other means (primary distance indicators can be used on their own because they give absolute luminosity distances).

The distance to the LMC can be determined from the Wilson-Bappu, RGB tip/supergiants, and variable star methods, giving a distance of $\sim$ 48 kpc (the same answer can be obtained by "going backwards" by taking the NGC 4258 maser distance as a baseline for other distance measures). The most accurate distance for an object, however, was found using the light echo of SN1987A off of a nearly perfectly circular ring of material (ejected during the SN progenitor's AGB phase) to determine a distance. The ring is inclined to us, and so while light from the SN should impact all sections of the ring at the same time, due to the inclination and light delay we see different sections of the ring light up at different times. The light delay can be used to determine the physical size of the ring, while its angular size is easily observed, and from these two measurements we can obtain a distance of 51.8 ± 3 kpc to the SNR.

The distance to Andromeda has been measured with variable stars, surface brightness fluctuations, RGB tip/supergiants, the Wilson-Bappu effect and measurements of eclipsing binaries; the current accepted distance is $\sim$ 770 kpc.

The distance to the Virgo Cluster can be determined by using a large number of these methods, including variable stars, novae, luminosity function methods, surface brightness fluctuations, the Tully-Fisher and $D - \sigma$ relations and Type Ia supernovae, applied to its component galaxies; all methods agree that it is $\sim$ 17 Mpc away.

2.5. *Question 5*

QUESTION: What evidence is there that most galaxies contain nuclear black holes? How do those black holes interact with their host galaxies?

For most observational purposes, a black hole (BH) is defined as a mass-concentration whose radius is smaller than the Schwarzschild radius r_S of its corresponding mass. This value is about $r_S \sim 10^7$ km ~ 15 $R_\odot$ for the SMBH in the Galactic centre (GC). Since we cannot even resolve the supermassive black hole (SMBH) in our own galactic centre (which spans 10^{-5}") in order to show that other galaxies contain SMBHs we must find indirect evidence of a massive compact object at the centres of galaxies.

For many galaxies, we can detect stellar kinematics around the galactic centre. The radius of influence of a black hole is defined as

$$r_{BH} = \frac{GM_{BH}}{\sigma^2} \approx 0.4 \text{ pc} \left(\frac{M_{BH}}{10^6\ M_\odot}\right) \left(\frac{\sigma}{100 \text{ km/s}}\right)^{-2} \tag{57}$$

where the inherent velocity dispersion of the material being affected by the black hole is σ. The equivalent angular scale is

$$\theta_{BH} = \frac{r_{BH}}{D} \approx 0.1'' \left(\frac{M_{BH}}{10^6\ M_\odot}\right) \left(\frac{\sigma}{100 \text{ km/s}}\right)^{-2} \left(\frac{D}{1 \text{ Mpc}}\right)^{-1}. \tag{58}$$

The presence of an SMBH inside r_{BH} is revealed by an increase in σ; $\sigma \propto r^{-1/2}$. If the centre of the galaxy rotates, we would also expect $v \propto r^{-1/2}$. The challenge here is to obtain stellar kinematics at resolutions substantially smaller than θ_{BH}; moreover, the kinematics of stars can be complicated, and the observed values of σ and v depend on the distribution of orbits and on the geometry of the distribution.

Nevertheless, we have made measurements of the stellar kinematics of nearby normal galaxies, and have found what appear to be $\sim 10^7$ $M_\odot$ highly concentrated masses in approximately 35 of them. In general, the kinematics are incompatible with the r^{-2} isothermal profile of a star cluster; it is especially well-constrained for our own galaxy. In some cases, water masers orbiting very close to the galactic centre were used to determine the existence of an accretion disk around the point masses.

For our own MW galactic centre, we can track individual stars, and we have found stars with proper motions exceeding 5000 km/s. Combining σ and the trajectories of individual stars, we are able to determine that the enclosed mass $M(r) \approx 4 \times 10^6$ $M_\odot$ down to ~ 0.01 pc. The closest approach of a star to the object is even smaller, 0.0006 pc. This is not within the Schwarzschild radius (3×10^{-7} pc).

We also have active galactic nuclei, and we know that the radio lobes in some of these AGN reach $\gtrsim 1$ Mpc, indicating that the central power source is long-lived. Luminous QSOs have luminosities of up to $L \sim 10^{47}$ erg/s. Combining these two numbers gives us $E \sim 10^{61}$ erg. Lastly, luminosity variations in some AGN can be up to 50% on the timescale of days. This means that $R \lesssim 0.0008$ pc. The radio lobes are fairly straight, indicating that the generator must be a very stable rotator. In many AGNs, X-ray observations of a (both special and general) relativistically broadened Fe line indicate emission in the inner regions of an accretion disk, with a radius of just a few Schwarzschild radii.

All these observations are not proof of the existence of a SMBH in these galaxies, because the sources from which we obtain the kinematic evidence are still too far away from the Schwarzschild radius. The conclusion of the presence of SMBHs is rather that of a missing alternative; we have no other plausible model for the mass concentrations detected. We may postulate an ultra-compact dark star cluster with a density profile steeper than r^{-4}, but such a cluster would evaporate within 10 Myr due to frequent stellar collisions. The case for AGN being due to SMBHs is stronger. No source of energy other than accretion onto a black hole can realistically explain so much energy being generated in a volume of space so small[10]. Moreover, the Fe line is entirely consistent with an accretion disk that extends down to a few Schwarzschild radii.

Black holes affect their host galaxies in a number of ways, primarily through the outflows they generate when the accrete:

1. Black holes can quench star formation in the galaxy by generating very strong outflows (i.e. quasar jets) during rapid accretion (Page et al. 2012; Faucher-Giguere & Quataert 2012). These outflows expel the ISM. It is possible that this is why there is a relationship between the mass of the bulge and the mass of the SMBH (Page et al. 2012).

2. Closely related to the previous point, black holes limit the growth of massive galaxies. The current luminosity function cutoff L_* corresponds to a halo mass considerably lower than the mass scale above which the abundance of dark matter halos is exponentially cut off. We also know that massive galaxies were already in place at high redshift, and so should have grown considerably since then via cooling flows. The solution is to invoke AGN - they heat the cooling, infalling material, heating it up and preventing it from forming stars. (Black holes therefore both expel and heat matter, both of which quench star formation.)

3. More dormant black holes can blow x-ray cavities filled with relativistic plasma in the centres of galaxies (Sijacki et al. 2007).

4. Black holes (AGN feedback) may solve cosmic downsizing.

5. Black holes tend to fling hypervelocity stars out of the galaxy through 3-body interaction. A binary system passing very close to the black hole may result in one star being flung into the black hole, and the other star obtaining an enormous velocity to conserve momentum. A few of these stars (generally early types, as that is the population of stars near the Galactic centre) have been seen in the MW.

[10] Black holes have an accretion efficiency of ∼6 - 29%; a star at the Schwarzschild radius has $E_{\text{grav}} \sim 10^{53}$ erg, so to power an AGN requires some 10^{-6} $M_\odot$/yr worth of accretion.

2.5.1. *How are SMBHs formed?*

See Sec. 5.4.1.

2.5.2. *What correlations are there between the properties of the SMBH and the host galaxy?*

One of the most famous is the correlation between black hole mass and the velocity dispersion of the bulge,

$$M_{\rm BH} \propto \sigma^4 \tag{59}$$

This can actually be used to predict the mass of the SMBH based on stellar kinematics of an elliptical or spiral bulge. Since the SMBH is unlikely affecting the kinematics of the entire bulge/elliptical, this is probably a feature of galactic evolution.

Black hole mass is also proportional to the luminosity of the bulge ($M_{\rm BH} \propto L_{\rm bulge}^{1.11}$, equivalent to $M_{\rm BH} \propto M_{\rm bulge}^{0}$.9). As noted earlier, this is possibly due to the black hole quenching star formation early in the life of the galaxy.

2.6. *Question 6*

QUESTION: Define and describe globular clusters. Where are they located? What are typical ages of globular clusters. How is this determined?

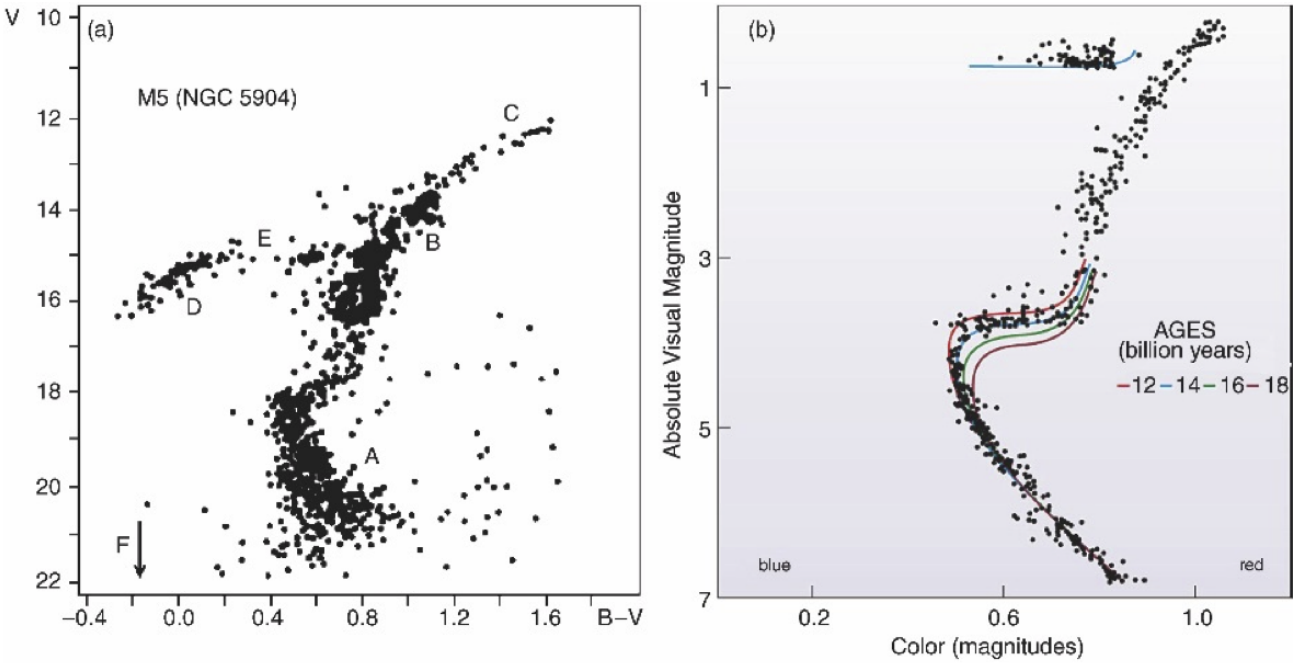

FIG. 33.— Left: CMD for globular cluster M5. The different labels denote evolutionary stages: A is the main sequence; B is the red giant branch; C is the point of helium flash; D is horizontal branch; E is the Schwarzschild gap in the horizontal branch; F denotes white dwarfs (below the arrow). The turn-off point is located where the main sequence turns over to the red giant branch. Right: CMD of the globular cluster 47 Tucanae compared to isochrone models parameterized by age (solid coloured lines). The ages plotted in this diagram should actually be decreased by 2 Gyr since those shown here were constructed from incorrect cluster distance measures. From Emberson (2012).

This question is related to Sec. 3.6 and Sec. 3.16, and much information is likely repeated here. Also, see my notes on the difference between a GC and a dSph.

Globular clusters (GCs) are spherical collections of gravitationally bound stars that orbit the core of a galaxy. They are highly spherical (with $b/a \sim 0.95$ on average) and comprised of $10^5 - 10^6$ (generally) metal-poor stars tightly bound by gravity. The average stellar density in a GC, 0.4 pc^{-1} is much larger than that of open clusters, and in the centres of GCs the density can rise up to 1000 pc^{-1}. The half-light radius of a GC is around 10 pc.

GCs cannot be considered collisionless systems, since their relaxation time ($10^8 - 10^9$ yrs) is significantly shorter than their age (10^{10} yrs). This explains why they are almost spherical: thermalization has resulted in the equipartition of stellar velocities along all three axes, and any departure from spherical symmetry is due to a relatively small amount of angular momentum possessed by the cluster, which lends to rotational support. Their density profiles are well fit to a King profile (an isothermal sphere with a fast drop-off to keep total mass finite).

GCs are ubiquitous: they can be found in the halos of almost all galaxies of sufficient mass (including many dwarf galaxies) are known to have GCs, with the Milky Way having about 150. For our galaxy, they come in two different populations: the metal poor population ([Fe/H] < - 0.8) is spherically distributed, while the more metal rich population ([Fe/H] > - 0.8), with the notable exception of 47 Tucana, is primarily distributed near the galactic thick disk, and may actually be associated with it (both populations having a scale height of 1 kpc and similar rotations about the galactic centre). GCs can be found from 500 pc to 120 kpc from the galactic centre, though the majority of them are within 40 kpc of the galactic centre. It is not clear whether or not the field stars and some GCs form a unified population of halo members.

GC formation history is still not well-understood. They are believed to have formed from the collapse and fragmentation of a single molecular cloud. As a result, the stars within the cluster formed with essentially identical compositions and within a relatively short period of time. In the MW, tight superclusters like Westurland 1 may be their precursors.

The stellar population of a GC is roughly coeval (though there are exceptions, like NGC 2808, that contain multiple generations of stars), i.e. all stars in the cluster have the same age. This allows the stellar population to be fit to an isochrone, a snapshot in time of the evolution of a line of coeval stars. Since isochrones are time-dependent, the age of the best-fitting isochrone gives an estimate of the mass of the GC (there are various methods of doing this; see (Binney & Merrifield 1998, Ch. 6.1.4)). Fig. 33 gives two examples.

The mean age of the oldest globular clusters around the MW is 11.5 ± 1.3 Gyr, though absolute ages are often greatly affected by errors on GC distances and reddening (De Angeli et al. 2005). The dispersion in ages of metal-poor MW GCs is about 0.6 Gyr De Angeli et al. (2005). Higher-metallicity clusters are on average 1 - 1.5 Gyr younger, though a rare few are younger stil De Angeli et al. (2005).

An interesting feature evident in the colour-magnitude diagrams of young globular clusters is the Hertzsprung Gap. This is the paucity of stars located between the turn-off point of the MS and the RGB. This feature arises because of the rapid evolution that occurs just after leaving the MS. The evolution of this stage is largely damped for low mass stars (1.24 $M_\odot$) and so this gap does not appear for old globular clusters where the turn-off point is located at low mass scales. Another interesting feature seen in some globular clusters is the existence of a group of stars, known as blue stragglers, that can be found above the turn-off point (i.e., still on the MS). Although our understanding of these stars is incomplete, it appears that their tardiness in leaving the MS is due to some unusual aspects of their evolution. The most likely scenarios appear to be mass exchange with a binary star companion, or collisions between two stars, extending the star's MS lifetime.

2.7. *Question 7*

QUESTION: What is the X-ray background and how is it produced?

This information is mostly from my own notes.

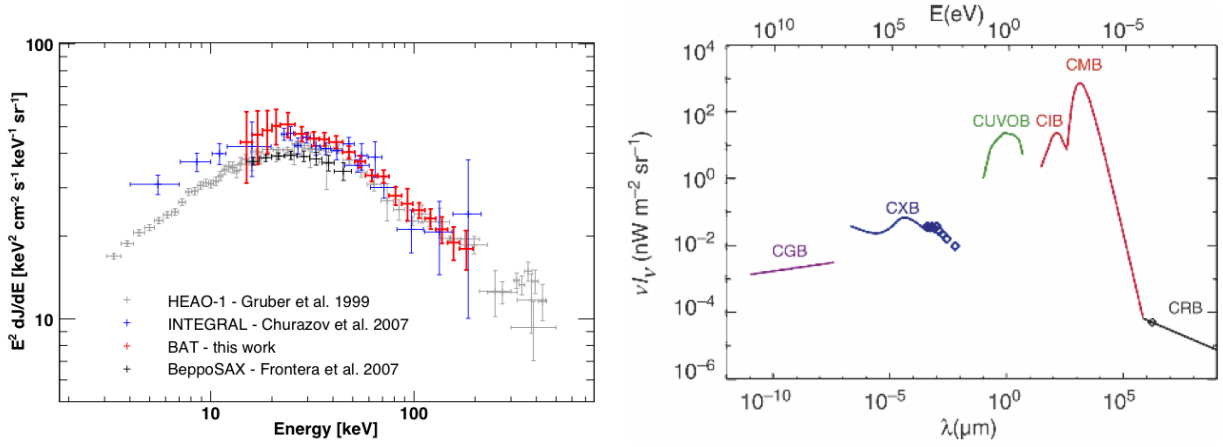

FIG. 34.— Left: comparison of SED measurements of the CXB. Here $E^2 dF/dE$, which is dimensionally equivalent to λF_λ. Right: SED of cosmic background radiation, plotted as νI_ν versus λ. Besides the CMB, background radiation exists in the radio domain (cosmic radio background, CRB), in the infrared (CIB), in the optical/UV (CUVOB), in the X-ray (CXB), and at gamma-ray energies (CGB). With the exception of the CMB, probably all of these backgrounds can be understood as a superposition of the emission from discrete sources. The energy density in the CMB exceeds that of other radiation components. From Emberson (2012).

In the 1960s and 1970s, rocket and satellite-based x-ray observatories began discovering strong galactic and extragalactic x-ray sources. In addition to these sources, however, they also detected an isotropic x-ray background that (at first) did not seem to be caused by any known source - this background has come to be known as the cosmic x-ray background (CXRB, or CXB). It was hypothesized early on that the background was due to a large number of

unresolved point-source AGNs, but AGN spectra have power laws $S_\nu \propto \nu^{-0.7}$ (S_ν is the source function) while the CXB can roughly be parameterized by $I_\nu \propto E^{0.3}\exp(-E/E_0)$, where $E_0 \approx 40$ keV.

Later, deep (i.e. able to distinguish faint sources) ROSAT observations showed that at least 80% of CXB emission in the 0.5 - 2 keV band can be attributed to discrete sources (predominantly AGN, but also starburst and normal galaxies, and galaxy clusters at large distance); later XMM-Newton observations would confirm this. Because of their spectral index AGN could not account for the CXB at high energy even if they contributed 100% of 0.5 - 2 keV CXB emission. Instead, the high energy CXB could be dominated either by enshrouded AGN (whose intrinsic spectrum would be modified by interactions with the shrouding material) or by pockets of superheated gas in the IGM with energies $kT \approx 30keV$ (a popular hypothesis, since the CXB spectrum, subtracted by the AGN contribution from ROSAT, gave a very hard (i.e. flat) spectrum reminiscent of thermal bremmstrahlung). This same gas, however, would inverse Compton scatter CMB photons, producing distortions to the CMB blackbody. Hot IGM gas was therefore eliminated as a high-energy CXB contributor when COBE observations showed negligible blackbody distortion.

More recent observations from Chandra in the 2 - 10 keV range have shown at least 75% of the CXB at this energy range can be resolved as descrete x-ray sources, most of which are AGN with strong self-absorption. AGN with strong self-absorption are Type 2 AGN, which have only been detected with Chandra (indicating why the high-energy CXB mystery was not resolved until its launch).

It is currently believed that a large population of AGN are believed to have column densities larger than $N_H \approx 1.5 \times 10^{24}$ cm^{-2}, corresponding to $\tau = N_H\sigma_T \approx 1$ (where σ_T is the Thomson scattering cross-section). Known as "Compton-thick" AGN, they are extremely difficult to detect, as much of the x-ray spectrum is downscattered by Compton scattering. At very high energies, these systems can still be detected, and form part of the CXB. Since telescope resolution is poor, the CXB then currently forms the primary means by which we can constrain the space density of Compton-thick AGN. Ultimately, however, discrete sources must be found, and their summed emission compared to the CXB; recent and near future hard x-ray surveys, as well as attempts to find high-redshift AGN in the IR (highly shrouded AGN will have their black hole UV emission reprocessed into the IR) will likely provide the discrete source densities needed to explain the high-energy CXB.

2.7.1. *What determines the emission spectrum of an AGN?*

See Sec. 2.13.

2.7.2. *Are there backgrounds at other wavelengths?*

Fig. 34 shows a large number of different cosmic backgrounds. As discrete sources were not removed from the figure, discrete sources such as galaxies and quasars are included in all the backgrounds.

The cosmic infrared background (CIB) was first discovered in the far-IR by COBE. It, like the CXB, is most likely due to a large unresolved population of discrete sources (such as high-redshift starburst galaxies). Unfortunately the CIB is very difficult to detect due to instrument cooling limitations and the fact that the CIB is much fainter than the galactic infrared background. As of 2006 most of the sub-mm CIB has been determined to be due to dusty star-forming regions and other discrete sources, while the ISO satellite has managed to resolve about 10% of ~ 2 μm emission as discrete sources.

Indeed, the other cosmic backgrounds are likely entirely due to discrete sources, as well, since we do not expect high-energy backgrounds to have cosmological origin. The CRB is likely due to a combination of the CMB and radio-loud galaxies, the CUVOB entirely due to stars, and the CGB from compact objects, AGN and cosmic ray-ISM interactions.

See Sec. 2.14.

2.8. *Question 8*

QUESTION: Describe the currently accepted model for the formation of the various types of galaxies. What are the lines of evidence to support this model?

As noted from Sec. 1.4, in concordance cosmology halos of lower mass form first, and massive halos form later. Since these massive halos will have smaller halos embedded in them, the formation of a massive halo can be thought of as due to the merging of smaller halos - this is known as hierarchical structure formation, or the "bottom-up" scenario of structure formation. If stars formed in these smaller halos, the mergers would be visible.

The formation of spiral galaxies can easily be explained by this picture. Suppose that gas in a halo can cool sufficiently for stars to form (see the cold flows follow-up for discussion). Cooling is a two-body process ($\propto \rho^2$), so the most efficient coolers will be dense clumps. These dense clumps will have some non-zero angular momentum, and will therefore settle onto a disk, and thus a small disk galaxy is formed. This process may look like a small-scale version of the "monolithic collapse" hypothesis of how the MW formed (Sec. 3.16). Spiral arms are formed by density perturbations in the disk, which generates waves of star formation in the galaxy. The galaxy can then grow by merging with other galaxies. We note, however, that there is no reason to simply expect mergers to continue building the disk (one can imaging various scenarios in which the total angular momentum of the system is much smaller than the spin angular momentum of each galaxy), and disks can be warped and destroyed by merger. Semi-analytical models of galaxy formation suggest that the bulge of a spiral is first formed through a major merger, and then the disk is formed through gas-rich minor mergers and accretion of cold flows from the IGM.

Several proposals have been made on how ellipticals can form. It is possible that ellipticals form by truncated monolithic collapse. In the MW monolithic collapse scenario, early star formation lead to the creation of the galactic halo and globular clusters. If all stars were to form this quickly, then an elliptical would be formed before any of the gas could settle onto a disk. It is not obvious how this could occur, however.

Currently ellipticals are generally thought to be formed out of major mergers of disk galaxies. The merger would destroy the disks of both initial galaxies, and shock the gas, generating a starburst. Due to the violence of such mergers, stars and gas are ejected from the system, and remaining gas becomes strongly heated, quenching further star formation after the starburst. Because the stars within ellipticals are generally old, it is suspected that most mergers that form ellipticals are gas-poor ("dry"), in order to minimize the effect of the starburst.

Evidence for this model:

- We see a plethora of small, irregularly shaped blue galaxies at high redshift, as well as substantial numbers of galaxy mergers[11].
- Ellipticals tend to show signs of complex evolution that might be produced by mergers.
- The model provides a partial explanation for the Butcher-Oemler effect (clusters contain more blue galaxies at high redshift)[12].
- The model also provides an explanation for the morphology-density relation, where ellipticals are much more abundant in the centres of dense clusters, while spirals dominate in less dense clusters and in the periphery of dense clusters.
- Detailed simulations of merging disk galaxies have shown they can produce ellipticals. Simulations indicate that dry mergers can preserve the fundamental plane (i.e. if, pre-merger, the merging galaxies obey the fundamental plane, then so will the merger remnant).
- High-luminosity ellipticals tend to have boxy isophotes and are slowly rotating. Low-luminosity ellipticals tend to be more rapidly rotating and have disky isophotes. This can be explaind by dry-mergers generating massive ellipticals and more gas-rich mergers between late-type galaxies generating lower-luminosity ellipticals (early-type galaxies tend to be more massive than late-type galaxies).

2.8.1. *What are cold flows?*

This information is from the introduction and conclusion of Brooks et al. (2009).

The classic model for galaxy formation, loosely describe above, has the dark matter violently relaxing and virializing very quickly during the non-linear collapse phase. This process shock-heats the infalling baryons to virial temperatures as well. Given some kind of density profile, a cooling radius can be calculated, and in the interior gas radiates away thermal energy while maintaining its angular momentum. Recent numerical studies have shown evidence to the contrary: that much of the infalling baryonic matter is never shock-heated to virial temperatures, and flows onto a cold disk without out the need for substantial dissipation. Brooks et al.'s simulations find that early disk growth can largely be attributed these cold flows, even if most of the inflowing gas is shock-heated to virial temperatures. This gas does heat up somewhat from loss of its own gravitational potential energy, but still substantially below virial temperatures. For an MW mass galaxy, only $\sim$25% of the accreted gas comes from acquiring other galaxy halos, and so most of the stars are formed out of cold flow material. This picture allows large disks to be built at much higher z than the classical model.

2.8.2. *What is feedback?*

Shortly after the formation of stars, the most massive of them explode as supernovae, which reheat the ISM around them. This reduces the cooling rate of the gas (and also blows some gas away). AGN and stellar winds from OB associations provide similar "feedback" mechanisms that help regulate star formation. Without this, the amount of star formation predicted by numerical simulations would far outpace the rate observed in nature. Indeed, massive galaxies would likely be much more massive, and luminous, without feedback, a fact suggested by the L_* turnoff translating to a much smaller halo mass than the M_* turnoff predicted by evolution of the matter power spectrum.

The combined efforts of AGN and starburst feedback mechanisms are also responsible for maintaining the high ionization level in the IGM, which serves to maintain a high kinetic and excitation temperature in the IGM. Large halos are more affected by their own internal radiation output than intergalactic ionizing radiation. For smaller halos, however, this effect is important, and suppresses star formation as well as lowers the baryon fraction (by making baryons escape the potential well). Since these small halos are also disproportionately affected by their own feedback mechanisms (SNe are not less powerful because they are in less massive galaxies!). This is a potential solution for the "missing satellites problem", where there are far fewer small galaxies in the universe than there should be given a Press-Schechter DM halo distribution.

[11] This is buoyed theoretically: $D_+(z)$ evolution suggests that cosmic evolution slows substantially after z = 1, meaning the rate of mergers slows.

[12] The other part of the explanation is that ram-pressure stripping removes gas from galaxies over time.

2.8.3. *What is downsizing?*

While there is significant evidence that the hierarchical structure formation model works quite well (ex. galaxies exist out to high redshift, but superclusters are only abundant at $z \lesssim 1$; a large population of small blue galaxies exist, along with merging events, at high redshift), there are also contradictory effects. It appears that large galaxies were already in place at high redshift, and that over time star formation has primarily shifted to smaller galaxies. Though this issue has yet to be resolved, it is possibly also due to the feedback mechanisms describe earlier. Suppression of star formation in small DM halos until after quasars "turn off", for example, might explain why stars are only forming in small galaxies today.

2.9. *Question 9*

QUESTION: Describe three different methods used in the determination of the mass of a galaxy cluster.

There are actually five ways of determining the mass of a galaxy cluster:

1. **Stellar light**: If we measure the luminosity and spectra of the galaxy cluster (supposing we knew the distance using, say, the surface brightness meethod), we could attempt to determine the stellar populations in the galaxies. This, combined with mass-to-light ratios derived from other clusters, could give us the total mass of the cluster. This requires that we determine the M/L for other clusters before using this technique.

2. **The galactic velocity dispersion**: the crossing time of a cluster $t_{\rm cross} \sim R/\sigma$ is much smaller than the age of the universe, which means we would expect the cluster to be virialized (if it could not virialize, it would not be in hydrostatic equilibrium and would fall apart). We can then say

$$\sum_i m_i v_i^2 = \sum_{i<j} \frac{G m_i m_j}{r_{ij}}, \tag{60}$$

   Substituting in $\langle v^2 \rangle \equiv \frac{1}{M}\sum_i m_i v_i^2$ and $r_G \equiv M^2 \left(\sum_{i<j} \frac{m_i m_j}{r_{ij}}\right)^{-1}$., we obtain,

$$M = \frac{r_G \left\langle v^2 \right\rangle}{G}. \tag{61}$$

   We now note that because the relaxation time of the cluster is small (Sec. 3.4), the system is also thermalized, meaning that $\langle v^2 \rangle = 3\sigma_v^2$ and $r_G = \frac{\pi}{2M^2}\left(\sum_{i<j} \frac{m_i m_j}{R_{ij}}\right)^{-1}$. This gives us

$$M = \frac{3\pi R_G \sigma_v^2}{2G}. \tag{62}$$

   We therefore can use the velocities of the galaxies to determine the mass of the cluster. Due to small-number statistics, anisotropic velocity distributions and projection effects can lead to systematic errors when using this method.

3. **X-ray emission of cluster gas**: most (5/6) of the mass of a cluster is located in hot ($T \sim 10^7$ K) gas. This diffuse gas also exists between clusters, and in this context is known as the intercluster medium (ICM). This gas is also viralized, which means that

$$\frac{k_B}{\mu m_p}\left(\rho_g \frac{dT}{dr} + T\frac{d\rho_g}{dr}\right) = -\frac{GM(<r)\rho_g}{r^2}. \tag{63}$$

   Measuring the radial profile of temperature and density, therefore, gives us a measure of $M(<r)$. This is done by measuring the projected surface brightness at different bands. As the method of emission is thermal bremmstrahlung, we can fit this data to a model of free-free emissivity in order to constrain both density and tempeperature.

4. **Gravitational lensing**: this method is particularly powerful because gravitational light deflection is independent of the state of the deflecting matter, and we therefore have to make a minimum of assumptions.

   Firstly, we assume that the gravitational field producing the lensing is weak. This is a good assumption since the strength of a gravitational field can be quantized using the virial theorem: if a mass distribution is in virial

equilibrium then $v^2 \sim \Phi$ where Φ is the gravitational potential. Weak fields are then characterized by $v^2/c^2 \ll 1$ and with $v \sim 1000$ km/s for galaxies in clusters this is well justified. Next, we assume that the deflecting mass has a small extent along the line-of-sight, as compared to the distances between the observer and the lens and the lens and the source. A system satisfying this configuration is known as a geometrically thin lens. Since clusters are typically 8 Mpc across, this assumption is justified for sufficiently distant clusters.

If we approximate both the background object and lens as points, we can use Eqn. 214 to determine masses. Since neither the lens nor the mass distribution will be symmetric, the analysis needed is more complicated. In general, lensing models are constructed and fitted to observed arcs. The more images spread over a greater area, the better the precise distribution of masses in the lens can be determined.

5. **The Sunyaev-Zel'dovich effect**: described in Sec. 2.18.

Exactly what galaxy clusters are is described in Sec. 2.16.

2.9.1. *What are cooling flows?*

The outer regions of clusters have very long cooling times, so our assumption of hydrostatic equilibrium for determining the temperature and density of hot gas, above, is valid. Near the centres of galaxies, however, the free-fall time and cooling timescale are of the same order of magnitude. Since cooling (equivalently the free-free emissivity) is a function of ρ^2, the collapsing process is a runaway. These "cooling-flows" have been observed at the centres of massive clusters, in the form of a sharp peak of X-ray emission.

2.10. *Question 10*

QUESTION: What is the density-morphology relation for galaxies? How is that related to what we know about the relationship between galaxy density and star formation rates in galaxies?

The density-morphology relation is the observation that in denser environments, such as galaxy clusters, there is a higher ratio of elliptical galaxies versus spiral compared with field galaxies. This is true between different clusters, as well as within a cluster: the central regions of a cluster have a high fraction of elliptical galaxies, while the outskirts have a relatively lower fraction. See Figs. 35 and 36. The density-morphology relation is also seen in galaxy groups, though for these groups, the increase in the fraction of S0 galaxies as a function of redshift is much more dramatic than in clusters.

This relationship between morphology and density can (see Sec. 2.1) be translated into a colour-density relation. Galaxies exist in two groups: the blue clump and the red sequence. Since earlier-type galaxies are redder, we expect the red sequence to dominate over the blue clump at the centres of clusters, and the blue clump to dominate the red sequence in low-density regions. This is indeed the case.

Fig. 36 suggests a physical cause for the effect. For $R \gtrsim R_{\rm vir}$, the fraction of the different galaxy types remains basically constant. In the intermediate regime, $0.3 \lesssim R/R_{\rm vir} \lesssim 1$, the fraction of S0 galaxies strongly increases inwards, whereas the fraction of late-type spirals decreases accordingly. Below $R \lesssim 0.3R_{\rm vir}$, the fraction of S0 galaxies decreases strongly, and the fraction of ellipticals increases substantially. This can be explained in the following way: in the intermediate regions of clusters, a variety of interactions between galaxies and their environment, including ram-pressure stripping, encounters with other galaxies and gas evaporation, turn spiral galaxies into S0 galaxies. Deeper inside the cluster, the merger rate becomes very large due to the high galaxy density, which efficiently generates ellipticals out of other galaxies.

This relation suggests that star formation in a galaxy is a function of the galaxy's environment: namely, star formation of galaxies in dense environments is either halted or accelerated, in order for the gas within these galaxies to be quickly be depleted. As a result, the stellar populations in these galaxies look much redder than those of their peers in the outskirts of the cluster. We can identify a number of effects:

- **Ram pressure stripping**: caused by the fast motion of a galaxy through the ICM, which causes a pressure front to build up at the leading edge of the galaxy. This pressure front heats gas within the galaxy, and, if it is strong enough, also ejects gas from the galaxy.
- **Galaxy strangulation**: occurs as a galaxy falls into the cluster for the first time. The gravitational potential of the cluster creates tidal effects that launch gas out of the galaxy.
- **Galaxy interactions**: these tend to disturb or radically alter the morphologies of galaxies. Interactions can tidally induce starbursts, generate warps and bars in disks, and eject mass. A long series of slow interactions are believed to the main driver for turning spirals into S0 galaxies through the thickening of their disks and the triggering of starbursts in their bulges. Fast interactions are nicknamed "galaxy harrassment".
- **Galaxy mergers**: these are more extreme versions of interactions. They tend to both eject gas (through heating it and pulling it into tidal tails) and trigger starbursts. Indeed, in dense environments, repeated mergers tend

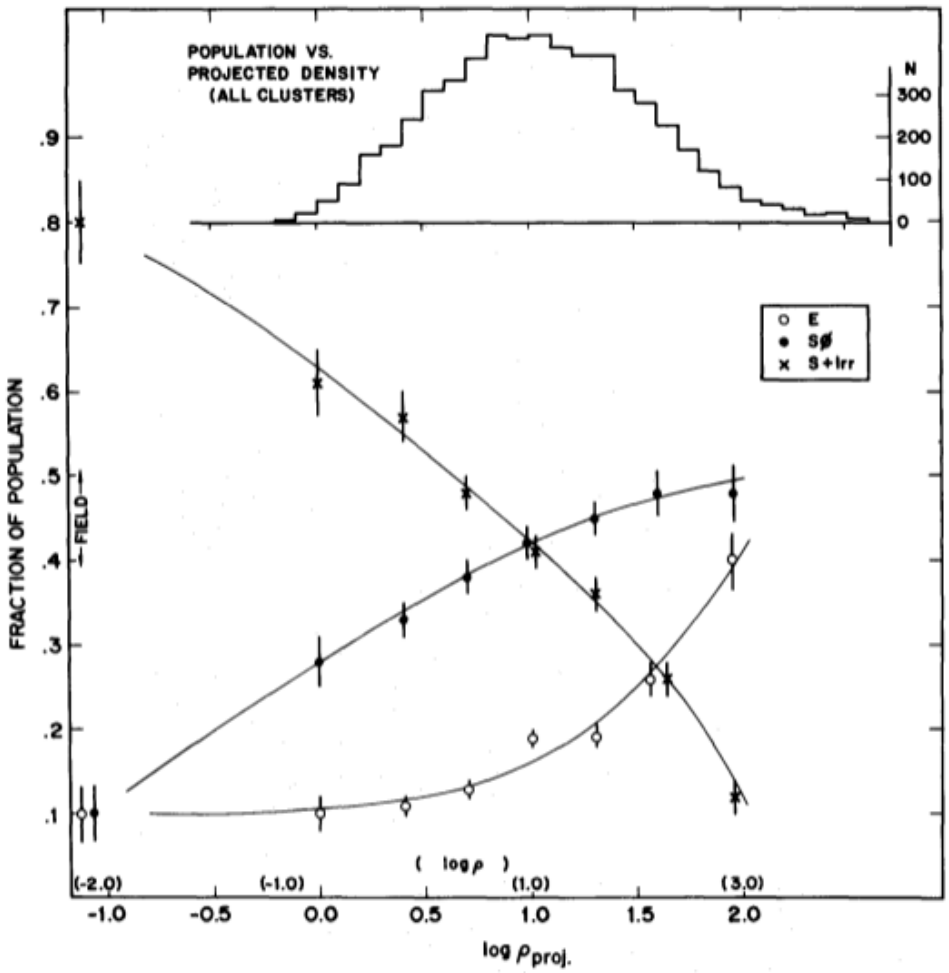

FIG. 35.— Fraction of E, S0 and S galaxies as a function of projected galaxy density (units of galaxies/Mpc^2). From Emberson (2012).

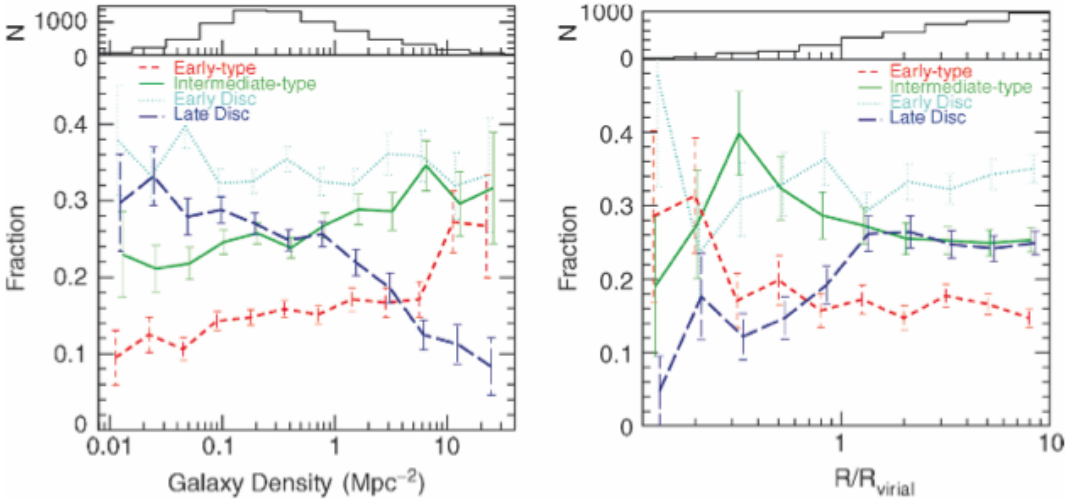

FIG. 36.— Left: same as Fig. 35, except from SDSS data. Note the point at which the lines begin to curve over is around 1 galaxy/Mpc^2. "Intermediates" are mainly S0 galaxies, and "early disk" Sa galaxies. Right, morphology fractions for clusters, as a function of radius from the centre of the cluster. From Emberson (2012).

to transfer most of the kinetic energy of the bulk motion of galaxies into kinetic energy of stars and gas. This is the reason why the velocity dispersion of very luminous (and therefore more massive and denser) clusters is lower than that of less luminous clusters.

2.11. *Question 11*

QUESTION: Draw the spectral energy distribution (SED) of a galaxy formed by a single burst of star formation at the ages of 10 Myrs, 2 Gyrs, and 10 Gyr.

This is mostly from my own notes.
See Fig. 37. The three necessary snapshots are described below:

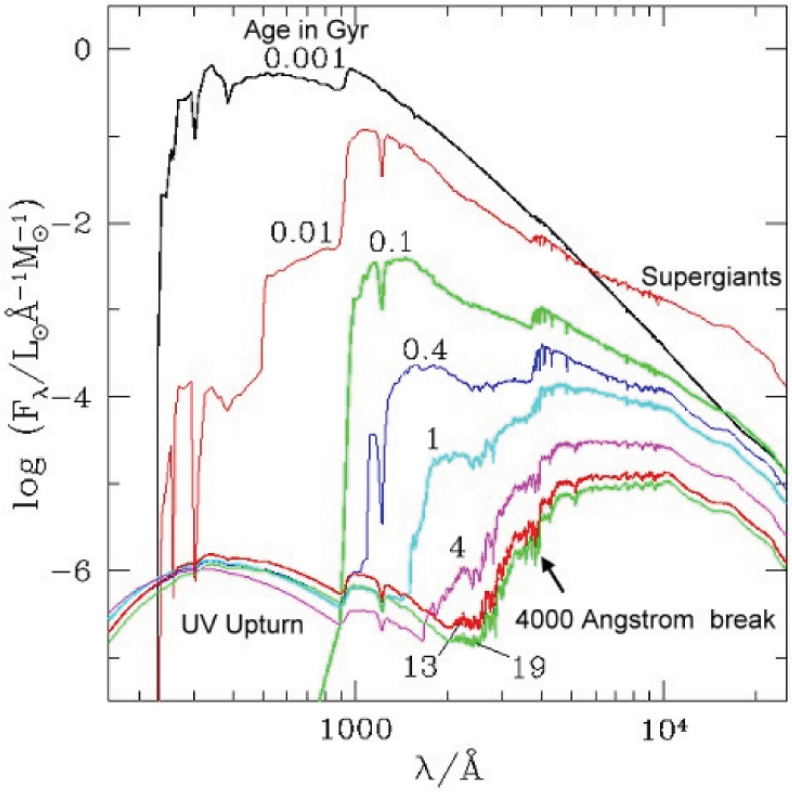

FIG. 37.— Plot of the cumulative spectral emission from a population (normalized to 1 $M_\odot$) born in a single burst of star formation at differing ages (given here in Gyr). Population has solar metallicity and uses a Salpeter IMF. From Bruzual A. & Charlot (1993), their Fig. 4.

- **10 Myr**: right after the burst of star formation, the spectrum resembles the 0.001 Gyr curve, and peaks in the UV due to the abundance of O and B stars in the young stellar population. These stars live for less than 10^7 years, and therefore by 10 Myr move off the main sequence and form red supergiants, resulting in a huge drop in UV emission, and a corresponding rise in NIR emission. A and late-type B type stars dominate the spectrum, and their combined emission peaks at around 1500 Å. At 912 Å there is a Lyman break, past which photoionization of H atoms creates significant absorption. This feature can be seen as early as 10 Myr, and becomes less prominent by 1 Gyr as fewer and fewer UV photons are created by the stellar population.

- **2 Gyr**: by 1 Gyr, most of the NIR emission comes from red giants. Meanwhile, absorption lines, in particular the break at 4000 Å, are becoming more prominent. The 4000 Å break is due to the accumulation of absorption lines of ionized metals in stars; it gets larger with age (due to decreasing temperatures corresponding to increasing opacity) and metallicity. The 4000 Å break is near the 3646 Å Balmer break, which marks the termination of the hydrogen Balmer series and and is strongest in A stars[13]. This break does not change strength monotonically with age, peaking in strength for a stellar population of 0.3 - 1 Gyr. Also starting at about 0.3 Gyr, UV radiation from the population again increases, this time due to the cooling of WDs created from low-mass main sequence evolution. WDs still feature a Lyman break, since they have thin atmospheres of H. At 2 Gyr the last A-stars are dying, and late-type A stars and F stars dominate the spectrum; their combined emission peaks at around 4000 Å.

[13] Note that both the Lyman and Balmer breaks can be used to determine the redshift of a galaxy, making them extremely important to high-z studies.

- **10 Gyr**: not much has changed since 2 Gyr. Late-type G stars and K stars dominate the spectrum, which now peaks at around 6000 Å.

These snapshots are created using population synthesis, which requries knowledge of the stellar initial mass function and stellar evolution. Of the two, the form has much more uncertainty.

2.11.1. *Describe population synthesis.*

Let $S_{\lambda Z(\tau)}(t)$ be the spectral emission, normalized to 1 $M_\odot$, of a stellar population (determined using an initial mass function; see Sec. 3.1) with zero-age metallicity $Z(\tau)$ and age t. We note that $\tau = t_{age} - t$, where t_{age} is the time since the first stars were born in the entire population [14]. Multiplying this with the star formation rate $\psi(\tau)$ (ψ is defined such that star formation starts at $\tau = 0$) gives us $\psi(t_{age} - t)S_{\lambda Z(t_{age}-t)}(t)$ the spectra emission from all stars in the population of age t. The spectral luminosity of a stellar population with age t_{age} can then be found using:

$$F_\lambda(t_{age}) = \int_0^{t_{age}} \psi(t_{age} - t)S_{\lambda,Z(t_{age}-t)}(t)dt. \tag{64}$$

Practically, such an integral is done numerically. Moreover, finding both ψ and S are difficult to achieve, and under/overestimating the formation of even a single class of stars can significantly change F_λ.

H II emission, the primary contributor of line emission to galaxies, is generally also added to spectra. After about 10^7 yrs the emission from gas nebulae becomes negligible.

There are very few places in the universe where one can view an isolated stellar population created from a single burst of star formation - certainly, galaxies look more complex. A more realistic star formation history is

$$\psi(t) = \tau^{-1}\exp(-(t - t_f)/\tau)H(t - t_f), \tag{65}$$

where H is the Heaviside function and t_f is the time of star formation. τ, the characteristic duration of star formation, as a large effect on the spectrum, because continuous star formation replenishes the supply of hot blue stars that tend to dominate the spectrum. For the same reason, an impromptu burst of star formation can significantly change the overall shape of the spectrum.

It is practically far more difficult to obtain spectra of galaxies than photometric observations, especially in deep-field surveys. Quite often, then, theoretically calculated spectra are reduced to colours by multiplying them with the transmission curves of colour filters.

Currently population synthesis can explain the colours of present-day galaxies, but not uniquely - a number of possible star formation histories all fit empirical data. Moreover, there is an age-metallicity degeneracy: contours of constant colour in an age/metallicity plot have the form

$$\frac{d\ln(T)}{d\ln(Z)} \approx -\frac{3}{2}, \tag{66}$$

i.e. smaller values of Z result in bluer looking spectra and a smaller mass-to-light ratio. Intrinsic dust absorption will also redden galaxies.

2.11.2. *What do the spectra of real galaxies look like?*

From our analysis above we would expect that younger stellar populations have blue spectra, strong emission lines, small 4000 Å breaks and weaker absorption lines. Indeed, this is what we see in Fig. 38. Different galaxy types form a continuum from irregulars, which are dominated by blue continuum emission and emission lines, to E0s, which have prominent absorption lines below 5000 Å (in particular the 4000 Å break is prominent), no emission lines, and has red continuum emission.

2.12. *Question 12*

QUESTION: What are Lyman-Break Galaxies and how do we find them?

High-redshift galaxies are very faint, and therefore difficult to resolve spectroscopically. As a result, it was nearly impossible until recently to actually conduct surveys for high-redshift galaxies. A major breakthrough in finding high-z objects is the Lyman-break method, which allowed the discovery of substantially redshifted galaxies through broad-band photometry[15].

[14] This can easily be seen: the first stars in the entire population to be born had metallicity $Z(\tau = 0)$, but are *currently* age t; stars that are just being born have age $t = 0$, but have metallicity $Z(\tau = t_{age})$.

[15] Narrowband photometry had also been used, specifically to look for the Ly-α emission line. This proved difficult because no one knew how faint galaxies are at $z = 3$, and how relatively the Ly-α line would be (Schneider 2006, pg. 357).

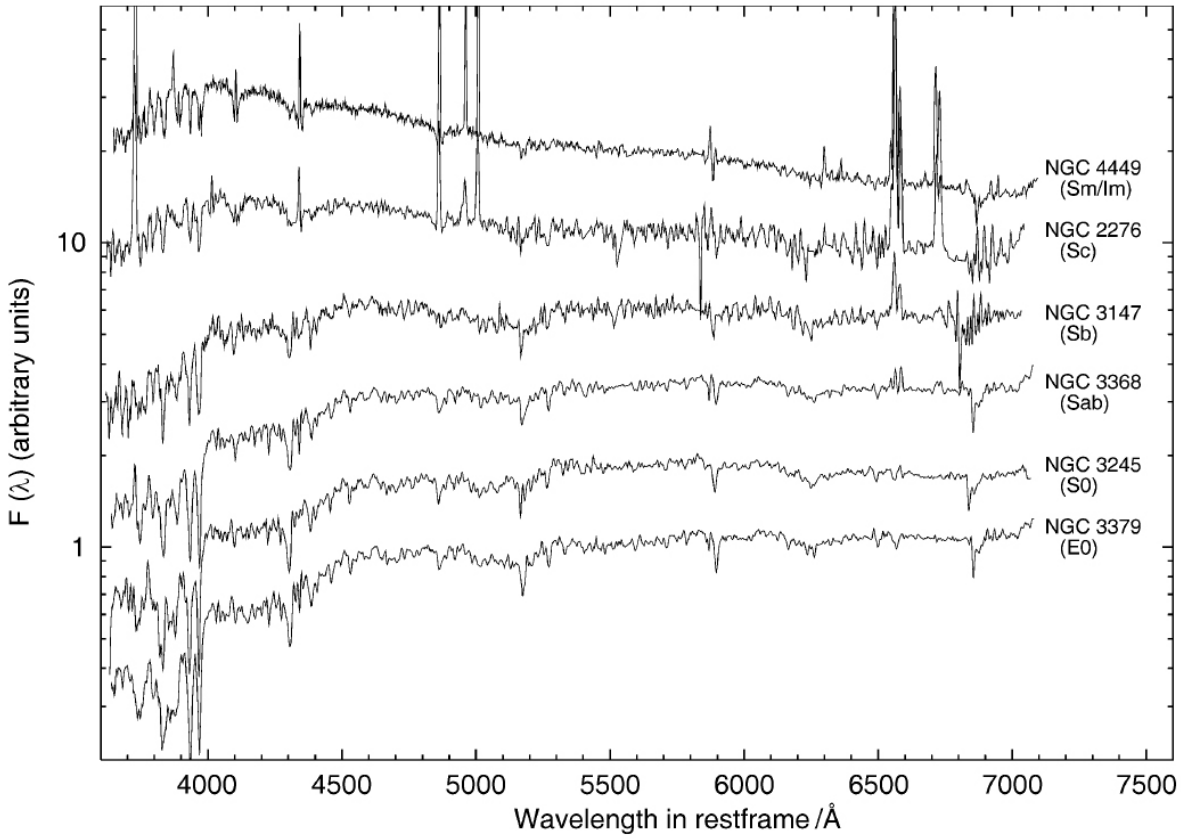

FIG. 38.— Spectra of galaxies of different types, ranging from very late (Irr) to very early (E0). Flux is plotted in arbitrary normalized units, and **if my understanding is correct** each spectrum has been artificially normalized to fit on the plot. From Schneider (2006), his Fig. 3.50.

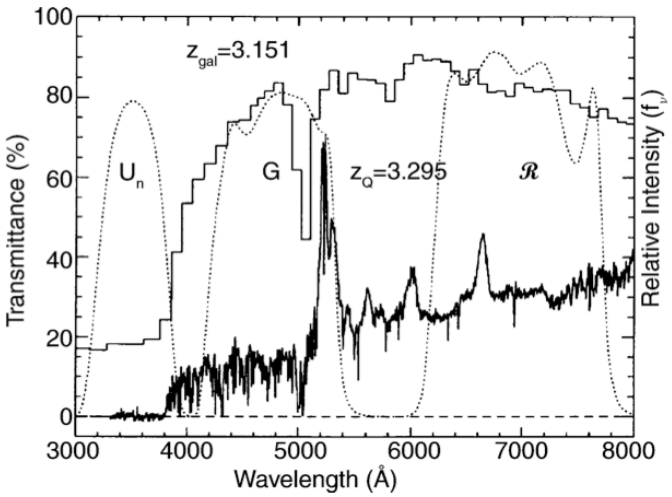

FIG. 39.— Principle of the Lyman-break method. The histogram shows the synthetic spectrum of a galaxy at $z = 3.15$, while the solid spectrum belongs to a QSO at slightly higher redshift. The dotted lines represent three different broad-band filteres. Both galaxies would be seen prominently in G and R, but not in U_n; equvialently $G - R$ would be be negative, while $U_n - G$ would be positive. From Schneider (2006), his Fig. 9.2.

Since H I is abundant and its ionization cross section is large, we expect photons with $\lambda < 912$ Å have a low probability of escaping from a galaxy without being absorbed. Intergalactic absorption also contributes (as evidenced by the Ly-α forest), absorbing photons with rest-frame wavelengths beyond Ly-α. From this, we expect that galaxies, especially high-redshift galaxies where the IGM may be partly neutral, are far less bright at wavelengths below $912(1 + z)$ Å, where z is the redshift of the galaxy.

We therefore can formulate an observing strategy to detect high-redshift galaxies (Fig. 39). We choose several

broad-band filters λ_i (three at minimum) with differing central wavelengths. If we find a source which can be seen in all filter images $i+1$ and beyond, but is not visible in all filter images before and including i, then we would suspect that we have found a galaxy at redshift z, such that $\lambda_i \lesssim 912\,\mathring{A}(1+z) \lesssim \lambda_{i+1}$. The z we find is known as a "photometric redshift". We can equivalently compare colours between filters, and would find that the colour would be very red between filters i and $i+1$. To make sure this is truly a break, we can also test to see if the colour between filters $i+1$ and $i+2$ is relatively blue. The galaxy candidate detected is known as a Lyman-break galaxy (LBG) or drop-out.

This method has been used to detect galaxies out to $z \sim 10$.

2.12.1. *What have we learned about high-z galaxies from LBGs?*

Since we require the colour between filters $i+1$ and $i+2$ to be blue, we tend to select for high-z galaxies undergoing active star formation.

Surveys for LBGs can be used to determine their spatial distribution, and therefore their clustering, as a function of redshift. Comparing this distribution to the distribution of dark matter halos at the same redshift. We find LBGs tend to correspond to high-mass dark matter halos, and tend to be assembled in proto-clusters. LGBs tend to have strong, blueshifted absorption lines, which could be evidence of feedback outflows from the many SNe generated by high star formation rates.

2.12.2. *Are there other "breaks" that could be used to detect galaxies?*

The Lyman-break technique is a special case of a general method of estimating the redshift of an object using drop-out in a photometric band. Though less strong than the Ly-α break, the 4000 Å break can also be used for photometric redshift measurements (this break is particularly strong in early-type galaxies). In general, a series of standard spectra can be convolved with the response of different colour filters, thus producing a template that can be used with photometric surveys. In general, photometric redshifts are estimates - to get an accurate redshift detailed spectroscopic follow-up must be performed.

2.12.3. *Can we find Lyman-break galaxies at low redshifts?*

The Lyman-break at low redshifts lies in the UV. The recent influx of GALEX data has been quite useful in the search for low-redshift analogs to LBGs. The result was the discovery of two populations: one composed mostly of normal late-type spirals, and the other composed mainly of compact galaxies with few stars, but the same star formation rate as much larger galaxies. Correspondingly, their stellar populations are more metal-poor. These compact galaxies are likely related to LBGs.

2.13. *Question 13*

QUESTION: Draw a spectrum of a high-redshift quasar. What do quasar emission lines typically look like? Explain what we see in the spectrum at rest wavelengths bluer than 1216 Å.

Quasars belong to a class of objects known as AGN, which is described in more detail in Sec. 2.15. Due to the angle at which we view their cores, the majority of observed luminosity comes from the compact, parsec-sized core of the galaxy.

The spectrum of a quasar can be found in Figs. 40 and 41. A number of notable features exist:

- The continuum spectrum of a quasar at long wavelengths can be parameterized by

$$S_\nu \propto \nu^{-\alpha}. \tag{67}$$

  $\alpha \approx 0.7$ for quasars in the radio regime, indicating optically thin synchrotron emission, though if one looks at the compact core α is much closer to zero due to synchrotron self-absorption.

- There is a substantial increase in emission over this power law (Fig. 41) in the IR, optical and UV. These are due to thermal emission of the various heated components of the AGN (warm dust in the IR, the accretion disk in the UV).

- The optical and UV spectra of quasars are dominated by numerous strong and very broad emission lines. The wavelengths are what we would typically expect in an H II region. The broadening is due to Doppler shifting of the gas rapidly orbiting around the black hole at several hundred r_s with speeds $v \sim 10^4$ km/s. The gas residing at these distances from the SMBH constitute a broad-line region (BLR) for line emission. Quasars also show narrow line emission, with Doppler widths suggesting $v \sim 500$ km/s. This is due to gas in the narrow-line region (NLR) for line emission, which extends to up to 100 pc from the centre of the AGN. These narrow lines are not due bulk velocities of the gas, but rather heating: AGN jets heat this region up to 1.6×10^4 K.

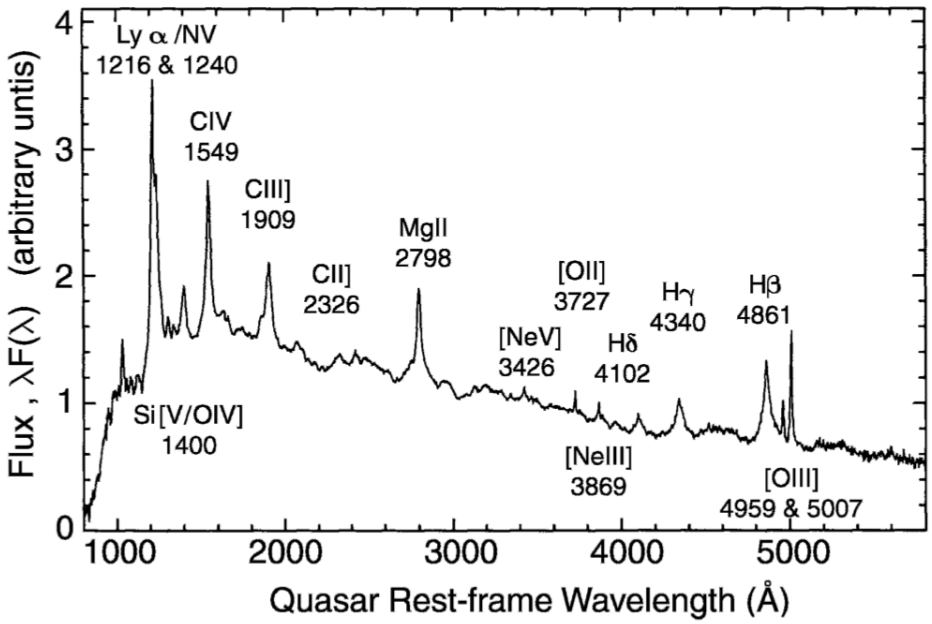

FIG. 40.— Combined spectrum of a sample of 718 individual quasars transformed into rest wavelengths of the sources. The most prominent emission lines are marked. From Emberson (2012).

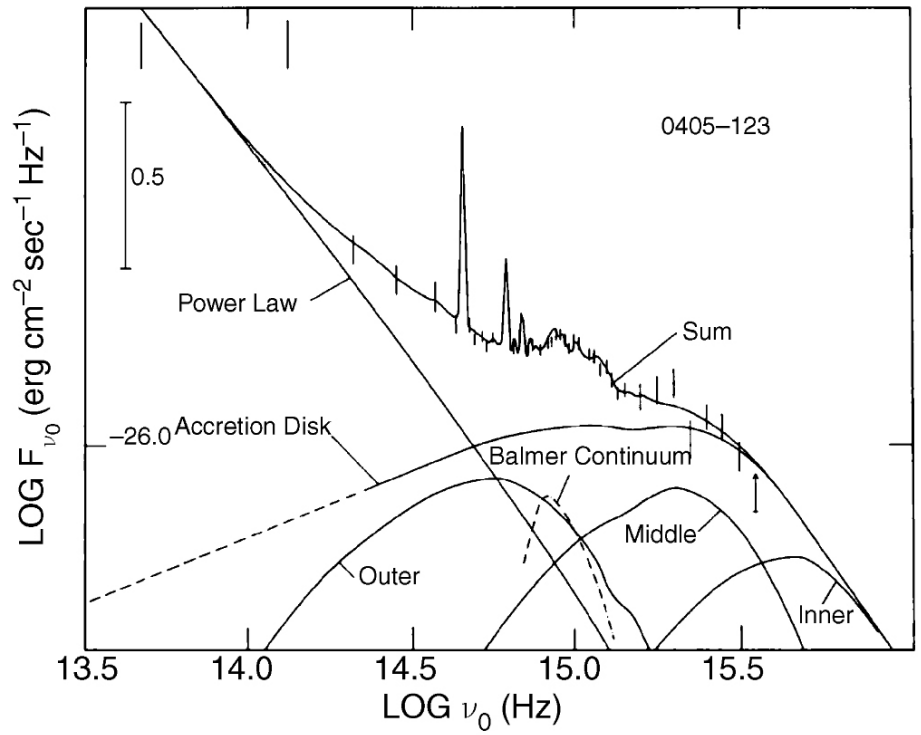

FIG. 41.— Spectrum of QSO PKS0405-123 ("sum") broken up into multiple labelled pieces. The accretion disk continuum emission is further broken down into the outer, middle and inner disk contributions. The higher frequency bump on the accretion disk is the "big blue bump". The Balmer continuum is the continuum created by a forest of H line emission from $n = 2$ to various higher states. Line emission is from the BLR and NLR. The "power law" is due to accretion jet synchrotron radition. From Schneider (2006), his Fig. 5.20.

- By the x-rays, the intrinsic spectrum of the AGN drops back down to a power law with $\alpha \approx 0.7$. Most of the emission in the soft x-rays comes from the inner accretion disk, while most of the harder x-rays come from a non-thermal source (hence the power law).

- Emission short of the rest-frame 1216 Å of the AGN may be absorbed, both by pockets of H I in the AGN itself, as well as the IGM on its way to us. This produces the Ly-α forest, and the lines in the forest can either be narrow, of intermediate width (due to Lyman Limit Systems) or broad (Damped Lyman Limit Systems). If a substantial

amount of neutral H I exists in between the quasar and us, we will obtain severe damping at wavelengths above $1216\,\mathring{A}(1+z)$ (where z is the redshift of the quasar); this is known as a Gunn-Peterson trough.

2.13.1. *What is the cause of the various emission features of AGN?*

AGN intrinsic emission is shaped by the AGN's various different regions. Thermal emission from the accretion disk provides the broad IR to UV continuum that peaks in the "big blue bump" in the UV (this bump is usually not directly observed; its empirical existence is extrapolated from UV and soft x-ray observations). For blazars, the optical continuum is dominated by power-law continuum emission, likely due to synchrotron radiation from the accretion jets. Warm dust contributes a second bump in the mid-IR.

The broad IR to UV excess is caused by emission from different parts of the accretion disk. Suppose a parcel of mass dm at a radial distance r from the black hole drops by distance dr. Taylor expansion of the gravitational potential energy gives us $dE = GMdmdr/r^2$. If half (from the virial theorem) of this liberated energy is expelled as heat (radiation), this gives $dL = GM\dot{m}dr/2r^2$. We can assume the disk is a blackbody, radiating from the top and bottom surfaces of the disk, in which case $dL = 2 \times 2\pi r dr \sigma_{\rm SB} T^4$. Equating the two dLs together gives us (see Sec. 5.13)

$$T(r) = \left(\frac{3GM\dot{m}}{8\pi\sigma_{\rm SB}r^3}\right)^{1/4}. \tag{68}$$

We can rewrite this (see Schneider pg. 187) as

$$T(r) = \left(\frac{3c^6}{64\pi\sigma_{\rm SB}G^2}\right)^{1/4} \dot{m}^{1/4} M^{-1/2} \left(\frac{r}{r_{\rm S}}\right)^{-3/4}. \tag{69}$$

From this we can easily see that for a large enough range of r we obtain a large range of T. Emission from the accretion disk can then be described, at least to first order, by a superposition of blackbodies (see Fig. 41).

Overlaid on top of this broad thermal emission is a power law, ostensibly from synchrotron radiation. Because of beaming effects, the strength of this power law is highly dependent on viewing angle. The spectral index will also change with viewing angle - $\alpha \sim 0$ is seen with blazars, while $\alpha \sim -1$ quasars, and so on.

AGN spectra also contain broad and narrow emission lines. No broad forbidden transition lines are observed, indicating that atoms emitting broad emission lines only have time to perform allowed and semi-forbidden transitions before suffering a collision with another atom - from this we obtain a density on the order of $n \sim 10^9$ cm^{-3} for the gas clouds doing the emitting. From line strength ratios, we know the clouds are at T $\sim 2 \times 10^4$ K, and, from line-broadening, they are travelling at around 10^4 km/s. Comparing the emission measure (photons per unit volume of the emitting gas) with the line strengths we note that these clouds are fairly small. The space covered by this collection of clouds constitutes the "broad line region" (BLR). Reverberation mapping studies [16] of this region show that these clouds occupy a significant range of radii from the supermassive black hole, and consequently different distance clouds have different emission properties (in accordance with their temperature). Not very much is known about the kinematics of these clouds, or why their high temperatures do not result in them disintegrating (that could either be due to them constantly being replaced, or due to magnetic, pressure or gravitational stabilization.

The narrow (emission) line region (NLR) rotates at a significantly slower rate ($\sim$ 400 km/s) than the BLR, and the abundance of forbidden transitions indicates a much lower density ($\sim 10^3$ cm^{-3}); line strength ratios give T $\sim 1.6 \times 10^4$ K. The region extends out to 100 pc (and is not homogeneous throughout), and because of this large size can be resolved - the NLR is cone-shaped, rather than spherically symmetric.

X-ray emission comes from the innermost part of the accretion disk, indicated by the extremely short time periods for variability. Spectrally, the emission can be characterized as $S_\nu \propto \nu^{-0.7}$ to first order; to second order, this functional form does not account for the big blue bump at low energies, and the spectrum flattening out at high. In addition there are a number of x-ray emission and absorption lines. For details on these lines, see Sec. 2.15

Some AGN have large H column densities that enshroud the AGN along our line of sight (these are known as Seyfert 2 and type-2 QSOs; they are likely the same objects as their type 1 counterparts, seen from different viewing angles). X-rays can photoionize H, and because photoionization absorption decreases with increasing frequency (i.e. energy), lower energy emission is attenuated while higher energy emission is not as affected. These objects are consequently very difficult to detect in the soft x-rays, and type-2 QSOs have only recently been found by (higher resolution hard x-ray observations from) Chandra and XMM-Newton.

2.14. *Question 14*

QUESTION: Sketch the SED from the radio to gamma of extragalactic radiation on large angular scales. Describe the source and emission mechanism for each feature.

[16] Reverberation mapping uses the coupling between the AGN SMBH and the dense clouds to determine the extent of the BLR. Due to light lag the coupling will naturally have a delay, which can be measured by careful timing of variations in the AGN continuum emission versus the strengths of different lines. Different lines will show different couplings, since hotter clouds will tend to have more high-energy emission lines.

Kissin (2012) gives good information on this topic.

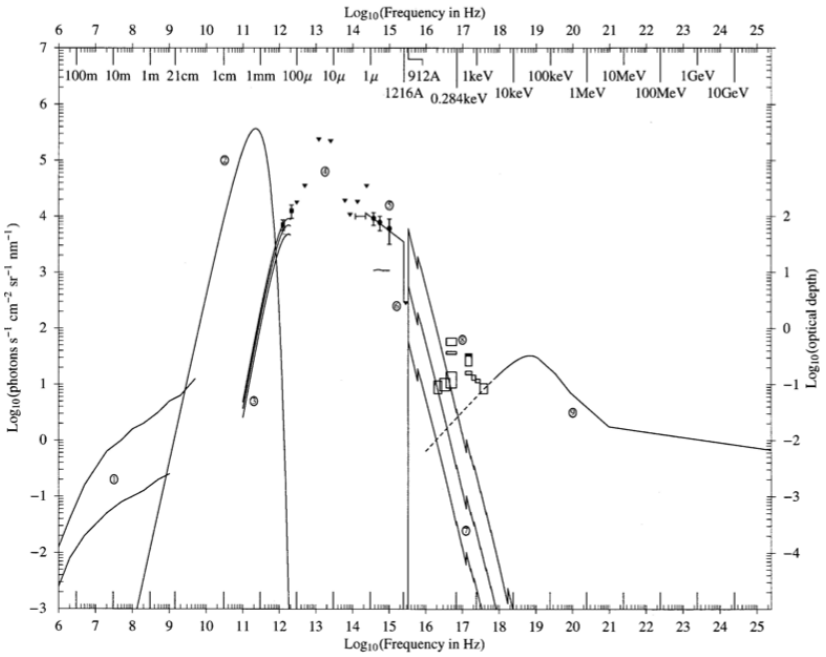

FIG. 42.— The background radiation SED of the universe: (1) radio, (2) CMB, (3) FIRAS excess, (4) DIRBE background (points with error bars), (5) optical background, (6) UVB, (7) the ISM photoionization optical depth (right-hand scale) for 10^{19}, 10^{18}, and 10^{17} H atoms cm^{-2}, (8) soft X-ray background, and (9) high-energy background. From Emberson (2012).

The cosmic background is due to a combination of a uniform background and a large number of point sources. It can be seen in Fig. 42, but a better, labelled picture can be found in Fig. 34, right. The cosmic background has numerous peaks and troughs, each of which pointing to a different emission mechanism and source.

The cosmic background is split up into:

- **The cosmic radio background** (CRB): follows a power law in its spectral flux density, with $\alpha \approx 0.6$. This suggests synchrotron radiation. Since normal galaxies are not strong radio emitters, and diffuse IGM would substantially inverse Compton scatter the CMB (not seen), it is likely this emission is due to quasars, radio supernovae and star-forming regions in galaxies. Current estimates (from counting discrete sources) suggest quasars and radio supernovae are minor contributors. To account for the rest of the CRB, star-forming regions would produce a much larger far-IR background than observed. This suggests another, hitherto unknown, component of the CRB, or a fundamental shift in the nature of the sources at higher redshift.
- **The cosmic microwave background** (CMB): this is due to photons escaping during last scattering at $z \approx 1100$. It has been redshifted from its original $\lambda \approx 1$ μm to today's 1.1 mm. The CMB is a near-perfect blackbody centred at 2.73 K.
- **The cosmic infrared background** (CIB): traditionally difficult to observe because of strong IR absorption in the atmosphere and emission from the MW, the CIB is likely thermal emission from heated dust in regular galaxies at low redshifts, LIRGs ($10^{11}\ L_\odot \lesssim L_{\rm IR} \lesssim 10^{12}\ L_\odot$) at $z \sim 1-2$ and ULIRGs ($L_{\rm IR} \gtrsim 10^{12}\ L_\odot$) at $z \gtrsim 2$. Obscured AGN provide a minor component as well.
- **The cosmic optical background** (COB): traditionally difficult to see because of airglow and zodiacal light. Most of it is likely thermal emission from ordinary stars in galaxies with $z < 10$. AGN accretion, supernovae and particle decay make up minor components.
- **The cosmic ultraviolet background** (CUVB): difficult to see for the same reasons as for the COB, as well as reflection of UV light from stars off of the ISM. Extragalactic UV is actually quite low. The main contributor is likely UV light scattered off the ISM in other galaxies, and hot IGM (thermal bremmstrahlung), though some speculate that redshifted Ly-α emission from recombination might contribute somewhat.
- **The cosmic x-ray background** (CXRB): discussed in Sec. 2.7.
- **The cosmic γ-ray background** (CGB): produced by a combination of blazars (beamed synchrotron) and cosmic ray/ISM interactions in star forming galaxies (pion production). Dark matter annihilation may also be a source.

2.14.1. *Are there non-EM backgrounds?*

The cosmic neutrino background is predicted from neutrino decoupling at $z \sim 10^{10}$. Because they were never heated by electron-positron annihilation at the end of lepton pair production, their current temperature, ~ 1.9 K, is even lower than that of the CMB. Because neutrinos at these energies have tiny cross-sections, they are nearly undetectable.

The cosmic gravitational wave background is predicted from a combination of discrete sources (inspiralling binaries, mergers, supernovae) and primordial GWs.

2.15. *Question 15*

QUESTION: What are AGNs? Describe different observational classes of them and how they may relate to each other.

This is mostly from my own notes.

An active galactic nucleus is a small region at the centre of a galaxy that shows highly luminous non-thermal emission that is strong over a broad range of the EM spectrum (often clear from radio to gamma). This is significant because if one were simply superimpose the spectra of the stellar population in these galaxies, the resulting spectrum would generally only stretch from 4000 Å to 20,000 Å (modulo recent star formation, which would add more UV, and dust extinction, which would bump up the far-IR). Moreover the luminosity of these central regions are often on the same order of magnitude as all other sources of radiation in the galaxy, making the source of this luminosity unlikely to be dense stellar populations.

The various classes of AGN are unified by the idea that all AGN emission is ultimately caused by accretion of material by supermassive black holes (SMBHs). Observational differences between AGNs can be explained due to orientation and line-of-sight extinction between their respective SMBHs and observers.

Galaxies are said to have AGN based on the following properties:

- Variability across the spectrum, with amplitude and frequency of variation increasing with respect to time.
- Extended radio emission, often in the form of jets.
- Continuum emission broadly described by a single power law ($S_\nu \propto \nu^{-\alpha}$). Emission is often highly polarized (in the core not as much, but in the extended regions up to 30%)- this fact, combined with the spectral indicies, suggests relativisitc synchrotron emission. This component decreases with weaker AGN.
- Extremely broad emission lines: the broad-line emission regions of AGN cores often have line FWHMs of $\sim$ 10000 km/s, while the "narrow'-line region line FWHMs are still around $\sim$ 100 km/s. These components become difficult to observe with weaker AGN.
- Low local space density: Seyferts make up $\sim 5\%$ of spirals, while quasars are extremely rare locally.

AGN contain the following subclasses:

- **Seyfert galaxies** are ordinary spiral galaxies that also appear to have very luminous nuclear regions with much broader emission lines than expected from normal spirals.
  - **Seyfert 1 galaxies** contain both broad and narrow lines. Their overall spectra resemble those of QSOs, and Seyfert 1 and weaker QSOs form a continuum (with the only difference between the two classes being their total luminosity; QSOs are 21 magnitudes brighter than Seyfert 1s).
  - **Seyfert 2 galaxies** contain only narrow lines. Seyfert 1 and 2 galaxy sub-classes form a continuum; Seyfert 1.5 or 1.8 galaxies are sometimes referred to.
- **Radio galaxies** are ordinary ellipticals that have particuarly strong radio emission. They spectroscopically resemble radio-loud Seyferts.
  - **Broad-line radio galaxies** (BLRGs) contain both broad and narrow lines. A continuum exists between BLRGs and quasars.
  - **Narrow-line radio galaxies** (NLRGs) contain only narrow lines. A continuum exists between NLRGS and BLRGs.
- **Quasars** are point-like in the optical (originally misidentified as galactic stars), with a very blue spectrum and broad emission lines. They tend to have high redshift, meaning they are extremely bright, especially for their size.
  - **Radio-loud quasars** (or quasi-stellar radio sources; QSRs) have strong radio emission.

  - **Quasi-stellar objects** (QSOs) are identical to quasars in the optical, but have very little radio emission. Found (on purpose) by photometric surveys looking for quasars in the optical. A continuum exists between quasars and QSOs, leading to the two terms sometimes being used interchangeably.

- **Blazars** are very similar to OSOs, except they appear to have very rapid (often over days) and greatly varying optical emission. This emission is also significantly more polarized (several % polarized) compared to that of QSOs ($\lesssim 1\%$).
  - **Optically violent variables** (OVVs) are QSOs that show substantial variation in the optical over very short periods of time.
  - **BL Lacertae objects** (BL Lacs) resember OVVs (fast, extreme variation in the optical, polarized emission) but have spectra completely devoid of emission lines. During their epochs of low luminosity, some BL Lacs will actually appear to morph into OVVs, suggesting a continuum between OVVs and BL Lacs.

- **LINERs** (low ionization nuclear emission regions) are not traditionally considered AGN (they, for example, are left out of Schneider), but are still characterized by line strengths that are hard to reproduce with a superposition of stellar spectra, combined with a slight UV excess. LINERs appear to be lower-luminosity versions of AGN, and there perhaps is a continuum between LINERS and Seyferts or radio galaxies. They are fairly common in the universe: over half the all local spirals show LINER activity.

See Fig. 43.

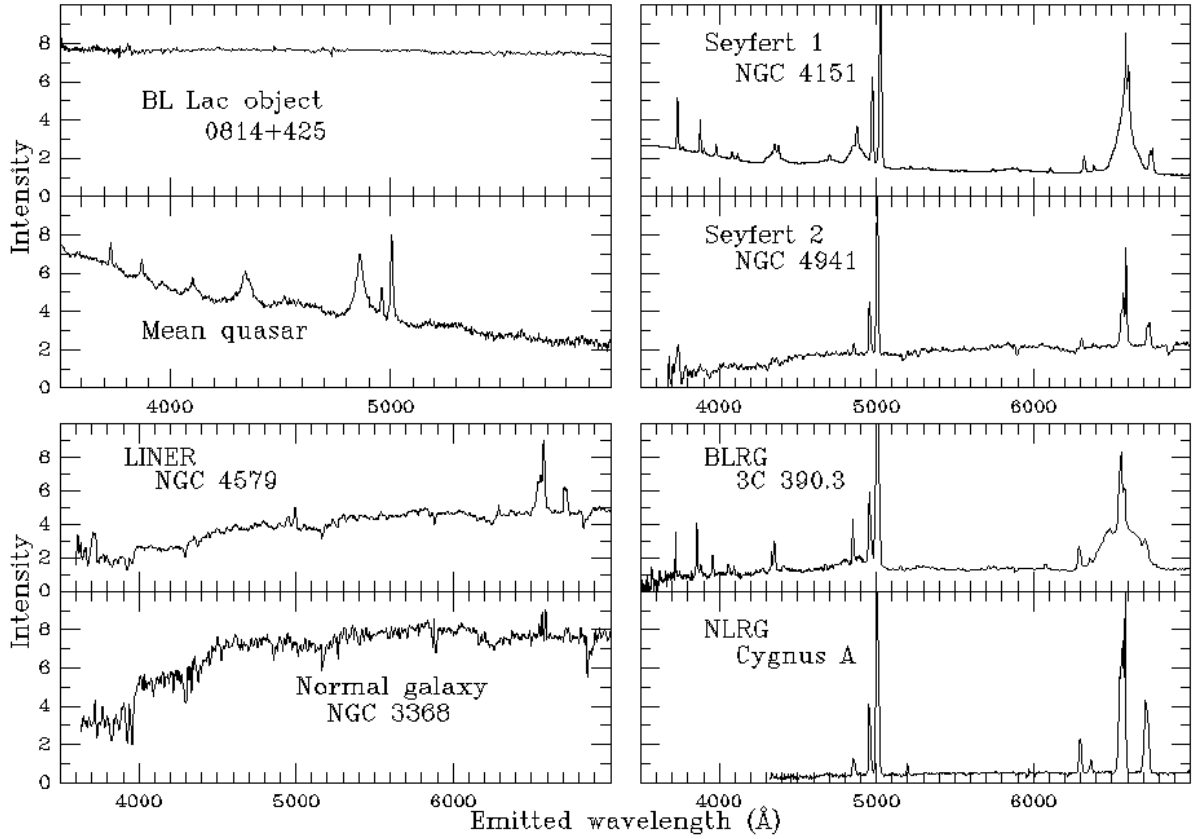

FIG. 43.— A comparison of optical ($\sim$ 300 - 600 nm) spectra of various classes of AGN. From Keel (2002).

Since AGN are all dominated by their SMBHs, it is to be expected that all AGN can be described by some combination of accretion rate ratio $\dot{m}/\dot{m}_{\rm Edd}$ and mass of the black hole M. M can explain the transition from Seyfert Is to QSOS, and BLRGs to quasars.

Discussion above also suggest a high degree of anisotropy in emissions, which suggests that absorption and scattering can explain the transition between certain AGN types. This is further reinforced by polarization measurements, which show narrow-line AGN will often also have fainter, highly polarized broad line emission hidden within their SEDs (which can be isolated using a polarizer). This suggests that the difference between a narrow-line radio galaxy or Seyfert and their broad-line counterparts is the obscuration of the broad-line emitting regions by dust and gas. We would also expect "type 2" quasars and QSOs. Chandra and XMM Newton recently discovered QSOs enshrouded

in regions with high hydrogen column density (as a result soft x-rays suffer high amounts of extinction); these QSOs contribute to the cosmic x-ray background issue described earlier. ULIRGS (ultra-luminous infrared galaxies) may also be type-2 QSO candidates, as they have similar luminosities but emit almost entirely in the IR.

Relativistic beaming affects AGN jets. If we suppose rest frame emission is isotropic (and for a cloud of electrons it should be) and the jet moving along some direction at velocity v, an observer viewing the jet at an angle ϕ with respect to the jet direction of motion will see the emission boosted by:

$$\left(\frac{1}{\gamma(1-\beta(1-cos\phi)}\right)^{2+\alpha}, \tag{70}$$

where $\beta = v/c$ and α is the rest frame emission spectral index. This boosting greatly increases the flux we recieve from an incoming jet, and equivalently decreases the flux we recieve from a jet moving away from us. This explains why, in general, only one jet can easily be detected [17]. This beaming is what makes blazars so bright, and fact that blazars can be divided into OVVs and BL Lacs is due to the strong dependence on angle near $\phi = 0$. Synchrotron (as noted earlier) is also polarized, which explains why blazars have polarized emission. Relativistic beaming also amplifies changes in luminosity, so any minute change in the AGN (which can occur quickly) equals a large change in the resulting emission.

AGN jets plow through their host galaxies, eventually dispersing into enormous radio lobes. It is believed that the synchrotron from the extended lobes is due to electrons being accelerated to relativistic speeds by shock-heating of the ISM and IGM by the jets.

How to connect radio loud and radio quiet systems is not entirely resolved, since there is more than one way to unify AGN. Schneider suggests that it may be due to galactic morphology (radio galaxies are ellipticals, while Seyferts are spirals). While there is a link between galactic and AGN luminosity, it is not obvious what morphology has to do with anything. Another proposal (also seen in Schneider, and Fig. 44 is suggestive of this as well) is that radio-loud/quiet has something to do with spin of the black hole - since accretion jets are created by the winding of magnetic fields, this is an understandable relation.

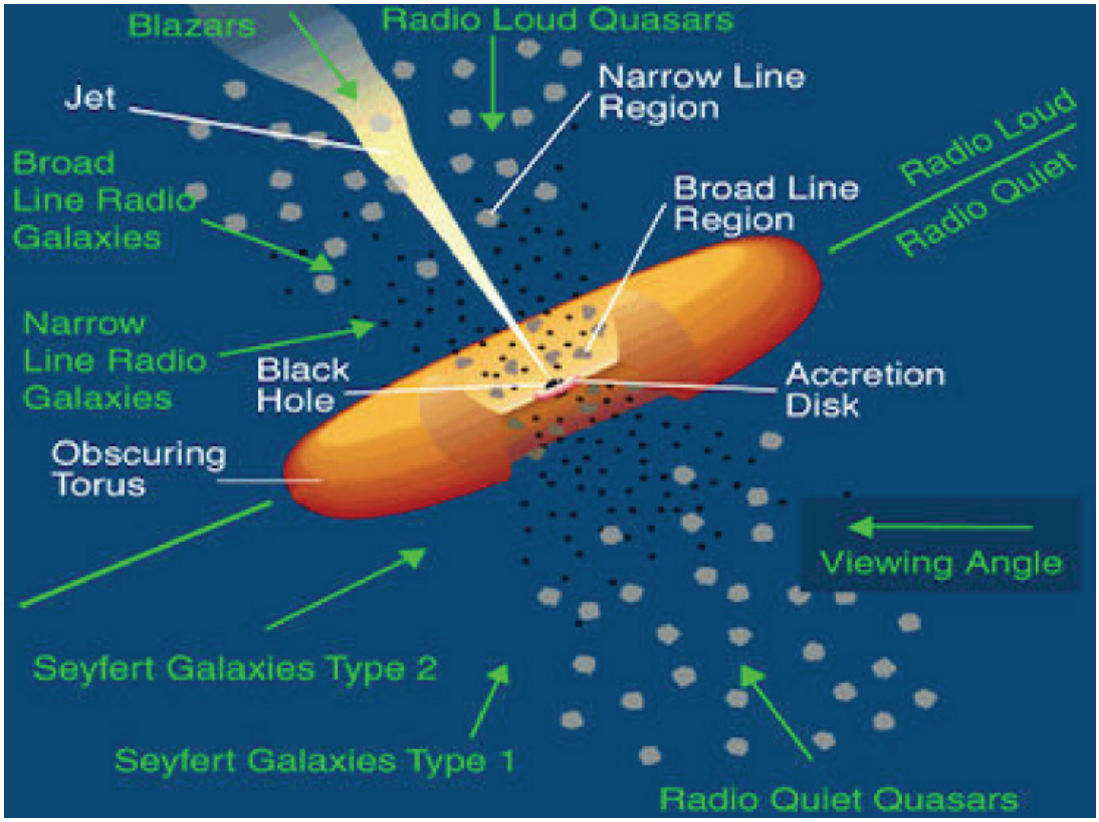

FIG. 44.— Scheme for unifying AGN. The progression from narrow-line galaxies to strong line to quasar/QSO to blazar depends on viewing angle. The difference between Seyferts and radio galaxies is explained as a matter of a lack of synchrotron emitting accretion jets (perhaps due to the spin of the SMBH?). From Abraham (2011a).

[17] The same is true at the kpc scale; the (still relativistic) jet coming toward us is far brighter than the jet going away. This is corroborated by the fact that in cases where two jets are seen, the faint jet appears to have undergone additional Faraday rotation compared to the bright jet, indicating it went through more of the galactic ISM to reach us.

2.15.1. *Where are most AGN located?*

Quasars and QSOs in particular have a space density vs. redshift distribution that peaks at $z \sim 3.3$, and tails off on either side. The reason for this is likely galactic evolution of some kind (ex. at higher redshifts there are fewer large black holes, at closer redshifts these black holes are not fed as well).

2.15.2. *What evidence is there that AGN are powered by supermassive black holes?*

This is covered in other questions as well, but we will treat it in detail here.

The size and structure of the jets/lobes of AGN requires an enormous power source that "points" in a constant direction for a time on the order of 10^7 yr. Bright AGN have luminosities up to 10^{47} erg/s; if this were constant for 10^7 yr 10^{61} erg is required, an enormous source of energy. AGN luminosity may change significantly on the timescale of hours; since information propagates at the speed of light global variations in emitted power must be the result of changes in a very compact source ($R \approx 10^{15}$ cm if we use $\Delta t = R/c$, where Δt is the variation timescale.

It is essentially impossible for this nuclear source to come from, for example, a dense star cluster. Fusion, at best, can derive $0.008 m_{\rm P} c^2$ worth of energy per nucleon - this translates to $> 10^9$ $M_\odot$ required if one assumes an AGN lifetime of $> 10^7$ yr. The Schwarzschild radius of a 10^9 $M_\odot$ system is $R_{\rm S} \approx 10^{15}$ cm, the same order of magnitude as the compact source radius derived above (and much larger than the $R_{\rm S}$ of a 10^6 $M_\odot$ SMBH). Strong gravitational effects cannot, then, be avoided even if a black hole is not assumed.

The only other reasonable power source is gravitational potential energy via accretion. Accretion onto a black hole, for example, has an efficiency ϵ (i.e. in ϵmc^2) of anywhere between 6% (for non-rotating BHs) and 30% (BH with maximum allowed angular momentum). Luminosity from steady-state accretion is Eddington luminosity limited; if we assume $L = L_{\rm Edd}$, we obtain, by inverting Eqn. **??**, a mass estimate of 10^6 $M_\odot$ for Seyferts and 10^8 for QSOs. Such a large mass in a region so small would require that the mass be contained in a supermassive black hole (SMBH); this conclusion is also consistent with the fact that using gravitational potential energy would still require the dumping of 10^9 $M_\odot$ into a tiny region of space.

An SMBH power source is also consistent with a number of other observations. Firstly, the "big blue bump" mentioned in a previous question can be explained as blackbody radiation from the superheated accretion disk. The bulk motion of radio jets is also highly relativistic (from direct observations and the requirement for shock-heated electrons to produce synchrotron radiation in the lobes), and in many astrophysical processes the velocity of ejected material is on the same order of magntude as the escape velocity of the system. Also, relativistic jets suggest escape from a black hole. Rotating SMBHs naturally act like gyroscopes, allowing them to create straight jets. X-ray emission line profiles can also be explained by emission conditions near the SMBH event horizon; see Fig. 45.

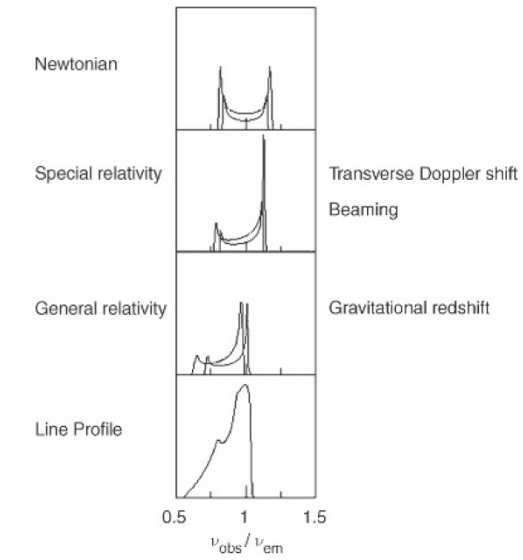

FIG. 45.— Line emission from near the event horizon of the SMBH, and how the emission is modified by both special relativity (the emitter is moving relativistically) and general relativity (the emitter is near the SMBH event horizon). From Schneider (2006), his Fig. 5.16.

2.16. *Question 16*

QUESTION: What are galaxy clusters? What are their basic properties (eg, mass, size). List and explain three ways they can be detected.

Galaxies are not uniformly distributed in space, but instead show a tendency to gather together into gravitationally bound collections known as galaxy groups and galaxy clusters. The transition between groups and clusters of galaxies is smooth. The distinction is made by the number of their member galaxies. Roughly speaking, an accumulation of galaxies is called a group if it consists of $N \lesssim 50$ members within a sphere of diameter $D \lesssim 2$ Mpc. Clusters have $N \gtrsim 50$ members and diameters between 2 and 10 Mpc. Typical values for the mass of a cluster are $M \sim 10^{14}\ M_\odot$ for massive clusters, whereas for groups $M \sim 10^{13}\ M_\odot$ is characteristic, with the total mass range of groups and clusters extending over $10^{12}\ M_\odot \lesssim M \lesssim 10^{15}\ M_\odot$. Alongside the cluster there is a substantial amount of intercluster medium (diffuse gas), and due to the massive gravitational potential a cluster this gas is hot, and the galaxies have high pecular velocities (~ 1000 km/s). In fact, on average galaxies make up only a few percent of the total mass of a cluster, while ICM makes up about 10% and dark matter about 90% (Wikipedia 2012e).

Methods of detecting galaxies include:

- **Looking for clustering in the optical**: pioneered by Abell in the 1950s. Clustering is detected by requiring a large number (in Abell's case, ≥ 50) of galaxies in a magnitude range ($m_3 \leq m \leq m_3 + 2$, where m_3 is the apparent magnitude of the third brightest galaxy in the cluster) within a certain angular radius ($\theta = 1.7'/z$). Abell determined redshift by assuming that the 10th most luminous galaxy in any given cluster has the same luminosity as one in any other. While Abell was limited by the sensitivity of photographic plates, modern surveys such as SDSS can probe up to significant redshifts using this method. Redshifts are then determined photometrically for each galaxy to disentangle projection effects and determine true 3D clustering.

- **X-ray emission**: hot gas inside clusters have characteristic thermal bremsstrahlung luminosities of 10^{43} erg/s in the x-ray alone, spread over a size of several Mpc. X-ray surveys looking for such emission are not mired by projection effects like optical surveys are (i.e. redshifts are unecessary).

- **The Sunyaev-Zel'dovich effect** (SZ effect): inverse-Compton scattering of the CMB by hot electron gas. This effect is not redshift-dependent, and therefore is extremely useful (if it can be measured, which is difficult) for determining the locations of high-redshift clusterse. See Sec. 2.18.

- **Weak gravitational lensing**: weak lensing is often used to map the matter distribution of known clusters, or to look for cosmic shear (Sec. 1.15), but it can also be used to find clusters. The advantage to using lensing is that it traces the true mass of the cluster; if a cluster contained very few stars and gas, it could still be detected by lensing. Lensing is subject to projection effects: it may not always be reasonable to assume only one lens along the line of sight.

- **Red cluster sequence** (RCS): early-type galaxies belonging to a cluster have very similar colours, only weakly depending on their luminosity (this is a metallicity effect - more massive galaxies are more metal-rich). These colours become bluer at higher redshifts - this correlation is so good that redshift can actually be determined using it. This relationship exists because the stars contained in the centres of clusters finished forming soon after the birth of the universe, and star formation has been minimal since. Strips of galaxies in CMDs of galactic surveys therefore indicate clusters. See Fig. 46.

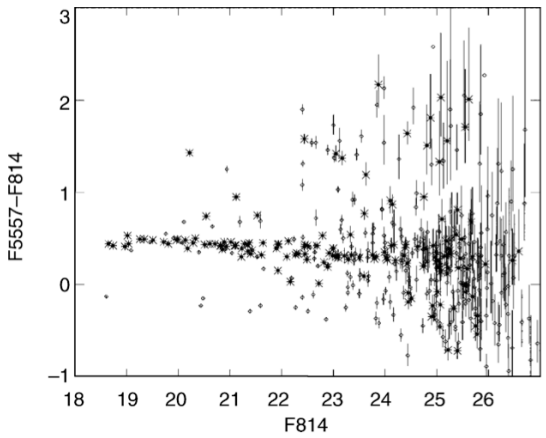

FIG. 46.— Colour-magnitude diagram of the cluster of galaxies Abell 2390, observed with the HST. Star symbols represent early-type galaxies, identified by their morphology, while diamonds denote other galaxies in the field. The red cluster sequence is clearly visible. From Emberson (2012).

2.16.1. *How are high-z cluster found?*

RCS and the SZ effect are largely redshift-invariant, and therefore work well to find clusters at high z. It can also be assumed that AGN track clustering, and deep imaging using narrow-band photometry centred around $1216\,\mathring{A}(1+z)$ (to pick out rest-frame Ly-α emission from galaxies) have found high-z clusters.

2.17. *Question 17*

QUESTION: Describe and give results from simulations of large scale structure in the universe. What role do they have in understanding the formation of large scale structure and Galaxy formation? What are their limitations?

Simulations have, in recent years, been instrumental in understanding large-scale structure formation in the universe. This is because, as a whole, gravitational dynamics is far too complicated to be explored fully with analytical or semi-analytical methods. In recent years, computers have achieved enough processing power for hydrodynamics to also be accounted for, and modern smoothed-particle hydrodynamics (SPH) simulations are routinely used to simulate the evolution of large-scale structure in the universe.

In a typical SPH code, we insert a series of particles, representing dark matter, gas and stars, into a box of comoving length L with periodic boundaries (often, dark matter alone is used, but modern simulations will also add in gas and stars). Each particle represents a certain mass, and so density is easily obtained by determining the number of neighbours within a certain volume around one particle. The choice of L limits the scale of the largest features the simulation can resolve (and so generally $L > 200h^{-1}$ Mpc). The periodic boundaries means that that we assume one side of the box connects to the opposite side (and particles that pass through one side of the box reappear on the other side) - in this way, we can estimate the forces due to objects outside the box. Forces are generally calculated using Newtonian physics, with the addition of a softening length (i.e. $F = GM^2/(r+a)^2$, where a is the softening length) to prevent artifical close-encounters between particles (since in reality we do not expect our matter to be composed of massive point sources). Calculating forces on a particle p is a process that takes an amount of time $\propto N^2$, where N is the number of particles. As a result, methods such as the Barnes-Hut tree or the particle-particle-mesh are used. Both methods bin masses at large distances from p[18]. Pressure forces can be calculated by using the particles as an irregular mesh to determine bulk properties of the fluid; the number of particles used in this mesh define a "smoothing length" - detailed variations in fluid properties cannot be captured at smaller resolutions. The initial conditions to the simulation can be determined by placing particles in accordance with a Gaussian random field with the power spectrum $P(k,z)$.

We can clearly see some limitations to the physical model described above:

- We have to make approximations to calculate gravity with reasonable speed. We also have to define a smoothing length to prevent gravity calculations from diverging. We have to interpolate over multiple particles to determine hydrodynamic forces. All these effects collectively set resolution limits on our simulation, and we cannot determine details more fine than this limit. If these details affect evolution on larger scales, we will not capture the evolution, either.
- A related problem is that we have defined set masses for our particles, which are much larger than the true individual masses of dark matter particles. This sets a mass limit - we cannot describe associates of less massive objects.
- We cannot discern phenomena that occur only on scales larger than L; if these affect smaller scales, we will not capture them.
- We generally use Newtonian dynamics. This is often fine, since typical speeds in simulations are $<< c$, and typical densities are below those where GR would significantly change things.
- Perhaps most importantly, it is extremely difficult to simulate magnetohydrodynamics, turbulence, feedback, radiation transport, and other physics (that generally act on small scales) in these simulations. The effects of them can be characterized by *ad hoc* insertions and simplifications, but this may generate biases in the simulations.

These simulations have shown that matter tends to form Mpc-spanning high density filaments that link together to form a web of material across the universe. Between them are large voids with densities much lower than the cosmological average.

[18] The tree code determines the nearest neighbours to a particle, which generates a network of neighbours from which closer and more distant objects to p can be determined; the particle-particle-mesh method determines the distribution of masses by binning masses into a 3D grid, and then taking a fast Fourier transform. This gives a smoothed mass distribution that can replace interparticle force calculations at large distances.

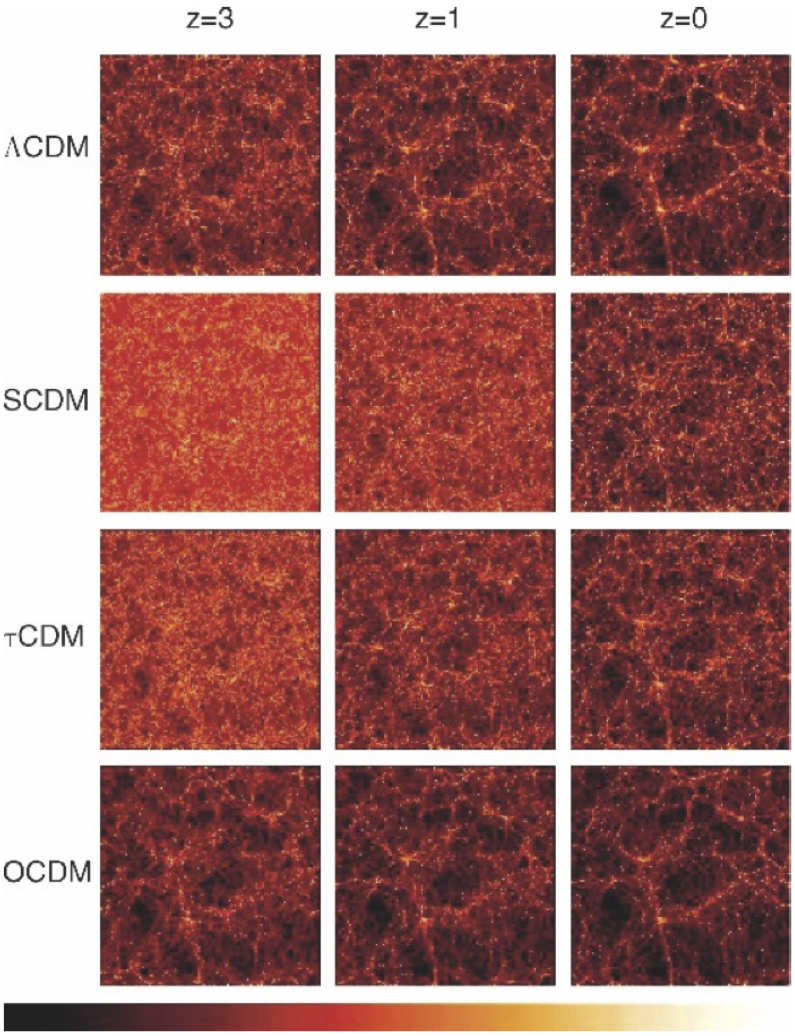

FIG. 47.— Simulation by the VIRGO Consortium using 256^3 particles in box of side length 240 Mpc/h for different cosmological models: $\Omega_{\rm m} = 0.3$, $\Omega_\Lambda = 0.7$ (ΛCDM), $\Omega_{\rm m} = 1.0$; $\Omega_\Lambda = 0.0$ (SCDM and τCDM); $\Omega_{\rm m} = 0.3$, $\Omega_\Lambda = 0.0$ (OCDM). The two Einstein-de Sitter models differ in the shape of their power spectra. Simulation parameters were chosen so that large scale structure at $z = 0$ is statistically identical between them - their differences lie at high redshift. From Emberson (2012).

On large scales simulations have allowed us to test ΛCDM cosmology against other universes. Fig. 48 shows that varying cosmological parameters will result in statistical differences between the matter power spectrum (and therefore correlation between galaxies) of the universe. This can be tested against observations of high-redshift galaxy clustering, and results have borne out in favour of ΛCDM.

Large scale simulations can also be used to test the Press-Schechter model for dark matter halo growth (by binning halos found in simulations by mass); simulations to date have shown Press-Schechter to, as a whole, be quite accurate, with numerically derived fitting formulae being only slightly different. This is astonishing, since Press-Schechter does not take into account non-linear evolution, while simulations do. Press-Schechter tended to underestimate the number of high-mass halos, and overestimate the number of low-mass ones. These very high mass halos collapse early in the simulations, and often continue to grow until they become the centres of massive galaxy clusters. If these halos represent quasars, they would support the claim that rare objects such as quasars can be formed in a ΛCDM universe.

Simulations can also be used to determine the two-point correlation function of galaxies, and how well this function traces the dark matter correlation function. As it turns out, there is a deviation at small scales.

On smaller scales, simulations have shown that dark matter halos have a universal density profile. Averaged over spherical shells, they resemble the Navarro-Frenk-White (NFW) profile:

$$\rho(r) = \frac{\rho_s}{(r/r_s)(1 + r/r_s)^2}. \tag{71}$$

The two free parameters ρ_s and r_s work together to determine both the total mass (inside r_{200}, the radius at which $\rho(r) = 200\rho_{\rm crit}$) and how concentrated the halo is. Other simulations have slightly different profiles, possibly due to their treatment of small scales. There is yet no analytical explanation for this profile; moreover, observations of stellar kinematics disagree with the NFW profile near the centres of galaxies (this is known as the "cusp-core war"). Another

issue with small scale numerical simulations is that for a given halo, a large number of sub-halos appear embedded in it. So far, despite efforts with microlensing, halo substructure within a galaxy has not been observationally confirmed. Moreover, one interpretation of these sub-halos is that they are satellite galaxies. If this were the case, the number of satellite galaxies around the MW and other local galaxies is far lower than expected from simulations (the "missing satellite problem").

2.18. *Question 18*

QUESTION: What is the Sunyaev-Zel'dovich effect and where is it detected? What are some challenges in using SZ measurements to determine the values of the cosmological parameters?

The Sunyaev-Zel'dovich (SZ) effect occurs when high-energy electrons (often from the ICM, which, as we have noted in other questions has temperatures around 10^7 K; high-energy electrons can also be found in the diffuse IGM, and near the MW) inverse Compton scatter low-energy CMB photons to higher energies. Since this scattering is isotropic, the number of photons headed toward us from the CMB remains fixed; it is the energy of each photon that differs. See Fig. 48.

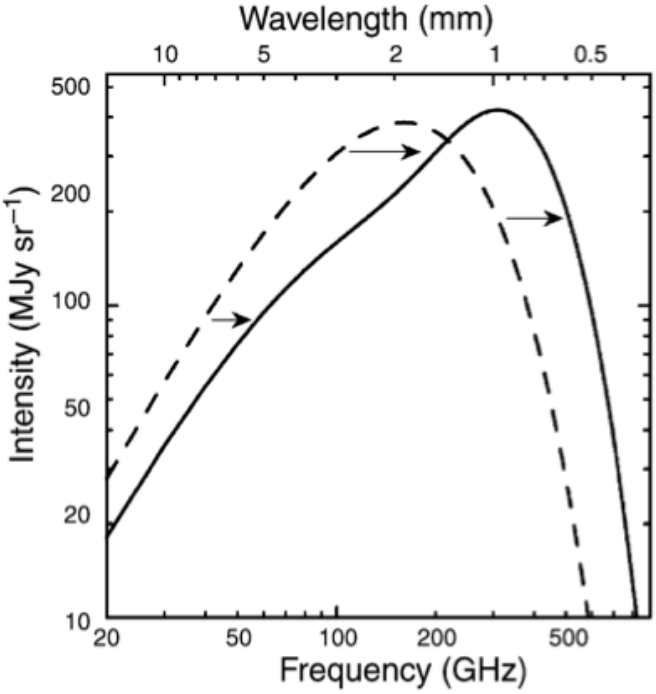

FIG. 48.— The influence of the SunyaevZeldovich effect on the cosmic background radiation. The dashed curve represents the Planck distribution of the unperturbed CMB spectrum, the solid curve shows the spectrum after the radiation has passed through a cloud of hot electrons. The magnitude of this effect, for clarity, has been very much exaggerated in this sketch. From Emberson (2012).

In the Rayleigh-Jeans domain of the CMB, the change imposed by the SZ effect can be parameterized by

$$\Delta I_\nu^{RJ} = -2yI_\nu^{RJ} \tag{72}$$

where y is the Compton-y parameter, $y \propto T_g n_e L$ (very roughly; see Eqn. 6.45 of Schneider (2006)), where L is the transverse length of the hot gas. Suppose we can also determine the x-ray emission intensity I_X, which is $\propto L n_e^2$. We can therefore eliminate n_e, and obtain

$$\Delta I_\nu^{RJ} \propto I_\nu^{RJ}\sqrt{LI_X}. \tag{73}$$

If the cloud of hot gas is spherical, then its radius, $R = L$. Interferometry in the submillimetre can now spatially resolve the SZ effect, which allows us to determine the angular diameter distance

$$d_A = \frac{R}{\sigma} \approx \frac{L}{\sigma} \propto \frac{\Delta I_\nu^{RJ}}{I_\nu^{RJ}}\frac{1}{I_X\theta} \tag{74}$$

d_A can then be used to determine the Hubble parameter, and therefore the expansion history of the universe, if the

redshift of these clusters is also known. At small redshift d_A is a function of the deceleration parameter, which allows us to constrain Ω_m and Ω_Λ.

Since the SZ effect is proportional to the electron number density, if we also knew a cluster's total mass, we could measure the baryon fraction of the cluster. SZ surveys could therefore determine Ω_b/Ω_m.

Difficulties with using this method include:

- The SZ effect is a weak signal ($\Delta T \sim 1mK$) that dominates the CMB only at small angular scales associated with secondary anisotropies.
- The SZ effect can only be identified using multiple band photometry, which is costly in the sub-millimetre. The combination of this and the above issue make SZ surveys time and resource-intensive.
- Assumptions made about the ICM itself, such as its column density and temperature structure, will change the results we obtain.

There are enough inherent inaccuracies with this method that it cannot compete with CMB angular fluctuations coupled with BAO in determining H_0. This method, and other lower-accuracy methods, are more useful as consistency checks.

2.18.1. *What are SZ surveys conducted with?*

This information is from Wikipedia (2012f).

The first attempted measurements were with single-dish telescopes in the 1970s. Recent efforts to conduct full-scale SZ surveys have been done with the Planck satellite, the South Pole Telescope (SPT) and the Atacama Cosmology Telescope (ACT).

2.18.2. *What is the kinematic SZ effect?*

The kinetic SZ effect is caused by Doppler shifting of CMB photons that Thomson scatter off ionized gas clouds possessing some non-zero bulk velocity relative to the CMB rest frame. This leads to temperature fluctuations corresponding to hotspots in the CMB if the ionized gas is moving toward the observer, and cold spots if the gas is moving away. The kSZ maintains the blackbody shape of the CMB, and is dependent on the peculiar velocities in the IGM rather than temperature of the gas. As a result, the kSZ can be used to trace the overall structure of the IGM. **IS THIS POSITION OR ALSO VELOCITY?** The kSZ is much weaker than the thermal SZ effect described above. Recently, ACT made the first detection of kSZ.

3. GALACTIC ASTRONOMY (INCLUDES STAR FORMATION/ISM)

3.1. *Question 1*

QUESTION: What is a stellar Initial Mass Function (IMF)? Sketch it. Give a couple of examples of simple parametric forms used to describe the IMF.

The stellar initial mass function is the function describing the number of stars of mass M formed per unit mass (sometimes per unit volume or area) of a patch of stars, denoted $\phi = \frac{dN}{dM}$. The function is normalized so that $\int_{m_l}^{m_h} M\phi(M)dM = 1$ $M_\odot$, i.e. one solar mass of stars is formed per solar mass of star-forming material. Sometimes, the cumulative IMF is also used, which describes the number of stars formed per unit mass of molecular cloud below or above a certain mass. The IMF assumes no star die-off (hence the "initial"), but is cumulative, i.e. $n = \int \phi dM$ is the number density of stars *that have ever formed* per solar mass of stars *that have ever formed* (Chabrier 2003). Under conditions when the IMF does not change with time, then the mass distribution, per Solar mass that goes into forming stars, of stars formed at a particular time t can also be represented by the functional form of the IMF, since the two will differ only by a scaling factor (Chabrier 2003). If the distribution of star masses being born changes over time, however, the present-day IMF would take into account the entire history of starbirth in the galaxy.

The classic Salpeter IMF is

$$\phi(M) = AM^{-2.35} \tag{75}$$

where A is a constant of integration. This is useful and well-known for stars above 0.5 $M_\odot$ (from looking at clusters), but terrible for stars below this number ("bottom-heavy"). Most modern IMFs used have at least two regimes (one above and one below $\sim$1 $M_\odot$).

The MS79, Scalo and KTG93 IMFs in Fig. 49 were based on galactic disk measurements, which cannot be used to accurately infer the high-mass end because of the complicated SFH of the galaxy ("top-heavy"). Measured IMFs within star clusters generally give a shallower IMF close to the Salpeter value. The Kennicutt, Kroupa01, BG03 and Chabrier IMFs in Fig. 49 are the best bet for reasonable mass-to-light ratios and galaxy colors. The BG03 analysis favoured a slope shallower than Salpeter at the high-mass end based on constraints from local luminosity densities and cosmic star-formation history; IMFs with high-mass slopes steeper than Kennicutt's were ruled out as a universal IMF.

The Chabrier IMF, commonly used today, is a lognormal distribution:

Kroupa et al. (2011) claims that since no stochastic process exists in star formation, the IMF should not be thought of as a probability distribution.

3.1.1. *What is the IMF useful for?*

The IMF is an integral part of population synthesis, which is a fundamental aspect of a very, very large number of fields. These include galaxy formation and evolution, transient frequency prediction, cosmology (especially reionization) and dark matter searches. It is also used for understanding star formation (i.e. the end result of all star formation is to determine how to generate a ZAMS population consistent with nature; or, in reverse a good handle on the IMF can constrain star formation theories).

3.1.2. *Is there a universal IMF?*

Theoretically one does not expect the IMF to be universal for any possible molecular cloud condition. There are two simple reasons, given by Kroupa et al. (2011):

- The Jeans mass of a low metallicity cloud is *ceteris paribus* larger than the Jeans mass of a high metallicity cloud. This is due to the fact that metals act as an important source of opacity for a molecular cloud, and thus metal-poor clouds are more inefficient at cooling.
- Adams & Fatuzzo reject the Jeans mass argument, claiming that a turbulent cloud has no preferred Jeans mass. They instead suggest that the zero-age main sequence mass of a star is set by the balance of the accretion rate onto the protostar and its circumstellar disk, and feedback from the star. As it turns out, in low metallicity environments photon-gas coupling is not as strong, decreasing the effective amount of feedback and resulting in higher-mass stars being formed. An increased temperature from reduced cooling alos increases the soundspeed, which increases the accretion rate.

In the most extreme case of metal-free population III stars, the IMF is expected to be extremely top-heavy, with average stellar masses in the hundreds of $M_\odot$.

Strong observational evidence, however, does not yet exist for IMF variation. The first tentative evidence came with De Marchi et al. 2007, which found that higher metallicity, less densely concentrated globular clusters had a paucity of low mass stars, contrary to what would be expected from stellar evaporation. The most reasonable explanation is that because metal-rich gas couples better to radiation, gas ejection in metal-rich GCs is more efficient than in metal-poor GCs. Also, metal-poor clusters form more compactly, while metal-rich clusters form more loosely due to

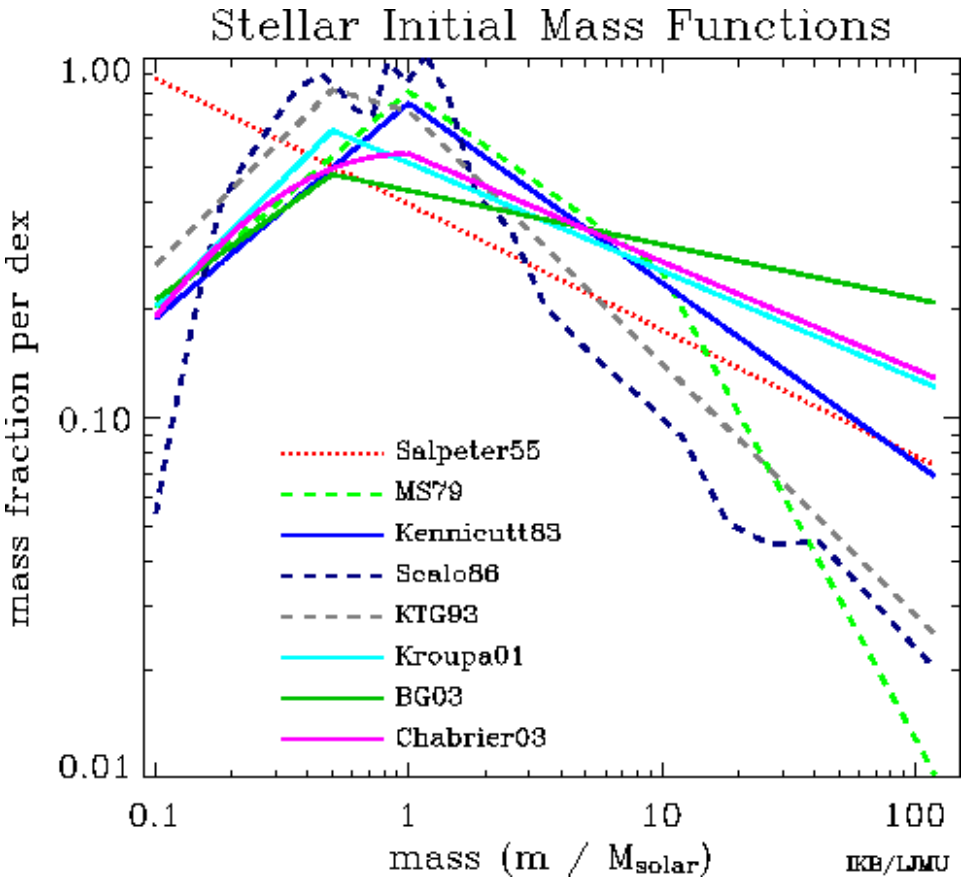

FIG. 49.— Comparison of a select sample of IMFs, normalized so their integral from 0.1 to 120 $M_\odot$ is 1. Note that the y-axis is $dn/d(logM)$, and not dn/dM, which accounts for why the more modern functions actually have a positive slope. From Baldry (2008).

increased fragmentation, making metal-rich clusters less gravitationally bound. These two effects combined suggest that to generate enough feedback for gas expulsion in metal-poor GCs, a top-heavy IMF, predicted by theory, is required. While De Marchi et al. formulated their conclusions from observations of 20 MW GCs, a survey of 200 GCs in Andromeda gives similar results. Work done on ultra-compact dwarves yield similar conclusions (Kroupa et al. 2011, pg. 131 - 132). Observations of entire galaxies suggests the Milky Way's IMF should be top-light if the stellar IMF were to remain invariant.

Note also that there are a large number of ways to "hide" changes to the IMF, including uncertainties in stellar evolution, external forces (such as tides) that make the IMF unobservable in clusters (see below) and instrumental limitations - it is, for example, difficult to observe the light from stars less massive than 0.5 $M_\odot$ in globular clusters.

3.1.3. *Why are the lower and upper limits of the IMF poorly understood compared to that of the middle (several $M_\odot$ stars)? What constraints are there?*

IMFs are generally created by observing the luminosity of individual stars, or the integrated luminosity of a region of space, and inferring masses from stellar evolution models. Direct mass determination can be used (ex. in the case of binary systems), but generally do not provide IMFs of good accuracy (Chabrier 2003, pg. 11). Therefore, uncertainties in stellar evolution translate to uncertainties in the IMF. Even with an IMF that does not change with time, a measurement of a field of stars (or even clusters, since clusters tend to eject stars and do not completely form before their most massive stars die) necessarily measures the present-day mass function (PDMF, below). This is referred to, tongue-in-cheek, as the "IMF Unmeasurability Theorem" by Kroupa et al..

We do not yet understand how extremely massive stars form. While both star formation and hydrostatic equilibrium become problematic past the Eddington luminosity, and runaway pulsational instability has been estimated to very roughly become catastropic at ~ 100 $M_\odot$, stars with ~ 150 $M_\odot$ have been observed (Kroupa et al. 2011). The lack of more massive stars in many star forming regions suggests that a drastic dropoff in $\phi(M)$ exists at $\sim 150 - 300$ $M_\odot$, though the exact number is not well constrained (Kroupa et al. 2011). Note that massive stars in close-in binaries can merge to form even more massive stars (Kroupa et al. 2011). (Star clusters will, of course, have limits set by the mass of the molecular cloud they were birthed from, and the star forming efficiency.)

The substellar IMF has, until recent years, been extremely difficult to probe because of low detection efficiency for M-dwarfs, and a complete lack of detection of brown dwarfs (BDs). In more recent years, observations of BD populations, along with N-body simulations, have shown that brown dwarfs are a distinct population from low mass stars, likely evidence of a formation history different than that of stars (Kroupa et al. 2011). The BD IMF, which

roughly connects with the stellar IMF at the H-burning limit, has a shallow slope of about -0.3, as opposed to the -1.3 of low mass stars (see Fig. 22 of Kroupa et al.).

3.1.4. *What's the difference between a field and stellar cluster IMF?*

Assuming the IMF is universal, there should be no difference between the IMF of a cluster and the IMF of a field stars.

In open clusters, differences arise due to intra-cluster dynamics such as the ejection of a large fraction of the cluster population, along with residual gas, early in a cluster's lifetime, and the density-dependent disruption of primordial binaries (Kroupa et al. 2011, pg. 13-14).

Globular clusters may also experience this gas expulsion, which might ejected less massive stars, early in their lives. GCs additionally have a tendency to sink massive stars to their centres, and eject lighter stars (this is known as evaporation), leading to a flatter mass function today than the cluster's IMF (Kroupa et al. 2011, pg. 70). (This obviously is not a problem for open clusters, which are not gravitationally bound and dissipate long before the several Gyr it takes for this effect to become significant.)

Gas expulsion and secular effects do *not* constitute changes in the IMF - they are changes in the PDMF of the clusters. These effects, however, make it so that the IMF never corresponds to any distribution of stars in the cluster at any time.

3.1.5. *How do you determine an a present-day mass function (PDMF) from an IMF?*

We define $C(M,t)dMdt$, where $C(M,t)$ is the creation function, as the number of stars per volume/mass formed that have masses from M to $M+dM$ during time t to $t+dt$. This means that, assuming an area of space has been forming stars since time τ_F, the minimum and maximum stellar mass that can be formed is fixed, and no stars have died, the current number density of stars is

$$n = \int_{M_{\min}}^{M_{\max}} \int_0^{\tau_F} C(M,t)dtdM. \tag{76}$$

If we integrate along M, we obtain the birth rate

$$B(t) = \int_{M_{\min}}^{M_{\max}} C(M,t)dM. \tag{77}$$

Following Chabrier (2003) we define $b(t) = \frac{B(t)}{\frac{1}{\tau_F}\int_0^{\tau_F} B(t)dt} = \frac{B(t)\tau_F}{n}$, noting that $\int_0^{\tau_F} b(t)dt = \tau_F$. Now, we shall suppose the creation function can be separated into two components, the initial mass function $\phi(M)$ and the birth rate B:

$$C(M,t) = \frac{\phi(M)B(t)}{n} = \phi(M)\frac{b(t)}{\tau_F}. \tag{78}$$

Integrating over time from 0 to τ_F gives us the initial mass function

$$\phi(M) = \int_0^{\tau_F} C(M,t)dt. \tag{79}$$

This is where the "all stars ever formed" comes from. Note that because of time separability, the mass function from a single period of star formation varies from the IMF only by a scaling factor - hence, the mass function of a single period of star formation is proportional to the IMF.

We now add in star death to the equation - a star of mass M only lasts for a lifetime $\tau(M)$ on the main sequence. This means that the total number of stars from mass M to $M+dM$ to be around currently is

$$\Phi(M)dM = dM\int_{\tau_F-\tau(M)}^{\tau_F} C(M,t)dt = dM\frac{\phi M}{n}\int_{\tau_F-\tau(M)}^{\tau_F} B(t) = dM\frac{\phi(M)}{\tau_F}\int_{\tau_F-\tau(M)}^{\tau_F} b(t)dt. \tag{80}$$

This equation defines $\Phi(M)$, the present day mass function. Note that $\Phi(M) \propto \phi(M)B(t)$.

3.2. *Question 2*

QUESTION: Describe the orbits of stars in a galactic disk and in a galactic spheroid.

The orbits of stars in the galactic disk are nearly circular, perturbed by some minor radial and z-axis oscillations. The orbits of stars in the spheroid (i.e. the Galactic halo) are randomly oriented, not restricted to travelling on a disk, and can be highly elliptical orbits that plunge through the disk. Orbits in the bulge are similar. See Fig. 50.

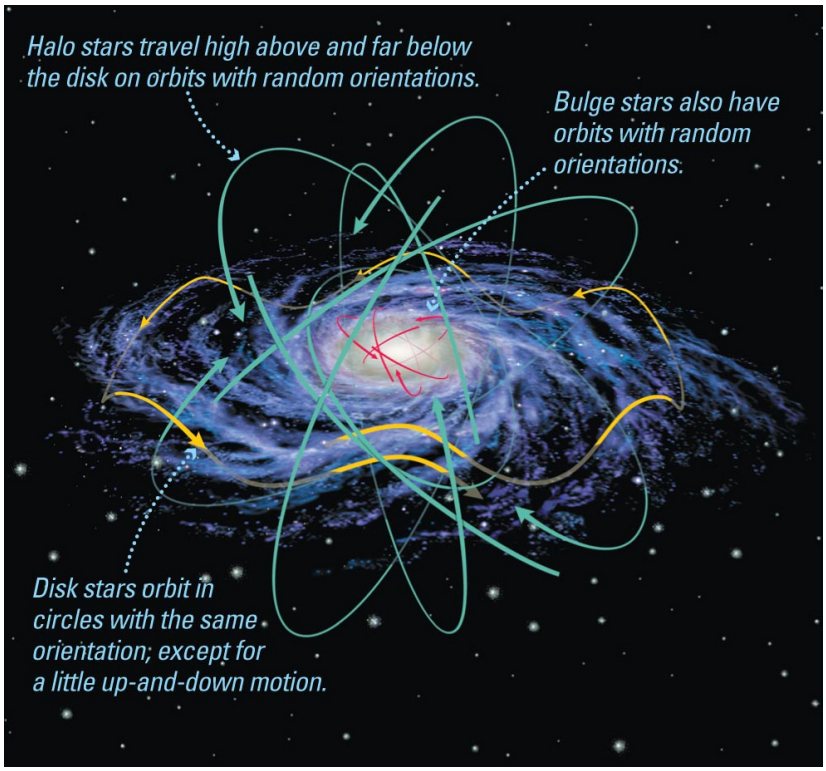

FIG. 50.— Characteristic stellar orbits in different regions of a standard disk galaxy. Green orbits are for halo stars, yellow are for disk stars, and red are for bulge stars. The disk stars orbit in roughly circular orbits, with some vertical oscillitory motion, while the bulge and halo stars do not have a standard inclination, and have orbits whose orientations precess. The figure is labelled, though fairly unhelpfully, as this is from a book for non-science students. From Bennett et al. (2007), their Fig. 19.2.

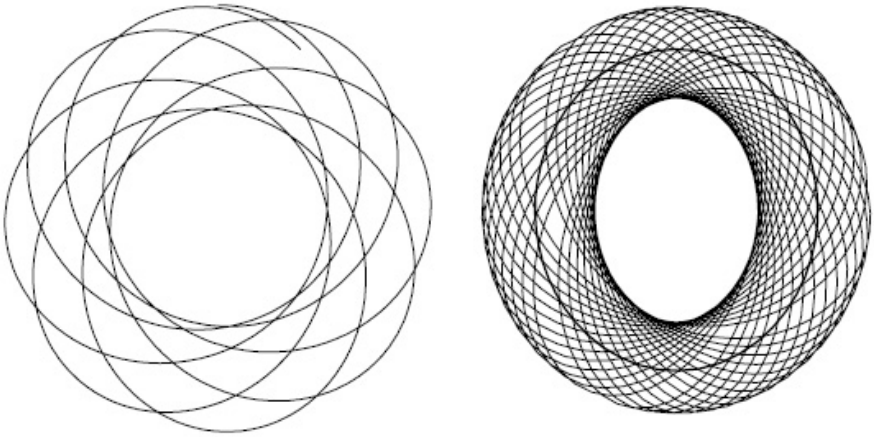

FIG. 51.— Left: a typical non-closed orbit in a spherically symmetric potential (the shape is commonly referred to as a "rosette"). Right: a typical orbit in an axisymmetric potential, commonly called a "loop" orbit. Both orbits were determined numerically. From Binney & Tremaine (2008), their Figs. 3.1 and 3.8.

3.2.1. *What Orbits are Allowed in an Axisymmetric Potential?*

Orbits within a spherical potential will in general not be closed, and will trace out a rosette (see Fig. 51, left panel). Stars confined to the equatorial plane of an axisymmetric galaxy have no way of perceiving that their potential is not spherically symmetric, and they will trace out rosette. A way of describing a nearly circular rosette is using the epicycle formulation. There exists no general orbital solution for an axisymmetric potential. If we restrict the potential of our galaxy to that of a flattened spherical potential, then, qualitatively, orbits exactly on the $z = 0$ will trace out a rosette. If the orbit carries the star slightly off the $z = 0$ plane, the rosette is augmented with an oscillitory motion in the z-direction (see section below). If the orbit is significantly above the $z = 0$ plane, then it can achieve a significant range of z-values.

3.2.2. *How did each population of stars gain their particular orbit?*

Halo and disk stars form two stellar populations that were birthed under different conditions.

Thin disk formation involves a dark matter halo forming gravitationally and then virializing; infalling baryonic matter then begins collecting in this potential. Since this matter loses copious amounts of gravitational potential energy, the baryons are shock-heated to high temperature. Radiative dissipation (ex. through bremsstrahlung) reduces the thermal motion of the baryons during and after collapse. Due to conservation of angular momentum, the cooling gas forms a centrifugally supported thin disk. Individual star formation will vary between galaxies, but for our galaxy the observational finding that most stars are moving at their circular motion velocity is consistent with most stars being formed after the formation of the disk. (Thick disk formation is not yet well understood; it potentially can come from perturbations of the galactic disk from infalling satellite galaxies.)

Modern heirarchical formation simulations suggest that halo stars come from two distinct populations. Outer halo stars likely originate in satellite galaxies that were accreted by the parent galaxy early in its history (indeed, the distribution of stars in the halo bears little resemblance to the distribution of satellite galaxies long after the creation of the halo). Individual tidal tails are short lived as identifiable structures, and as they overlap and cross (as there need not be any common axis of accretion for protogalaxy mergers), a halo is formed. Since they retain the orbital characteristics of their progenitor population, they may have highly eccentric orbits. The inner halo is comprised of a roughly equal mix of stars obtained through merger, and stars that formed either *in-situ* within the halo early in the galaxy's history, or in the disk, migrating out due to interactions with satellite galaxies. The *in-situ* star formation would occur not due to monolithic collapse of the proto-galactic cloud (as was once thought), but due to gas rich mergers early in the galaxy's history.

3.2.3. *Orbits in Elliptical Galaxies*

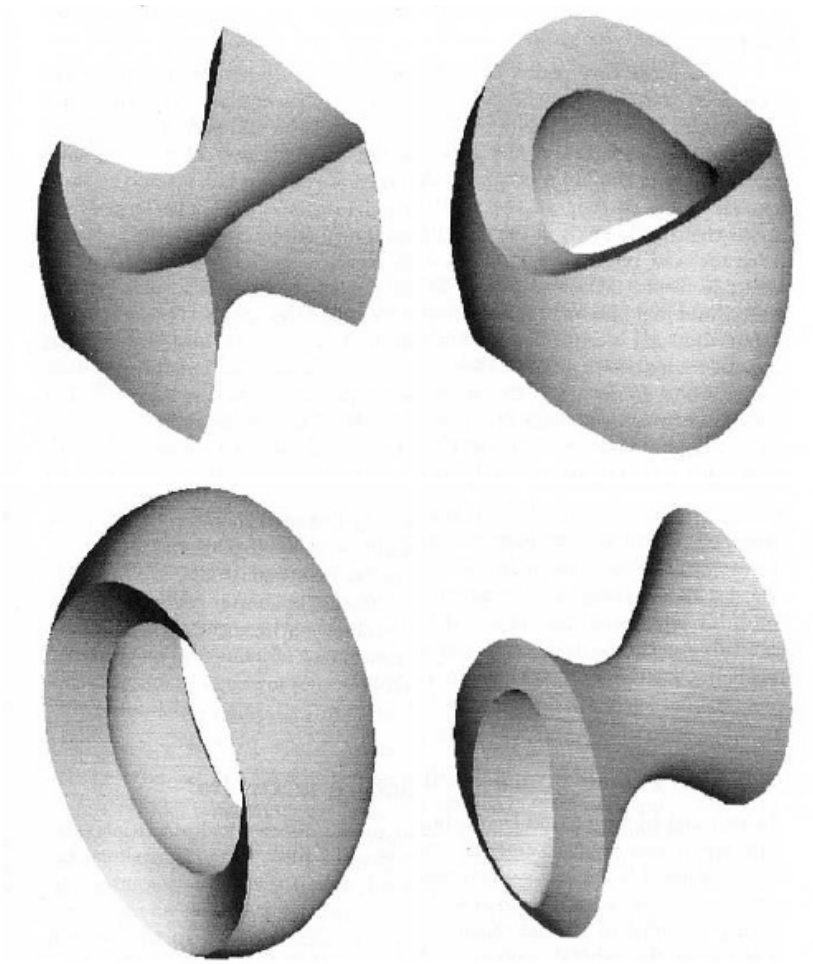

FIG. 52.— Four examples of orbits found numerically in the triaxial potential common to elliptical galaxies. The upper left orbit is a box orbit, the upper right a short-axis tube orbit (resembling a loop orbit from axisymmetric potentials), the lower left an inner long-axis tube orbit, and the lower right an outer long-axis tube orbit. From Binney & Tremaine (2008), their Fig. 3.46.

Elliptical galaxies are in general triaxial, with a potential that can be thought of as a spherically symmetric potential flattened along two axes. They therefore admit a wider range of orbits than the approximately planar, precessing, rosette loop orbits found for oblate spheroidal potentials. Elliptical orbits come in four general types: box and three types of tube orbits. Box orbits have no particular sense of circulation about the centre, and their angular momenta

average out to zero. The three-dimensional shape they trace out in space resembles a distorted box (see Fig. 52). Tube orbits do have a sense of circulation, and therefore have an axis of rotation they will never approach, i.e. all tube orbits trace out a shape with a hole in the middle. The most prominent tube orbits are those with the potential's shortest axis as their axis of rotation; these are slightly distorted versions of the loop orbits in axisymmetric potentials. The other two types of tube orbits are outer and inner long-axis orbits, which are oriented along the long axis of the potential (see Fig. 52). Orbits with the potential's intermediate axis as their axis of rotation are unstable. The population of short and long-axis tube orbits is be different.

Triaxial potentials also admit chaotic, or irregular, orbits. These orbits can be created by adding a cusp or core to the potential, or by adding a central black hole - all these features tend to deflect orbits, adding random motions to their trajectories. Addition of a cusp also increases the number of orbits trapped in resonances. Irregular orbits have a tendency to wander through phase space, in a process that is still not well-understood known as Arnold diffusion. Arnold diffusion is thought to possibly have significant effects on galactic evolution.

The tensor virial theorem requires that velocity dispersion in a triaxial system be larger parallel to the longest axis. Black holes have the tendency to make velocity distributions isotropic, robbing the triaxial elliptical of thermal support. If the potential loses its triaxiality as a result, stellar orbits will acquire a non-zero pericentric distance (for spherical potentials this should be obvious), robbing the black hole of material.

3.2.4. *What are the different populations in the galaxy, and what are their ages and metallicities?*

See Sec. 3.16

3.2.5. *What is the spheroid composed of (globular clusters?)?*

See Sec. 3.16

3.3. *Question 3*

QUESTION: Every now and then a supernova explosion occurs within 3 pc of the Earth. Estimate how long one typically has to wait for this to happen. Why are newborn stars likely to experience this even when they are much younger than the waiting time you have just estimated?

Two distinct populations of stars exist that can generate a supernova: stars greater than ∼8-10 but less than ∼50 solar masses end in core-collapse supernovae, and stars from ∼2 - 8 solar masses can form WDs, which may eventually generate thermonuclear supernovae if they exist in binaries. We will assume that the number of stars so massive as to generate a pair-instability supernova.

The rate at which supernovae occur (the death rate of stars) is approximately the same as the birth rate of stars, so we can use the IMF and the MW star formation rate (assume this rate is constant for the entire lifetime of the MW) of ~ 1 $M_\odot$ per year.

The number of stars between 8 and 50 solar masses per solar mass of stars formed is (using Salpeter):

$$f_{8-50M_\odot} = \frac{\int_8^{50} \phi(M) dM}{\int_{0.1}^{120} M\phi(M) dM} \approx \frac{1}{143}. \tag{81}$$

The number of stars between 2 and 8 solar masses in a binary per solar mass of stars formed is:

$$f_{2-8M_\odot,\text{bin.}} = \frac{2\int_2^8 \phi(M) dM}{3\int_{0.1}^{120} M\phi(M) dM} \approx \frac{1}{36}, \tag{82}$$

where the extra factor of $\frac{2}{3}$ is due to $\frac{2}{3}$ of stars forming in binaries. Only a fraction of these (say, 10%) end up in close-in binaries, reducing the fraction further to $\frac{1}{360}$.

The Milky Way thin disk has a radius of ~ 15 kpc and a height of 300 pc. This means the fraction of the MW our 3 pc neighbourhood covers is 5×10^{-10}, and assuming star formation is uniformly spread throughout the MW, this means 2×10^{11} years between SNe events.

3.3.1. *What errors are in your analysis?*

While core-collapse occurs for single stars (on their own), and while stars in binaries might be somewhat affected by their companions, by and large our estimated value should be reasonably accurate. Our estimate for thermonuclear supernovae is likely an overestimate, since it assumes all CO WDs lead to explosions. The amount of star formation in our local neighbourhood is also much smaller than that for star clusters, since most young, hot stars that are liable to end in CC SNe are born within clusters. This is also why a star within a cluster is more likely to be in the vicinity of an SN long before the estimated time given here.

There are no known supernova remnants within 3 pc of Earth.

3.4. *Question 4*

QUESTION: Galactic stars are described as a collision-less system. Why?

Stars do not physically collide with each other - in the MW disk there are about 0.3 stars/pc^3, and the radius of an average star is some 2×10^{-8} pc. The mean free path for a collision, then, is $\lambda = \frac{1}{n\sigma} \propto \frac{1}{nR^2}$, netting us about 10^{15} pc, longer than the observable universe. Considering the average speed of a star in the MW disk is about 10^{-4} pc/yr, we will have to wait a very long time before a head-on collision occurs. Even if we include major deflections due to close passes between stars as collisions, R should not increase by more than three orders of magnitude, and so the time between direct collisions is still exceedingly long. Note that this is not the case in globular cluster cores, where n is about 10^4 stars/pc^3 and a low velocity dispersion means that gravitational focusing effects increase σ by up to 10^5 (Raskin et al. 2009).

A system is considered collisional when gravitational two-body interactions between its constitutent members becomes a significant factor in its secular evolution. The force on a star with mass m experiencing a glancing deflection from a field star, also of mass m, is

$$F_\perp = \frac{Gm^2}{b^2 + x^2} \cos\theta = \frac{Gm^2 b}{(b^2 + x^2)^{3/2}} = \frac{Gm^2}{b} \left(1 + \left(\frac{vt}{b}\right)^2\right)^{-3/2} \tag{83}$$

Integrating this over time gives us

$$\delta v = \frac{1}{m} \int_{-\infty}^{\infty} F_\perp dt = \frac{2Gm}{bv}. \tag{84}$$

Note that the total deflection becomes large ($\delta v \sim v$) when $b = b_{\text{lim}} \sim 2GM/v^2$ - the approximation breaks down in this regime.

The surface density of stars is approximately $N/\pi R^2$, where R is the radius of the galaxy. The star therefore suffers $\delta n = \frac{N}{\pi R^2} 2\pi b db = \frac{2N}{R^2} b db$ collisions with impact parameter b each time it passes through the galaxy. The mean change in v^2 (the mean change in v is 0) is

$$\Sigma \delta v^2 \approx \delta v^2 \delta n = \left(\frac{2Gm}{bv}\right)^2 \frac{2N}{R^2} b db. \tag{85}$$

The integral over all impact parameter is then

$$\Delta v^2 = \int_{b_{\text{min}}}^{b_{\text{max}}} \Sigma \delta v^2 \approx 8N \left(\frac{Gm}{Rv}\right)^2 \ln\left(\frac{b_{\text{max}}}{b_{\text{min}}}\right) \tag{86}$$

To reasonable approximation (since we are using logarithms), $b_{\text{max}} \approx R$ and $b_{\text{min}} = b_{\text{lim}}$. Lastly, we assume $v^2 \approx \frac{GNm}{R}$, which is a good approximation for real galaxy orbital speeds. This gives

$$\frac{\Delta v^2}{v^2} \approx \frac{8 \ln\left(b_{\text{max}}/b_{\text{min}}\right)}{N} \tag{87}$$

The number of collisions it takes for the change in v^2 to reach v^2 is the inverse of $\Delta v^2/v^2$. The time it takes is $t_{\text{relax}} \approx n_{\text{relax}} t_{\text{cross}} = n_{\text{relax}} R/v$ (there is one "collision" per crossing time, since our calculation gives the deflection Δv^2 per pass through the field of stars), which gives us the relaxation time. We can estimate $b_{\text{max}}/b_{\text{min}} = R/b_{\text{lim}}$ by using the approximation $N \approx Rv^2/Gm$ (the same formula we used to estimate the orbital speed). From all this we obtain

$$t_{\text{relax}} \approx \frac{0.1N}{\ln N} t_{\text{cross}} \tag{88}$$

We interpret this relaxation time as the time before which collisions must be accounted for in the system.

For an ordinary, reasonably sized galaxy, the relaxation time is far in excess of the age of the universe. This allows us to use the collisionless Boltzmann equation to describe the dynamics of stars in large galaxies. This include ellipticals, which are held up by "thermal" pressure (i.e. velocity dispersion σ), but are not themselves thermalized (if they were thermalized, equipartition would result in them losing the σ anisotropy that gives them their elliptical shape) (Schneider 2006). Other systems, such as galactic centres, globular clusters, open clusters and galaxy clusters, however, are collisional systems and thermalize to some degree over the course of their lives.

3.4.1. *What stars are collisional?*

As noted above, galaxy centres, globular clusters and open clusters are all collisional systems. This owes to the fact that they have very short crossing times and relatively few numbers of particles. Galaxy clusters actually have very long crossing times, but there are so few galaxies in each cluster that the relaxation time becomes quite small.

A perhaps counterintuitive issue is the fact that the relaxation time becomes greater as N increases, rather than decreases - one would expect that a huge number of objects, given the same volume, would increase the chances for collisions. The reason for this is the coupling of R and v in Eqn. 88 from the assumption that $N \approx Rv^2/Gm$, which, rewritten as $v^2 = GNm/R$, is a statement that each star is virialized with respect to the potential it is in. Higher speeds mean that there is less time for glancing collisions to impart velocity, resulting in a large number of collisions necessary to thermalize the system. This formulation breaks down, however, in the limit where $b_{\rm lim} \sim 2GM/v^2$ becomes commonplace, since under those conditions major deflections become significant to the system, and thermalization becomes more efficient as a result.

(Note that fixing the number of objects N and compressing the volume without decreasing velocity (i.e. shortening $t_{\rm cross}$) does decrease the relaxation time, as expected.)

3.4.2. *Gas is collisional. Why?*

The important difference between a gas and a self-gravitating system like a galaxy or galaxy cluster is that gases do not have any long range forces. The only timescale, therefore, that affects gas is the thermalization timescale, by which through collisions a gas acquires a uniform temperature. In astrophysical systems this translates to local thermodynamic equilibrium only - on global levels gravity, radiation and magnetic fields will also have strong effects on the gas.

3.5. *Question 5/6*

QUESTION: Given that only a tiny fraction of the mass of the ISM consists of dust, why is dust important to the process of star formation? The ISM mainly consists of hydrogen and helium which are very poor coolants. How, then, do molecular cloud cores ever manage to lose enough heat to collapse and form stars? Why are H and He such poor coolants?

The average density of the universe ($\sim 10^{-30}$ g/cm^3) is about 30 orders of magnitude less density than the centre of the average star ($\sim 10^3$ g/cm^3), and the average dark matter particle can only collapse to the point of forming a virialized halo with a characteristic scale of several tens of kpc across. Unlike dark matter, however, baryons can emit their energy through radiative dissipation.

For optically thin hot gas, the strength of radiative dissipation is dependent on the number of line transitions multiplied by the frequency at which they can occur. For this reason, metal-polluted gas (especially polluted with complex molecules) does a much better job of cooling than pristene, metal-free gas. Fig. 53 shows the cooling curve of hot plasma past the ionization temperature of H at about 10^4 K. While two orders of magnitude increased heating can be achieved by populating the gas with metals, pristene H II still can act as an effective coolant.

This picture changes below (Fig. 54) 10^4 K - in the dense H_2 regions that will form stars, metals and dust play an essential role in cooling the gas to 10 K. H_2 has a weak, non-permanent dipole moment, which can only perform electric quadrupole transitions ($J = 2 \rightarrow 0$). These transitions are both slow and have a large ΔE, and as a result is not only easily suppressed by collisional de-excitation when the gas is $> 10^4$ g/cm^3, but cannot cool the gas below ~ 100 K (because not enough of the gas will be high-energy for an appreciable fraction of molecules to enter the $J = 2$ excited state). As a result, pristene H_2 will loiter at 100 K and 10^4 g/cm^3 indefinitely.

The addition of metals and dust allows for a greater number and diversity of possible atomic transitions. In metal rich gas at temperatures below 100 K, singly ionized carbon (C^+) is the primary coolant of H_2 below ~ 100 g/cm^3, CO (which has a significant dipole moment) the primary coolant from ~ 100 - $\sim 10^5$ g/cm^3, and FIR thermal emission from dust grains coupled to the gas the primary coolant past $\sim 10^5$ g/cm^3. As inferred from Fig. 54, H_2 itself contributes a negligible amount of the total cooling, with metals boosting the rate by more than two orders of magnitude.

Once the cloud is collapsing, cooling also governs the fragmentation of the cloud. During collapse, the Jeans mass $M_J \propto T^{3/2}\rho^{-1/2}$ decreases for an isothermal cloud and increases for an adiabatic cloud. This means if the entire collapse were adiabatic, which should be the case if cooling is highly inefficient, fragmentation is suppressed.

3.5.1. *How does H_2 form?*

H_2 forms in two ways. The most efficient means is grain catalysis, where H I is trapped in the potential wells of a dust grain. When two bound H atoms meet, they combine to form an H_2 molecule, and the energy liberated (from an increased negative binding energy) in the reaction is enough to eject the H_2 from the surface of the dust grain.

A far more inefficient process (but essential for forming molecular hydrogen in pristene gas) is

$$\begin{aligned} H + e^- &\rightarrow H^- + \gamma \ (\text{“radiative association''}) \\ H^- + H &\rightarrow H_2 + e^- + KE \ (\text{“associative detachment''}). \end{aligned} \tag{89}$$

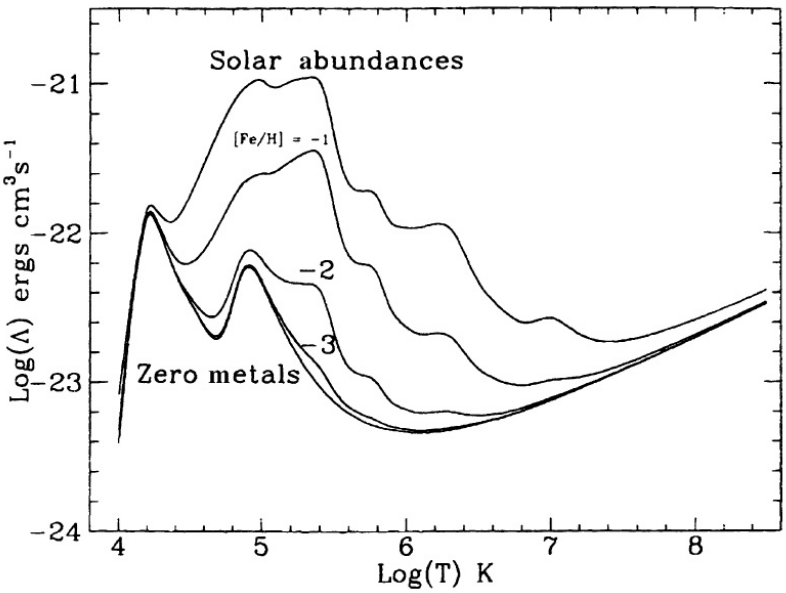

FIG. 53.— The cooling curve (in units of luminosity density) for a 1 particle/cm^3 astrophysical plasma at various temperatures. Different curves represent different metallicities. At high temperatures, thermal bremsstrahlung dominates over all other radiation processes, while at lower temperatures recombination becomes dominant. For the primordial, zero-metallicity curve, the two humps are due to ionization/recombination radiation of H (at $\sim 10^4$ K) and singly-ionized He (at 10^5 K). Below 10^4 K all hydrogen is assumed to be in its neutral state. From Longair (2008), his Fig. 16.2.

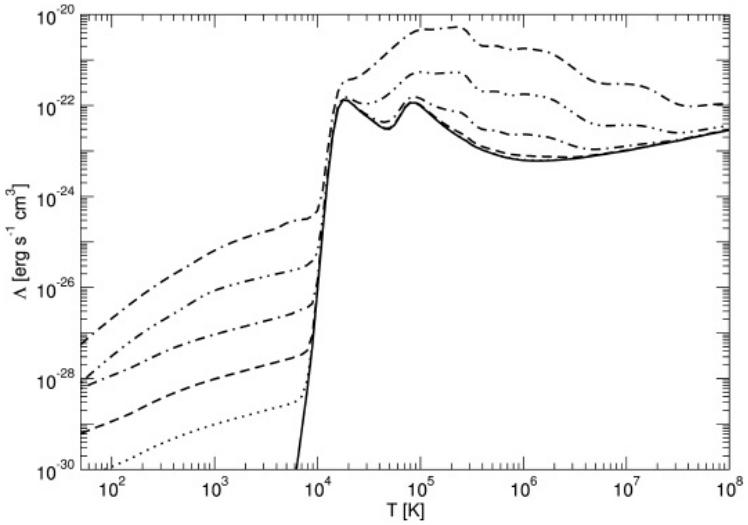

FIG. 54.— A cosmic cooling curves, indicating the difference between metal-free hydrogen (solid line) and metal-enriched hydrogen ($10^{-3}Z_\odot$ (dotted), $10^{-2}Z_\odot$ (dashed), $10^{-1}Z_\odot$ (dot-dashed), $Z_\odot$ (dot-dot-dashed), and $10Z_\odot$ (dot-dash-dashed)) at $n_H = 1$ cm^{-3}. H_2 cooling has not been included, but Smith et al. gives $\Lambda = 10^{-32}$ g/cm^3 at 100K and $n_H = 1$ cm^{-3}, contributing less than 1% to the overall cooling. From Smith et al. (2008), their Fig. 2.

This process is limited by the popuation of H^- ions, which are easily destroyed.

3.5.2. *Why is H_2 necessary for star formation?*

In short, it is not necessary at all. Simulations have shown that if H_2 formation is artificially suppressed, C^+ becomes the dominant cooling mechanism for H I at low temperatures until densities become high enough that dust FIR emission becomes significant. The reason H_2 is associated with star forming regions is because both H_2 formation and star formation needs shielding from the interstellar radiation field (ISRF) (star formation requires shielding in order for gas to cool). When properly shielded, cold, dense gas preferentially forms stars, and preferentially turns into H_2, creating a correlation between the two.

3.5.3. *How do Population III stars form?*

Simulations have suggested that gravitational instability of a ~ 100 K, 10^4 g/cm^3 cloud will eventually drive it to collapse. Cooling governs fragmentation, and with very little cooling, the collapsing cloud does not fragment at all,

resulting in the formation of massive stars.

It is not obvious, once gravitational instability sets in, what causes the cloud to collapse - an adiabatic gas equation of state has $\gamma = 5/3$, meaning that a gas cloud that cannot cool is stable against collapse.

3.6. *Question 7*

QUESTION: What's the difference between a globular cluster and a dwarf spheroidal galaxy?

The following are similarities between GCs and dSphs:

- Similiar range in luminosities and number of stars
- Metal-poor
- Satellites of larger galaxies
- Spheroidal in shape (though dSphs have much greater ellipiticity)

The following are differences:

- dSphs are spheroidal, while GCs are nearly spherical
- dSphs are heavily dark matter dominated while GCs have almost all their mass in stars
- dSphs have complex star-forming histories, evidenced by a spread in metallicities, while most GCs do not
- dSphs are formed independently, while GCs are formed in star forming regions in their parent galaxies
- dSphs are much larger than GCs
- dSphs are often more luminous and larger
- dSphs have longer relaxation times than GCs (as they are larger and have many more dark matter "particles")

3.6.1. *What observational differences are there between GCs and dSphs?*

a number of these features are difficult to use for classification. Size and luminosity, for example, cannot seem to be used to distinguish between the two as a continuous M_v vs. r_h relationship exists from GCs to dSphs. The presence of dark matter, particularly for dim dSphs, is the most reliable distinguishing feature between the two classes of object, but radial velocity curves are not always easily obtained. Using the spread in metallicities is likewise time-consuming. A somewhat more reliable method than appealing to M_v vs. r_h is to use ellipticity - according to van den Bergh (2008) $\sim 85\%$ of dSphs have ellipticities $((a-b)/a)$ greater than 0.25, while almost no GCs do.

3.6.2. *What is a Galaxy?*

As it turns out, no official definition of a galaxy exists, as noted by (Forbes & Kroupa 2011). This problem is exacerbated more by ultra-compact dwarf galaxies (UCDs) than dSphs because UCDs are compact enough that using size/stellar density is problematic, and it is more difficult to determine dark matter fractions of compact galactic objects. (Forbes & Kroupa 2011) suggest a definition of a galaxy that includes i). the system is gravitationally bound and ii). contains stars. Since this would include globular clusters, an additional criterion is needed. Requiring complex star formation may not remove massive GCs from being called galaxies (though since these might be stripped cores of dwarf galaxies, that might be apt). Requiring high concentrations of dark matter will remove tidal dwarf galaxies (formed from the collapse of tidal tails) and possibly some dwarf ellipticals from being called galaxies. Requiring the galaxy have its own system of globular clusters would remove many low-stellar mass galaxies. If a relaxation time greater than the Hubble time is used, everything that currently considered a galaxy would remain so, and everything currently not would also.

One additional problem is that such observational classifications do not directly take in the histories of the objects - Ω-Cen, for example, is often classified as a globular cluster (and would be under some of the above classifications), but is likely the nucleus of a disrupted dwarf galaxy.

3.7. *Question 8*

QUESTION: The stars in the solar neighbourhood, roughly the 300 pc around us, have a range of ages, metallicities and orbital properties. How are those properties related?

As star formation progresses in a galaxy, the ambient environment is polluted with metals, which are then incorporated into future generations of stars. In this manner, as the galaxy ages, newly formed stars become more and more metal rich. This age-metallicity relationship is roughly an inverse relation, though it flattens 5 Gyr ago, indicating that the regional star-formation rate was much lower (combined, possibly, with the addition of gas poor material raining onto the disk) before 5 Gyr ago. The burst of star formation at 5 Gyr may be due to mergers and other interactions with other galaxies in the Local Group (Marsakov et al. 2011). Likewise, while in a galaxy gas is collisional and radiative, stars are collisionless and non-radiative (they radiate via nuclear fusion and do not tap into their reservoir of orbital kinetic energy). As a result, gravitational kicks obtained during scattering events (scattering off of GMCs, accretion of satellite galaxies, gas, etc.; see below) tend to increase the velocity dispersion of stars, which then have no means to re-equilibrate or radiate away the excess energy. There should then be an age-σ relation, whereby older stars have higher velocity dispersions. To first order this relation is roughly linear, though in our own galaxy it appears to be a power law (or set of power laws) with exponent < 1.

These simple relations are complicated by the fact that star formation and mergers are inhomogeneous in both space and time, and stars will migrate out of their place of birth over time. The overall star formation and merger history of the galaxy will determine the functional forms of the age-metallicity and age-σ relations, while the spread in any given region of the galaxy will be due to migration of different stellar populations. There are, for example, populations of old, metal-rich stars in our solar neighbourhood.

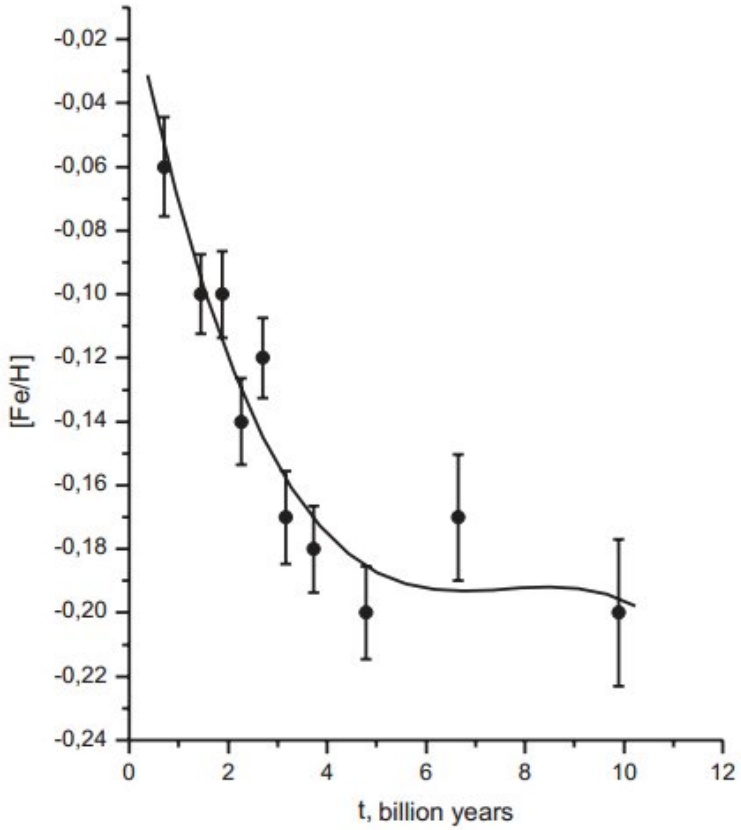

FIG. 55.— Age/average metallicity relation for thin disk stars in the Geneva-Copenhagen catalogue that are within 70 pc of the Sun. Note that for very old stars the metallicity dispersion becomes very high. From Marsakov et al. (2011), their Fig. 6.

3.7.1. *How does velocity dispersion increase over time?*

The stellar population of the thin disk, initially with a low velocity dispersion (because they were formed from low velocity dispersion gas), can have its dispersion boosted by scattering stars against external bodies such as GMCs embedded in the thin disk, black holes from the halo or "dark clusters" (the latter two have no observational evidence

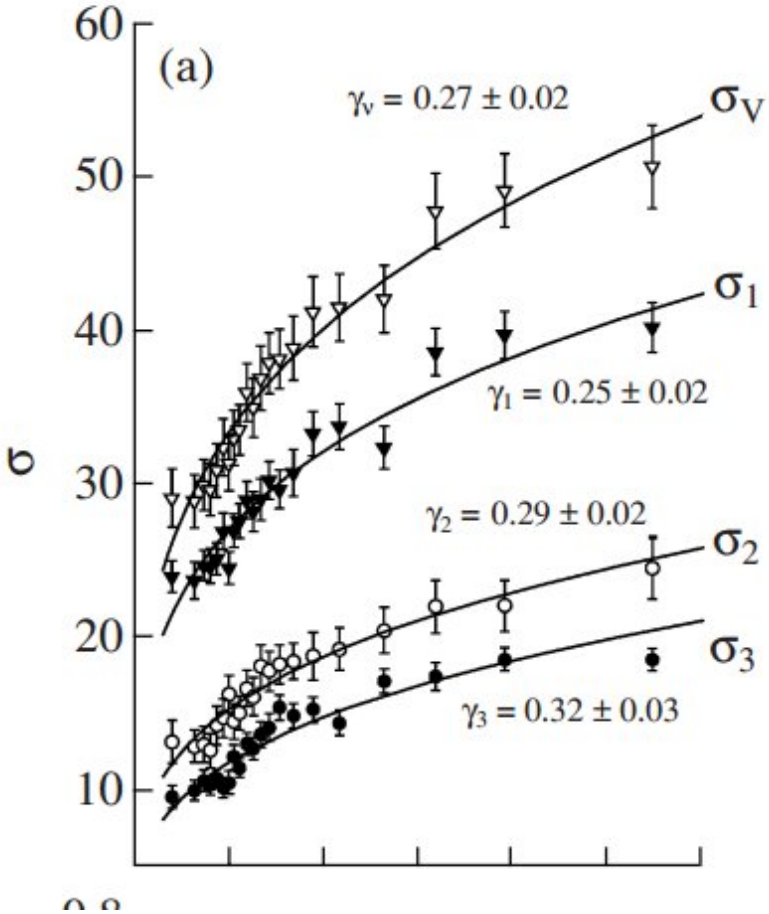

FIG. 56.— Age/velocity-dispersion relation for thin disk stars in the Geneva-Copenhagen catalogue. Stellar streams are included, but excluding them does not change the relationship significantly. 1, 2 and 3 correspond to U, V and W, while V corresponds to the total dispersion. The x-axis is time, in units of 2 Gyr. From Koval' et al. (2009), their Fig. 3.

to their name) (Seabroke & Gilmore 2007). Minor mergers with dwarf galaxies and Lindblad resonances with the Galactic bar may also provide kicks to a stellar population (Seabroke & Gilmore 2007).

Depending on which mechanism dominates over the others, the velocity dispersion vs. time relation for the galaxy will be different. To complicate matters different processes increase the dispersion along the U (galactic radial coordinate), V (azimuthal angle) and W (perpendicular to the disk plane) directions by different degrees. Scattering off of GMCs and spiral structure (De Simone et al. 2004 have shown spiral structure scattering to dominate over scattering from GMCs) for example, should be a less effective heating process for σ_W over time because GMCs and spiral structure are localized on the thin disk (Seabroke & Gilmore 2007). Once a star's epicyclic amplitude becomes large enough that the star crosses more than one spiral arm as it radially oscillates, spiral structure heating along σ_U and σ_V may also saturate (though this is uncertain). Heating from minor and major mergers does not have such restrictions.

It is also possible to envison creating an age-σ relation without heating stars after they form. Kroupa (2002), for example, suggest evaporating star clusters are the progenitors of the the current high-σ star population (House et al. 2011). It is also possible that older stars were born in hotter disks. This could be due to a turbulent ISM, where the turbulence itself may be due to cold flows, SNe feedback or mergers. A related complication is whether a distinction needs to be made between the thin and thick disks. The high-dispersion thick disk could come about from the kinematic heating of younger thin disk stars, but could also have been formed back when the ISM was more turbulent, or could even be formed out of stars from merged satellite galaxies.

The current observational data of velocity dispersion and age in the Solar neighbourhood is insufficiently detailed to pick out a single channel as the correct evolutionary history of the MW disk. While some evidence exists that the age-σ relation is a power law with no cut-off, it seems most likely (due to a lack of mechanisms that can increase σ_W) that minor mergers (and possibly major mergers) played a significant role in increasing σ_W until about 3 Gyr ago, after which GMCs became the primary means of scattering.

3.7.2. *Why is the mean [Fe/H] not -∞ at the birth of the MW?*

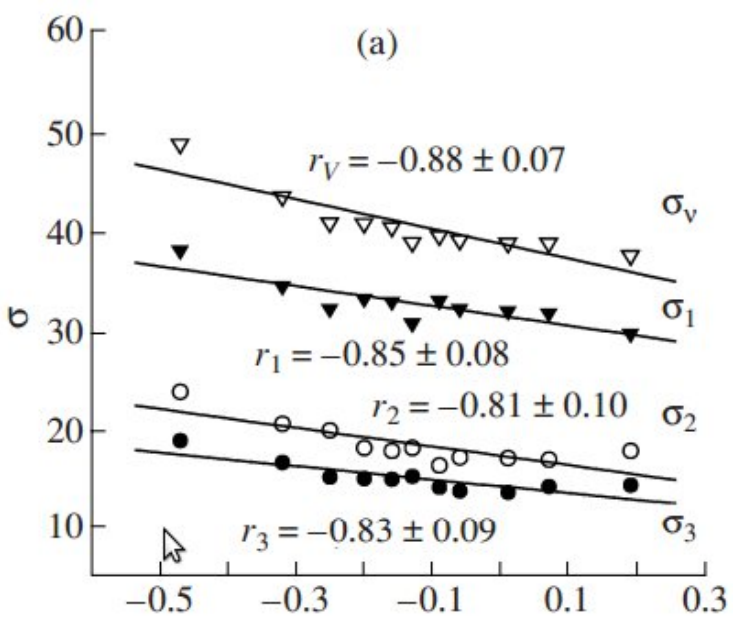

FIG. 57.— Velocity-dispersion/metallicity relation for thin disk stars in the Geneva-Copenhagen catalogue. Stellar streams are included. For σ, 1, 2 and 3 correspond to U, V and W, while V corresponds to the total dispersion. The x-axis is [Fe/H] metallicity. From Koval' et al. (2009), their Fig. 4.

Stars that formed in the early MW have a relatively high average metallicity ([Fe/H] $\approx$ -0.2), but have also a significant dispersion of metallicities. The high average metallicity is likely due to the fact that regions of the ISM with the highest metallicity have the greatest rate of star formation. The fact that these metals exist at all in the ISM suggests an era early in the evolution of our galaxy that produced substantial amounts of metals and when low-mass star formation was simultaneously suppressed. The dispersion of metallicities suggests the ISM was highly inhomogeneous early in the life of the MW, and that over time mixing homogenized the MW, reducing the dispersion in star metallicities.

3.8. *Question 9*

QUESTION: What are the main sources of heat in the interstellar medium?

A good reference for this is (Dyson & Williams 1997, pg. 32 - 45), as well as scattered passages throughout Draine (2011).

The main sources of heat in the ISM include

- **Photoionization**: a photon is absorbed by a bound electron, and the electron is given enough energy to be liberated from the atom:

$$A + h\nu \rightarrow A^+ + e^-. \tag{90}$$

  This free photoion contributes its thermal energy to the gas via collisions with ions and other free electrons. Contact with ions may result in recombination, contact with bound electrons may result in the bound electron becoming excited, then re-radiating the excitation energy, and contact with free electrons produces bremsstrahlung (nearly elastic collisions are, therefore, the most desired, since they transfer the least energy into heat). Nevertheless, overall heating of the cold ISM is achieved through photoionization of C, Si and Fe, and overall heating of H II regions through H ionization. Overall, 2.1 eV is deposited into the ISM per ionization for a gas of cosmic abundance.

- **Photodissociation of H_2**: photons of 11.2 eV $\leq h\nu \leq$ 13.6 eV may dissociate molecular hydrogen via

$$H_2 + h\nu \rightarrow H + H. \tag{91}$$

  When H_2 absorbs a photon, it is promoted via permitted transition to the first or second excited electronic state (with some vibrational and rotational, or "rovibrational" substate). 85% of the time, this excited state will, via electric dipole transition (the state itself has an electric dipole, while ground state H_2 and the excited rotational

states mentioned in H_2 cooling do not) move back to a ground state (with some excited rovibrational substate). 15% of the time, however, the system falls on the vibrational continuum, and H_2 dissociates to form two H atoms. 0.4 eV is released per dissociation. In the case of decay to a rovibrational excited electronic ground state, the remaining energy can be harvested by collisional de-excitation for ~ 2 eV **is this per dissociation?**.

- **Cosmic ray excitation and ionization**: high energy (non-thermal) protons and electrons ejected from a distant energetic source are known as cosmic rays. Interactions between cosmic rays and bound electrons can result in the ejection of an energetic secondary electron, i.e.

$$p^+ + H \rightarrow p^+ + H^+ + e^-, \tag{92}$$

  where p^+ can be substituted with e^- for electron cosmic rays. The mean kinetic energy of the newly liberated electron is ~ 35 eV, regardless of the speed of the incoming cosmic rays (as long as they are fast compared with the Bohr speed, $c/137$). The liberated may further ionize other atoms and promote H, H_2 and He into excited states, which then radiatively de-excite (resulting in no heating). Interactions between the liberated electron and other free electrons result in scattering. In a completely neutral medium, about ~ 0.2 of the cosmic ray's energy eventually goes into ISM thermal energy (the rest is lost by radiation), while in a completely ionized medium, most of the cosmic ray's energy is lost through elastic collisions with other free electrons, resulting in all of the electron's energy being retained.

- **X-ray photoionization**: x-rays emitted by compact objects or the hot ISM may ionize hydrogen, which in general results in very energetic free electrons, which cause more secondary ionization and heating than cosmic ray ionization. Since the H cross-section for x-ray absorption is small, x-ray penetration into sources optically thick to UV/visible/IR photons is significant, and allows for heating of shielded ISM. The He and metal cross-section for x-ray ionization is much larger, and therefore ionization of ISM pollutants contributes significantly to the overall heating from x-rays. The x-ray background is sparse compared to cosmic rays, and as a result, despite greater heating per ionization, x-ray photoionization is much less significant than cosmic ray ionization.

- **Dust photoelectric emission**: when an energetic photon is absorbed on a dust grain, it may excite an electron and allow the electron to escape the dust grain. This process requires > 8 eV photons, and is most effective for small grains, such as PAHs, since electrons have an easier time of escaping if they are liberated at the surface of the grain, and smaller grains have a larger surface area to volume ratio (Pogge 2011, Ch. 5). This mechanism is the dominant source of heating in the diffuse neutral ISM.

Less important sources of heat include collisional and frictional heating through damping of MHD waves and shock-waves (this occurs in regions of considerable gas flow). Transient events (stellar superwinds, supernovae, etc.) will also heat their surroundings.

3.9. *Question 10*

QUESTION: Draw an interstellar extinction curve (ie, opacity), from the X-ray to the infrared. What are the physical processes responsible?

Interstellar extinction is a wavelength-dependent process by which light, which would otherwise travel unimpeded from a source to an observer, is scattered by the interstellar medium.

The two physical processes that determine the nature of the interstellar extinction curve are scattering by dust (a classical electrodynamic effect with significant quantum corrections for more complex molecules) below the Lyman limit (912 Å) and photoionization of atoms above (Ryter 1996). We can describe classical scattering through Mie Theory. Let us assume dust is spherical, and therefore has a cross-section $\sigma_g = \pi a^2$, where a is the sphere's radius. Mie assumed an elastically scattered incoming electromagnetic plane wave, and using a scattering formalism (similar to that presented in Ch. 11 of Griffiths' Introduction to Quantum Mechanics). The resulting scattering cross-section σ_λ is some multiple of σ_g[19], an proportional to (Carroll & Ostlie 2006; Hahn 2009)

$$\begin{aligned} \sigma_\lambda &\propto \frac{a^6}{\lambda^4} \ (\lambda \gg a) \\ \sigma_\lambda &\propto \frac{a^3}{\lambda} \ (\lambda \approx a) \\ \sigma_\lambda &\propto a^2 \ (\lambda \ll a) \end{aligned} \tag{93}$$

[19] We may define the dimensionless parameter $Q_\lambda = \frac{\sigma_\lambda}{\sigma_g}$, where σ_λ is the scattering cross-section.

The work necessary to determine these proportionalities is from classical electromagnetic theory, and is beyond the scope of this document. The common analogy used is of water waves attempting to cross a set of vertically oriented cylindrical barriers. If the physical size of the wave (equivalent to its wavelength) is much larger than the object, the waves will be negligibly affected. If the physical size of the wave is about the same as the cylinders, diffraction becomes significant. If the physical size of the wave is much smaller than the cylinders, the cylinders will physically block the wave from propagating. This result also shows that most scattering is in the Mie regime or above - the λ^{-4} dependence of Rayleigh scattering makes it difficult to affect anything.

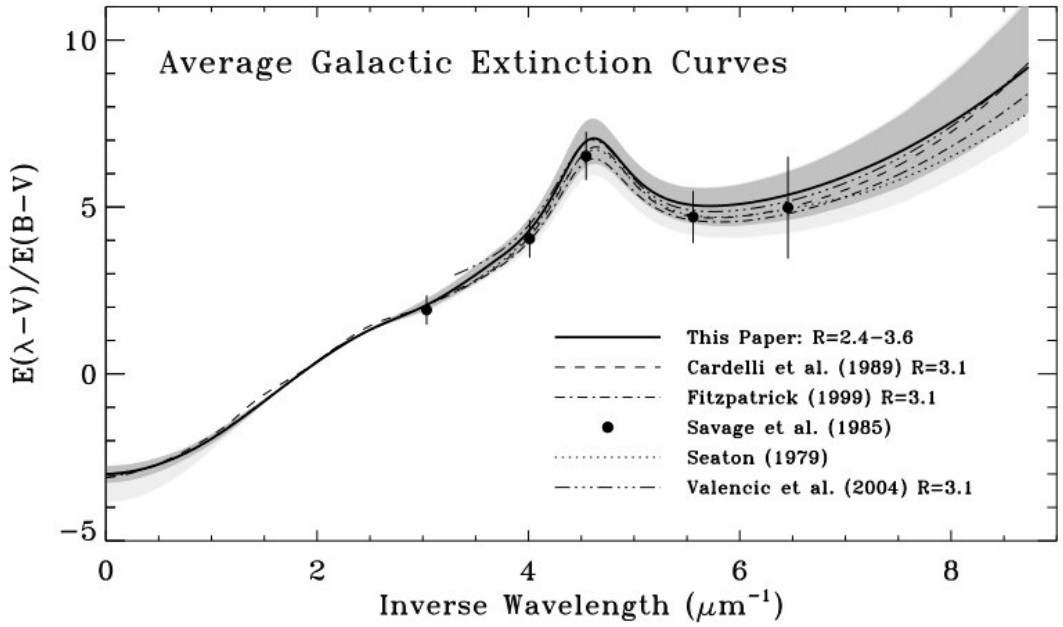

FIG. 58.— Galactic mean extinction curve from $\sim$ 10 μm to 100 nm, based on colour excess. The black line is the mean result of Fitzpatrick & Massa (2007), and the dark grey shaded area is the variance. Other lines are from other works, for comparison. From Fitzpatrick & Massa (2007), their Fig. 9.

The extinction curve can be parameterized using a ratio of extinction A_λ to some reference extinction (ex. A_V), colour excess ($\frac{A_\lambda - A_V}{A_B - A_V} = \frac{E(\lambda-V)}{E(B-V)}$) or cross section. Fig. 58 plots the interstellar extinction curve from the near-IR to 100 nm. While most of the curve roughly reproduces Mie scattering wavelength dependency, suggesting that dust is primarily composed of $\lesssim 0.01$ μm particles, so that scattering up to the UV limit is always in the regime $2\pi a/\lambda < 1$ (Draine 2011, pg. 240), there are significant deviations as well. The most prominent are (Carroll & Ostlie 2006; Draine 2011, Ch. 23):

- **The 2175 Åbump:** since this feature is a bump, it strongly suggests a form of resonance scattering, rather than scattering as predicted from Mie theory. The origin of the bump is not entirely understood, but candidate scattering sources include graphite (though it is not known how graphite would form in the ISM) and polycyclic aromatic hydrocarbons (also responsible for a series of molecular emission bands from 3.3 to 12.7 μm).
- **The 3.4 μm bump:** this is due to C-H stretching in hydrocarbons.
- **The 9.7 and 19 μm bumps:** these are caused by absorption bands at 9.7 and 18 μm from the stretching of the Si-O molecular bond and the bending of Si-O-Si bonds in silicates, respectively. Neither this bump nor the 3.4 μm bumps can easily be observed in Fig. 58; see Fig. 21.1 and 23.2 in Draine.
- **Diffuse Interstellar Bands (DIBs):** exist throughout the optical and IR. None have convincingly been identified.
- **Dark cloud absorption features:** O-H stretching (3.1 μm) and features caused by CO (4.67 μm), CH_3OH (3.53 μm) and CO_2 (15.2 μm) only appear in dense molecular clouds, possibly because shielding is needed for these compounds to survive.

At 912 Å, hydrogen is ionized, and as a result the extinction curve shoots up dramatically. The ground-state photoionization cross section of H and one-electron ions follows the proportionality $\sigma_{\rm ph} \propto \nu^{-3}$ for ν greater than the ground state energy of the electron E_i and $\sigma_{\rm ph} = 0$ for ν smaller than E_i - this hard cut-off at E_i is known as an

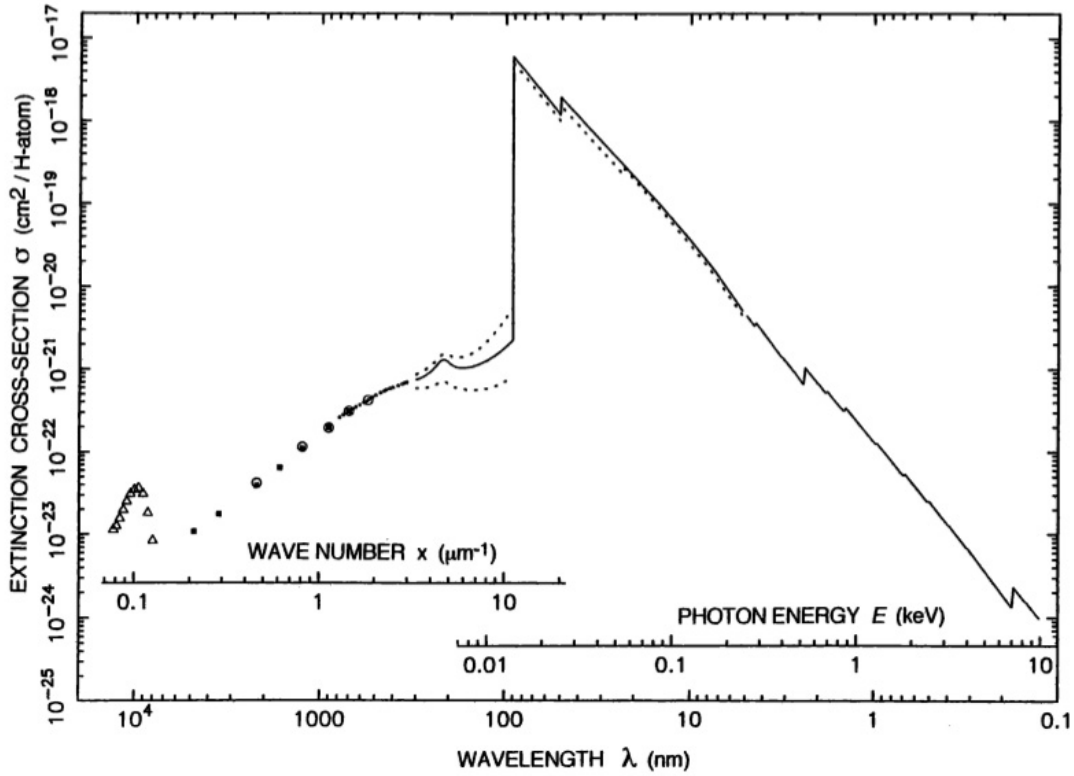

FIG. 59.— Galactic mean extinction curve from $\sim$ 10 μm to 0.1 nm, based on extinction cross-section per H atom in the ISM. 5% ionization is assumed. To the left of 912 Åthe dotted lines represent the extremes in variation from looking at differet regions of the sky. To the right of 912 Åthe dotted lines represent the curve if 20% ionization is assumed. This curve includes the ionization of H (the peak at 912 Å) and multi-electron atoms (the peaks above 912 Å). From Ryter (1996), their Fig. 1.

"absorption edge". For H_2 and more complicated systems the ground-state photoionization cross-sections are more complicated and do not have to begin with an absorption edge; they, however, do have an absorption edge at the E_i of their lowest energy electron (in the 1s shell) (see Draine, pg. 129 - 130). The $\sigma_{\rm ph}$ of multi-atom elements at their lowest energy electron absorption edge can be up to a factor of 10^4 larger than the $\sigma_{\rm ph}$ of H at the same photon energy. As a result, these additional peaks are quite prominent in Fig. 59 despite there being much less metal in the ISM than H. Increasing the ionized fraction in the ISM lowers extinction due to photoionization.

3.9.1. *Given a cross-section, how would you calculate the amount of extinction?*

The equation of radiative transfer is

$$\frac{dI_\nu}{ds} = -\alpha_\nu I_\nu + j_\nu \tag{94}$$

where j_ν is the emission coefficient and $\alpha_\nu = n\sigma = \kappa\rho$ is the absorption coefficient. If we assume light scattering into the line of sight can be neglected (which is not the case for reflection nebulae) then scattering can be subsumed into α. If we also neglect emission (which does occur for dust in the far-IR), then

$$I_\nu = I_{0,\nu} \exp(-\int_{s_0}^{s} \alpha_\nu ds') \tag{95}$$

$\tau_\nu = \int_{s_0}^{s} \alpha_\nu ds'$, and therefore $I_\nu / I_{0,\nu} = e^{-\tau_\nu}$. Now,

$$A_\nu = -2.5 \log_1 0(I_\nu / I_{0,\nu}) = 2.5\tau_\nu \log_1 0(e) = 1.086\tau_\nu. \tag{96}$$

Extinction is therefore directly related to the optical depth.

3.9.2. *What percentage of the light is absorbed (rather than scattered) by the light?*

3.9.3. *Why is dust scattering polarized?*

ISM scattered light is polarized to a few percent, and this polarization is dependent on both wavelength and position in the sky. This suggests that interstellar dust is not completely spherical, on average, and locally has a preferred direction of orientation. This is likely caused by weak local magnetic fields, which tends to force dust grains to rotate

with their long axes perpendicular to the magnetic field (Draine 2011, pg. 242). Polarization goes down in the UV, suggesting that the population of dust particles that do cause scattering are of size ∼0.1 μm; light with $\lambda < 0.3$ μm will geometrically scatter at the "geometric limit", where both polarization modes suffer the same extinction. Dust that is smaller than ∼0.1 μm must then not be aligned.

3.9.4. *What is the grain distribution of the ISM?*

Currently it is impossible with observational data to uniquely constrain the grain size and shape distribution. A model assuming that dust consists of amorphous silicate, carbaneceous material and PAHs, along with some assumed shape, can replicated both the interstellar extinction curve and the polarization of scattered light. Fig. 60 shows three possible models.

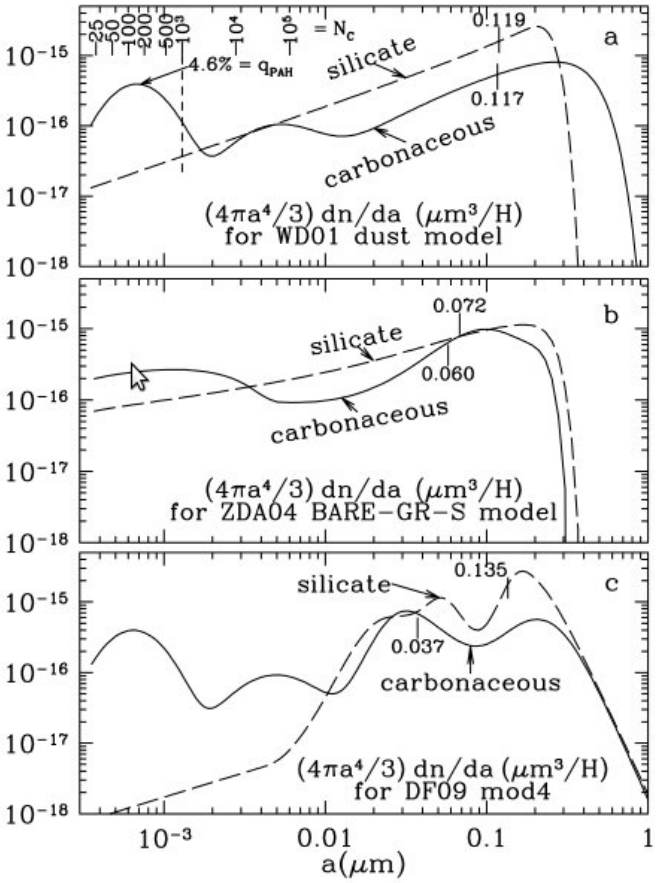

FIG. 60.— Size and composition distributions of dust that reasonably replicate the ISM extinction curve and polarization. The models come from a). Weingartner & Draine 2001, b). Zubko et al. 2004 and c). Draine & Fraisse 2009. The tick mark indicates the half-mass radius for each kind of grain. From Draine (2011), his Fig. 23.10.

3.10. *Question 11*

QUESTION: What is dynamical friction? Explain how this operates in the merger of a small galaxy into a large one.

This information comes from (Binney & Tremaine 2008, Sec. 8.1), (Carroll & Ostlie 2006, pg. 1001 - 1005) and (Schneider 2006, pg. 235 - 236).

Dynamical friction is a means by which the kinetic energy of stellar systems colliding with each other can be transferred into the kinetic energy of their constituent particles, i.e. thermal energy. Suppose two galaxies (or other comparable system made of constituent particles) are colliding with each other. As they pass through each other, their constituents gravitationally attract each other, but because both galaxies are moving relatively quickly, by the time the constituents react to the gravitational pull and clump, both galaxies have moved. As a result, overdensities form in the wake of each galaxy, creating a net drag on both galaxies. This drag lowers the group velocity of the stars in each galaxy, but at the same time it increases the peculiar velocities of individual stars.

This effect is most easily seen in the case of a minor merger between a small system of mass M and a large system made of constituent particles of mass $m \ll M$. We approximate the small system as a point mass with velocity $\vec{v}_M$, and the large system as an initially uniform density field of particles with mass m. Under these circumstances the phase space diffusion of M simplifies to an acceleration proportional to the Laplacian of the velocity distribution of the field particles and pointed in the opposite direction of $\vec{v}_M$, i.e.

$$\frac{d\vec{v}_M}{dt} = D\Delta v \tag{97}$$

This can be expanded using the equations derived in Ch. 7 of Binney & Tremaine. If we further assume an isotropic velocity distribution for the field particles, then we obtain Chandrasekhar's dynamical friction formula,

$$\frac{d\vec{v}_M}{dt} \approx -\frac{16\pi^2}{3} G^2 M m \ln(\Lambda) \frac{\vec{v}_M}{v_M^3} \int_0^{v_M} v_a^2 f(v_a) dv_a, \tag{98}$$

where Λ is the "Coulomb logarithm" ($\Lambda \approx \frac{M_a}{M}\frac{R}{R_a}$), subscript a represents properties of the field particle system and f is the distribution function of field particles. We now assume that the distribution function is a Maxwellian with velocity dispersion σ. Under these conditions, and Eqn. 98 reduces to

$$\frac{d\vec{v}_M}{dt} = -\frac{4\pi G^2 M \rho \ln(\Lambda)}{v_M^3} \left[\mathrm{erf}(X) - \frac{2X}{\sqrt{\pi}} e^{-X^2} \right] \vec{v}_M, \tag{99}$$

where $X = \frac{v_M}{\sqrt{2}\sigma}$. In the limit of $v_M \ll \sigma$, $\mathrm{erf}(X) = \frac{2}{\sqrt{\pi}}\left(X - X^3/3\right)$ and $e^{-X^2} \approx 1 - X^2$, and so

$$\frac{d\vec{v}_M}{dt} = -4\pi G^2 M \rho \ln(\Lambda) \sqrt{\frac{2}{9\pi}} \frac{\vec{v}_M}{\sigma^3}, \tag{100}$$

and the limit of $v_M \gg \sigma$

$$\frac{d\vec{v}_M}{dt} = -\frac{4\pi G^2 M \rho \ln(\Lambda)}{v_M^3} \vec{v}_M, \tag{101}$$

which is in the same form as Eqn. 26.1 of Carroll & Ostlie, $\frac{\vec{f}_d}{M} \approx C(v_M, \sigma)\frac{G^2 M \rho}{v_M^2}\hat{v}_M$. The various terms have simple motivations: the force depends obviously on the mass density of background stars, since a greater density would entail a greater gravitational force on M. M factors into the acceleration (and therefore M^2 factors into the force) since a larger mass will stir up more of a wake of particles. In the case where $v_M \gg \sigma$, a factor of two $\vec{v}_M$ increase means making the gravitational deflections between M and the field particles more glancing, halving the impulse given to field particles. Since the field particles now take twice as long to create an overdensity, M is twice as far away, and the gravitational pull goes down by a factor of four. In the case where $v_M \ll \sigma$, more velocity creates a greater density gradient between the space "in front" of M and the space "behind" M.

Let us calculate the inspiral time of a satellite galaxy into its much more massive parent, what is known as a "minor merger". The dark matter halo of a (large) galaxy can roughly be approximated by the density gradient $\rho(r) = \frac{v_M^2}{4\pi G r^2}$. In this scenario, $v_M \ll \sigma$, so we use $\frac{\vec{f}_d}{M} \approx C(v_M, \sigma)\frac{G^2 M \rho}{v_M^2}\hat{v}_M$, and we substitute in $\rho(r)$ to obtain $f_d = C(v_M, \sigma)\frac{GM^2}{4\pi r^2}$. A torque $r f_d$ is being applied on the satellite due to dynamical friction, which will lead to a loss in $L = M v_M r$. For a flat rotation curve $v(r) \approx v_M$, and we therefore obtain

$$M v_M \frac{dr}{dt} \approx -rC\frac{GM^2}{4\pi r^2}, \tag{102}$$

(assuming σ is constant in space). Solving this first-order ODE from some initial radius $r = r_i$ to $r = 0$ gives us

$$t_c = \frac{2\pi v_M r_i^2}{C(v_M, \sigma) GM}. \tag{103}$$

If we plug in $t_c = T_{\mathrm{gal}}$, the age of the galaxy, and invert for r_i, we can determine the maximum distance satellites could have been accreted from via dynamical friction. A refined version of this estimate (Binney & Tremaine, pg. 650) that includes an estimate for tidal stripping of the satellite during its infall (see below) finds that a satellite galaxy in a circular orbit around the Milky Way at 30 kpc will have merged with the MW in 10 Gyr.

3.10.1. *Under what conditions does the above formulation of dynamical friction hold?*

Both satellite galaxies and globular clusters can be accreted by dynamical friction, and despite the fact that they are not point masses, and approximations were made in the derivation sketch above (most importantly we neglected the gravity an overdensity of field particles would have on their neighbouring field particles), the dynamical friction

equations for a point mass in a field of much less massive particles gives a good approximation to the true dynamical friction felt by satellites and GCs (see Binney & Tremaine, pg. 646 - 647 for detailed arguments). A further refinement can be made (Binney & Tremaine 2008, pg. 649 - 650) to account for tidal stripping of the galaxy or cluster during its infall. From such a calculation, the LMC and SMC will merger with the MW in about 6 Gyr (note that Andromeda will merger with the MW in 3).

Dynamical friction is also the reason supermassive black holes cannot linger in the outer regions of a galactic potential, but rather must approach the galactic centre and form a binary with the SMBH already there - for an MW-like spiral the inspiral time is ~ 3 Gyr (Binney & Tremaine 2008). Unlike for isolated stellar black hole binaries, after a supermassive black hole binary is formed dynamical friction continues to act on each black hole, which may eventually drive the SMBH binary close enough that gravitational radiation takes over as the primary angular momentum loss mechanism.

Dynamical friction also affects galactic bars (Binney & Tremaine 2008).

3.10.2. *Under what conditions does it not?*

The physical motivation behind dynamical friction is the same used to transfer a system's kinetic energy into thermal energy in general and is applicable to system interactions in general. The formalism describe above, however, does not work under all conditions.

Major mergers, or mergers between two stellar systems of comparable size, cannot be described by the particular dynamical friction formalism described above (for obvious reasons - the "field" is the same size as the object!). Evidence of major mergers exist in the form of elliptical galaxies, which are much more common in the centres of clusters, where more mergers are expected.

During the major merger of two galaxies, gravitational deflections between individual stars will be rare (the relaxation time is far longer than the merger time) but the significant changes in the overall potential of the system will result in transferral of the kinetic energy of the galaxies moving with respect to each other into the kinetic energy of individual, randomly moving stars - a tranferral of galactic kinetic energy into thermal energy (Binney & Tremaine 2008). Gas is collisional, and so during the merger shock compression from colliding gas flows will serve to greatly enhance the star formation rate of both galaxies.

A simple random-walk argument for collisions suggests that major mergers should occur once every 10^6 Gyr, and until the 1970s most astronomers did not believe mergers played a major role in galaxy evolution (Binney & Tremaine 2008). This is incorrect for two reason. First, the true size of a galaxy is the size of its dark matter halo, which will merger with other halos. Once this occurs, dynamical friction brings the baryonic components of the galaxies to the centre of the merging halos. Second, the peculiar velocities of galaxies are not randomly oriented, but in general oriented toward each other - this is a consequence of the fact that these peculiar velocities were originally generated by inter-galaxy gravitational attraction.

High-speed encounters between two galaxies of roughly equal size (such as the first pass of a major merger) will not produce significant slowing through dynamical friction, but will increase the overall kinetic energy of each galaxy (Carroll & Ostlie 2006). Here, "high-speed" refers to the interval during with mutual gravity between the two galaxies is significant being much shorter than the crossing time within either galaxy (Binney & Tremaine 2008). As the galaxies virialize once again, this increase in kinetic energy will serve to puff up each galaxy as a whole, or eject gas and stars each galaxy (the Cartwheel Galaxy is an extreme example of a rapid head-on collision; its circle of star formation comes from an expanding ring of material ejected from an interaction with another galaxy) (Carroll & Ostlie 2006).

Extended bodies will also tidally strip streams of gas and dust off of each other. This tidal stripping is the cause of "polar-ring" galaxies (disk galaxies with a ring of material significantly deviated from the plane of the disk) and "dust-lane ellipticals" (ellipticals with a dust disk) (Carroll & Ostlie 2006).

3.11. *Question 12*

QUESTION: Sketch the SED, from the radio to Gamma, of a spiral galaxy like the Milky Way. Describe the source and radiative mechanism of each feature.

The SED of a galaxy is composed of the combined SEDs of the emitting bodies within that galaxy. For spiral galaxies, the majority of their emission is from the infrared to the optical, and is caused, directly or indirectly, by stars. Minor contribution comes from more exotic objects such as x-ray binaries and cosmic rays, but because these contributions are in regions (radio and gamma) where stars, hot gas and dust do not emit in great amounts, they also disproportionately affect the SED. The full breakdown:

- **Radio**: synchrotron emission from relativistic electrons interacting with ISM magnetic fields (exceptionally strong near SNR), and free-free radio emission from H II regions. The H I 21 cm line is also of note.
- **Millimetre/sub-millimetre**: line emission from cold molecular gas (such as CO).
- **Far-IR**: Warm dust blackbody emission - responsible for the 100 μm bump in Fig. 61 upper left panel.

- **Mid-IR**: Polycyclic aromatic hydrocarbon (PAH) and other complex organic molecule emission - responsible for the complex forest of lines between the two bumps in Fig. 61 upper left panel.
- **Near-IR**: Cold star blackbody emission.
- **Optical**: Star blackbody emission and line emission from H II regions - responsible for the optical bump in Fig. 61 upper left panel. Significant absorption also exists at these wavelengths - this light is reprocessed into mid- and far-IR radiation.
- **Ultraviolet**: Blue star blackbody emission. Emission lines from metals.
- **X-rays**: Hot, shocked gas and accretion onto compact objects.
- **Gamma rays**: Cosmic ray/ISM interactions, accretion onto compact objects.

Fig. 61 gives a visual description of νF_ν (a metric for spectral energy, since $E = h\nu$) for four different types of galaxies.

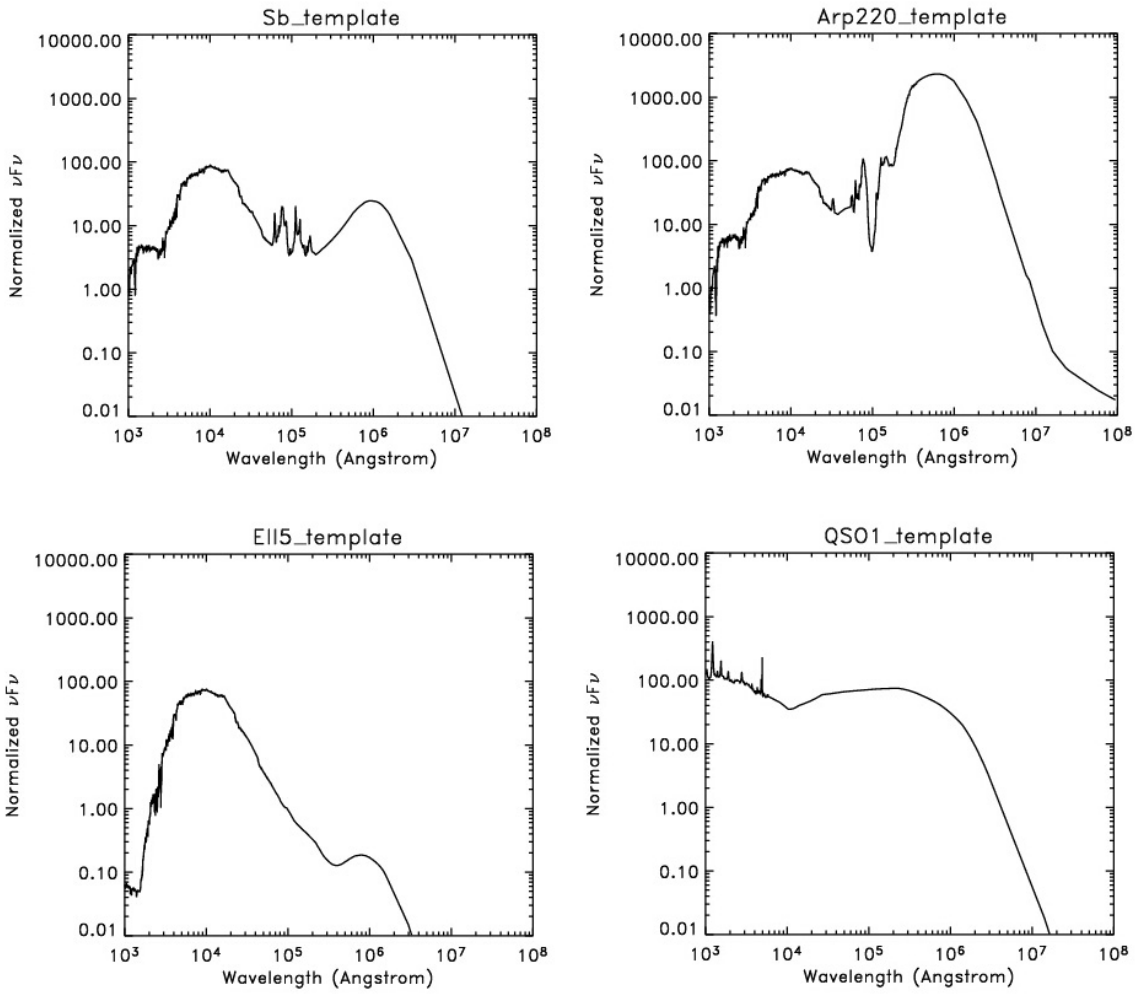

FIG. 61.— A comparsion between the SWIRE templates of an Sb SED (representative of the Milky Way), a 5 Gyr old elliptical, a ULIRG (Arp 220) and a type-1 QSO. Based on data from Polletta (2006).

3.11.1. *How about for different galaxies?*

Elliptical galaxies tend to have a large number of older stars, and very little gas and dust. Because of this, the elliptical SED is to first order a single blackbody. ULIRGS, on the other hand, have copious amounts of dust, and are therefore tremendously bright in the far-IR. QSOs have a nearly flat emission spectrum over a large portion of their

emission due to most of their emission coming from thermal emission of the accretion disk and synchrotron emission from the relativistic jet of the central AGN.

The same basic SED building concepts for the MW can be used for any galaxy. A dSph, for example, would look roughly equivalent to an elliptical due to having mostly stars (though of varying populations) and very little gas and dust.

3.12. *Question 13*

QUESTION: How many stars does one expect to find within 100 pc of the Sun? If all stars are distributed evenly across the galaxy, how many of these will be B spectral type or earlier? How many of these are younger than 100 Myrs?

The Milky Way contains some $N = 10^{11}$ stars that are roughly distributed in a disk of radius $R_{\rm disk} \sim 10$ kpc and thickness of $h_{\rm disk} \sim 1$ kpc. Assuming a uniform distribution of stars (not a terrible approximation considering that open clusters are gravitationally unbound, there are about 10^6 stars within 100 pc of the Sun. This is roughly consistent with there being about 1 star/pc in the Solar neighbourhood, which is indeed the case.

To determine the number that are B stars, we need to consider the evolutionary history of the Milky Way. Suppose we assume a Salpeter initial mass function $\phi(M,t) = \phi(M) \propto M^{-2.35}$ that stretches from the H-burning limit at $\sim 0.1 M_\odot$ to $100 M_\odot$ and does not change with time, and a constant star formation rate ψ, the present-day mass function (PDMF) is given by $\Phi(M)$, where

$$\begin{aligned}\Phi(M) &= \frac{\int_{t_{\rm MWbirth}}^{t_{\rm now}} \phi(M,t)\psi(t)H_{\rm death}(t_{\rm now}-t,M)dt}{\int_{0.1}^{100}\int_{t_{\rm MWbirth}}^{t_{\rm now}} \phi(M,t)\psi(t)H_{\rm death}(t_{\rm now}-t,M)(t)dtdM} \\ &= \frac{\phi(M)\min(T,T_{\rm gal})}{T_{\rm gal}\int_{0.1}^{M_c}\phi(M)dM + \int_{M_c}^{100}\phi(M)T(M)dM} \end{aligned} \tag{104}$$

where $t_{\rm MWbirth}$ is the time of the birth of the Milky Way, $H_{\rm death}(t_{\rm now}-t,M)$ is a Heaviside step function to represent the death of stars of a certain mass, $T(M)$ is the star's age, and M_c is the mass of a star with a lifespan the same as the age of the Milky Way. The best way to think of this equation is that the top times dM is proportional to the number of stars in question, while the bottom is proportional to the number of stars ever formed in a clump of the Milky Way, and both top and bottom have the same coefficient of proportionality. We have removed the Heaviside step function by noting that it basically nullified the integrand for all values of $t_{\rm now} - t > T(M)$, and that ϕ and ψ are, in our case, time-independent. Then $\int_{t_{\rm MWbirth}}^{t_{\rm now}} H_{\rm death}(t_{\rm now}-t,M)dt = t_{\rm now} - (t_{\rm now} - T(M)) = T(M)$. M is in solar masses. Further simplification comes from the fact that main sequence lifetime $T \propto M/L$, and we known that $L \propto M^{-\sim 3.5}$, giving us $T = 10{\rm Gyr}\, M^{-2.5}$. We then get

$$\Phi(M) = \frac{M^{-4.85}}{\int_{0.1}^{M_c} M^{-2.35}dM + \int_{M_c}^{100} M^{-4.85}dM} \tag{105}$$

noting that all normalizations wash out because they appear in both the numerator and denominator. We can assume $t_{\rm now} - t_{\rm MWbirth} \approx 10$ Gyr (Chabrier 2003), giving us $M_c = 1$. Appendix G of Carroll & Ostlie (2006) gives B stars to be $\sim 3\ M_\odot$ or above. All this gives

$$\begin{aligned} f &= \frac{\int_3^{100} M^{-4.85}dM}{\int_{0.1}^{1} M^{-2.35}dM + \int_1^{100} M^{-4.85}dM} \\ &= \frac{-\frac{1}{3.85}(100^{-3.85} - 3^{-3.85})}{-\frac{1}{1.35}(1-0.1^{-1.35}) - \frac{1}{3.85}(100^{-3.85}-1)} \\ &= 2.35\times 10^{-4} \\ &\approx \frac{\frac{1}{3.85}3^{-3.85}}{\frac{1}{1.35}0.1^{-1.35}} = 2.3\times 10^{-4} \end{aligned} \tag{106}$$

$$\tag{107}$$

Since there are 10^6 stars in our Solar neighbourhood, this gives about 200 B stars within 100 pc of the Sun.

To determine the number of stars that are 100 Myr or younger (these may or may not be long lived, so long as they

were born recently), we modify Eqn. 104 slightly:

$$f = \frac{0.1\int_0 .1^{M_{c2}} \phi M dM + \int_{M_{c2}}^{100} \phi(M) T(M) dM}{10\int_{0.1}^{M_c} \phi(M) dM + \int_{M_c}^{100} \phi(M) T(M) dM} \tag{108}$$

where M_{c2} is the mass of a star with a lifespan of 100 Myr and all times have been converted to Gyr (meaning $T(M) = 10M^{-2.5}$). From $T = 10\text{Gyr}\, M^{-2.5}$, $M_{c2} = 6.31 \approx 6$, so

$$\begin{aligned} f &= \frac{0.1\int_0 .1^{6} M^{-2.35} dM + 10\int_{6}^{100} M^{-4.85} dM}{10\int_{0.1}^{1} M^{-2.35} dM + 10\int_{1}^{100} M^{-4.85} dM} \\ &= \frac{-\frac{0.1}{1.35}(6^{-1.35} - 0.1^{-1.35}) - \frac{10}{3.85}(100^{-3.85} - 6^{-3.85})}{-\frac{10}{1.35}(1 - 0.1^{-1.35}) - \frac{10}{3.85}(100^{-3.85} - 1)} \\ &= 1.03 \times 10^{-2} \\ &\approx \frac{\frac{0.1}{1.35} 0.1^{-1.35}}{\frac{10}{1.35} 0.1^{-1.35}} = 1 \times 10^{-2} \end{aligned} \tag{109}$$

$$\tag{110}$$

One in a hundred stars is younger than 100 Myr.

3.12.1. *State the assumptions in this problem.*

A number of assumptions have gone into this problem:

- We used the Salpeter IMF for simplicity. The Salpeter, unfortunately, greatly overestimates the amount of stars less massive than 1 $M_\odot$, a fact first pointed out by Miller & Scalo (1979). Chabrier's current galactic disk IMF further lowers the number of stars less than 1 $M_\odot$.
- We assumed the IMF was independent of time. From 3.1, we know this to potentially be untrue.
- We assumed a constant star formation rate. The cosmic star formation rate density (SFRD) can observationally be determined by observing the luminosity of galaxies at different redshifts - since short-lived massive stars contribute disproportionately to the total luminosity of a galaxy, making luminosity a function of the star formation rate (it is also common to look only at the UV luminosity, since this is where the largest change will occur). From such observations we know that the cosmic star formation rate is *not* constant, and increases past $z = 1$, turns over, and decreases after $z = 3$ out to reionization (Shim et al. 2009; Cucciati et al. 2012). The exact turning point around $z \sim 2$ is still not particularly well constrained (Cucciati et al. 2012). See Fig. 62.
- Stellar mass loss is not accounted for.
- No binary systems are assumed. Binaries allow for the possibility of mass transfer between stars, which can produce very odd looking stars.

3.13. *Question 14*

QUESTION: Describe what happens as a cloud starts to collapse and form a star. What is the difference between the collapse and contraction stages? What happens to the internal temperature in both? When does the contraction phase end, and why does the end point depend on the mass of the object?

Most of this information comes from (Kippenhahn & Weigert 1994, Ch. 27 - 28) and (Carroll & Ostlie 2006, Ch. 12)

The criterion for hydrostatic collapse is that $P \propto \rho^\Gamma$ where $\Gamma \leq 4/3$. Because clouds with efficient cooling are roughly isothermal, meaning $\Gamma = 4/3$, we expect collapse to be inevitable once a cloud becomes unstable to collapse (see below for a more precise definition). The resultant collapse is isothermal, and occurs on a free-fall time, Eqn. 111.

$$t_{\text{ff}} = \sqrt{\frac{32\pi}{G\rho}}. \tag{111}$$

This number is approximately several million years. Since the free fall timescale near the centre of the cloud is shorter than that near the outskirts, the core of the cloud collapses faster. From Kramer's rule, that opacity $\kappa \propto \rho T^{-7/2}$, the

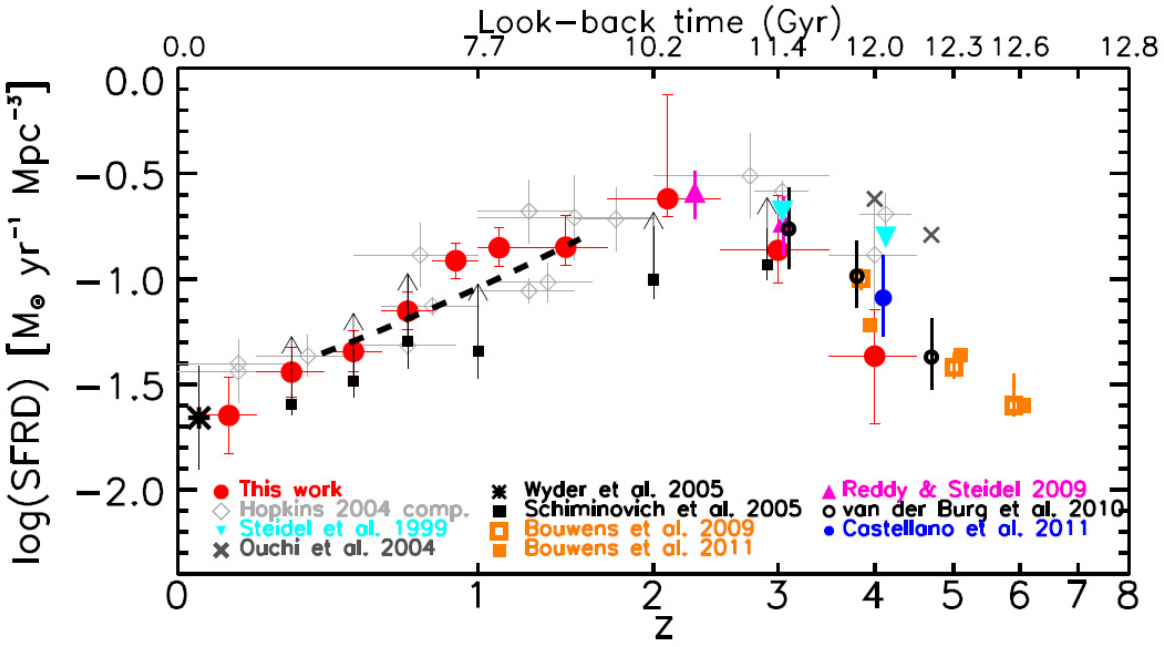

FIG. 62.— The cosmic star formation rate out to $z = 8$, as measured by a large number of deep, high-redshift galaxy surveys. From Cucciati et al. (2012), their Fig. 5 (which has more information on the specific objects observed for each survey).

core of the cloud must at some point become optically thick. This results in the equation of state becoming adiabatic, and since an equilibrium γ of 7/5 is stable to collapse, the core pressure increases until the core is in quasi-hydrostatic equilibrium; this core is now known as a protostar. Thermal energy escapes as radiation, and so without any avenue of (rapid) emission, we now have $\gamma = 7/5$ and $\rho \propto T^{5/2}$ (from diatomic gas ideal equation of state), and the star greatly heats up and pressure support increase. In the outer regions of the cloud matter is still falling in on a free-fall timescale, resulting in a shockwave forming at the surface of the of the protostar.

When the temperature reaches approximately 1000 K, the dust within the developing protostar vaporizes and the opacity drops. This substantially decreases the optical depth, resulting in a smaller photosphere, and therefore higher effective temperatures.

When temperatures reach past 2000 K, dissociation of H_2 begins. A fraction of the work done by collapse will go into ionizing H_2 rather than heating the system, and γ briefly drops below 4/3, resulting in another collapse. Once this is completely finished, $\gamma = 5/3$, and equilibrium is re-established. A second shock front forms at the surface of this core, which quenches the first shock, above. Soon after equilibrium is reached, deuterium fusion occurs ($^2_1\mathrm{H} + p \rightarrow ^3\mathrm{He} + \gamma$), and while it does contribute to the total luminosity, it does not stop the collapse (as there is too little deuterium).

Fig. 63 shows the collapse of the cloud on a ρ-T diagram, and the collapse of a number of different clouds on the HR diagram.

Eventually, the accretion timescale increases significantly, due to the thinning out of the disk. At this point, the mass of the protostar becomes fixed (to first order), and subsequent evolution occurs on a Kelvin-Helmholtz timescale as the protostar contracts - this evolution is charted in Fig. 64. From virial theorem, $T \propto R^{-1}$, and of course $\rho \propto R^{-3}$, which nets us $T \propto \rho^{1/3}$. This is significant because the degeneracy line $\rho T \propto P_{\mathrm{ideal}} \approx P_{\mathrm{degeneracy}} \propto \rho^{5/3}$ have $T \propto \rho^{2/3}$, and so it is possible for contraction to eventually be stopped by degeneracy before nuclear fusion begins. This occurs for objects less than $\sim$0.08 $M_\odot$, who never see significant amounts of fusion. They are known as brown dwarfs.

During the contraction phase, heavy elements ionize, leading to a population of H^- ions in the upper atmosphere. The resulting large opacity drives convection, which extends deep into the star. As luminosity decreases **WHY???**, the convective star's effective temperature grows slightly, as it is near the Hayashi line. Meanwhile, the central temperature is increasing, which for $\gtrsim$1 $M_\odot$ stars leads to a more radiative core and allowing energy to escape more readily into the convective envelope, increasing luminosity. At the bottom of the downward Hayashi track, the central temperature becomes hot enough to begin nuclear fusion. For $M_\odot$ stars the first two steps of the PP I chain, and the CNO reactions that turn ^{12}C into ^{14}N dominate the nuclear energy production, while for less massive stars only the PP I chain is ignited. Over time these reactions begin to dominate over gravitational contraction as an energy source, and the luminosity increases substantially. The CNO cycle establishes a steep temperature gradient, and causes the core to expand slightly, which drops the luminosity and temperature slightly (in Fig. 64 this occurs near the dotted line). In $M_\odot$ stars ^{12}C is eventually exhausted, while the core temperature becomes hot enough to ignite the entire PP chain. In more massive stars ^{12}C reaches its equilibrium value on the main sequence as the temperature becomes hot enough to ignite the entire CNO cycle. Either way, the star moves onto its zero-age main sequence (ZAMS) position on the HR diagram.

A 1 $M_\odot$ star takes approximately 40 Myr to complete the contraction phase, the same order of magnitude as given by the Kelvin-Helmholtz time of the star.

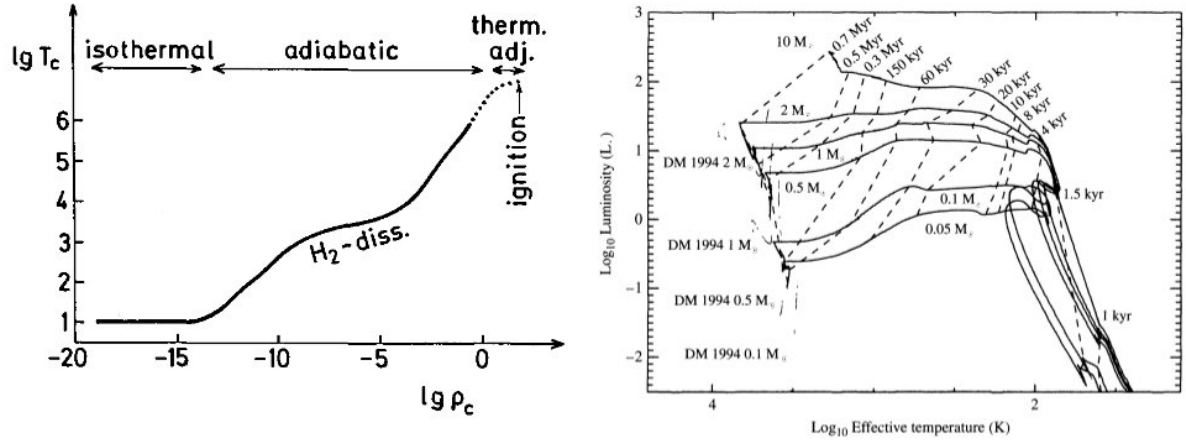

FIG. 63.— Left: evolution of a collapsing cloud on a ρ-T diagram. Adiabatic collapse begins due to increased opacity, while H_2 dissociation reduces the temperature increase. Adiabatic collapse ends when the thermal adjustment time is smaller than the mass accretion timescale. During the left/upward travel near 100 K, most of the luminosity is accretion-powered. Once the flat region is reached, accretion has diminshed enough that most of the energy comes from deuterium fusion and gravitational contraction (Kippenhahn & Weigert and Carroll & Ostlie disagree on which dominates). The beginnings of evolution along the Hayashi line can be seen on the far left. From Kippenhahn & Weigert (1994), their Fig. 27.3, and Carroll & Ostlie (2006), their Fig. 12.9.

3.13.1. *What complications are there?*

If we plug the density of a GMC core into the free-fall time equation, we get of order only 10^5 yrs; dense molecular cores, however, are commonly observed. This suggests we are missing important aspects of cloud collapse.

Neglecting the spherical symmetry approximation, the cloud may have initial velocity, rotation and turbulence. We neglect radiative transfer as well. Importantly, we neglect magnetic fields. Since magnetic fields resist a change in flux, fields buoy clouds against collapse. If magnetic fields are the only pressure against collapse, the critical mass needed for collapse is estimated to be

$$M_B \approx 70 M_\odot \left(\frac{B}{1\,\mathrm{nT}}\right)\left(\frac{R}{1\,\mathrm{pc}}\right)^2. \tag{112}$$

If the cloud Jeans Mass exceeds this value, then the cloud is "magnetically supercritical" and can collapse.

A molecular cloud is neutral, and retains most of its magnetic buoyancy through coupling between neutral and charged particles. Since the coupling is not absolute, it is possible for neutral matter to drift away from ions in a process called ambipolar diffusion. Suppose the neutral material has a net velocity $\vec{v}$ such that the force due to collisions between charged particles (with velocity 0) and neutral particles is balanced out by the Lorentz force. The charged material can then prevent itself from moving, while the neutral material moves through it - this allows for a relative reduction in magnetic flux within a collapsing cloud.

Other magnetic effects include Parker Instability, which arises from the presence of magnetic fields in a plasma in a gravitational field, wherein the magnetic buoyant pressure expels the gas and causes the gas to move along the field lines. Magnetic fields can expel angular momentum from a collapsing cloud, helping it collapse further.

Observationally, dense molecular cores are in hydrostatic equilibrium, suggesting that the very beginning of a collapse is quasi-static. They are also quite small ($\sim$0.1 pc), and physical size is not taken into account in our model above.

Massive star formation may become so luminous during their pre-main sequence evolution that they interfere with their own accretion through stellar feedback. This has led to suggestions that massive stars form not through the channel above, but through mergers of multiple small protostars, or through an accreting protostellar disk.

3.13.2. *What triggers the collapse?*

The collapse is triggered by density perturbations to a cloud in an unstable equilibrium. Consider a cold, thin isothermal cloud, assume spherical symmetry, and rope off a section of it. From the virial theorem with an external P,

$$\xi E_i + E_g = 4\pi R^3 P_{\mathrm{ext}} \tag{113}$$

$\xi = 2$, and $E_i = 3MkT/\mu m_H$ for ideal gases, while $E_g = -\frac{3}{5}GM/R$ for an isodensity cloud. If we take the derivative of P with respect to R, we obtain a characteristic maximum in P_{ext} at R_m. For lower radii, $dP_{\mathrm{ext}}/dR > 0$, while for

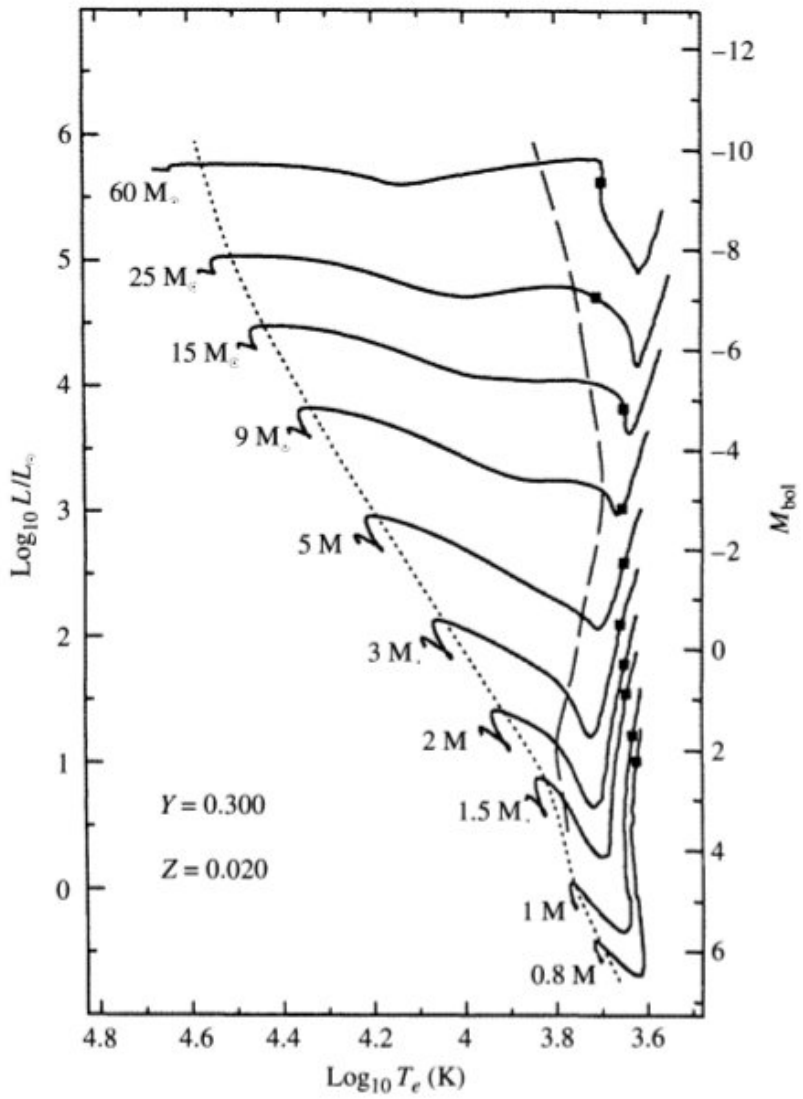

FIG. 64.— A series of classical pre-main sequence evolutionary tracks. Evolution is from the right to the left. The square on each track indicates where deuterium fusion begins (it is inconsistent with where deuterium fusion begins during the collapse phase), the long-dashed line represents when the envelope switches from being convective to radiative, and the dotted line represents when the core becomes convective. From Carroll & Ostlie (2006), their Fig. 12.11.

higher radii, $dP_{\rm ext}/dR < 0$ (pressure can be negative). Therefore, at $R > R_m$, a compression will result in an increase in pressure (a stabilizing effect) while for $R < R_m$, a compression will result in a decrease in pressure (a destabilizing effect). States where $R < R_m$, then, are unstable to collapse. R_m is known as the Jeans mass, and is given by

$$R_m = R_J = \sqrt{\frac{45kT}{16\pi G\mu m_H \bar{\rho}}}, \tag{114}$$

where $\bar{\rho} = \frac{3M}{4\pi R_m^3}$. A similar derivation exists where $2E_i < -E_g$ (i.e. $P_{\rm ext} = 0$) is the condition for unstable equilibrium, which gives a 15% deviation from the value calculated here. Note that all stars have $P_{\rm ext} = 0$, which corresponds to an $R < R_m$ (or $R = R_m$ for the $2E_i < -E_g$ derivation), and therefore an isothermal star is doomed to collapse. The physical interpretation of this derivation is that if a mass M of a certain temperature and composition can be squeezed into a sphere of radius R_J or less, it will collapse. Reversing to obtain M_J

$$M_J = \frac{4\pi}{3}\bar{\rho}R_J^3 = \frac{45}{16}\left(\frac{5}{\pi}\right)^{1/2}\left(\frac{kT}{m_H G\mu}\right)^{3/2}\left(\frac{1}{\bar{\rho}}\right)^{1/2} \tag{115}$$

The interpretation for this is given a cloud of a certain temperature and density, if it exceeds mass M_J it will collapse given a push.

3.13.3. *What causes these overdensities?*

In general, density perturbations in the ISM are caused by turbulence (Nguyen 2012). Turbulence can generated by any of the following: magnetorotational instability in galactic disks, spiral arm instabilities, accretion of extragalactic gas, colliding flows in molecular clouds, and expanding HII regions (Nguyen 2012).

3.13.4. *Why do D and Li fusion occur before H fusion? Do they change this cycle?*

Deuterium and lithium fusion both occur before stars reach the ZAMS. D burning was described above as occuring near 2000 - 3000 K sometime during the accretion to homologous contraction transition phase. Li burning (both ^{6}Li and ^{7}Li have large cross-sections for proton capture) occurs at around 2×10^6 K, which is achieved just before the primary H-burning nuclear processes begin to dominate (Nelson et al. 1993). Since there is very little of either element in protostars, burning Li and D do not change significantly the overall evolution of protostars onto the main sequence. Objects destined to become brown dwarfs can actually burn D (for masses $\gtrsim$0.013 $M_\odot$) and Li ($\gtrsim$0.06 $M_\odot$), but doing so does not stop them from passing the degeneracy line and cooling.

3.13.5. *How does fragmentation work?*

Consider the Jeans mass, Eqn. 115, which is $M_J \propto T^{3/2}\rho^{-1/2}$. If the equation of state is isothermal, then $M_J \propto \rho^{-1/2}$. If however the equation of state is adiabatic, then $PV^\gamma \propto P^{1-\gamma}T^\gamma$ is a constant, giving $P \propto T^{5/2}$, which, combined with $P \propto \rho T$ (ideal gas law), gives us $T \propto \rho^{2/3}$ and $M_J \propto \rho^{1/2}$.

What this means is that during the initial collapse, the Jeans mass decreases (as it decreases with increasing density), allowing local regions of collapsing material to themselves collapse independently, leading to fragmentation of the cloud. Once the cloud becomes optically thick, however, fragmentation must end, since the Jeans mass now increases with density. A rudimentary means (Kippenhahn & Weigert 1994, pg. 254 - 255) of calculating the M_J at which the cloud collapses is to equate the energy loss rate (gravitational potential energy divided by the free fall time) with the Stefan-Boltzmann law (since a blackbody maximizes the amount of light the object can emit). This gives us (assuming a 1000 K cloud at the time the system becomes optically thick) M_J is of order 1 $M_\odot$.

3.14. *Question 15*

QUESTION: Sketch the rotation curve for a typical spiral galaxy. Show that a flat rotation curve implies the existence of a dark matter halo with a density profile that drops off as $1/r^2$.

If we approximate the movements of disk stars as circular and on the plane of the galactic disk, then uniform circular motion gives (Schneider 2006, pg. 100 - 101)

$$v^2(r) = GM(r)/r. \tag{116}$$

If dark matter did not exist, we may use $M(r) = M_{\rm lum}$ and assume some mass to light ratio to convert surface brightness to mass. While the M/L can indeed be adjustable (one can either determine it by determining the stellar population in a region of the galaxy and the finding the M/L from stellar evolution models, or simply use M/L as a fitting parameter), the rotation curves of spiral galaxies are, at large r, flat, which is wholly incompatible with any scaled M/L as we shall see below.

The dark matter can be determined by (Schneider 2006, pg. 100 - 101)

$$M_{\rm dark}(r) = \frac{r}{G}(v^2(r) - v_{\rm lum}^2(r)) \tag{117}$$

Suppose mass is distributed in our galaxy with a $\rho = A/r^2$ profile. From $dM/dr = 4\pi r^2 \rho dr$, $M(r) = 4\pi A r$. Plugging this into Eqn. 116 gives us a flat velocity profile.

The reason this flat velocity curve implies the existence of dark matter is because observationally the rotation curve remains flat to as large an r as H I kinematics ($R_{\rm max} = 30h^{-1}$ kpc) or satellite galaxy orbital statistics ($R_{\rm max} = 100h^{-1}$ kpc) allow us to probe (Schneider 2006, pg. 101). Since for an $1/r^2$ density profile $M \propto r$, the only thing limiting the halo mass is the extent of the halo. This, however, is not the case for the density of stars in the galactic halo, which has $\rho \propto r^{-3.5}$. The stellar halo therefore has a definite maximum mass (assuming the central density is a finite constant) and cannot produce a flat velocity profile out to very large radii. We therefore must supplement luminous matter with a large body of dark matter.

Fig. 65 shows a number of rotation curves for different spiral galaxies. The decomposition of the rotation curve of NGC 3198 into a disk and halo component is not unique, and in this case it has been assumed that the entirety of the innermost rotation curve is due to luminous matter (this is known as the "minimal mass halo").

The innermost region of the rotation curve has velocity $v \propto r$, indicating rigid body rotation due to a $M(r) \propto r^3$ (i.e. constant density). This suggests that the bulge and inner halo have a large constant mass density that dominates over the mass contribution of the disk and dark matter halo (see Carroll & Ostlie (2006), pg. 919, Fig. 24.28).

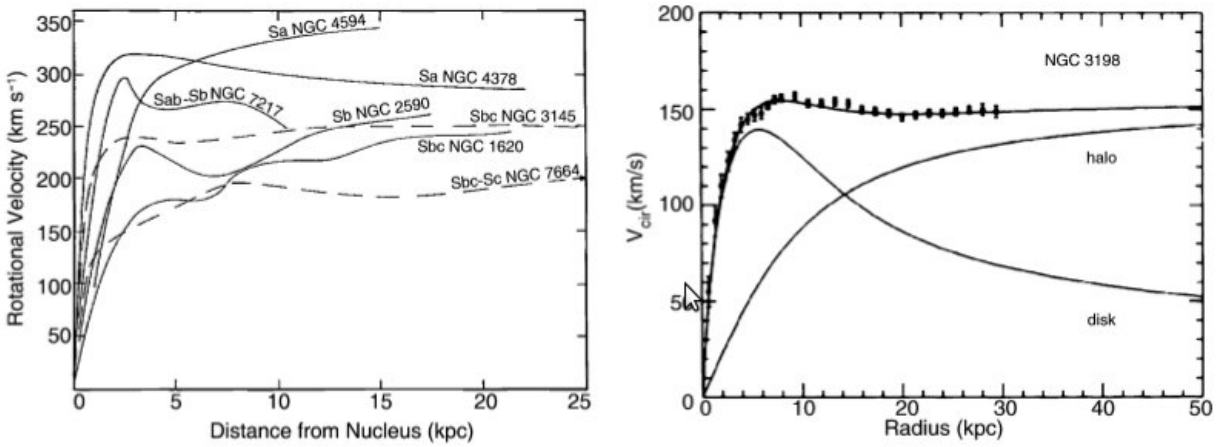

FIG. 65.— Left: the rotation curves of a number of spiral galaxies, all of which asymptote to some roughly constant value at large radii. Right: the rotation curve of NGC 3198, and a decomposition of that curve into a disk and halo component assuming a minimal mass halo. From Schneider (2006), his Figs. 3.15 and 3.16.

Cosmological simulations by Navarro et al. (1996) give the following profile for a dark matter halo:

$$\rho(r) = \frac{\rho_0}{(r/a)(1+r/a)^2} \tag{118}$$

Despite no longer diverging because of the asymptote at the centre, the NFW profile is still divergent if $r \to \infty$, and so an outer limit for the disk must be defined.

3.14.1. *How do halo properties scale with the properties of luminous matter in late-type galaxies?*

This information is from Schneider (2006). More luminous galaxies tend to have more massive halos. For the characteristic values of the various Hubble types, one finds $v_{\max} \sim 300$ km/s for Sa's, $v_{\max} \sim 175$ km/s for Scs, whereas Irrs have a much lower $v_{\max} \sim 70$ km/s. Given the same overall luminosity, $v_{\max}$ is higher for earlier types of spirals, reflecting an increase in bulge mass. The overall shape of the rotation curve does not change, indicating that dark matter (with a common density profile) dominates spiral galaxy rotation curves after several kpc.

3.14.2. *How do we observationally determine the rotational profiles of galaxies?*

Observations of the rotation of our own Galaxy are done mostly by using long-wavelength (to sidestep extinction) gas tracers, or, with stars very close to the Sun, by measuring the Oort constants. In the region of the galaxy interior of our own orbit, this can be done using the "tangent point" method if we assume orbits are circular (they should be due to dissipation) - measuring maximum Doppler shift along one line of sight will give us the rotational velocity of the galaxy at one radius (see (Schneider 2006, pg. 62)). In the region exterior to our galaxy, this is mainly done by looking at the radial velocity of objects with known distance measurements.

The rotation profiles of other galaxies are much easier to observe. Doppler shift observations of the H I disk and the stars, combined with inclination information obtained by disk inclination, allow us to probe the rotation profiles of spirals and lencticulars. Early-type galaxies often lack an H I disk, and astronomers instead rely on observing stellar kinematics, and then binning large swaths of the galaxy together to increase signal-to-noise (integral field spectrographs are useful here, since they can create 2D spectral maps of the galaxy, which can then be binned). Alternatively, the motions of other objects, such as planetary nebulae and satellite galaxies, or the temperature of embedded hot gas can be used (Schneider 2006, pg. 101).

3.14.3. *What is the difference between late and early type rotation curves?*

While late-type galaxies get most of their pressure support from rotation, early-type galaxies obtain much of their support from the velocity dispersion (equivalent to a "thermal pressure" in a gas). The observational means of determining this support is absorption line spectroscopy in order to determine the line-of-sight velocity distribution (LOSVD), which can be written out as the velocity v, (Gaussian) dispersion σ and a series of Gauss-Hermite moments starting with h_3.

A good first order measurement for the support against gravity is the RMS velocity $v_{\rm RMS} = \sqrt{v^2 + \sigma^2}$. Much like spiral galaxies, the $v_{\rm RMS}$ along the major axis of an early-type galaxy could be compared to the expected $v_{\rm RMS}$ from a set (or fit-for) M/L and the surface brightness profile of the galaxy. $v_{\rm RMS}$, however, assumes that the absorption line profile is a Gaussian, or, equivalently, that the distribution of stellar kinetic energies is isotropic. In general this is not true, and results in the "mass-anisotropy degeneracy" (Binney & Merrifield 1998, pg.). As an example (Fig. in Binney & Merrifield (1998)), let us define a cylindrical coordinate system for an early-type galaxy with $\hat{z}$ along the galaxy's minor axis. Suppose σ is greater along the $\hat{\theta}$ direction than the $\hat{r}$ direction. An σ measurement of the line of sight crossing the galactic centre would be sensitive to σ_r, but not σ_θ, while the reverse would be true for a line

TABLE 1
Thermal phases of the ISM, from Table 1.3 of Draine (2011) with additional information from elsewhere in Draine and from Ferrière (2001) and (Pogge 2011, Ch. 1). The total mass of hydrogen in the galaxy ($R < 20$ kpc) that is locked in the ISM 4.9×10^9 $M_\odot$, and the total mass of helium is 1.8×10^9 $M_\odot$ (Draine 2011, pg. 5). Ferrière (2001) gives slightly the slightly different number of 1×10^{10} $M_\odot$ (including He), though her mass fractions are similar. Molecular clouds is separated by density into diffuse and dense H_2 in Draine (2011). The fraction of the ISM that is in H I is 0.6 (Draine 2011, Table 1.2), and the CNM and WNM fractions are from (Draine 2011, pg. 331).

Phase	Temperature (K)	Density cm^{-3}	f_V	f_M	Heating Mechanisms	Cooling Mechanisms
Molecular clouds (H_2)	10 - 50	10^2 - 10^6	0.001	0.17	Dust photoelectrons, cosmic rays, starlight	Fine structure (especially CI), CO line emission
Cool H I (CNM)	100	30	0.01	0.24	Dust photoelectrons, cosmic rays, starlight	Fine structure
Warm H I (WNM)	∼5000	0.6	0.4	0.36	Dust photoelectrons, cosmic rays, starlight	Fine structure, optical lines
H II (WIM)	10^4	0.3 - 10^4	0.1	0.23 for all H II	H and He photoelectrons	Optical lines, fine-structure, bremsstrahlung
H II coronal gas (HIM)	$\gtrsim 10^{5.5}$	~ 0.004	∼0.5		SN/stellar wind shock-heating	Adiabatic expansion, X-ray emission

of sight that crossed through the galactic outskirts. To break this degeneracy, the higher order h_3 and h_4 moments, which probe these anisotropies, must be fit for (Binney & Merrifield 1998, pg.). To do this, however, requires good S/N, and as a result extensive long-slit spectroscopy or integral field spectroscopy must be performed (Weijmans et al. 2009).

3.15. *Question 16*

QUESTION: What thermal phases are postulated to exist in the interstellar medium? Describe the dominant mechanism of cooling for each phase.

Most of the information in this section comes from a combination of Draine (2011), (Pogge 2011, Ch. 1) and Ferrière (2001).

Interstellar matter accounts for ∼10 15% of the total mass of the Galactic disk. It tends to concentrate near the Galactic plane and along the spiral arms, while being very inhomogeneously distributed at small scales. Roughly half the interstellar mass is conned to discrete clouds occupying only ∼1 2% of the interstellar volume. These interstellar clouds can be divided into three types: the dark clouds, which are essentially made of very cold (T ≈ 10 20 K) molecular gas and block off the light from background stars, the diffuse clouds, which consist of cold (T ≈ 100 K) atomic gas and are almost transparent to the background starlight, except at a number of specic wavelengths where they give rise to absorption lines, and the translucent clouds, which contain molecular and atomic gases and have intermediate visual extinctions. The rest of the interstellar matter, spread out between the clouds, exists in three different forms: warm (mostly neutral) atomic, warm ionized, and hot ionized, where warm refers to a temperature of $\sim 10^4$ K and hot a temperature of $\sim 10^6$ K.

Mixed in with all but the very hottest phases is interstellar dust. Dust grains are an important source of interstellar extinction, gas-phase element depletion, sites of interstellar chemistry, etc. This is solid-phase rather than gas-phase material. Dust grains range in size from a few microns down to macromolecular scales (clumps of 50 - 100 atoms or less).

The interstellar medium is in general not in any sort of thermal equilibrium. Nevertheless, it can be classified into a number of different thermal phases, characterized by differing thermodynamic properties. A summary of these properties can be found in Table. 1. Draine also cites "cold stellar outflows" as a component of the ISM, but this is not echoed in the other two works.

The temperature of the ISM is given by the Maxwellian distribution of velocities established by ISM gas and dust particles colliding elastically with one another (this is known as the "kinetic temperature", distinguishable from the effective temperature obtained by observing line strengths when the ISM is not in LTE) (Pogge 2011, Ch. 1). Since kinetic energy is distinct from internal energy, in order to [gain/lose] energy the ISM must [absorb/emit] a photon, [followed by/preceded by] a collisional [de-excitation/excitation] to bring a particle [out of/into] an excited state. In this manner, the kinetic energy is coupled with the internal energy and therefore the radiation field. In local thermodynamic equilibrium (LTE), the radiation field is completely coupled to the matter, meaning that a single temperature can describe them both. In non-LTE (i.e. low-density) conditions, the distribution of excited states is governed by collision and radiative rates, and may either be described by departure coefficients $n_{\rm state}/n(LTE)_{\rm state}$ or by an "excitation temperature". An extreme example of a non-thermal distribution of excited states is an inverted population, which may exist due to optical or collisional pumping, and produces a maser or laser from stimulated emission.

Which cooling mechanisms dominate, therefore, is dependent on which process:

- Collide frequently
- Results easily in excited states
- Can be reached at the medium's particular kinetic temperature
- Results in photon emision before collisional de-excitation
- Emits a photon that can most easily escape the system

3.15.1. *Molecular Gas*

Diffuse molecular gas is similar to the cool H I clouds, but with sufficiently large densities and column densities so that H_2 self-shielding allows molecules to be abundant in the cloud interior. Dense molecular gas is composed of gravitationally bound clouds that have achieved $n_H \gtrsim 10^3$ cm^3 . These clouds are often "dark" - with visual extinction AV $\gtrsim 3$ mag through their central regions. In these dark clouds, the dust grains are often coated with "mantles" composed of H_2O and other molecular ices. It is within these regions that star formation takes place. It should be noted that the gas pressures in these "dense" clouds would qualify as ultrahigh vacuum in a terrestrial laboratory.

Molecular gas is primarily composed of H_2, but H_2 is not directly observable in the radio [20]; this feature of H_2 emission is also what prevents it from being a good coolant. Instead CO rovibrational line emission (particularly 2.6 mm) acts as the primary tracer.

As noted in Question 3.5, the primary means of cooling for an H_2 region is C^+ fine structure line emission at low densities and reasonably high temperatures ($\Delta E/k$ for the transition is 92 K, or 168 μm). At other densities and temperatures, CO rovibrational line emission contributes the majority of cooling.

Photodissociation Regions (PDRs) are the warm, partially ionized surfaces of molecular clouds. The UV radiation field impinging upon the cloud may be either a discrete UV source, like O or B stars, or the diffuse interstellar radiation field (ISRF). The principal ionic state in a PDR is C^+, which is why they are sometimes referred to as "C^+" or "C II" regions. They get their name from the fact that this is the region in which H_2 is dissociated by UV photons into H I. Since dust and molecular gas are closely related (much of the dust in the Galaxy is in molecular clouds, since significant shielding is required for both to exist (Glover & Clark 2012)), PDRs are also significant sources of FIR emission from hot dust. PDRs are heated by photoelectric heating from dust grains and UV emissions from the interstellar radiation background. They are cooled by emission of collisionally excited FIR fine structure lines, including C^+ or [C II] (158 μm) and [O I] (63 μm), CO rovibrational transitions, collisionally excited NIR H_2 rovibrational lines and collision of molecules with dust grains. If the surface of the PDR is hot enough to be partially ionized, NIR/optical lines from partially ionized metals will also contribute.

3.15.2. *Cold Neutral Medium*

Cold neutral hydrogen (H I) gas is distributed in sheets and filaments occupying by volume ∼1 - 4% of the ISM with temperatures of ∼80 - 100 K and densities of ∼30 - 50 cm^{-3}. The main tracers are UV and optical absorption lines seen towards bright stars or quasars, and 21 cm line emission (and absorption if a bright object is in the background). Indeed, mapping of the 21-line emission works to pick out both the warm and cold neutral medium: narrow peaks seen in both emission and absorption are due to the CNM, while broader features seen in emission only are due to the WNM. The CNM is approximately in pressure equilibrium with its surroundings.

Cooling is accomplished primarily through the [C II] (158 μm) fine structure transition, and from the [O I] (63 μm) fine structure transition at higher temperatures (see Fig. 66). The critical densities for [C II] and [O I] are $\sim 4 \times 10^3$ cm^3 and $\sim 10^5$ cm^3, respectively, implying that collisional deexcitation of these levels is unimportant in the diffuse ISM of the Milky Way.

3.15.3. *Warm Neutral Medium*

Warm neutral atomic hydrogen occupies $\sim 30 - 40$ % of the volume of the ISM, and is located mainly in photodissociation regions on the boundaries of HII regions and molecular clouds. It has characteristic temperatures of ∼5000K and densities of ∼0.5 cm^{-3} and is traced by H I 21-cm line emission. The WNM is approximately in pressure equilibrium with its surroundings.

WNM cooling is generally the same as CNM cooling. In the 1000 K range of temperatures [O I] emission generally is a better coolant than [C II]. As 10^4 K is approached, H atom excited states become populated, and Ly α becomes a significant coolant.

3.15.4. *Warm Ionized Medium*

Diffuse gas with temperatures of 6000 - 12000 K, and densities ~ 0.1 cm^{-3}, occupying about $\sim$ 10 - 25 % of the volume of the ISM. While primarily photoionized (it requires about 1/6th of all of the ionizing photons emitted by the Galaxy's O and B stars), there is some evidence of shock or collisional ionization high above the plane of the Galaxy.

[20] It can be detected as absorption lines in the UV/optical, but extinction in dense molecular clouds prevents this from being a useful probe (Ferrière 2001).

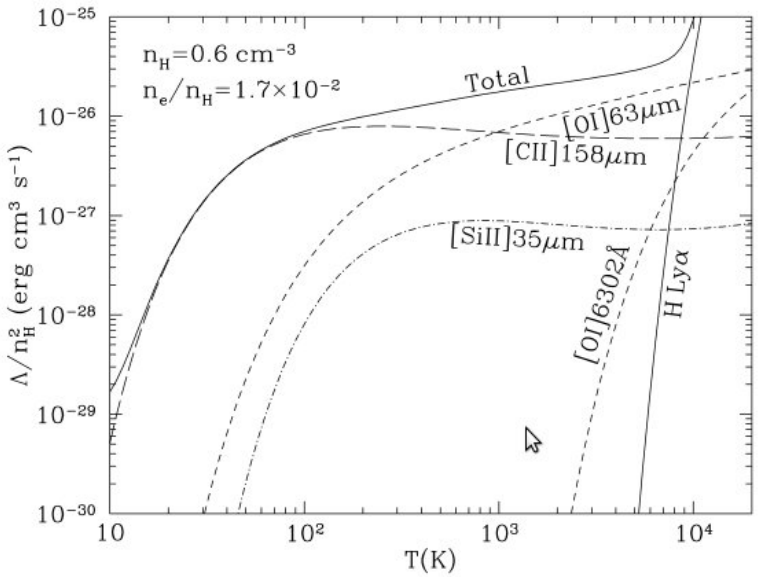

FIG. 66.— An H I cooling curve for $n_H = 0.6$ cm^{-3} and $n_e/n_H = 1.7 \times 10^{-2}$. From Draine (2011), his Fig. 30.1.

It is traced by low-surface brightness Hα (6563 Å) emission. Nearly 90 % of the H^+ in the Galaxy resides in the WIM, with the remaining 10 % in the bright high-density H II regions that occupy a tiny fraction of the ISM. The WIM is associated with regions of recent massive-star formation and planetary nebulae (the ejected outer envelopes of AGB stars photoionized by the cooling WDs at their centres).

The UV, visible and IR spectra of H II regions are very rich in emission lines, primarily collisionally excited lines of metal ions and recombination lines of H and He. H II regions are also observed at radio wavelengths, emitting radio free-free emission from thermalized electrons and radio recombination lines from highly excited states of H, He, and some metals.

The WIM cools through recombination radiation, where an electron recombines with an ion and therefore no longer contributes its kinetic energy to the medium, and free-free emission, where free electrons accelerate off of each other to produce bremmstrahlung. Most importantly, however, metal-polluted H II regions are generally at temperatures at which metals such as O may be collisionally excited (this is difficult to achieve with He or He^+, as their first excited states are highly energetic). Table 27.2 of Draine lists principal collisionally excited lines. In metal-depleted H II regions (0.1 × the metallicity of the Orion Nebula) the cooling rate from collisionally excited (forbidden) line emission is several times larger than the radiative recombination and free-free rates (which are roughly comparable). For an Orion-metallicity cloud it is about an order of magnitude.

The balance between heating and cooling provides, given a certain density, the temperature of an H II region. In H II regions photoionization provides the primary means of heating the gas, and balancing that against the cooling mechanisms described above, and assuming $n_H = 4000$ cm^{-3} nets us ∼8000 K.

3.15.5. *Hot Ionized Medium (Coronal Gas)*

Hot, low-density gas heated by supernovae, with temperatures $> 10^{5.5}$ K and very low densities of < 0.004 cm^{-3} occupy ∼50 % of the ISM. The vertical scale height of this gas is ∼3 kpc or larger, so it is sometimes referred to in the literature as the hot corona of the galaxy. This hot gas is often buoyant and appears as bubbles and fountains high above the disk. Its primary tracers are absorption lines seen towards hot stars in the far-UV (e.g., O IV, N V, and C IV) in gas with T ≈ 10^5 K, and diffuse soft X-ray emission from gas hotter than 10^6 K.

Coronal gas is primarily heated and collisionally ionized by the violent interaction between the ISM and rapidly expanding gas from supernovae. At high temperatures, Compton heating becomes important as well. Coronal gas cooling is dominated by neutral H and He collisional excitation, with collisional excitation of He^+ becoming important at 4×10^5 K. UV collisionally excited lines (mainly forbidden lines at lower and allowed lines at higher temperatures) also play a role. At extremely high temperatures, free-free continuum radiation, appearing at EUV and soft X-ray energies, dominates.

3.15.6. *Why Do the Phases of the ISM Have Characteristic Values At All?*

Simply put, phases of the ISM exist when the various phenomena that affect ISM temperature and density balance out and form a stable equilibrium. H II regions, for example, only exist in the vicinity of a strong UV source, such as a young cluster or association. Molecular hydrogen, on the other hand, can only form in regions of high self-shielding, where the amount of dissociation radiation and heating are minimized.

As an example, consider the Two-Phase Model for describing why there is a warm and cold neutral medium (pg. 341 - 343 of Draine and Ch. 1 of Pogge). The generalized loss function L is defined as the net cooling subtracted by

the net heating per unit volume. We may calculate equilibria between heating and loss ($L = 0$) given some average density of the medium and the need to exert a certain pressure. At this equilibrium, if $\partial L/\partial S < 0$, the equilibrium is unstable (heating is an increase in entropy, and so if $\partial L/\partial S > 0$ entropy is increased for some reason, the net cooling effect will, over time, bring the system back to equilibrium).

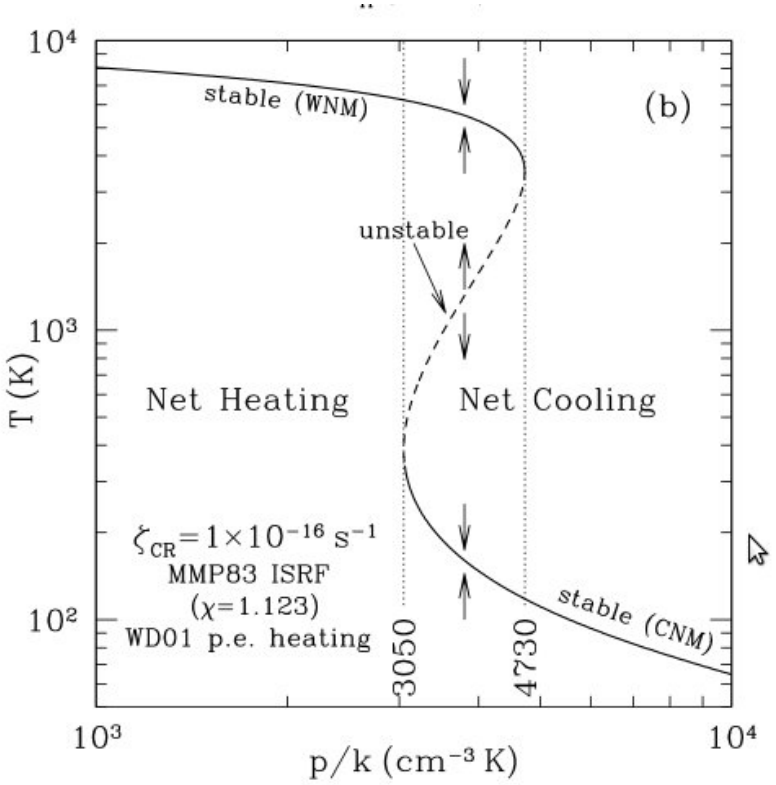

FIG. 67.— Steady state temperature as a function of thermal pressure. Since p is proportional to the inverse of Γ/nT, left of the line, $\Lambda < \Gamma/nT$ the system will heat up until it reaches equilibrium, and right of the line $\Lambda > \Gamma/nT$ resulting in net cooling - the direction of heating/cooling is shown by the vertical arrows. Either a temperature or a pressure perturbation in the dashed-line region results in movement away from the equilibrium. From Draine (2011), his Fig. 30.2.

Suppose loss is due to collisionally excited line emission. Then $L = n^2\Lambda - n\Gamma$ (Λ is cooling from collisions, Γ is heating). Fixing nT for an ideal gas means fixing the pressure, and for both dust photoions and cosmic rays Γ is independent of the ISM properties, for a given p, i.e. a given nT, we can find the temperature at which $\Lambda = \Gamma/nT$ (Pogge, Ch. 1, pg. 21). Therefore the heating/cooling balance can also be plotted in pressure-temperature space (another way of putting it is that given a pressure p, we find equilibria points). This is plotted in Fig. 67. As noted in the figure caption, the dashed-line region represents an unstable equilibrium, resulting in two unjoined regions of equilibrium - hence, a two-phase ISM.

The fact that more than two phases exist in reality is a reflection of the fact that additional physics must be included, such as H_2 grain catalysis and H I ionization. The standard picture of the ISM is a three-phase model (where cold includes molecular clouds and warm includes H II regions) - the hot phase is the coronal gas, shock heated by SNe bubbles, though this model is likely an oversimplification. See Ch. 1, pg. 23 - 24 of Pogge for details.

3.16. *Question 17*

QUESTION: Characterize the stellar populations in the following regions: i) the Galactic bulge ii) the Galactic disk, outside of star clusters iii) open star clusters iv) globular clusters v) a typical elliptical galaxy.

Unless otherwise noted, the information in this section comes from Carroll & Ostlie (2006), Ch. 24.2 and Schneider (2006), Ch. 2.3. Information on open clusters comes mainly from Binney & Merrifield (1998), Ch. 6.2.

Table 24.1 neatly summarizes much of the information required for this question. It is reproduced here as Fig. 68.

3.16.1. *The Galactic Disk*

The disk is roughly 40 - 50 kpc in diameter, with an ellipticity of about 0.9. It is composed of two major components. The younger think disk has a scale height of ∼350 pc and contains star formation regions. The thin disk dust and

Disks			
	Neutral Gas	Thin Disk	Thick Disk
M (10^{10} $M_\odot$)	0.5[a]	6	0.2 to 0.4
L_B (10^{10} $L_\odot$)[b]	—	1.8	0.02
M/L_B ($M_\odot/L_\odot$)	—	3	—
Radius (kpc)	25	25	25
Form	e^{-z/h_z}	e^{-z/h_z}	e^{-z/h_z}
Scale height (kpc)	< 0.1	0.35	1
σ_w (km s^{-1})	5	16	35
[Fe/H]	$> +0.1$	-0.5 to $+0.3$	-2.2 to -0.5
Age (Gyr)	$\lesssim 10$	8[c]	10[d]

Spheroids			
	Central Bulge[e]	Stellar Halo	Dark-Matter Halo
M (10^{10} $M_\odot$)	1	0.3	190^{+360}_{-170}[f]
L_B (10^{10} $L_\odot$)[b]	0.3	0.1	0
M/L_B ($M_\odot/L_\odot$)	3	~ 1	—
Radius (kpc)	4	> 100	> 230
Form	boxy with bar	$r^{-3.5}$	$(r/a)^{-1}(1+r/a)^{-2}$
Scale height (kpc)	0.1 to 0.5[g]	3	170
σ_w (km s^{-1})	55 to 130[h]	95	—
[Fe/H]	-2 to 0.5	< -5.4 to -0.5	—
Age (Gyr)	< 0.2 to 10	11 to 13	~ 13.5

[a] $M_{\rm dust}/M_{\rm gas} \simeq 0.007$.

[b] The total luminosity of the Galaxy is $L_{B,\rm tot} = 2.3 \pm 0.6 \times 10^{10}$ $L_\odot$, $L_{\rm bol,tot} = 3.6 \times 10^{10}$ $L_\odot$ ($\sim 30\%$ in IR).

[c] Some open clusters associated with the thin disk may exceed 10 Gyr.

[d] Major star formation in the thick disk may have occurred 7–8 Gyr ago.

[e] The mass of the black hole in Sgr A* is $M_{\rm bh} = 3.7 \pm 0.2 \times 10^6$ $M_\odot$.

[f] $M = 5.4^{+0.2}_{-3.6} \times 10^{11}$ $M_\odot$ within 50 kpc of the center.

[g] Bulge scale heights depend on age of stars: 100 pc for young stars, 500 pc for old stars.

[h] Dispersions increase from 55 km s^{-1} at 5 pc to 130 km s^{-1} at 200 pc.

FIG. 68.— Properties of different components of the Milky Way. From Carroll & Ostlie (2006), their Table 24.1.

H_2 gas lane (sometimes called the "young thin disk") has a scale height of ∼90 pc, while H I has a scale height range from 160 to 900 pc depending on distance from the centre of the galaxy. The less well-defined thick disk likely has a scale height of ∼1000 pc and is an order of magnitude less dense than the thick disk. Combined, the two disks have roughly the following profile:

$$n(z,R) = n_0(e^{-z/350} + 0.085e^{-z/1000})e^{-R/h_R} \tag{119}$$

where all values are in pc and h_R, the disk scale height, is $> 2.25 \times 10^3$ pc.

As a general rule, thick disk stars are more metal poor than thin disk stars. Popular means of measuring metals include Z, the mass fraction of the star composed of elements heavier than He, and the metallicity, [Fe/H], defined as

$$[Fe/H] = \log_{10}\left[\frac{(N_{Fe}/N_H)_{\rm star}}{(N_{Fe}/N_H)_\odot}\right]. \tag{120}$$

Metallcity values in our galaxy range from 0.6 to -5.4, and more negative numbers generally represent older stars, though as noted in Sec. 3.7 significant challenges exist in relating metallicity to stellar age. The metallicity range in the thin disk is from -0.5 to 0.3, and in the thick disk from about -0.6 to -0.4, though the dispersion is higher. This suggests the thin disk began forming about 8 Gyr ago, while the thick disk formed 10 - 11 Gyr ago.

Spiral arms are associated with the thin disk (since they form out of star-forming regions in the young thin disk). As a result, young OB associations are associated with the thin disk, and die before they become part of the thick disk.

3.16.2. *The Galactic Bulge*

The galactic bulge occupies the innermost ∼0.7 kpc of the galaxy, and has a scale height of 100 - 500 pc, depending on the age of the stellar population being measured. Its light follows roughly a Sersic profile of index 4, i.e. a de

Vaucouleurs profile:

$$I(R) = I_e \exp(-7.669((R/R_e)^{1/4} - 1)) \tag{121}$$

where I_e is the intensity at R_e, the radius enclosing half of the light from the bulge. Integrating $I(R)$ gives us the relation $L = 7.215\pi I_e R_e^2$.

The chemical enrichment of the bulge is strange: the range of metallicities ranges from -2 to 0.5, and suggest three distinct stellar populations: one with age just 200 Myr, another between 200 Myr and 7 Gyr, and a third older than 7 Gyr. The age-metallicity relation works in reverse, however - the most metal-rich stars are also the oldest, while the youngest stars have a range of metallicities. This may be due to rapid enrichment of the bulge through CC SNe early in its life combined with metal-poor gas inflow later in the evolution of the bulge. It is also possible that some metal rich stars are actually members of the inner galactic disk.

A bar extends outward from the bulge, and is responsible for perturbing the 3-kpc spiral arm outward.

3.16.3. *Open Star Clusters*

Open clusters are coeval[21] groups of stars that form out of a single giant molecular cloud. They are distributed close to the galactic thin disk. They consist of anywhere between a few to several thousand stars, with a spatial extent of several pc and a density ranging from 0.1 to 10^3 pc^{-3} (very low-density open clusters are known as "associations"). Unlike globular clusters, they do not have "simple", well-ordered structures, though King density profiles (see below) can provide reasonable fits. The relaxation time for open clusters is small enough that most older clusters are expected to be thermalized. Some evidence of this exists in the form of observations of mass segregation (if all objects have the same energy, more massive objects will have lower characteristic velocities and sink to the centre of the potential), though direct measurements of stellar kinematics contradict these observations (see Binney & Merrifield, pg. 388 - 389).

Open clusters often contain significant amounts of ISM gas and dust, as well as a population of hot, blue stars, both indicators of the youth of the average open cluster. The distribution of known open cluster ages peaks at ∼100 Myr and drops off sharply, though a few open clusters > 3 Gyr or older have been observed. Metallicity is generally around solar, but older clusters may be as metal poor as [Fe/H] ≈ -0.75 and younger cluster as rich as [Fe/H] ≈ 0.25.

Open cluster lifetimes are limited by close encounters with GMCs - only rich, densely populated clusters (which have significant gravitational potentials) and clusters born on the upper and lower fringes of the thin disk (where there are few GMCs) may avoid tidal destruction.

Since kinematic parallax can be used to determine their distance, and their stellar populations are coeval and fairly young, they (like their globular cluster cousins) are useful laboratories for stellar evolution.

3.16.4. *Globular Clusters*

The globular cluster system is connected to the Galactic halo. Stellar members of the halo are distinguished from members of the disk by a high velocity component perpendicular to the plane of the disk. As a whole, this galactic halo can be represented by either a density profile of

$$n(r) = n_0(r/a)^{-3.5} \tag{122}$$

where $n_0 = 4 \times 10^{-5}$ pc^{-3} (roughly 0.2% of the thin disk's midplane value) and $a \sim 1000$ pc. Halo field stars have been found as far as 50 kpc away, while GCs arguably extend out to 120 kpc. It is not clear whether or not the field stars and GCs form a unified population of halo members, as there is some evidence to suggest that the field stars occupy a flattened volume with $c/a \sim 0.6$, rather than the spherical volume occupied by GCs.

Globular clusters (GCs) are compact (∼ 10 pc), highly spherical ($b/a \sim 0.9$) collections of 10^5 - 10^6 metal-poor stars. GCs have a King density profile (an isothermal sphere with a fast drop-off to keep total mass finite; Binney & Tremaine, pg. 307 - 308), and the core stellar densities of GCs can be massive, up to 1000 pc^{-3}. Because they are old and coeval, most of the more massive stars have already evolved off the main sequence; from isochronal fitting of this "main sequence turnoff", GCs have an age range from 11 - 13 Gyr. The $\gtrsim 150$ GCs of our Milky Way come in two different populations: the metal poor population ([Fe/H] $<$ - 0.8) is spherically distributed, while the more metal rich population ([Fe/H] $>$ - 0.8), with the notable exception of 47 Tucana, is primarily distributed near the galactic thick disk, and may actually be associated with it (both populations having a scale height of 1 kpc and similar rotations about the galactic centre). GCs can be found from 500 pc to 120 kpc from the galactic centre, though the majority of them are within 40 kpc of the galactic centre. It is debatable whether or not GCs past 70 kpc were actually formed alongside the MW, or were captured from, or formed out of, satellite galaxies.

3.16.5. *(Ordinary) Elliptical Galaxies*

Ordinary elliptical galaxies, at least to first order, resemble the bulges of spiral galaxies. Their brightness profile is well-described by the de Vaucouleurs profile (Eqn. 121). Their total mass ranges from 10^8 - 10^{13} $M_\odot$ and their

[21] Open clusters are young enough that they cannot be considered purely coeval. A commonly observed mismatch between the age of the cluster as determined by the main sequence turnoff and as determined by low mass protostar cooling suggests low-mass stars form earlier than high-mass stars (Binney & Merrifield 1998, pg. 384).

physical extent ranges from 1 - 100 kpc (Carroll & Ostlie 2006, pg. 985). The stellar population in an average elliptical is red, and little (but non-zero) mass in gas and dust. Ellipticals hold their shape by velocity dispersion anisotropy rather than bulk rotational velocity - more orbits are oriented along the elliptical's shortest axis than along ather axes. While thermalization of the stellar population will lead to a roundening of the system (ex. with globular clusters), the relaxation times of ellipticals are an order of magnitude greater than their age. For lower mass ellipticals, rotational support due to an overall galactic spin also plays a significant role (these ellipticals also have more "disky" isophotes).

The centres of elliptical galaxies have redder stellar populations than their outskirts, suggesting either that metallicity is highest at the centre, or that the central stellar population is older[22] (Schneider 2006; Binney & Merrifield 1998, pgs. 193, 320). The same issue confounds our ability to determine an absolute metallicity for ellipticals. Efforts to disentangle the two effects have lead to the determination that most ellipticals formed their stars at z $\gtrsim$ 2 - 3 in a period of star formation inversely proportional to the mass of the galaxy (the most massive galaxies formed their stars in $\lesssim$ 1 Gyr) (Renzini 2006).

3.16.6. *What do these properties tell us about the formation and evolution of these objects?*

This information comes from Ch. 26.2 of Carroll & Ostlie (2006).

The earliest attempt at explaining the properties of these different regions of the MW, and other large late-type galaxies, was the Eggen/Lynden-Bell/Sandage homologous collapse model, where a massive proto-galactic cloud collapsed to form a single galaxy. In this model, the oldest (pop II) halo stars were formed when the cloud was nearly spherical and collapsing radially, which explains their current trajectories. The younger disk stars were formed after the kinetic energy of the infalling cloud was dissipated through thermalization of the cloud constituents, followed by radiation. This process is spherically symmetric, and therefore no angular momentum is lost; this angular momentum instead generates a disk. This disk is metal enhanced from the first generation of stars that exploded back during the period of radial collapse, and so new stars formed after disk formation are metal rich.

This model fails to explain a number of features of our galaxy: first, the halo has no sense of rotation at all (while the homologous collapse model posts the cloud as having some initial angular momentum). The spread of ages in the halo population is about 2 Gyr, an order of magnitude longer than the free-fall time of a giant proto-galactic cloud. Homologous collapse also would not explain a multi-component disk with varying ages.

The modern conception of the formation of an MW-like galaxy combines the homologous collapse model with a bottom-up process of gas-rich mergers. Following inflation, a series of (dark and baryonic) matter-energy overdensities formed that were bottom-heavy, i.e. 10^6 - 10^8 $M_\odot$ overdensities were much more common than 10^{10} $M_\odot$ overdensities. Following recombination, these overdensities began to attract one another while gas began to fall into the potentials of these overdensities. The initial collapse of these gas clouds (first on a free-fall timescale and then on a dissipative timescale) generated the first stars and clusters, and the first dwarf galaxies were born. The degree of star formation was different for different overdensities, so each dwarf galaxy had its own chemical history. As fragments combined to form larger galaxies, many of these dwarf galaxies were tidally destroyed, possibly leaving globular clusters that were at their cores. The stars and clusters form the modern-day halo of massive galaxies. At the centre of the budding, the potential well is deepest, and as a result significant chemical processing occurred, a possible explanation for the metal-rich stars in the MW bulge. The globular cluster mass distribution was also formed out of the merging process - massive GCs were towed into the bulge through dynamical friction, while lightweight GCs were destroyed through tidal disruption. Since there does not need to be any global sense of spin in the initial grouping of overdensities, the net angular momentum of the halo is negligible. Not all the overdensities merge with the galactic DM halo within a Hubble time, and the MW today features an extensive collection of dSph galaxies around it.

Since the relaxation time for stellar systems far exceeds the dissipation time for gas, the halo stars achieved fixed orbits (which they retain today), while the gas collapsed onto a thick disk with a temperature of $\sim 10^6$ K. A burst of star formation followed, which simultaneously polluted the ISM from [Fe/H] < -5.4 to [Fe/H] $\approx$ -0.5 and generated feedback to prevent further star formation. (An alternate theory postulates that the thick disk of our MW was actually formed cold, but a major collision with another galaxy 10 Gyr ago puffed up the disk.) The gas continued to dissipate and form new stars. Stars forming out of the thin disk provided some feedback, but as gas was depleted from the ISM, star formation slowed, and dissipation exceeded energy injection from feedback. The disk scale height eventually achieved the present-day scale height of the young thin disk.

Mergers continued after the formation of the galactic disk. New material streaming into the bulge from dwarf galaxies, or kicked into the bulge due to bar instabilities, generated new populations of metal-poor stars.

Ellipticals are believed to have formed in one of two ways: either through rapid star formation (the overdensities form stars so quickly that little gas remains by the time they merge to form the dark matter halo) or through major mergers of late-type galaxies. The latter hypothesis has support from N-body simulations and observed stellar kinematics, and neatly explains the morphology-density relation (elliptical galaxies are preferentially found near the centres of clusters). At the same time, however, it is possible that deep potential wells favour rapid star formation (and independently also favour clustering), and it is difficult to reconcile the existence of ellipticals with extensive globular cluster systems with the merger picture (unless those GCs are the result of the major merger, or subsequent minor merger).

3.17. *Question 18*

[22] This is known as the age-metallicity degeneracy, and is quantitatively given by Worthey (1994) as $\frac{d \ln t}{d \ln Z}_{\text{colours}} \approx -\frac{3}{2}$.

QUESTION: How can you determine the temperature of an H II region?

Most of the information in this section comes from (Draine 2011, Ch. 18).

There are three primary methods of determining the temperature of an H II region: nebular line diagonstics, the recombination continuum and dielectric recombination lines.

Line diagnostics use ions with two excited levels that are both "energetically accessible" at the temperatures of interest, but with an energy difference between them that is comparable to kT , so that the populations of these levels are sensitive to the gas temperature. The level populations are normally observed by their line emission.

Line diagnostics are also useful for determining the density of an H II region. To do this, ions with two or more "energetically accessible" energy levels that are at nearly the same energy are used, so that the relative rates for populating these levels by collisions are nearly independent of temperature. The ratio of the level populations will have one value in the low-density limit, where every collisional excitation is followed by spontaneous radiative decay, and another value in the high-density limit, where the levels are populated in proportion to their degeneracies. If the relative level populations in these two limits differ (as, in general, they will), then the relative level populations (determined from observed emission line ratios) can be used to determine the density in the emitting region.

During recombination, atoms with $E_K \geq 0$ become bound, meaning that the recombination continuum will periodically contain sharp drops, corresponding to the energy released when $E_K = 0$ electrons become bound (they are drops because $E_K < 0$ free electrons do not exist). The strength of the jump is related to the population of negligible KE electrons, and therefore is a function of the cloud density and temperature, or, more precisely, a function of the emission measure[23] multiplied by $T^{-3/2}$. To remove the emission measure dependence, we may use a recombination line as scaling, so long as the line has a different temperature dependence than the amplitude of the drop. For example, the drop most commonly used to measure temperature is the "Balmer jump" at $\lambda = 3645.1$ Å[24]. We can scale it with the $n = 11 \rightarrow 2$ ($\lambda = 3769.7$ Å) hydrogen recombination line, which is proportional to the rate of recombination onto levels $n \geq 11$, which turns out to be $\propto EM \times T^{-0.8}$. The ratio between the two measurements is $\propto T^{-0.7}$.

For some ions, it is possible to observe both collisionally excited lines and lines emitted following dielectronic recombination [25]. For example, electrons colliding with C IV can produce collisionally excited levels of C IV, but can also produce excited levels of C III by dielectronic recombination. Because the rate coefficients for collisional excitation and for dielectronic recombination will have different temperature dependences, the ratio of dielectronic lines to collisionally excited lines will be temperature-sensitive, and therefore useful as a temperature diagnostic. Examples of useful line ratios are C III 2297/C IV 1549 Å, O II 4705/O III] 1665 Å, and C II 4267/C III] 1909 Å.

3.17.1. *Details?*

Like hydrogen, a unified spin state for a multi-electron ion can be written in the form

$$^{2S+1}L_J \tag{123}$$

where S is the vector sum of the spin angular momenta and L the vector sum of the orbital angular momenta of all valence electrons, and $J = L + S$. L is often written in SPDF notation (i.e. S is $L = 0$, P is $L = 1$, and so on) (Wood 2011). Since these are conglomerate spins, there is no real reason that P must be higher energy than S, and so forth [26].

Atoms or ions with six electrons have $2p^2$ as their lowest configuration. Their ground state is ^{3}P, and the first two excited states are ^{1}D and ^{1}S. So long as the ^{1}S term is sufficiently populated (which occurs when $E/k \lesssim 7 \times 10^4$ K and the ion is sufficiently abundant), then there will be observable emission from both excited states. Since these levels are at very different energies, different temperatures will affect how the different states are populated. For example, consider the ion N II in Fig. 69, and assume that densities are below the critical density for both excited states. In this case, most N II ions are in their ground state, and as soon as a collision excites an electron, it will radiatively decay back to the ground state. The power going from state 4 to 3 is $E_{43} n_0 C_{04} \frac{A_{43}}{A_{43}+A_{41}}$, while the power going from state 3 to 2 is $E_{32} \left(n_0 C_{03} + n_0 C_{04} \frac{A_{43}}{A_{43}+A_{41}} \right) \frac{A_{32}}{A_{32}+A_{31}}$. The collisional excitation rate $C_{lu} \propto \frac{\Omega_{lu}}{g_l} e^{-E_{ul}/kT}$, where Ω_{lu} is the (roughly of order unity and weakly dependent on temperature) collision strength and g_l is the degeneracy of state

[23] The emission measure $EM \equiv \left[\frac{n_e n_p}{\kappa_{\mathrm{free-free},\nu}} \right]_T \tau_\nu$ is a measure of the column density of the cloud, scaled to how well it absorbs radiation (Draine 2011, pg. 96).

[24] The inverse of the Balmer jump can be used to determine the temperature of stellar atmospheres - because ionization of $n = 2$ electrons becomes possible with any photon with $\lambda < 3647$ Å, there is a sharp drop in the spectrum of stars at that wavelength. The magnitude of this drop is dependent on the number of $n = 2$ excited hydrogen atoms, which, since the stellar atmosphere is in LTE, is dependent on the Boltzmann equation, and therefore the temperature (Carroll & Ostlie 2006, pg. 247).

[25] Dielectronic recombination occurs when a free electron approaches an ion that contains one or more valence electrons. Instead of initiating a radiative recombination, the free electron transfers energy to one of the bound electrons, promoting the bound electron to an excited state and losing enough energy to become bound to the ion. The two excited electrons can then drop to the ground state, or exchange energy to re-ionize one of them. See (Draine 2011, pg. 151 - 153).

[26] For alkali atoms/ions, L and S individually do not contribute any changes in energy - the most prominent correction to the n-governed energy level is fine splitting, caused by a relativistic correction and spin-orbit coupling (J)

l. As a result, the ratio of line emission power is

$$\frac{j_{4\to3}}{j_{3\to2}} \propto \frac{e^{-E_{43}/kT}}{1+Xe^{-E_{43}/kT}} \tag{124}$$

where $X = \frac{A_{43}\Omega_{04}}{(A_{43}+A_{41})\Omega_{03}}$, and there is an obvious dependence on temperature. See pg. 205 of Draine for details. Fig. 70 shows the line ratio as a function of temperature and density.

Candidate $2p^2$ ions are C I, N II, O III, F IV, Ne V, and so on. C I is easily photoionized, and will have very low abundance in an H II region. The ionization potentials of F IV, Ne V, and so on exceed 54.4 eV, and we do not expect such high ionization stages to be abundant in H II regions excited by main-sequence stars with effective temperatures $kT_{\rm eff}$ ¡ 5 eV. This leaves N II and O III as the only $2p^2$ ions that will be available in normal H II regions. The same situation exists for $2p^4$ ions and above (O I, F II, and Ne III being appropriate temperature diagonstics for $2p^4$, P II and S III $3p^2$, and Cl II, Ar III, and K IV $3p^4$). Atoms/ions with seven (O II, F III, and Ne IV are usually used) and fifteen (S II, Cl III, and Ar IV are usually used) electrons have different energy levels, but the same principle applies.

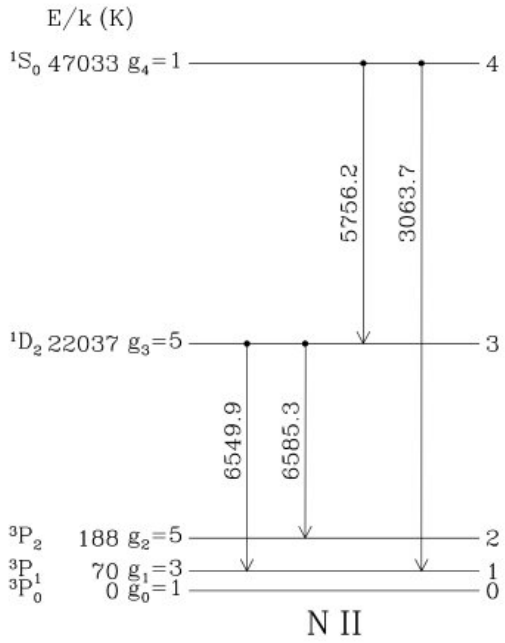

FIG. 69.— Energy level diagram for temperature diagnostic $2p^2$ ion N II. From Draine (2011), his Fig. 18.1.

Note that it is assumed collisional excitation is the only way to achieve an excited state. This is not true if a substantial contributor of excited states comes from recombination from the next ionized state.

3.17.2. *How is density determined?*

Density can be determined through collisionally excited optical/UV lines or near/far-IR fine structure lines.

Ions with 7 or 15 electrons have $2s^22p^3$ and $3s^23p^3$ configurations, with energy-level structures that make them suitable for use as density diagnostics: the ground state is a singlet $^4S_{3/2}$, and the first excited term is a doublet $^2D^o_{3/2,5/2}$. Let us consider OII (Fig. 71) and again assume low densities, so that a radiative decay immediately follows from a collisional excitation. The relative intensity of the $2\to0$ and $1\to0$ lines (3727.1 and 3729.8 Å; see Fig. 71) are, at low density

$$\frac{j_{2\to0}}{j_{1\to0}} = \frac{\Omega_{20}}{\Omega_{10}}\frac{E_{20}}{E_{10}}e^{-E_{21}/kT} \approx \frac{\Omega_{20}}{\Omega_{10}} \tag{125}$$

where the approximation comes from the fact that the fine-splitting energy is tiny, i.e. $E_{21} \ll kT$. At high densities, the levels are thermalized (LTE), meaning

$$\frac{j_{2\to0}}{j_{1\to0}} = \frac{g_2}{g_1}\frac{E_{20}A_{20}}{E_{10}A_{10}}e^{-E_{21}/kT} \approx \frac{g_2}{g_1}\frac{A_{20}}{A_{10}}. \tag{126}$$

These two expressions involve fundamentally different microphysics, and as a result have different values. The sensitivity range for the diagnostic is the transition between the low density to LTE regime, which usually spans an order of magnitude below to the same order of magnitude as the critical density.

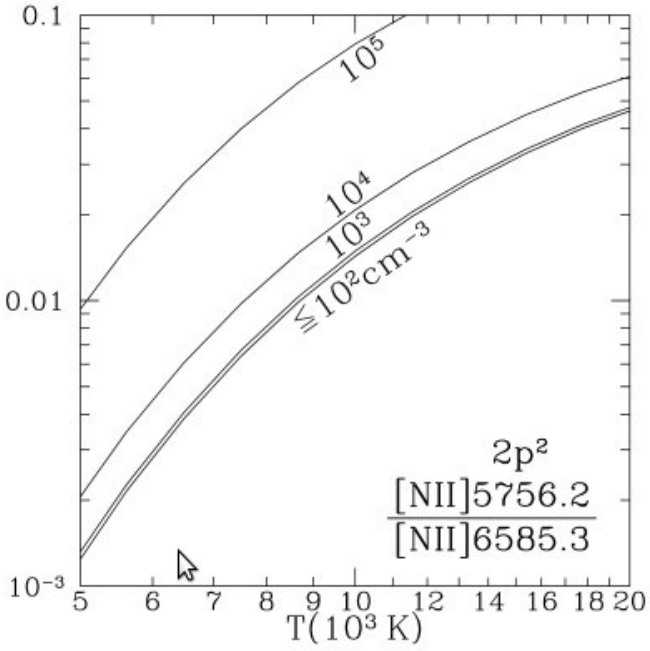

FIG. 70.— Line ratio as a function of temperature for temperature diagnostic $2p^2$ ion N II's 5756.2 and 6585.3 Ålines. This is the characteristic shape of a temperature line diagnostic. From Draine (2011), his Fig. 18.2.

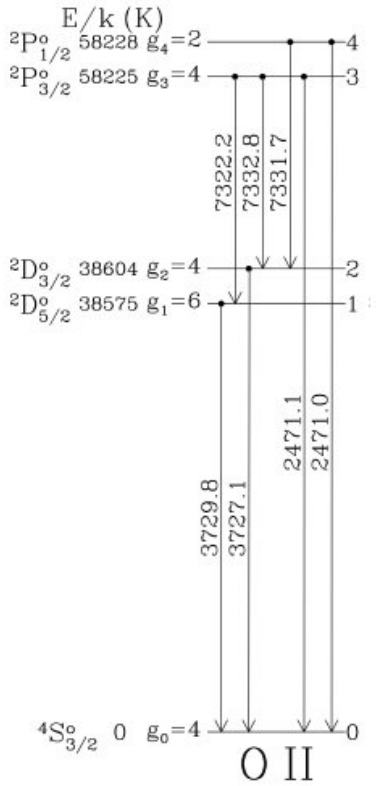

FIG. 71.— First five energy levels of the $2p^3$ ion O II. Of interest are the $2 \rightarrow 0$ and $1 \rightarrow 0$ (3727.1 and 3729.8 Å) lines, which can be used as a density diagnostic. From Draine (2011), his Fig. 18.3.

Ions with triplet ground states - in particular, the $^3P_{0,1,2}$ terms for np^2 and np^4 configurations - allow density determination from the ratios of mid-infrared and far-infrared fine-structure lines. Examples are the $2p^2$ ions N II,O III, and Ne V, the $2p^4$ ion Ne III and the $3p^2$ ion S III. If we treat the system as a three-level system (i.e. we do not care about decays from higher energy states), the intensity ratio $\frac{j_{2\rightarrow1}}{j_{1\rightarrow0}}$ has to first order no temperature dependence so

long as $E_{ul} \ll kT$, which is the case here (we need not worry about $2 \rightarrow 0$ transitions because $\Delta J = 2$ is a forbidden transition).

In both cases, collisional de-excitation of ions is dominated by electronion collisions, with rates that scale as $n_e\Omega_{ij}/\sqrt{T_e}$ ("e" stands for electron). If the collision strengths Ω_{ij} were independent of temperature, and the electron temperature is sufficiently high that E_{21}/kT_e, the line ratio would be dependent solely on $n_e/\sqrt{T_e}$, and a plot of the line ratio vs. $n_e/\sqrt{T_e}$ would bridge the low and high density extremes in the manner shown in Fig. 72. This is often not the case, and there is some dependence on T_e as well.

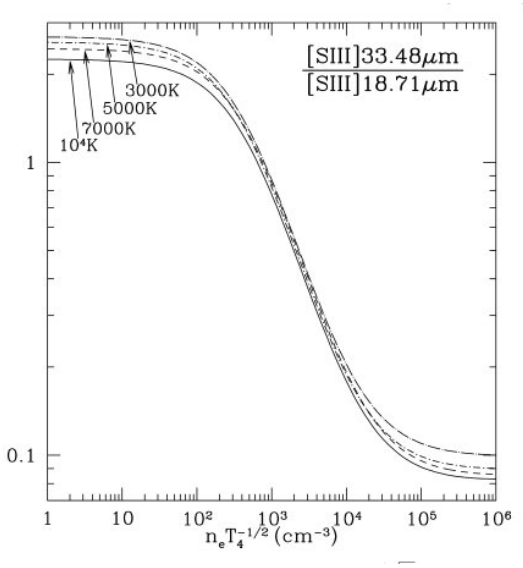

FIG. 72.— [S III] 33.48 μm / [S III] 18.71 μm as a function of $n_e/\sqrt{T_e}$ and T_e. This is the characteristic shape of a density line diagnostic. From Draine (2011), his Fig. 18.6.

Densities can also be determined using the "Balmer decrement" (the set of ratios that includes Hβ to Hα emission, Hγ to Hβ emission, and so on). At low densities these ratios are used to measure extinction, but at high densities the line ratios are affected by collisional effects, with systematic enhancement of the high-n levels relative to Hα and Hβ, and these line ratios can therefore be used to constrain the electron density when $n_e > 10^4$ cm^3.

3.17.3. *How is abundance determined?*

The abundance of He relative to H is determined from comparison of the strengths of radiative recombination lines of H and He in regions ionized by stars that are sufficiently hot ($T_{\rm eff} > 3.9 \times 10^4$ K) so that the He is ionized throughout the H II zone. The abundances relative to H of elements heavier than He can be inferred by comparing the strengths of emission lines excited by collisions between electrons and ions heavier than He to emission resulting from recombination of electrons with H^+, since the former is dependent on the density of these ions, while the latter is dependent on the density of H^+. There generally is also a significant dependence on temperature, which must be determined separately. Fine-structure lines can also be used to determine abundances of ions with fine-structure splitting of the ground state.

To determine the total abundance, one must sum over all important ion stages. In an H II region ionized by a B0 star, most of the oxygen will be O II, because there will be few photons above the O II $\rightarrow$ O III ionization threshold of 35.1 eV. However, in H II regions ionized by hotter stars, or in planetary nebulae, much of the oxygen may be present as O III.

Optical recombination lines resulting from radiative recombination of X^{+r+1} have similar temperature and electron density dependencies as H^+ recombination lines. The ratios of these lines, therefore, are a probe of elemental abundances that is not highly dependent on temperature.

3.17.4. *Can you use the same techniques to determine the properties of other ISM regions?*

It is difficult to determine properties of the ionized medium that are beyond the effective ranges of particular nebular diagnostics, as described above. Use of different species may help widen this range, but it is still theoretically possible to have an H II region that is too dense, for example, for diagonostics to probe.

H_2 itself does not have line emission at the temperatures and densities found in the ISM, but embedded CO does, and the ratio of peak intensities of CO lines can be used to determine the temperature of the cloud (Ferrière 2001, pg. 1037). The low-J levels are very optically thick SO HOW DO THEY CONTRIBUTE TO COOLING?, and if the lines are thermalized, measuring the line strength gives the excitation temperature, and hence the kinetic temperature of the H_2 (Pogge 2011, Ch. 5, pg 7). NH_3 can also be used (though it is optically thin). Molecular fine structure lines, such as those of CO, NH_3, CS, and HCO^+, generally have different critical densities, and therefore comparisons between line strengths can be used to determine density (Pogge 2011, Ch. 5, pg. 6 - 8). Ideally a set of multiple lines

from non-thermal ($n < n_{\rm crit}$) states of a single molecule are used, but cross-species comparisons may also be done. These diagnostics involve molecules that are not in LTE, and therefore detailed quantum-mechanical calculations are needed to interpret them. Density can be more simply estimated by observing extinction, translating it to a column density, and dividing it by the length-scale of the cloud. An extremely crude method involves obtaining a linewidth for the cloud, assuming that the velocity derived is the virial velocity needed to keep the cloud in equilibrium, and then dividing the virial mass by the volume.

A number of temperature probes exist for H I. In the radio regime, the classical method consists of observing the 21-cm line in absorption towards a bright radio continuum source; this, in conjunction with observations of the emission spectrum along a nearby line of sight allows one to measure the spin temperature of the H I (Roy et al. 2006). While the H I spin temperature, strictly speaking, characterizes the population distribution between the two hyperne levels of the hydrogen atom, it is often used as a proxy for the kinetic temperature of the gas, as it is tightly coupled to the kinetic temperature via collisions in high-density regions, and coupled through Ly α photons at low densities (Roy et al. 2006). UV observations of the Lyman and Werner bands of H_2, which determines the distribution between para and ortho H_2, a temperature-dependent quantity (Roy et al. 2006). H I column density can directly be probed by 21-cm line emission (Ferrière 2001, pg. 1039), but since a large fraction of H I is diffuse and spread throughout the ISM, only a rough estimate of the average density can be determined. Measurement of the fine-structure excitation of species such as C I and C II can be used to constrain the density and temperature in the H I (Draine 2011). For example, for a given gas composition (fractional ionization and H_2 fraction), temperature and density, we can theoretically calculate the ratio of C I ions that are in the $^3P_0$, $^3P_1$ and $^3P_2$ states; observationally determining these values allow us to pin down the temperature and density (or temperature and pressure) (Draine 2011, pg. 198 - 202).

Note that line width is not an accurate probe of ISM temperature, since turbulence as well as thermal motion can broaden lines (Ferrière 2001, pg. 1037).

4. STARS AND PLANETS (INCLUDES COMPACT OBJECTS)

4.1. *Question 1*

QUESTION: Sketch out an H-R diagram. Indicate where on the main sequence different spectral classes lie. Draw and describe the post main-sequence tracks of both low- and high-mass stars.

The Hertzsprung-Russell diagram is a plot of a measure of stellar luminosity versus a measure of stellar effective surface temperature. Observers have generally used absolute (or apparent, if all stars are known to be of a very similar radial distance from Earth) magnitude as a measure of luminosity, and colour as a measure of temperature (hence the oft-used term "colour-magnitude diagram", or CMD). Theorists generally just use luminosity and temperature. Figs. 73 and 74 show three examples of HR diagrams. Fig. 75 shows a more artistic depiction of an HR diagram.

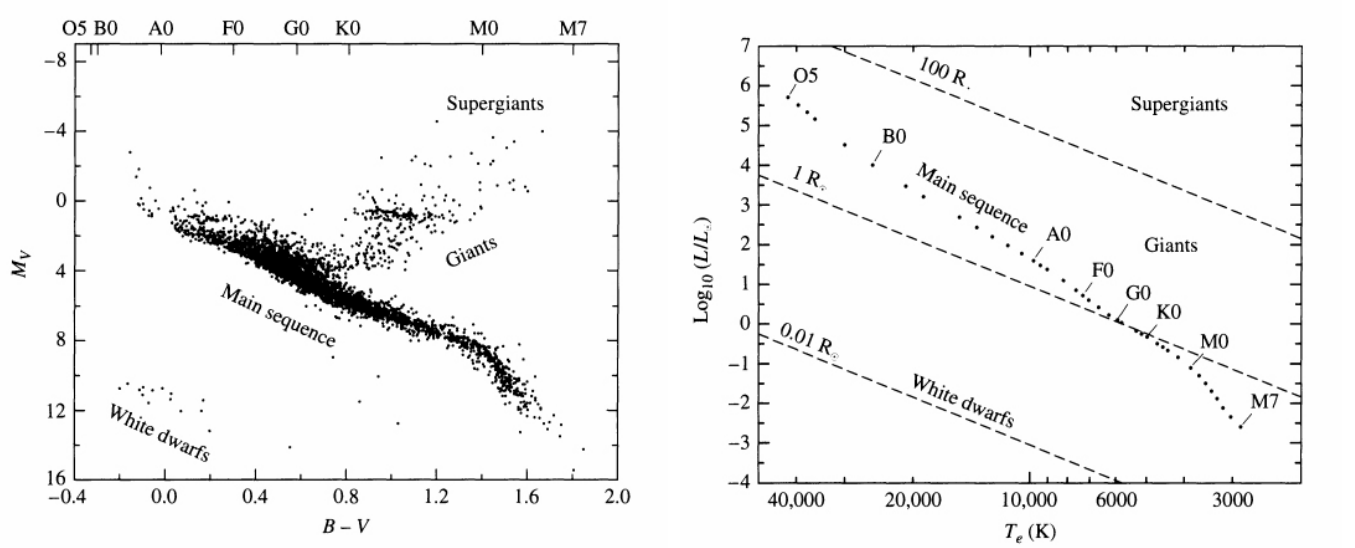

FIG. 73.— Two HR diagrams. Left - an observer's HR diagram of ∼3700 Hipparcos catalog stars, where populations with different luminosity classifications have been labelled. Right - a stellar evolution theorist's HR diagram, with different spectral types along the main sequence labelled. Contours of costant stellar radius, and contours From Carroll & Ostlie (2006), their Figs. 8.13 and 8.14.

Important features include:

- The **main sequence**: the population of stars burning hydrogen in their cores. Field HRDs will have most of their points along this line, because stars' nuclear timescales are at their longest. To first order, both temperature and luminosity are governed by the star's mass, and the slope of the main sequence can be recovered from R and L homology relations from zero-age main sequence stellar models (see Kippenhahn & Weigert (1994), Ch. 22.1). The main sequence stretches from ∼60 $M_\odot$ (stars more massive than this are exceedingly short-lived) to ∼0.08 $M_\odot$ (stars less massive than this never achieve steady core hydrogen fusion, and become brown dwarfs). The thickness of the main sequence is governed by higher-order processes, namely small changes stars make over the course of their lives on the main sequence and metallicity differences between stars of the same spectral type.

  The OBAFGKM(LT) Harvard spectral classification system is a spectal line-based method of differentiating stars. Table 2 describes physical properties and spectral features of different classes of stars. Each type/division of star can be divided into ten subdivisions, where 0 is most like the previous, hotter, division, and 9 most like the next, cooler, division (ex. A0 is the closest A subdivision to B, while A9 is the closest to F). Types L and T are used for substellar objects.

- The **horizonal branch** is a group of giant stars (the large clump of stars directly to the right of the main sequence and resting on the Luminosity Class III line, in Fig. 74) that have hydrogen shell and core helium burning.

- The **Hertzsprung gap** is the region directly to the right of the main sequence, and directly to the left of the horizontal branch, in Fig. 74. The lack of stars in this region is caused by rapid evolution of massive stars from the main sequence to the red giant branch.

- The **red giant branch** and and **horizontal branch** cannot be seen in these figures, but can be more easily spotted in HR diagrams of globular clusters.

- The **main sequence cutoff** is also a feature that can only be seen in stellar populations that are (roughly) coevol (born at the same time; ex. globular clusters) - it exists because, due to the population's age, some stars

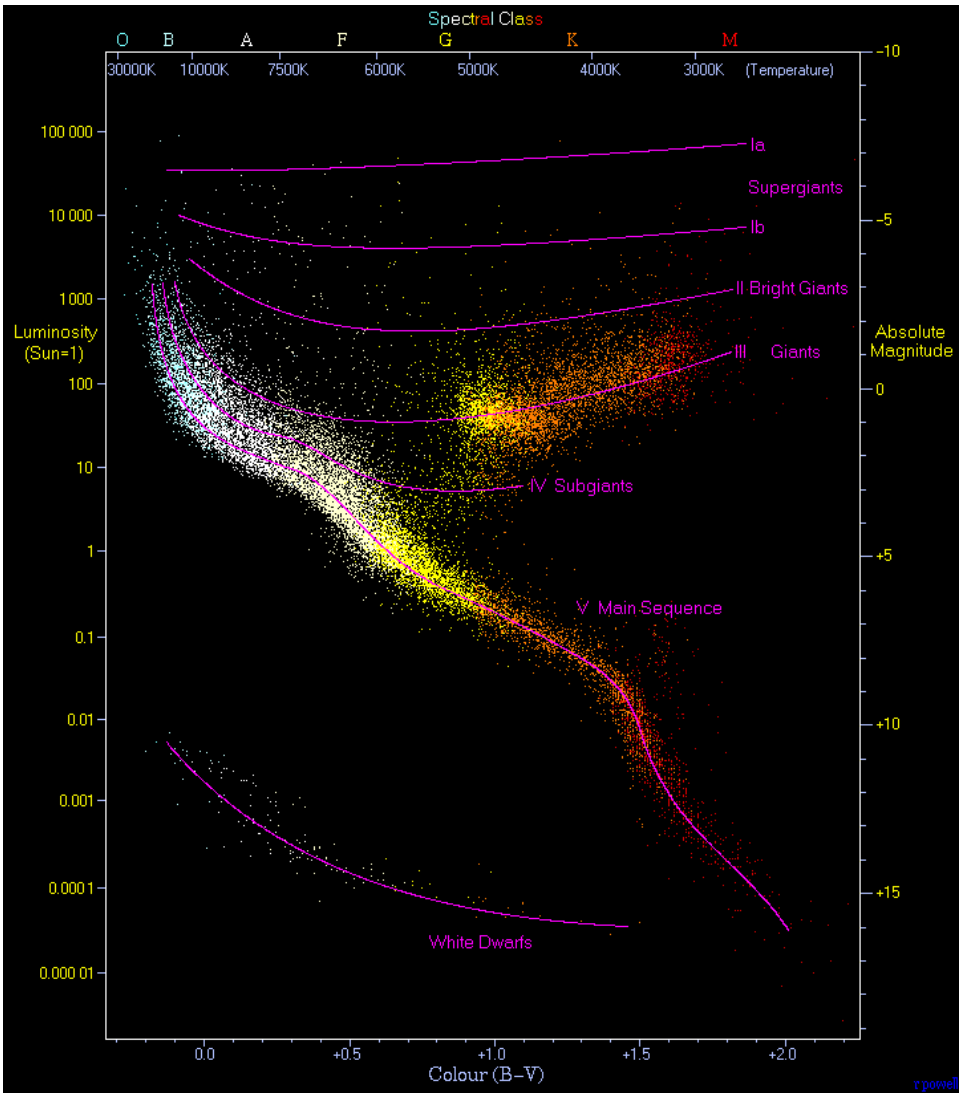

FIG. 74.— An observer's HR diagram composed of 23,000 stars from the Hipparcos and Gliese catalogues. Luminosity classes are labelled. From Wikipedia (2011c).

(those above the cut-off) have already evolved off the main sequence, while other stars (those below) have yet to. The cutoff is a useful way of determining the age of a coevol stellar population (known as spectroscopic parallax).

- The **white dwarf cooling curve** is the rough track formed by white dwarfs cooling. The white dwarf mass distribution is fairly sharply peaked (and He WDs cannot be made by single star evolution in a Hubble time); this implies that the vast majority of white dwarfs have the same initial mass, and therefore the same initial radius (from the mass-radius relationship of degenerate matter). The Stefan-Boltzmann equation then fixes the luminosity/temperature relation, resulting in a line.

In less massive stars, the core's hydrogen is depleted from inside out, while for more massive stars (due to convection) the core's hydrogen is depleted at once. As the hydrogen is exhausted, the star switches from core hydrogen burning to shell hydrogen burning, leaving an inert helium core left at the centre of the star. The core is roughly isothermal, and assuming roughly solar metallicity, the maximum stable mass of the core is 0.08 times the mass of the star - this is known as the Schonberg-Chandrasekhar limit, and its functional derivation is identical to the derivation for the Jeans mass in Sec. 3.13 (modulo adding "core" to the subscriptm, and a 3/5 in front of GM/R). Once this value is reached, the core either collapses on a thermal timescale until helium fusion begins, or another source of pressure becomes more important.

For low mass stars, evolution off the main sequence works like:

1. The core becomes degenerate before the Schonberg-Chandrasekhar limit is reached, a nuclear burning shell develops, and the star moves toward the Hayashi line on a nuclear timescale. During this time, luminosity remains constant while temperature decreases (Fig. 32.2 and 32.3 in Kippenhahn & Weigert). Something about

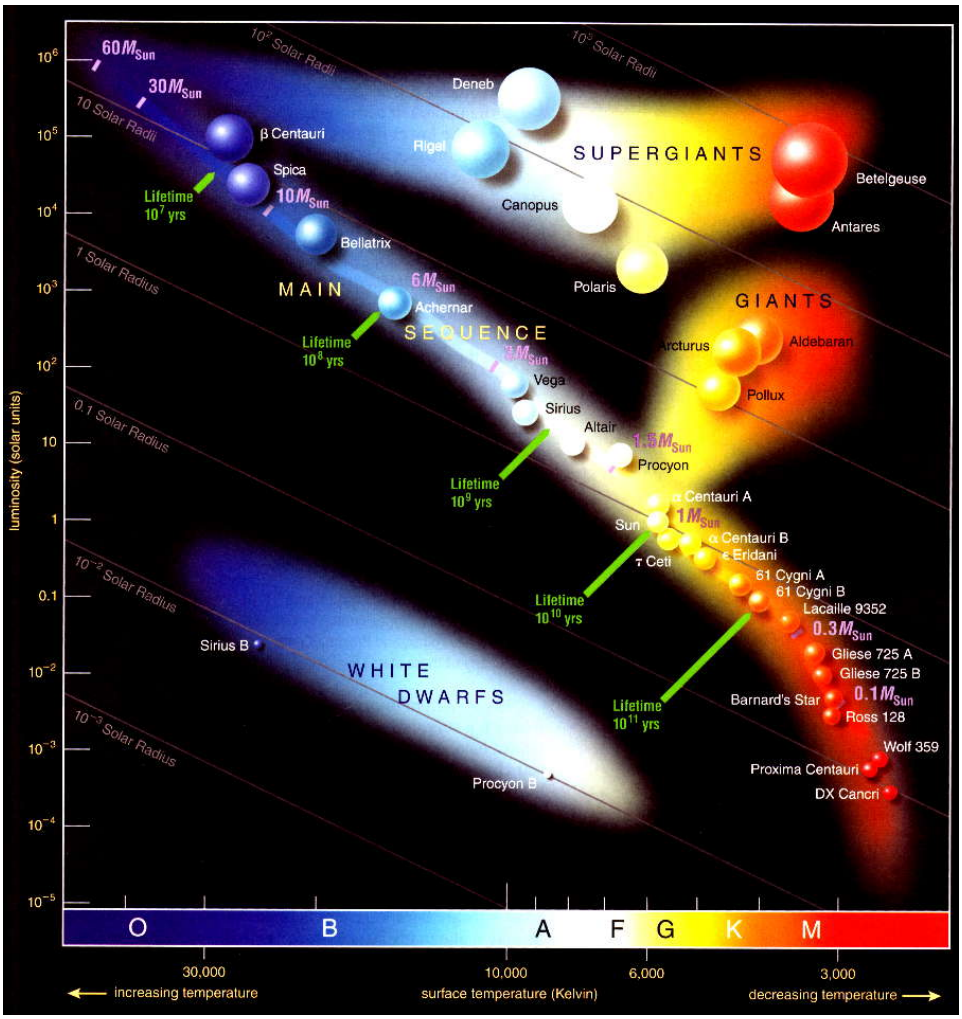

FIG. 75.— An artist's rendition of the HR diagram, with well-known stars depicted. From Nitschelm (2011).

TABLE 2
HARVARD SPECTRAL CLASSIFICATION OF MAIN SEQUENCE STARS.

Type	$T_{\rm eff}$ [K]	L $[L_\odot]$	Mass $[M_\odot]$	R $[R_\odot]$	Apparent Colour	Features
O	$\gtrsim 33{,}000$	$\gtrsim 10^5$	$\gtrsim 20$	$\gtrsim 10$	Blue	Strong He II absorption lines
B	$10{,}000 - 33{,}000$	$50 - 10^5$	$3 - 20$	$2.2 - 10$	Blue white	Strong He I absorption lines
A	$7{,}500 - 10{,}000$	$6 - 50$	$1.8 - 3$	$1.5 - 2.2$	White	Strong Balmer absorption lines
F	$6{,}000 - 7{,}500$	$1.5 - 6$	$1.05 - 1.8$	$1.1 - 1.5$	Yellow-white	Neutral metal absorption lines (Fe I and Cr I)
G	$5{,}200 - 6{,}000$	$0.6 - 1.5$	$0.8 - 1.05$	$0.94 - 1.1$	Yellow	Ca II and other neutral metal lines becoming stronger
K	$4{,}000 - 5{,}200$	$0.1 - 0.6$	$0.52 - 0.8$	$0.7 - 0.94$	Yellow-orange	Ca II H and K lines strongest
M	$\lesssim 4{,}000$	$\lesssim 0.1$	$\lesssim 0.52$	$\lesssim 0.7$	Orange-red	Dominated by molecular absorption bands, especially TiO and VO

the outer atmosphere, possibly an opacity runaway increase, results in the system becoming much more opaque, and the star balloons and cools.

2. Once the shell develops, the star runs up the RGB, straddling the Hayashi line. For a shell, $T_{\rm shell} \propto P/\rho \propto \frac{1}{\rho}\frac{GM_c\rho H}{R_c^2} \propto M_c/R_c$ (where shell scaleheight $H \propto R_c$), meaning that for a degenerate core $T_{\rm shell} \propto M_c^{4/3}$. Nuclear fusion is highly dependent on temperature, leading to an extreme dependence of L on M_c. The shell then thins

and heats. The star moves up the Red Giant Branch of the HRD. First dredge-up occurs soon after the base of the RGB is reached.

3. When the core reaches 0.45 $M_\odot$ (independent of the star mass because the shell temperature, which is core dependent, is what ignites He fusion), a He flash is triggered, which results in stable He burning at the core.

4. Helium flash evolution is not well understood, but following it, the star burns with stable core helium fusion, combined with hydrogen shell fusion. Its place on the HR diagram depends on its metallicity - solar metallicity stars end up as red clump stars, which stay near the Hayashi line, while lower metallicity stars end up on the horizontal branch.

5. When core He runs out, a degenerate C core will form, surrounded by an He burning shell. The star moves up the Asymptotic Giant Branch. Mass loss becomes important. During this time, thinning of the He shell leads to thermal pulses (since thin shell burning is inherently unstable).

6. Eventually the luminosity becomes high enough to drive a superwind. When the hydrogen-rich envelope reaches 1% of the star's mass, the entire star contracts, resulting in a fast journey toward the blue end of the HR diagram (at constant luminosity). The outgoing UV radiation is enough to ionize the surrounding material (ejected earlier from the star), creating a planetary nebula. What is left is a 0.55 - 0.6 $M_\odot$ CO white dwarf.

For intermediate mass stars (which start core He fusion non-degenerately, but do not start C fusion):

1. A non-degenerate isothermal core is generated at hydrogen exhaustion, and is built up by H shell burning. When the core exceeds the Schonberg-Chandrasekhar limit it contracts on a thermal timescale, and the star rapidly moves toward the Hayashi line. On the HRD, the star crosses the Hertzsprung Gap.

2. The star runs up the Hayashi line slightly.

3. The contraction stops when He is ignited non-explosively.

4. The He core expands slightly (H-burning stars do the same on the MS), reducing shell luminosity and partly reversing the process that made a red giant. The star moves along the "blue loop"

5. Once central He is exhausted, a non-degenerate CO core is produced, and reaches 0.08 the mass of the helium + CO part of the star, and the star once again moves rapidly to the Hayashi line. The He-burning shell, carried along by this contraction, becomes much more luminous ($T_{\rm shell} \propto M_c/R_c$ and L depends on T), driving another expansion of the outer regions of the star, and quenching H-fusion (by reducing the density). This creates a second dredge-up. The core stops contracting when it becomes sufficiently degenerate.

6. When the He shell approaches the extinguished H shell, the H shell is reignited.

7. The star moves up the Asymptotic Giant Branch. Mass loss becomes important. During this time, thinning of the He shell leads to thermal pulses (and a possible third dredge-up).

8. Eventually the luminosity becomes high enough to drive a superwind, which blows off a lot of mass (this is why a carbon flash never occurs). When the hydrogen-rich envelope reaches 1% of the star's mass, the entire star contracts, resulting in a fast journey toward the blue end of the HR diagram (at constant luminosity). The outgoing UV radiation is enough to ionize the surrounding material (ejected earlier from the star), creating a planetary nebula. What is left is a 0.55 - 0.6 $M_\odot$ CO white dwarf.

For massive stars:

1. A non-degenerate isothermal core more massive than 0.08 the mass of the star is formed at hydrogen exhaustion. The core subsequently contracts on a thermal timescale until helium fusion begins.

2. A carbon core is produced, and contracts on a thermal timescale until carbon fusion begins.

3. The process continues until an onion-like structure is achieved. An iron core is formed, and must begin to collapse to account for neutrino losses.

4. When collapse is sufficient to make iron fusion energetically preferred, the entire iron core collapses on a dynamical time, triggering a core-collapse supernova.

For massive stars mass loss (poorly understood) is significant, and may force the HRD track of a massive star to shift to the left due to the complete loss of the hydrogen envelope. Observationally, a nearly complete deficit of stars exist in the upper-right of the HR diagram (this is known as the Humphrey-Davidson limit), and mass loss is the likely cause.

4.1.1. *Describe protostellar evolution*

See Sec. 3.13 for details.

4.1.2. *What does this look like on a ρ-T diagram*

See Fig. 76. On a ρ-T diagram, the main sequence is located on the hydrogen ignition line, and evolution is generally toward high pressures and densities.

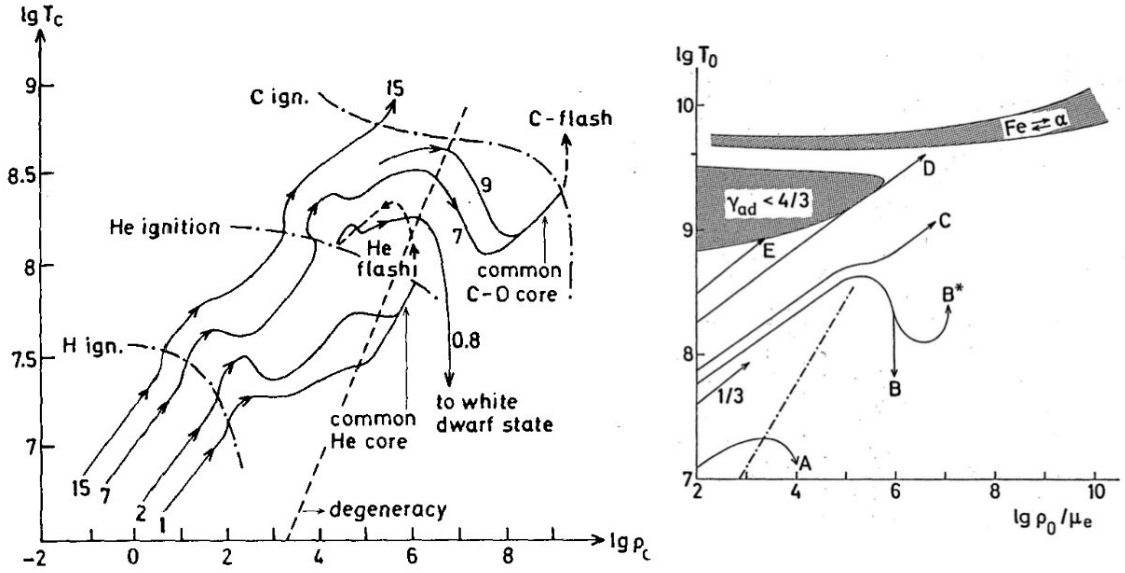

FIG. 76.— ρ-T diagram of stellar evolution. The left plot is a detailed subset of the right plot. More massive stars have tracks on the upper left of diagrams, while less massive on the lower right. $\gtrsim$120 $M_\odot$ stars may cross through the pair-instability region, losing pressure support and resulting in a pair-instability supernova. Note that the carbon flash does not occur due to mass loss. From Kippenhahn & Weigert (1994), their Figs. 33.6 and 34.1.

4.2. *Question 2*

QUESTION: Sketch a plot of radius versus mass for various "cold" objects, including planets, brown dwarfs and white dwarfs. Explain the mass-size relationship for rocky and gaseous objects.

Most of this information was derived by ourselves. It is consistent with the information presented in Chabrier et al. (2009).

The radius of an object is determined from a combination of the equation of hydrostatic equilibrium (with possible rotational correction), the energy generation equation, the opacity of the material, and the equation of state. In other words, the relevant equations are:

$$\begin{aligned} \frac{\partial m}{\partial r} &= 4\pi r^2 \rho \\ \frac{\partial P}{\partial r} &= -\frac{Gm\rho}{r^2} \\ \frac{\partial l}{\partial r} &= 4\pi r^2 \rho \left(\epsilon_{\rm nuc} - \epsilon_\nu - c_P \frac{\partial T}{\partial t} + \frac{\delta}{\rho} \frac{\partial P}{\partial t} \right) \\ \frac{\partial T}{\partial r} &= -\frac{GmT\rho}{r^2 P} \nabla \end{aligned} \tag{127}$$

For order of magnitude scaling relations, we can

$$
\begin{aligned}
M &\propto R^3\rho \\
P &\propto \frac{M\rho}{R} \propto \frac{M^2}{R^4} \\
L &\propto R^3\rho\epsilon \propto M\epsilon \\
T^4 &\propto \frac{\rho L}{R} \propto \frac{M}{R^4}L
\end{aligned} \tag{128}
$$

These scaling terms are generally useful for objects with spherical symmetry in hydrostatic equilibrium.

Rocky planets are held up by the lattice structure of rock, and to rough approximation can be assumed to be incompressible. Therefore, M is simply proportional to R^3, giving us

$$R \propto M^{1/3} \tag{129}$$

Brown dwarfs and white dwarfs are supported mostly by electron degeneracy pressure. For most white and brown dwarfs, the non-relativistic degenerate equation of state can be approximated by $P \propto \rho^{5/3}$. Combining this with the hydrostatic equilibrium and the fact $\rho \propto M/R^3$ gives us

$$R \propto M^{-1/3} \tag{130}$$

Near the Chandrasekhar mass, the Fermi sea becomes filled to the point that the most energetic electrons are relativistic. Under these conditions $P \propto \rho^{4/3}$. R, then, is roughly independent of mass.

Gaseous planets have degenerate cores, and extensive envelopes of non-degenerate gas. Since there should be a continuum of objects from Uranus/Neptune to 0.05 $M_\odot$ brown dwarfs (barring compositional differences which we do not take into account), we would assume that R is roughly independent of mass. A more physical explanation is that for gas giants, Coulomb repulsion between ions dominate over electron degeneracy, and since pressure is then dependent on the square of charge density (charge density is proportional to mass density), the equation of state can be approximated as $P \propto \rho^2$ (a polytrope of $n = 1$), which naturally results in radius being independent of mass. Detailed calculations (see Fig. 77) show this to indeed be the case, with the turnover point at around 3 Jupiter masses (Chabrier et al. 2009).

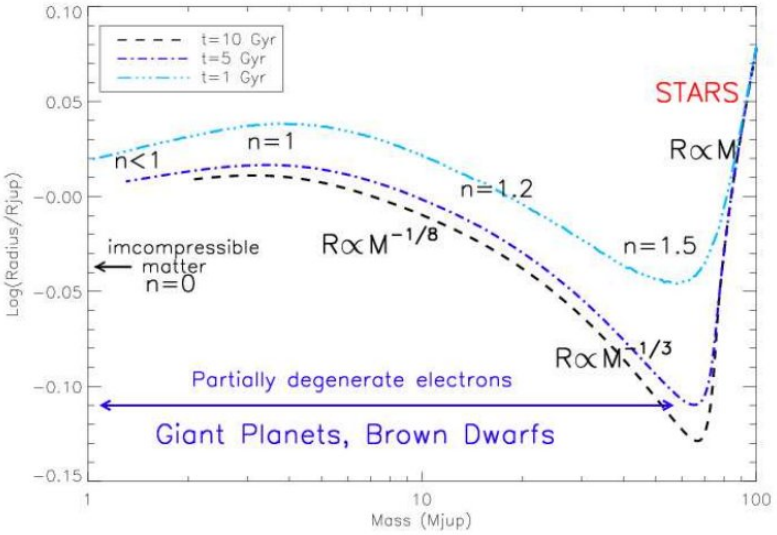

FIG. 77.— A plot of the mass-radius relation of objects from 1 to 100 Jupiter masses. Mass-radius relation regimes are labelled, as are their equivalent polytropic indices $n = 1/(\gamma - 1)$. Different lines represent different times during the evolution of the objects, which are assumed to be in isolation. Below 1 Jupiter mass, an $R \propto M^{1/3}$ relationship develops. At ~85 Jupiter masses, protostars are able to stably ignite the PP chain. From Chabrier et al. (2009), their Fig. 1.

4.2.1. *What other factors could change the mass-radius relation of an object?*

While the question asks for essentially zero temperature objects (all the forces above are loosely independent of temperature), in reality planets and brown dwarfs are not at zero temperature, especially if they are young and still being heated by gravitational contraction. The objects also rotate and, for brown dwarfs and low-mass stars, can be strongly magnetized, and both processes can act to puff the object up, or stymie convection (i.e. energy transport) through the object.

A detailed calculation of the radius of a gas giant must take into account a temperature gradient, changing elemental composition (i.e. where metals are located vs. where H/He are located) and changing phase of material throughout the star. Irradiation from the host star may also be important. Even taking all these complications into account, there is an ongoing mystery as to why certain planets, including a few at large distances from their parent planets, tend to be overly puffed up (Chabrier et al. 2009). Explanations range from increasing opacity to oscillatory convection to try and allow the planet to retain their primordial heat for longer periods of time (Chabrier et al. 2009).

Simulations of the mass radius relationship taking into account varying elemental compositions with respect to height can be found in Fig. 78.

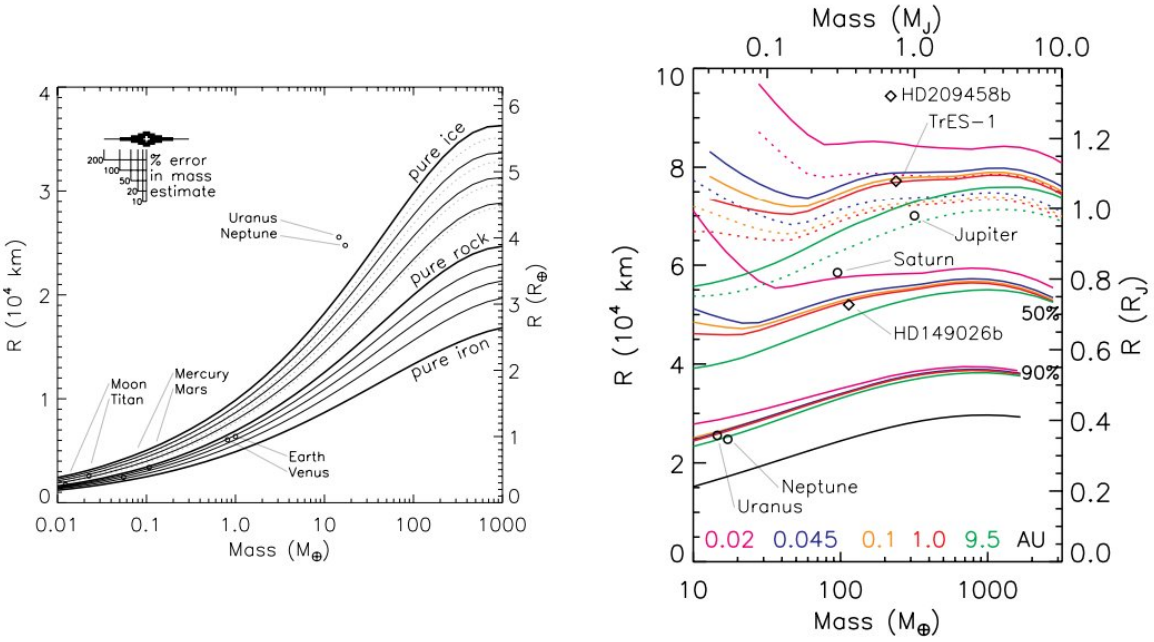

FIG. 78.— Left: mass radius relationships for planets of varying pure compositions. The top curve is pure ice, the middle pure rock (Mg_2SiO_4), and the bottom iron. The thin curves between the lines of pure material indicate mixtures of the material above and below. Error is shown on the upper left. Two orders of magnitude change in mass gives about a half order of magnitude change in radius, indicating a slightly shallower than $R \propto M^{1/3}$ relationship. Right: planetary radii at 4.5 Gyr as a function of mass. Models are calculated at 0.02, 0.045, 0.1, 1.0, and 9.5 AU and are color coded at the bottom of the plot. The black curve is for a heavy element planet of half ice and half rock. The groups of colored curves above the black curve is for planets that are 90%, 50%, 10% and 0% heavy elements. The open circles are solar system planets and the diamonds are extrasolar planets. From Fortney et al. (2007), their Figs. 4 and 7.

4.3. *Question 3*

QUESTION: Describe the physical conditions which lead to the formation of absorption lines in stars' spectra. What leads to emission lines?

When emission and absorption lines can be seen can be summed up by Kirchhoff's laws:

1. A hot, dense gas or hot solid object (can be generalized to any optically thick object in internal thermal equilibrium) produces a Planck continuum without any lines.

2. A diffuse gas produces emission lines. Emission lines are produced when an electron transitions from a higher energy state (bound or free) to a lower energy state (bound, or else the emission is not quantized). Transitions between bound states come in two types: allowed, which occur on a timescale of 10^{-8} s and follow selection rules calculated from time-dependent perturbation theory for electric dipole radiation (ex. for hydrogen $\delta l = \pm 1$ and $\delta m = 0, \pm 1$), and forbidden, which occur on much longer timescales (they will be allowed by higher-order multipole expansions of the electric and magnetic field; see Griffiths 2006, pg. 366). Since excited atoms are needed to generate emission lines, the gas must either be sufficiently hot kinetically so that collisions bring atoms into excited states, or be irradiated by photons at the appropriate wavelengths to bring atoms into excited states. The gas must also be tenuous enough that collisional de-excitation or multiple absorptions do not occur, as these will blur out the emission lines.

3. A colder, diffuse gas in front of a hotter continuous spectrum will produce absorption lines in that spectrum. These absorptions are due to electron transitions equivalent to the emission lines described above. Excited atoms spontaneously emit photons in random directions, and so the diffuse gas will have (generally weak) emission lines when viewed from other angles. The situation becomes complicated when a hot gas is in front of a hot body, in which case a combination of absorption and emission (sometimes at the same wavelength) will be seen.

For stars, the absorption lines are caused by scattering and absorption of the diffuse gas in the stellar atmosphere on top of the optically thick hot dense gas of the star. Another, more physical way of looking at the picture is to consider that when we view a star we are essentially looking at the last scattering/absorption surface, which, by the Eddington approximation, exists at an optical depth of $\tau \approx 2/3$. Atoms more readily absorb photons with energies close to the atoms' electron transition energies, and therefore the opacity κ_λ has resonant peaks near these transitions. The last scattering/absorption surface is less deep for photons near these resonant peaks than for other photons, and since temperature increases with depth into the star, when we look at the star at these resonant wavelengths we see a colder (i.e. less bright) spectrum than when we look at the star at other wavelengths[27]. This is the cause of absorption lines.

A similar explanation exists for solar limb darkening. When we observe the centre of a stellar disk, we are looking in the same direction as the radial direction of the star, but when we observe the edge of a disk, our line of sight has a perpendicular to the radial direction of the star. As a result, we are looking at a cooler, and therefore darker, region of the star if we observe its edge.

-Is there a middle ground between blackbody and emission lines? -Why do we not consider collisional de-excitation when calculating Einstein coefficients.

4.3.1. *How does radiative transfer occur inside a star?*

See (Carroll & Ostlie 2006), Ch. 9.2.

Consider a beam of photons with intensity I_λ passing through a cloud of material. Due to absorption and scattering of the light, photons can be removed from the beam (we will use "absorption" as the blanket term for this removal). The amount of absorption per unit mass is given by the absorption coefficient κ_λ, and

$$dI_\lambda = -\kappa_\lambda I_\lambda ds. \tag{131}$$

κ_λ can be due to a host of physical phenomena, including bound-bound transitions, photoionization, free-free absorption and electron scattering (Thompson/Compton). In stars later than F0 the majority of absorption comes from the H^- ion, as it can serve as the ionic component for free-free absorption (reverse bremsstrahlung requires two charged particles to conserve momentum and energy) and has a very low (0.754 eV, or 1640 nm) ionization threshold. Dust is also important in cool stellar atmospheres, since dust generally has a large number of molecular transitions. For earlier stars H-ionization and free-free absorption dominate opacity, and for O stars electron scattering becomes important. In a constant density medium with fixed κ_λ and no emission, the length over which I_λ drops to 1/e its initial value is known as the mean free path

$$l = \frac{1}{\kappa_\lambda \rho} = \frac{1}{n\sigma} \tag{132}$$

while the differential $\kappa_\lambda \rho ds = d\tau_\lambda$ is known as the optical depth (we have flipped sign to indicate that we are looking backward along the line of sight). In cases where detailed microphysics can be aggregated, we may use the Rosseland mean opacity, which is a harmonic average of κ_λ:

$$\frac{1}{\bar{\kappa}} = \frac{\int_0^\infty \frac{1}{\kappa_\nu}\frac{\partial B_\nu(T)}{\partial T} d\nu}{\int_0^\infty \frac{\partial B_\nu(T)}{\partial T} d\nu} \tag{133}$$

Rosseland bound-free and free-free opacities have the functional form $\bar{\kappa} \propto \rho/T^{3.5}$, which is known as a Kramers opacity law. The $\bar{\kappa}$ of H^- absorption is proportional to $\rho^{1/2}T^9$.

The plane-parallel approximation greatly simplifies the equation of radiative transfer, $\frac{dI_\lambda}{d\tau_\lambda} = I_\lambda - S_\lambda$, by assuming the atmosphere only changes along the $\pm\bar{z}$ direction. In this case, we can define a vertical optical depth $\tau_{\lambda,v}$, where $\tau_\lambda = \tau_{\lambda,v}/\cos\theta$, where θ is the angle between the line of sight and the z-axis; this nets us $\cos\theta \frac{dI_\lambda}{d\tau_{\lambda,v}} = I_\lambda - S_\lambda$. Integrating both sides with respect to Ω gives us (Carroll & Ostlie 2006, pg. 261)

$$\frac{dF_{\text{rad}}}{d\tau_v} = 4\pi\left(\langle I\rangle - S\right), \tag{134}$$

while integrating both sides by $cos\theta d\Omega$ nets us $\frac{dP_{\text{rad}}}{d\tau_v} = \frac{1}{c}F_{\text{rad}}$, which can be rewritten as $\frac{dP_{\text{rad}}}{dr} = -\frac{\bar{\kappa}\rho}{c}F_{\text{rad}}$. In a plane-parallel atmosphere, F_{rad} is a constant (σT_e^4) at equilibrium, netting us $P_{\text{rad}} = \frac{1}{c}F_{\text{rad}}\tau_v + C$ and the fact that

[27] The specific intensity I_λ, defined as the flux per unit steradian per unit wavelength, decreases, but not the total luminosity of the star. The deficit in intensity at resonant wavelengths is compensated by a larger total radiating surface.

$\langle I \rangle = S$. Lastly, we make the approximation that intensity only changes in the vertical direction. At any given point, then, $\langle I \rangle = \frac{1}{2}(I_{\rm out} - I_{\rm in})$ ("in" refers to lower z, while "out" higher), and from the general definitions of $F_{\rm rad}$ and $P_{\rm rad}$ (i.e. as integrals of I) we obtain their values as functions of $I_{\rm out} - I_{\rm in}$. From this, we find $C = \frac{2}{3c}F_{\rm rad}$. We know what $F_{\rm rad}$ is equal to (it is a constant), and for LTE $S = B = \frac{\sigma T^4}{\pi}$. Putting this all together finally gives us

$$T^4 = \frac{3}{4}T_e^4\left(\tau_v + \frac{2}{3}\right). \tag{135}$$

This equation is what sets the fact that the temperature we see is that of the photosphere where $\tau_v = 2/3$.

The angular dependence for limb darkening can be determined by assuming a plane-parallel atmosphere (so that τ_v can be used) and integrating the general equation for radiative transfer, then assuming some simple functional form for the source function as a function of τ_v.

4.4. *Question 4*

QUESTION: Why do some stars pulsate while some others do not? Consider Cepheids as an example.

This information comes from (Carroll & Ostlie 2006), Ch. 14.

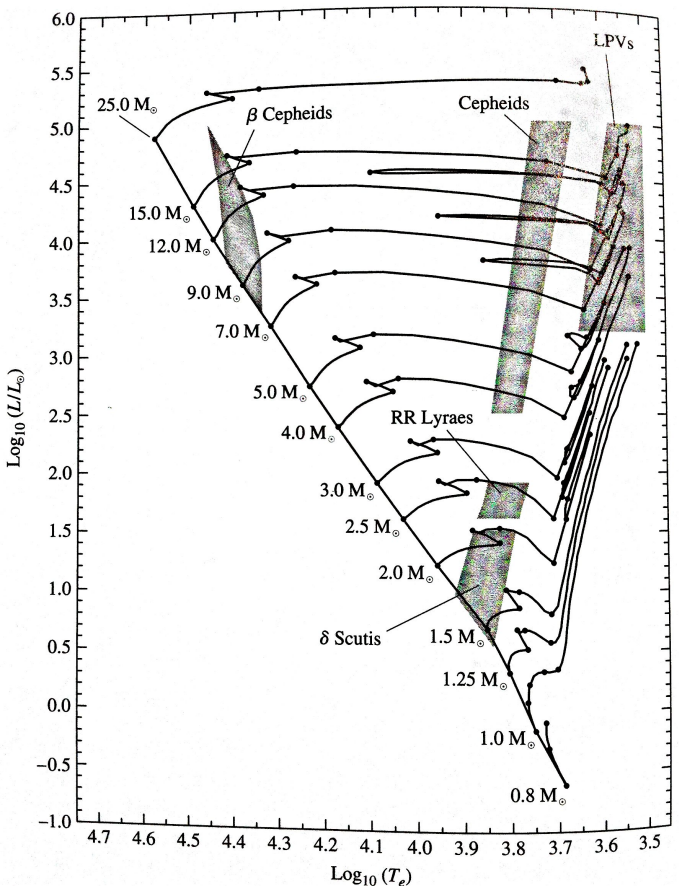

FIG. 79.— An HR-diagram with representative post-main sequence evolution tracks for stars of differing masses. Regions where pulsating stars can be found are shaded and labelled. From Carroll & Ostlie (2006), their Fig. 14.8.

Cepheids are a type of pulsating star, which in general tend to brighten and dim periodically. On the HR diagram a number of regions feature pulsational instability, most prominently the "instability strip", which passes from the WD line, through A and K MS stars to the AGB tip. Stars residing in the instability strip and above the main sequence include Cepheids and their W Virginis and RR Lyrae cousins reside. Since this region is not on the main sequence, stars that are currently Cepheids have only been so for a short period of their lives.

For stars to pulse, an instability in the stellar structure is needed as a driver, or else the vibrations would be damped on a Kelvin-Helmholtz timescale as heat leaks into the surroundings (Kippenhahn & Weigert 1994, pg. 408 - 409). If not damped, nuclear reactions (the "ϵ mechanism") can act as a spring - a compression of the core would lead to

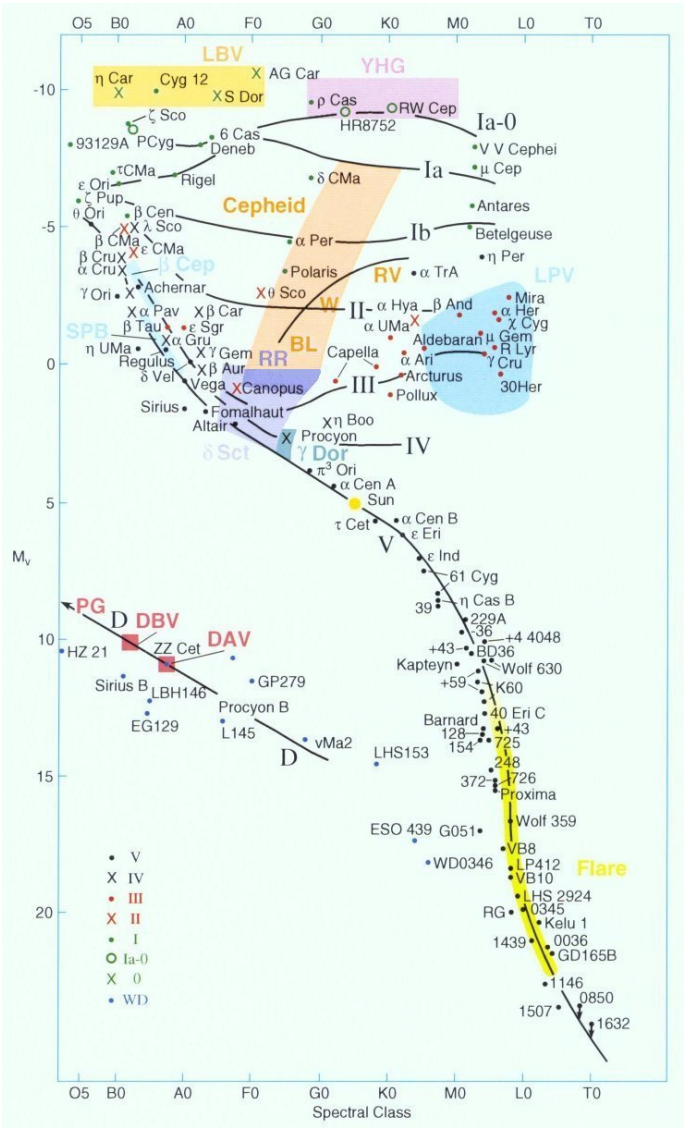

FIG. 80.— Similar to Fig. 79, but with more populations. LBV stands for luminous blue variables, YHG for yellow hypergiants, BL for BL Herculis, W for W Virginis, RV for RV Tauri, RR for RR Lyrae, SPB for slowly pulsating B stars, LPV for long period (Mira) variables, DAV for WDs with pure H atmospheres (also known as ZZ Ceti variables), DB for He atmosphere WDs, PG for PG 1159 stars and Flare for flaring M-dwarfs. From Kaler (2006).

increased nuclear reactions, driving an expansion that decreases nuclear reactions. The decrease again contracts the core, which increases nuclear reactions again. This instability may be important to preventing $\gtrsim$90 $M_\odot$ stars from forming.

Another method is for opacity to increase when a star is compressed - this would heat up and drive an expansion of the star, which would then decrease opacity, leading to less radiation retention, a cool down, and a contraction of the star. Generally this does not happen, since $\kappa \propto \rho T^{-3.5}$ by Kramers law (for adiabatic compression of an ideal monatomic gas $T \propto \rho^{\gamma-1} = \rho^{2/3}$), but in regions of partly ionized material, a compression will drive increased ionization, sapping some of the thermal energy generated by the compression and allowing κ to increase. Roughly speaking, for this to occur T must be proportional to $\rho^{2/7}$, meaning the system is roughly isothermal due to ionization (WHICH ALSO MEANS SYSTEM IS UNSTABLE TO COLLAPSE, BUT MIGHT NOT MATTER EXCEPT TO PREVENT THE EXPANSION FROM BEING HALTED DURING THE IONIZATION). Eventually ionization becomes less prominent, κ increases again, and drives an outward expansion. During the expansion, recombination reduces κ, allowing the envelope to cool, fall back down and restart the cycle. This is the reason why Cepheids pulsate.

Two ionization zones exist in stars: the first is a broad region where neutral H and He are ionized ($1 - 1.5 \times 10^4$ K) known as the hydrogen partial ionization zone, while the second is a more narrow region centred around 4×10^4 K where He is doubly ionized, known as the helium partial ionization zone. Proper driving of vibrational modes requires a sufficient amount of mass that undergoes the instability (the rest of the star damps such motions, since κ decreases from compression in those regions), and so these regions must be deep in the star. As a result, the surface temperature of the star cannot be $\gtrsim$ 7000 K. If the surface temperature is too low, the star skirts the Hayashi line and

TABLE 3
Pulsating Stars

Type	Range of Periods	Pop.	(Non-)Radial	Notes
Long-Period Variables (Miras)	100-700 days	I, II	R	RG or AGB stars
Cepheids	1-50 days	I	R	Produced by instability in partial ionization zones
W Virginis	2-45 days	II	R	Metal-deficient Cepheids with smaller luminosities
RR Lyrae	1.5-24 hrs	II	R	HB stars found in globular clusters. Used as standard candles
δ Scuti	1-3 hrs	I	R, NR	Evolved F stars found near the MS
β Cephei	3-7 hrs	I	R, NR	Similar to Cepheids, but due to iron partial ionization zone
ZZ Ceti (DAV)	100-1000 s	I	NR	Pulsating WDs on the instability strip

convection (which is more efficient with high compression) prevents the heat buildup due to compression required for the instability.

The helium partial ionization zone is more important than the hydrogen partial ionization zone. The hydrogen partial ionization zone generates a luminosity lag between minimum radius and maximum luminosity (see Carroll & Ostlie, pg. 498).

4.4.1. *What are the different classes of variable stars?*

Most variable stars are located on the instability strip in the HR diagram, and pulsate due to the κ-mechanism induced by partial He ionization. From most to least luminous, these members are Cepheids, W Virginis (pop II Cepheids four times less luminous), RR Lyraes (pop II horizontal branch stars, usually found in globular clusters), δ Scutis (and γ-Doraduses) and DBs (WD variables with He atmospheres). See Table 3 for details.

Some populations of variables rely on other mechanisms. Long-Period Varibles (Miras) pulsate due to the κ-mechanism induced by partial H ionization, complicated by dust formation, thermal pulses and mass loss. ZZ Cetis lie on the instability strip, but it is partial H ionization that drives their pulsation as well. PG variables are extremely hot versions of ZZ Cetis, and possibly pulsate due to partial metal ionization. Partial Fe ionziation is responsible for β Cepheis and slowly pulsating B stars (iron has many transition lines, making it significant despite being a minor pollutant in stars). Yellow hypergiants are semi-regular variables subject to huge dips as a result of dusty outbursts Kaler (2006). Flare stars pulse from reconnection/collapse of magnetic fields (Kaler 2006).

4.4.2. *Why do different stars have different pulsation frequencies? Describe the period-luminosity relation.*

We can obtain a rough estimate of the pulsation period by considering the adiabatic soundspeed

$$v_s = \sqrt{\frac{\gamma P}{\rho}} \tag{136}$$

If the density were roughly constant through the star (a horrifying assumption, but good for order of magnitude), $\frac{dP}{dr} = -\frac{4}{3}\pi G\rho^2 r$, which can be integrated to $P = \frac{2}{3}\pi G\rho^2(R^2 - r^2)$, where R is the star radius. We can then find, by determining the time it takes for sound to travel from $r = 0$ to $r = R$ and back, i.e. the period:

$$\Pi \approx \sqrt{\frac{3\pi}{2\gamma G\rho}}. \tag{137}$$

This is a fair approximation to the period/mean density relation (at least for radial oscillations - non-radial ones have longer periods). Gas giants are less dense than white dwarfs, and so have longer pulsation periods. This is the reason why, along the instability strip, more luminous stars have longer periods.

4.4.3. *What kinds of pulsations are there?*

Radial modes are radial standing waves (since $r = 0$ is tied down, the frequencies at which waveforms can propagate without destructively interfering are quantized). The fundamental mode involves the entire star expanding and contracting, while higher order modes have different regions of the star alternating in their expansion and contraction (see Fig. 14.9 and 14.10 in (Carroll & Ostlie 2006)). Radial modes disproportionately change the structure of the stellar outskirts, with the interior largely unchanged. Most classical Cepheids and W Virginis stars pulsate in the fundamental mode. RR Lyraes pulsate in the fundamental or first overtone, as (likely) do Miras.

We can mathematically describe radial pulsation by modifying the equation of hydrostatic equilibrium to include radial accelerations (i.e. by adding a $\rho\frac{d^2r}{dt^2}$ term after the $\frac{dP}{dr}$. While the full treatment of such a problem would involve heat transfer and non-linearity, we can simplify substantially by taking a first order perturbation of the equation, assuming some reasonable polytrope (combined with $\rho \propto M/R^3$ this allows us to determine $\frac{\delta P}{P} = -3\gamma\frac{\delta R}{R}$) to obtain a simple equation (Carroll & Ostlie 2006, pg. 501 - 502), $\frac{d^2(\delta R)}{dt^2} = -(3\gamma - 4)\frac{GM}{R^3}\delta R$, that has a sinusoidal solution.

More complex pulsation involve travelling waves that move across the star with a period of $m\Pi$[28] (where m is the

[28] This is because a non-radial pulsation is written as $Y_a^b(\theta,\phi)e^{-i\omega t}$; using a non-zero m nets us an expression that ends with $e^{im\phi - i\omega t}$. Analagous to $e^{ikx - i\omega t}$, the speed of propagation is ω/m.

azimuthal number), and can be described using spherical harmonics. Non-radial pulsations can either be mediated by pressure (p-modes) or surface gravity (f-modes, which only significantly shift the surface of the star). The pulsation period of p-modes can be estimated by the time it takes for a sound wave to travel one spatial wavelength $\lambda_h = \frac{2\pi r}{\sqrt{l(l+1)}}$, modulo the effect of moving pulsations with nonzero m. Deep inside the star, gravity can also act as a restoring force for non-radial modes known as g-modes, produced by competition between gravity and the buoyancy of piece of material. The physical process behind this oscillatory motion can be thought of as "failed convection", and so the material is physically "sloshing" back and forth at the Brunt-Väisälä frequency (see Carroll & Ostlie, pg. 507 - 508). G-modes involve significant movement in the deep interior of a star.

Pulsations are probes of stellar internal structure. The pulsation period above, for example, depends on average stellar density, while

4.5. *Question 5*

QUESTION: Define the terms "thermal equilibrium" and "hydrostatic equilibrium". How do they apply to stars? If a solar-type star is out of dynamical equilibrium, how long does it take to restore it? What is the time scale to restore thermal equilibrium?

Thermal equilibrium occurs when a system has uniform temperature throughout, such that the net transfer of thermal energy between any two regions in the system is zero. A stronger condition, thermodynamic equilibrium, also exists, where the object which thermal equilibrium, mechanical equilibrium, radiative equilibrium, and chemical equilibrium. Hydrostatic equilibrium is a condition in fluid mechanics where a volume of fluid is either completely at rest or moving with uniform and constant velocity - i.e. each fluid element feels no acceleration.

Stars are, in general, in hydrostatic equilibrium. Taking a spherically symmetric approximation, this means

$$\frac{dP}{dr} = -\frac{GM_r\rho}{r^2}, \tag{138}$$

i.e. a difference in (thermal, Coulomb, degeneracy, etc.) pressure holds a slice of material up against gravity. There are, however, many cases in which only quasi-hydrostatic equilibrium is maintained, such as with pulsating stars. At certain stages of their lives stars experience significant mass loss, during which they are not in hydrostatic equilibrium at all (since they are expelling mass). Stars are never in global thermal or thermodynamic equilibrium, as they do not have uniform particle or radiation field temperature throughout their entire structure, or the same level of ionization (since the H and He ionization temperatures are higher than the surface temperatures, but lower than the core temperatures, of many stars). The mean free path of a photon or particle is something like ten orders of magnitude smaller than the scale height of temperature variation $(T/\left|dT/dr\right|)$[29], and locally the internal, kinetic energies of the particles and the radiation field are coupled so that they can be defined by a single temperature. As a result, stars are in local thermal and thermodynamic equilibrium.

Stars out of hydrostatic equilibrium restore equilibrium on a dynamical timescale, while stars out of thermal equilibrium restore equilibrium on a thermal timescale. The dynamical time is derived from the equation of motion $\frac{1}{4\pi r^2}\frac{\partial^2 r}{\partial t^2} = -\frac{\partial P}{\partial m} - \frac{Gm}{4\pi r^4}$. Suppose we were to suddenly turn of the pressure. If we approximate the acceleration $\frac{\partial^2 r}{\partial t^2} \sim R/\tau_{\rm ff}^2$, where R is the stellar radius and $\tau_{\rm ff}$ is the free-fall time, we obtain $R/\tau_{\rm ff} \sim g$, or $\tau_{\rm ff} \sim (R/g)^{1/2}$. If we suddenly turn off gravity, a similar argument gives us $\tau_{\rm explosion} \sim R(\rho/P)^{1/2}$. These two terms give the timescale over which their respective processes work. If the star were near thermal equilibrium, we would expect the two timescales to be of the same order of magnitude. Taking $g \sim GM/R^2$ and the equation for $\tau_{\rm ff}$, we obtain the dynamical time

$$\tau_{\rm dyn} \sim \left(\frac{R^3}{GM}\right)^{1/2} \sim \frac{1}{2\sqrt{G\bar{\rho}}}. \tag{139}$$

This is the timescale over which a star moved out of hydrostatic equilibrium will restore it, and is of the same order of magnitude as the sound crossing time, Eqn. 137. For the Sun, it is about 27 minutes; for a red giant, 18 days, and for a WD, 4.5 seconds.

Assuming the star is approximated by an ideal gas, the timescale for thermal fluctuations to travel through a star is generally given by

$$\tau_{\rm KH} \sim \frac{c_V \bar{T} M}{L} = \frac{E_{\rm internal}}{L} \tag{140}$$

where the specific heat c_V is per unit mass. This happens to be the same timescale that Kelvin and Helmholtz gave for the "lifetime" of the Sun, before nuclear fusion was determined to be the primary source of stellar energy; hence,

[29] See Kippenhahn & Weigert, pg. 27 - 28 for estimates. The temperature gradient can be estimated by $(T(0) - T(R))/R$, while the mean free path by the mean density of the Sun, and $\kappa \approx 1$ cm^2/g.

this timescale goes by the name Kelvin-Helmholtz timescale. For the Sun this value is $\sim 10^7$ yrs, much longer than the dynamical time, but much smaller than the nuclear burning timescale of 10^{11} years.

The fact that the thermal adjustment time is usually[30] the Kelvin-Helmholtz timescale can be argued thusly: the photon diffusion process that takes the thermal energy of the star and turns it into blackbody radiation is a global thermal process, and therefore the KH timescale indicates how quickly the star as a whole can transfer energy from something hot (the stellar interior) to something cold (space). A more nuanced argument converts the radiative transfer (diffusion) equation to a form resembling the heat transfer equation, and then uses the thermal adjustment time in that equation to derive the equivalent adjustment time for stars. This time turns out to be, of order magnitude, the KH timescale.

4.5.1. *What is the Virial Theorem?*

Stars in hydrostatic and thermal equilibrium obey the virial theorem, which can be derived in the following manner. Begin with

$$\frac{\partial P}{\partial m} = -\frac{Gm}{4\pi r^4}. \tag{141}$$

Take the left side of Eqn. 141, and multiply by $4\pi r^3$, then integrate over mass shells:

$$\begin{aligned}\int_0^M \frac{dP}{dm} 4\pi r^3 dm &= \left[4\pi r^3 P\right]_0^M - 3\int_0^M P 4\pi r^2 \frac{dr}{dm} dm \\ &= -3\int_0^M \frac{P}{\rho} dm.\end{aligned} \tag{142}$$

The last line was obtained by using $\frac{dm}{dr} = 4\pi r^2 \rho$, $r(0) = 0$ and $P(M) = 0$. The right side of Eqn. 141 nets us

$$\int_0^M \frac{dP}{dm} 4\pi r^3 dm = -\int_0^M \frac{Gm}{4\pi r^4} 4\pi r^3 dm = E_{\rm grav}. \tag{143}$$

The virial theorem is then the statement

$$E_{\rm grav} = -3\int_0^M \frac{P}{\rho} dm \tag{144}$$

To obtain a more familiar formulation, we use the ideal gas to determine P: $PV = Nk_BT = \frac{M}{\mu} k_B T$, which gives $\frac{P}{\rho} = (c_P - c_V)T = 3(\gamma - 1)c_V T$, where c_V is the heat capacity per unit mass. If we define $E_{\rm internal} = c_V T$, we obtain

$$E_{\rm grav} = 3(\gamma - 1)E_{\rm internal}. \tag{145}$$

Setting $\gamma = 5/3$, for a non-relativistic monotonic gas, gives us the familiar $E_{\rm grav} = 2E_{\rm internal}$. Since any system at equilibrium will obey the virial theorem, we are afforded a powerful tool that can be used to determine how the internal energy of a system changes due to its changing shape.

4.5.2. *What is the instability criterion for stars?*

The instability criterion states that for a polytrope $P = K\rho^\gamma$ a star will be unstable to collapse if $\gamma \leq 4/3$. This can be shown reasonably easily. The scaling relation for hydrostatic equilibrium is $P \propto M^2/R^4$. The scaling relation for the polytrope is $P \propto M^\gamma/R^{3\gamma}$. Fixing mass, we require $P \propto R^{-4}$ and we have $P \propto R^{-3\gamma}$. If $\gamma \leq 4/3$, pressure cannot increase quickly enough to maintain hydrostatic equilibrium, and the star is unstable to collapse.

A more nuanced argument would resemble the effort to determine the Schönberg-Chandrasekhar mass or the Jeans instability - at a certain critical mass, the star will be unable to provide increased pressure support in the event of a momentary contraction. As a result, it is liable to collapse.

4.5.3. *What about the convective turnover time?*

The convective thermal transport timescale is orders of magnitude faster than the KH timescale. An upper limit can be found through the convective velocity, which is about $10^{-4} v_s$ for an $M_\odot$-star (see derivation in Ch. 10.5 of Carroll & Ostlie). This makes the convection timescale several orders of magnitude longer than the dynamical time, but much shorter than any other energy transport timescales in the star.

While convection can effectively transport energy through the interior of a star, the star as a whole will still have a thin radiative shell around its convective zone (even if the star is "completely") convective. Since the photon transport

[30] Usually, but not always! Consider the isothermal He core of an evolved star.

time in this shell is much longer than the time it takes to convect energy from the star's core, this radiative envelope sets the energy transport time of the entire star, given by the KH timescale[31].

4.6. *Question 6*

QUESTION: Define and describe Type Ia, Type Ib, Type Ic, and Type II supernovae.

Most of the information here comes from Carroll & Ostlie (2006) Ch. 15 and 18.5, and Arnett (1996) Ch. 13 (cited wherever used).

Supernovae are divided into the following observational subclasses:

- Type Ia supernovae do not exhibit any hydrogen lines in their spectra. They have strong Si II (615 nm) lines, and also lack He I lines (most prominently at 587.6 nm) (Pastorello 2012).
- Type Ib have no hydrogen lines, and strong He I lines, but lack Si II lines.
- Type Ic have no hydrogen lines, no He lines. They do have strong Ca II and O I lines (Pastorello 2012). The most luminous of these have broad-lined spectra and are sometimes associated with gamma-ray bursts (Pastorello 2012).
- Type II-P exhibit hydrogen lines, and have significant ∼50-day "bump" in the light curve following maximum light. The luminosity may still decline over the course of the bump, but at a much reduced rate. Typically, their spectra show broad H Balmer lines with prominent and quasi-symmetric P-Cygni proles (Pastorello 2012).
- Type II-L exhibit hydrogen lines, and have no 50-day bump (they instead have a linear increase in magnitude as a function of time following peak light). In general, the spectra are dominated by prominent and broad H lines, usually with the emission components dominating over the absorptions (Pastorello 2012)
- Type IIn show narrow H emission lines superimposed on broader components. Normally, the photometric evolution of SNe IIn is slower than that of other SN types (Pastorello 2012).
- Type IIb transition between II and Ib. Their spectra show simultaneously He I and (sometimes weak) H lines. The photometric evolution is fast and similar to that of type I SNe (Pastorello 2012).

Many of the lines described here have a characteristic P Cygni profile during the early phase of the supernova, which is an extended and blueshifted dip in the absorption, followed by a redshifted emission bump (see Fig. 82). The dip is from absorption due to blushifted outward-expanding material (as well as more stationary material expanding perpendicular to the line of sight), and the bump due to emmission from redshifted material. See Fig. 12.17 in Carroll & Ostlie for the emission variant of the a P Cygni profile.

Note that these subclasses are entirely observational in nature, and betray no information regarding the progenitor object or process. To determine, for example, nucleosynthetic yields, late-time spectra of the nebular phase of emission (i.e. after the entire supernova becomes optically thin) is required.

There are two different physical mechanisms for creating canonical supernovae: core collapse of a massive star and thermonuclear detonation of a white dwarf.

Core collapse occurs after a massive star developes an "onion-shell" structure of material (H burns to He burns to C burns to O burns to Si; to extract the same amount of luminosity from heavy element fusion requires much greater fusion rates, leading to short lifetimes (2 days for Si burning!)) with an iron core at the centre. Since iron fusion is not energetically viable, and the core loses energy through neutrino emission, the core contracts. Iron fusion itself may occur, but since it is energetically favourable for its products to dissociate, trans-iron elements are not fused in abundance. Eventually temperatures in the core reach 5×10^9 K, which is enough to destroy heavy nuclei in photodisintegration ($^{56}\mathrm{Fe} + \gamma \rightarrow 13\,^4\mathrm{He} + 4n$ and $^4\mathrm{He} + \gamma \rightarrow 2p^+ + 2n$). This reduces thermal support, furthering contraction. At even higher temperatures and densities electron capture, $p^+ + e^- \rightarrow n + \nu_e$, becomes energetically favourable. This release of neutrinos has five orders of magnitude greater luminosity than the luminosity of the star at the time. This robs the core of its main support, electron degeneracy pressure, and as a result a dynamical collapse becomes inevitable.

A white dwarf, normally inert, can be driven into thermonuclear detonation in several ways. The first is by compressing the white dwarf until pyconuclear fusion (fusion due to particles being so close-packed that the probability for quantum tunnelling increases) is ignited. This can be accomplished by driving the WD to close to the Chandrasekhar mass (equilibrium near-M_{ch} models have extremely high density), or by using a external shockwave to compress the interior. The second is by adiabatically compressing the WD until the core material becomes hot enough to ignite carbon fusion. In either case fusion ignited inside a degenerate object is a runaway process, since pressure does not increase with increasing temperature (the fusion timescale is far shorter than the KH timescale).

[31] The diffusion time will be shorter than if the star were completely radiative, since there is less matter for the photons to diffuse through, but that will be reflected in an increased luminosity, and the KH time still applies

It is the chemical composition of the atmosphere that differentiate different SNe. Type II supernovae must occur in hydrogen-rich objects, while type Ib and Ic in Wolf-Rayat stars that have lost their hydrogen and helium outer envelopes, respectively. Type Ia supernovae, because they occur in CO WDs, generate explosions devoid of both hydrogen and helium in their spectra, and have a large amount of intermediate mass (i.e. between He and Ni) material in their out envelopes, such as Si and Mg. See below for more.

Aside from these, a number of odd outliers exist, including underluminous supernovae, which lie between the maximum magnitude for novae, $M_v \approx -10$, to the minimum magnitude for canonical supernovae, $M_v \approx -16$, and ultra-bright SNe. The following comes from Pastorello (2012).

- Under-luminous SNe II-P, which look like bolometrically dimmed SNe II-P with especially narrow lines during the nebular emission phase (line broadening is a function of ejecta velocity). There is evidence these SNe are due to core collapse of 8 - 10 $M_\odot$ stars.

- SN 2008S-like transients, which look like weak SNe IIn (luminous blue variable outbursts have similar characteristics to very weak SNe IIn, and therefore SN 2008S-like transients can be considered a IIn/LBV outburst hybrid). They have peak magnitudes in the range between -12.5 and -14, type IIn-like spectra evolving with time to redder colors and showing the characteristic narrow emission lines. These are possibly electron capture SNe (core-collapse due to a degenerate core exceeding the Chandrasekhar mass).

- SN 2008ha and 2002cx-like SNe, which have narrow P Cygni features (indicating low expansion velocities) no H and He I features, and only very weak Si II and S II lines. It is possible these objects are failed SNe Ia that exploded as either pure deflagrations (for the brighter members of the class) or failed deflagrations (for the dimmer members). Pastorello speculates they may be related to SN 2005E-type calcium rich SNe.

- SN 2005E-like SNe, which are fast, low-luminosity type Ib SNe that produce very little ^{56}Ni have low ejecta masses and have moderate ejecta velocities, and spectroscopically show strong lines of He, Ca and Ti (SN 2005E itself had 0.28 $M_\odot$ of ejecta, 0.003 ^{56}Ni produced, and 1.1×10^4 km s^1 average ejecta velocity) (Zhu 2011). These explosions may be due to core collapse of $\sim$10 $M_\odot$ stars or the detonation of a thick He shell on top a CO WD that does not destroy the CO WD.

- Ultra-luminous SNe Ia, such as SN 2009dc, that can be explained by either more than 1.4 $M_\odot$ of material exploding (the merger of two WDs can momentarily produce a WD more massive than $M_{\rm ch}$, though this object is obviously short-lived) or a non-spherically symmetric explosion.

- Ultra-luminous SNe IIs, such as SN 2007bi, that show no evidence of CSM interaction (which could artificially boost the luminosity of a supernova by transferring expansion kinetic energy into heat). These may be due to pair-instability supernovae in massive stars.

Additionally, there are phenomena such as accretion-induced collapse (AIC, essentially a bare electron capture), SNe .Ia and electron capture supernovae that have been theoretically predicted but not yet definitively seen.

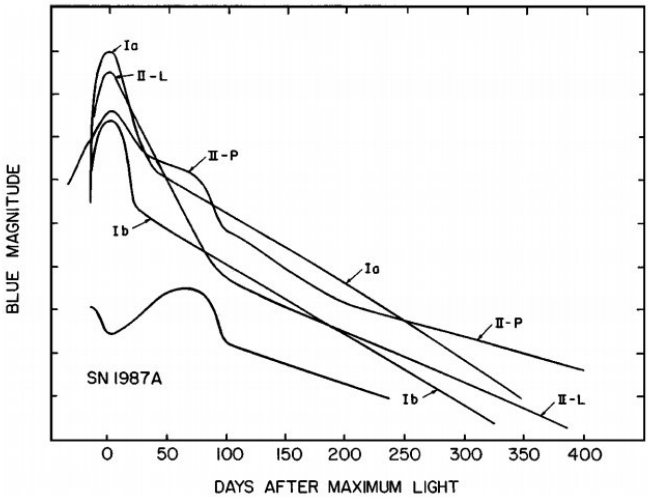

FIG. 81.— Schematic light curves for SNe of Types Ia, Ib, II-L, II-P, and SN 1987A. The curve for SNe Ib includes SNe Ic as well, and represents an average. For SNe II-L, SNe 1979C and 1980K are used, but these might be unusually luminous. Each tick is -0.5 magnitudes, and the tick nearest SN 1987A's maximum light is -15.5 mag. From Filippenko (1997), their Fig. 3.

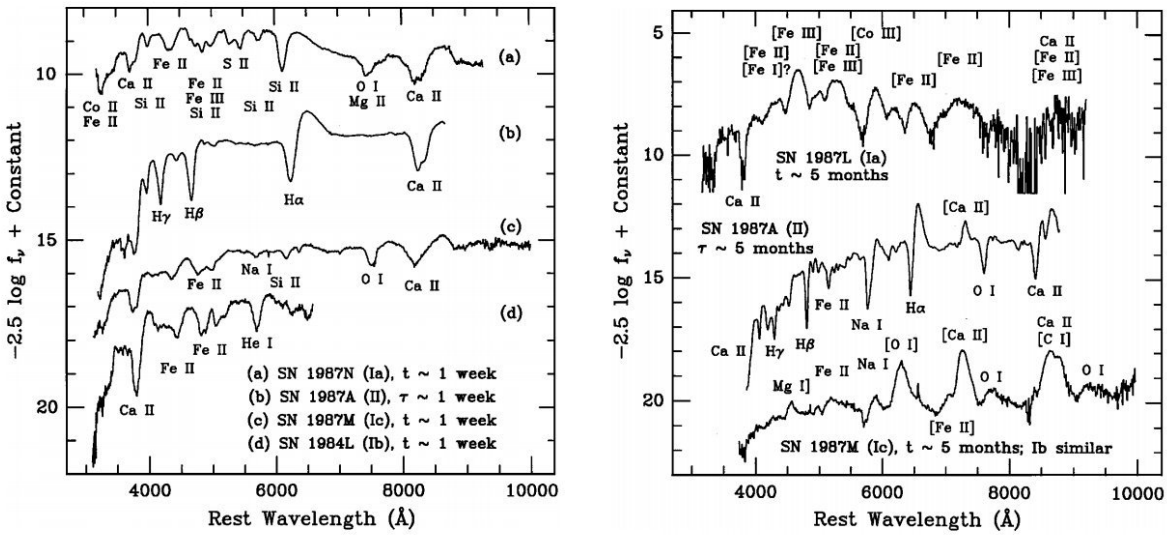

FIG. 82.— Schematic spectra of SNe, showing early-time distinctions between the four major types and subtypes. On the left are early-time spectra, and on the right late-time spectra. The parent galaxies and their redshifts (kilometers per second) are as follows: SN 1987N (NGC 7606; 2171), SN 1987A (LMC; 291), SN 1987M (NGC 2715; 1339), and SN 1984L (NGC 991; 1532). The variables t and τ represent time after observed B-band maximum and time after core collapse, respectively. The ordinate units are essentially "AB magnitudes". From Filippenko (1997), their Figs. 1 and 2.

4.6.1. *Describe the dynamical evolution of a supernova.*

During core collapse, initially the collapse is homologous, until the speed required for homologous collapse exceeds the sound speed in the outer layers of the core. The inner core therefore collapses (on a timescale of one second) down to about 50 km (from the size of the Earth!), with the remaining material in rapid (a decent fraction of c) free-fall. Collapse is stalled by the density of the inner core exceeding $\sim 8 \times 10^{20}$ g/cm^3, at which point the strong force between neutrons becomes repulsive and neutron degeneracy pressure can hold up the core. The sudden stalling of collapse leads to an outward propagating shockwave. Alternatively if the core is too massive, a black hole is produced, and the collapse is stalled at the Schwarzschild radius. This produces a weaker shock, or, for $\gtrsim 40$ M$_\odot$ stars, no shock at all (Belczynski et al. 2011). This shock drives into the outer core, creating a great deal more neutrinos through more photodisintegration of iron. The shock then stalls, with a large overdensity of neutrinos (a "neutrinosphere", from the photodisintegrations and the extremely hot proto-neutron star) contained below it (the neutrinos must interact strongly with matter this dense). These neutrinos superheat the medium below the stalled shock, which (somehow) triggers convection near the stalled shock, pushing infalling material back outward and allowing the shock to propagate outward (see Belczynski et al. for details).

The composition and mass of the envelope (ex. whether or not it is hydrogen free) are what determine whether or not the supernova is type I or type II. Type II supernovae are more common, and are due to massive red supergiants near the Hayashi line. Type Ib and Ic have suffered significant mass loss (hence their lack of hydrogen), and are likely due to the explosion of a WN and WC star, respectively[32]. As mentioned ealier, WD explosions should have no He or H, and significant amounts of intermediate mass materials, in agreement with the spectra of SNe Ia.

In WDs, an explosion is triggered by runaway nuclear fusion (i.e. fusion increases the temperature, which increases fusion rates). If the temperature becomes so high that the nuclear burning time exceeds the dynamical time of the star, an explosion results. In a pyconuclear fusion object must first be a subsonic deflagration, which allows the burning regions to expand somewhat to give the burning products seen in SNe Ia, and then a supersonic detonation, which gives high ejecta velocities. For an adiabatic compression object, a pure detonation is required for high ejecta velocities and correct burning products.

A supernova's velocity profile is imprinted by the outgoing shockwave of the explosion, and can be described roughly as a homology, which means that the outermost layers of material in the star are ejected the most quickly. SNe Ia, for example, have velocities ranging from 2×10^4 km/s to 5×10^3 km/s (Zhu 2011). The average velocity for type I SNe is 1×10^4 km/s, while for a type II it is 5×10^3 (Arnett 1996, pg. 418).

Eventually the expanding shell of material must come into contact with the surrounding C/ISM (all of this information comes from Draine (2011) Ch. 39.1). The interaction between the supernova and the ISM occurs in the following three phases (technically, by the optically thin phase an SN can be considered an SN remnant (SNR), but I will continue to use SN throughout):

[32] Wolf-Rayat stars (Carroll & Ostlie 2006, pg. 521 - 522) can be subdivided into WN (He and N absorption lines in its atmosphere, with a little H), WC (He and C) and WO (mostly O) types. This sequence of subtypes is an evolutionary sequence: WN stars have lost most of their hydrogen envelopes, with CNO processed material (mostly N) brought to the surface by convection. WC stars have lost their CNO processed material, leaving only He and 3-α producs. WOs have lost most of their He envelope, leaving only O and heavier elements.

1. **The ballistic phase**, when the SN encounters little resistance from the C/ISM because the SN is too dense and energetic. When the pressure from shocking the C/ISM exceeds the thermal pressure of the SN ejecta, a reverse shock propagates back through the supernova remnant. Eventually the amount of C/ISM material swept up is comparable to that of the SN ejecta, and the reverse shock becomes powerful. SN II Cas A is in the midst of this phase, 400 years after it exploded.

2. **The Sedov-Taylor phase** begins when the reverse shock reaches the centre of the SN. The outgoing SN can now be approximated by a point-explosion depositing energy E_0 in a uniform medium of density ρ_0. The integrated similarity solutions (Fig. 39.1 of Draine) show the outgoing explosion imprints a linear velocity structure onto the expanding medium swept up by the explosion. They also give the expanding medium a very top-heavy density structure, such that most of the mass is situated near the shock front. During the Sedov-Taylor phase, we can assume radiative losses to be minimal, and so as to conserve energy (though not mass, as more material is swept up in the shockwave). The Sedov-Taylor phase ends when radiative losses become significant ($\sim 1/3$ is assumed by Draine).

3. **The snowplow phase** begins when the SN outer layers have cooled sufficiently such that the thermal pressure of the SN is comparable to the thermal pressure of the C/ISM, briefly stalling the explosion shock. The inner layers, however, still have substantial overpressure and has not cooled radiatively, and therefore expands adiabatically while forcing the cool outer shell to do the same (at almost the same speed as the shock). The cool outer shell therefore sweeps up stationary matter as it moves outward (which is why this phase is called the snowplow phase). The speed of the outflowing material eventually slows down to below the sound speed, and the shockwave dissipates, becoming a sound wave. The SN stops expanding and fades away when its expansion velocity equals the average velocity dispersion in the ISM it is situated.

4.6.2. *What are the energy scales and peak luminosities of these supernovae? Describe their light curves and spectra.*

A core collapse supernova is driven by the gravitational potential energy released from the collapsing core, and generates about 10^{53} erg, with 99% of it in neutrinos, 1% in ejecta kinetic energy and $\lesssim$0.01% in photons. This amount is approximately the same as the gravitational binding energy of the neutron star, since the neutron star binding energy is orders of magnitude greater than the binding energy of the core preceeding it. It takes approximately 20 (for type I) - 40 (type II-P) days to reach a peak luminosity of 10^{43} erg/s (-19.5 bolometric magnitude) (read from Fig. 83). Generally optical filters are used to observe SNe (since they eject most of their electromagnetic radiation in the optical), and the peak brightness of a core collapse supernova is about $M_B = -17.5$. Following peak light, light curve decline for type I b/c is about 0.065 mag/day, and reasonably similar ($\sim$ 0.05 mag/day) for type II-L (II-Ps have plateaus that can last for several weeks). 40 days or so after peak light for type Ib/c, and 75 - 100 days after peak light for type II (depending on whether or not there is a plateau), a turnoff for the decay occurs, and decay becomes constant. The rate is about 0.01 mag/day for Ib/c and 0.007 - 0.01 mag/day for type II (type II values based on Fig. 15.8 of Carroll & Ostlie).

The subluminous $M_B = -15.5$ SN 1987a was a particularly dim SNe II; this is because instead of a red supergiant, the star was a blue supergiant and therefore much denser. More of the thermal energy from shock heating and nuclear decay had been converted into adabatic expansion, and peak light was CAUSED? by the recombination bump at around 80 days.

It takes about 20 days to reach a peak luminosity of 10^{43} erg/s (-19.5 bolometric magnitude). Peak brightness in B is $M_B \approx M_V \approx -19.3 \pm 0.03$. Following peak light, light curve decline is about 0.065 mag/day until about 40 days after peak light the decay rate drops and becomes a constant 0.015 mag/day. A relationship, known as the Phillips Relation, exists between peak luminosity and decay rate, and it is from this relation that SNe Ia can be turned into distance indicators.

A supernova's light curve is determined by several phenomena: the cooling of shock-heated material during recombination, recombination of shock-heated material, and radioactive decay of elements synthesized in the explosion.

Shock heating occurs as the shockwave from the SN explosion spreads throughout the entire star. The entire star is superheated as a result, and when the shock reaches the surface, it too is superheated, resulting in a temperature and luminosity spike - this is the first electromagnetic indication that an explosion has occurred (Arnett 1996, pg. 417 - 421). As the SN expands, the diffusion timescale decreases while the expansion timescale (R/v, where v is fairly constant after the initial outgoing shock) increases, and the photosphere and SN interior both begin to cool substantially. Roughly speaking, the expansion of the SN is adiabatic (since the kinetic energy of the ejecta dominates over the radiation bath), and as a result the SN cools as work is done to expand the ejecta (Arnett 1996, pg. 423). This is less of a problem for large, extended objects, since they must cool less to escape their own potentials, and for SN II shock heating actually contributes a substantial portion of the luminosity at peak light (Carroll & Ostlie, pg. 534 and Arnett, pg. 424). Small, compact objects must be powered by radioactive decay.

After several days, radioactive decay from ^{56}Ni (half-life 6 days) becomes the primary source of energy (core collapse SNe produce less of it, which is why they have lower peak luminosities). While the supernova is optically thick this energy simply adds photons to the photon bath inside the supernova, which eventually diffuse out of the expanding shell; when the supernova becomes optically thin the photons can escape more readily (though they are still downscattered into the optical from the gamma before escaping) (Arnett 1996, pg. 441). As a result, the slope of the very late time light curve is determined by the half-life of the primary radioactive isotope in the supernova (generally ^{56}Co, with a

half-life of 77 days) using the expression $\frac{dM_{\rm bol}}{dt} = 1.086\frac{\ln 2}{\tau_{1/2}}$ (recall M is in log while t is not, hence an exponential decay becomes a straight line).

Recombination produces the a recombination plateau, where the supernova enters a period of constant luminosity as layer under layer of the supernova reach the recombination temperature of its primary envelope species and become optically thin. The recombination plateau is most easily identified in type II-Ps, which have enormous hydrogen envelopes (H-recombination occurs at ~ 5000 K). The recombination bump is powered by two factors: the energy of recombination and, generally much more important, the inward movement of the recombination front, which allows for photons to more easily escape, and changes the temperature structure of the optically thick region of the SN to allow for faster diffusion (Arnett 1996, pg. 441).

A supernova's spectrum depends essentially on the composition of the surface of last scattering. An SNe Ia's early time spectra consists mainly of intermediate-mass elements such as O, Mg and Si, and gradually switches to Fe and radioactive CO lines as the SNe becomes progressively less optically thick - this is because ^{56}Ni and other peak-iron elements are produced primarily in the core of the SN, which does not become optically thin until late in the supernova's evolution (note that absorption lines gradually become emission lines as the SN becomes optically thin) (Filippenko 1997). A type Ib looks like a type Ia (sans Si lines) until He appears several weeks after peak light (Filippenko 1997). Late-time spectra are strong in Mg, Ca and O. SNe Ic spectra are similar to SNe Ib, except with no He lines (Filippenko 1997). This makes them appear superficially like SNe Ia, but the nebular spectra of SNe Ia consist of broad emission-line blends of many forbidden transitions of singly and doubly ionized Fe and Co. SNe Ic (and SNe Ib), on the other hand, are dominated by a few strong, broad, relatively unblended emission lines of neutral oxygen and singly ionized calcium, together with weaker lines of C I, Mg I, Na I, and other intermediate-mass elements (Filippenko 1997). For SNe II-P, early spectra are nearly featureless, with a weak H-α p-Cygni lines. Late spectra feature strong H-α emission, alongside weaker Fe and intermediate element emission (Filippenko 1997). SNe II-L have similar spectra evolution, except that H-α absorption (alongside emission in a p-Cygni profile) never appears, which may be due SNe II-L progenitors not having a substantial H envelope (Filippenko 1997).

CSM interactions can also "re-light" a supernova: the rapidly moving supernova ejecta can shock-heat the CSM, leading to transferral of kinetic energy into thermal energy.

Supernovae are a source of $< 10^{16}$ eV cosmic rays. These charged particles orbit along a magnetic field in the vicinity of a supernova blast, and as a result sees a centripetal Lorentz force alongside a centrifugal force from the explanding SNe. In this way, the particles can be accelerated to extremely high velocities before escaping.

4.6.3. *What is the association between supernovae and gamma-ray bursts?*

Gamma-ray bursts (GRBs) are $\sim 10^{-2} - 10^3$ s bursts of intense gamma radiation with energy budgets of order 10^{53} erg (in γ-rays alone!). There are two types of GRBs, short (< 2 sec) and long, and short bursts tend to be harder (i.e. greater energy per photon). Short-hard GRBs are associated with neutron stars merging with other neutron stars and black holes, while long-soft GRBs have been seen to be associated with particularly energetic (30 times more energetic in certain cases) SNe Ib/c.

Since the burst has an enormous energy budget, it is likely the relativistic beaming is involved. A Lorentz factor of $\gamma = 10^2$ would allow the total energy budget to be reduced by $\gamma^2 = 10^4$. In the case of long-soft GRBs, this jet is produced by the formation of a black hole[33], and the subsequent accretion of material from a debris disk around the black hole, which would produce relativistic jets. In the collapsar model, the jet is formed during the supernova, and must plow its way out of the infalling stellar envelope. In the "supranova" model, a rapidly rotating neutron star is formed during the supernova, which subsequently collapses (because it slows down THROUGH MAGNETIC BRAKING?) to form a black hole and debris disk. The jet is then launched weeks or months after the supernova.

4.6.4. *What are the nucleosynthetic processes involved?*

The information in this paragraph comes from van Kerkwijk (2012). Explosive fusion is the rapid fusion of material in an SN. In such reactions, p^+, n and α particles are of prime importance, since it requires significantly less energy to pass even an α particle through a Coulombic barrier than a heavy nucleus (Arnett 1996). α-particles are also more easily created than free p^+ and n because individual α particles are strongly bound. As a result, α particles play a DOMINANT? role in creating heavier nuclei. For high enough temperatures and densities, nuclear fusion becomes significantly more rapid than any other processes, and the fusion reactions bring the material up to nuclear statistical equilibrium (NSE). Since ^{56}Ni has the highest binding energy given equal numbers of neutrons and protons, it is preferentially made[34] and dominates the nuclear ashes of SN (β-decay of ^{56}Ni into ^{56}Fe occurs on a much longer timescale). When temperatures/densities are insufficient, NSE is not reached and intermediate mass products are produced. This occurs in type II SNe (BECAUSE THEY'RE LESS DENSE?) and the lower-density regions of WDs (Zhu 2011).

To fuse with high-Z nuclei, charged particles have to cross/tunnel through a large potential barrier. Neutrons have no such issue, however, allowing for neutron capture:

$$ {}^{A}_{Z}X + n \rightarrow {}^{A+1}_{Z} X + \gamma, \tag{146}$$

[33] This occurs if the collapsing core has mass $\gtrsim 2.2$ M$_\odot$. A rotating neutron star may have mass up to ~ 3 M$_\odot$.

[34] If heavier elements are made, it is energetically favourable for the product to dissociate, and there are enough nearby photons to do it.

the product of which may or may not be stable against beta decay,

$$^{A+1}_{Z}X \rightarrow ^{A+1}_{Z+1}X + e^- + \bar{\nu}_e + \gamma. \tag{147}$$

If the neutron capture rate is slow compared to the half-life of beta decay, the neutron capture reaction is said to be a slow process, or s-process. This tends to produce stable nuclei, either directly if the product is stable, or through beta decay to a stable product (the s-process may produce stable products - it is still considered an s-process if potential beta decay is fast compared to neutron capture). If the opposite is true, the neutron capture reaction is said to be a rapid process, or r-process. The s-process can occur in stars and explosions. The r-process tends to occur in regions of high neutrino flux, namely core-collapse supernovae (and nuclear bombs). Both processes are responsible for $A > 60$ nuclei.

The mark of supernovae can be seen in the chemical compositions of stars. Li, Be and B, for example, are not commonly ejected in supernovae, and there is a deficit of these elements in the Sun[35]. C, N, Si, Ne and Fe have peaks because they are stable α-particle rich nuclei and are preferentially generated by massive star and SNe nucleosynthesis, and then ejected by SNe.

4.6.5. *What is the spatial distribution of supernovae?*

This information is from (Filippenko 1997, Sec. 2.3), updated slightly from information in Zhu (2011).

The spatial distribution of supernovae reflect the spatial distribution of their progenitors. SNe II, Ib, and Ic have rarely been seen in early-type galaxies. They are generally in or near spiral arms and H II regions, implying that their progenitors must have started their lives as massive stars. SNe Ia, on the other hand, occur in all types of galaxies, including ellipticals, though they are most common in spirals (there is no strong preference for spiral arms). SNe Ia probably come from intermediate-age (∼0.1 0.5 Gyr), moderately massive stars (4 7 $M_\odot$) stars, consistent with their progenitor being a CO WD.

4.7. *Question 7*

QUESTION: Small asteroids are usually odd shaped, while larger celestial bodies are round. The dividing line occurs at around 200 km. Explain what this is determined by. Using the same logic, can you estimate the tallest mountain that can stand on Earth and Mars, respectively?

The following information and calculation comes from Hartmann (2005).

Asteroids are irregular shaped, while rocky planets round, because below a certain mass, the material strength of the rock that makes up the body dominates over gravity. If, however, the central presure exceeds the material strength of the rock, the rock will deform, either by plastic flow or fracture, as a result of failure of the normal elastic properties of solid rock. Higher internal temperatures also favour deformation (since the material becomes more malleable) onto equipotential surfaces.

We can perform a back of the envelope calculation to determine the dividing line between round and irregular objects. Most irregular objects are in the imcompressible regime for materials, meaning ρ is approximately constant. If we assume $\rho \approx \bar{\rho}$, i.e. use the average density as an estimate of the true density, we may then easily integrate Eqn. 138 to obtain

$$P_c = \bar{\rho}^2 \frac{2\pi G}{3} R^2, \tag{148}$$

where we have assumed $P_s = 0$. We may rewrite this to obtain

$$R = \frac{1}{\bar{\rho}} \sqrt{\frac{3P_c}{2\pi G}}. \tag{149}$$

Now, rocky planetismal material has a strength and density of $\sim 2 \times 10^8$ N/m^2 and 3500 kg/m^3. This gives us $R \approx 340$ km (Hartmann obtains slightly different results because inexplicably he considers force at a point $0.2R$ from the centre). Iron-cored planetismal material has a strength of $\sim 4 \times 10^8$ N/m^2 and 8000 kg/m^3. This gives us $R \approx 210$ km. Real-world statistics give a radius of around 2 - 300 km being the transition zone between the population of almost entirely spherical objects and the population of objects with a huge range of possible a/b and a/c (i.e. triaxial axis lengths) (see pg. 177 of Hartmann).

The known asteroid population spans some five orders of magnitude in size, up to a diameter of about 1000 km Emberson (2012). As a consequence, it happens to include the transition from the shape regime dominated by material strength, to that controlled by self-gravitational forces. Most small asteroids are irregular rocky boulders, generated by catastrophic fragmentation events, while the largest ones instead look like regularly-shaped small planets. However,

[35] Comparing the Sun's Li and Be abundances to those of meteorites, we find comparable amounts of Be but a much less Li on the Sun. Presumably convection brings surface material far enough so $T \approx 3 \times 10^6$ K, which burns Li but not Be, but this disagrees with stellar models, which never have convection go down that far - this is known as the "solar lithium problem".

there are exceptions to this rule, due to variations in self-gravity, material properties, and collisional history. In the intermediate transition range, in particular, a variety of shapes and surface morphologies are present Emberson (2012).

Observations of fast-rotating asteroids (FRAs) larger than $R \approx 100$ m find a sharp cutoff of the spin rate that corresponds well to the the rate that circular motion balances against gravity Emberson (2012). This suggests that FRAs larger than 100 m are "rubble piles" rather than monolithic objects that would have additional tensile support from the material they are made of Emberson (2012). The lack of tensile strength does not mean that rubbles piles are in hydrostatic equilibrium - friction and finite particle size effects prevent the bodies from forming equilibrium shapes unless they are so massive that gravity overcomes them Emberson (2012).

A similar argument can be made for the maximum height of a mountain on Earth or Mars. The pressure at the base of the mountain is $P = \rho g h$, which should be equal to the material strength of rock. Surface rock is rarely entirely compsed of iron, so let us assume $\sim 2 \times 10^8$ N/m^2 and 3500 kg/m^3. The surface gravity on Earth is 9.80 m/s^2, while on Mars it is 3.71 m/s^2 (Wolfram 2012). This gives us $h = 6$ km on Earth, and $h = 15$ km on Mars. If we assume mountains are cylinders, these will be the maximum heights. If we instead assume mountains are cones, and that the mean height of the mountain (given by $\frac{\int_0^{R_{\rm xy}} 2\pi h r dr}{\pi R_{\rm xy}^2}$, where $h = h_0 - \frac{h_0}{R_{\rm xy}} r_{\rm xy}$) $\bar{h} = \frac{1}{3} h_0$ must equal $h = 6$ km on Earth and $h = 15$ km on Mars (since the mountain holds itself up as a whole rather than as individual vertical skewers) then the maximum height of a mountain will be $h = 18$ km on Earth and $h = 45$ km on Mars.

Additional pressure for mountains can come from the upthrust pressure from the rock below (ultimately due to movement in the mantle). Mountains on Earth also have atmospheres on top of them, but this pressure is negligible (P at sea level is 10^5 N/m^2). Rapid rotation may lower the effective g (both by adding a centrifugal force and by increasing radius from the planet's centre), but on Earth this is a 1% effect.

The true height of Mount Everest is 8.8 km, and the height of Olympus Mons on Mars is 21 km (Wolfram 2012).

4.7.1. *How are asteroids related to Solar System formation?*

Asteroids and other planetismals are the remnants of planetary formation, bodies that were not incorporated into a planet during the formation of the Solar System, and were too large to be swept away by the T-Tauri phase of the Sun (which removed gas and dust not coalesced into larger bodies) (Carroll & Ostlie 2006). It is also possible that they are the rock mantle (asteroids) and volatile crust (comets) remnants of a giant impact event during Solar System formation (Cole & Woolfson 2002). Once the giant planets (particularly Jupiter) formed, they began to scatter off of planets in glancing gravitational encounters, giving them enough velocity that an encounter with another asteroid would lead to a fragmentation collision rather than a accretive one. As a result, planetismals within the orbit of Neptune could never gather together to form a planet. Instead, their fates are either to collide with planets, become captured into satellite or resonant orbits, or fragment further. Only material beyond ~ 43 AU, where the Kuiper Belt is, can escape gravitational perturbation.

4.7.2. *How does rotation change things?*

It cannot. Since rotation provides uniform support along one plane only, it can only support one kind of deformation (ellipsoidal deformation from a perfect sphere).

4.8. *Question 8*

QUESTION: Why are low mass stars convective in their outer envelopes while high mass stars are convective in their inner cores?

This information is from Ch. 6 of Kippenhahn & Weigert (1994).

Convective transport of energy occurs when a "hot" (high-entropy) region exchanges mass with a "cold" (low-entropy) region. Once this occurs, the convecting blobs of matter will dissolve into their new environments, which results in a net exchange of energy from the hot to the cold region. Convection is a difficult process to quantify in detail, especially in stars. The onset of convection, however, requires a dynamical instability to a thermal perturbation. The criterion for when this may occur is known as the Ledoux Criterion. We shall assume uniform composition for the star, in which case the Ledoux Criterion simplifies to the Schwarzschild Criterion,

$$\nabla_{\rm rad} > \nabla_{\rm ad}. \tag{150}$$

(This is wholly equivalent to saying $\left(\frac{dT}{dr}\right)_{\rm rad} < \left(\frac{dT}{dr}\right)_{\rm ad}$.) If this criterion is satisfied, then an adiabatically expanding piece of material in pressure equilibrium with its surroundings will always be hotter, and therefore less dense, than its surroundings. The gradient as a result of radiative diffusion alone is given by:

$$\nabla = \frac{3}{16\pi acG} \frac{\kappa l P}{m T^4} \tag{151}$$

In high mass stars, the central temperature is sufficiently high to light the CNO cycle. Because the CNO cycle has a T^{20} temperature dependence, the energy generation becomes very centrally peaked (i.e. l swiftly increases from

$r = 0$ outward), and because Eqn. 151 is dependent on l (and also T^{-4}, but l has a much greater dependence), $\nabla_{\rm rad}$ becomes large, exceeding $\nabla_{\rm ad} \approx 0.4$ for an ideal monatomic gas (combine $PV^\gamma =$ constant with $PV = Nk_BT$) in the star's core. At even higher masses, the star becomes increasingly radiative, and from Sec. 4.12.3 we find that $\nabla_{\rm ad}$ for a purely radiative star is 1/4, meaning that the convection zone extends further out from the core. At around a solar mass, the pp chain becomes dominant and therefore the star's core is no longer radiative.

In low mass stars, the κ term in Eqn. 151 becomes important - in the cold outer regions of low-mass stars opacity jumps by several orders of magnitude, which drives $\nabla_{\rm rad} > \nabla_{\rm ad}$, creating an outer convection zone. For even lower mass stars this outer convection zone dips into the core, and stars with $M < 0.25$ M$_\odot$ are completely convective.

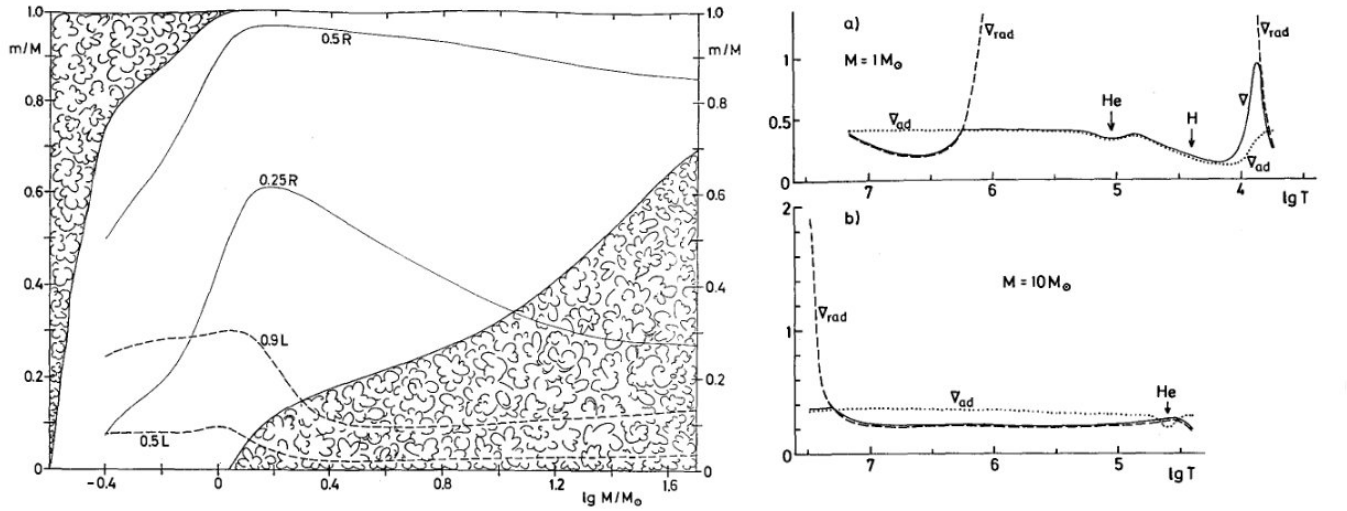

FIG. 83.— Left: stellar structure as a function of mass. Cloudy regions represent convection zones, solid lines represent contours of constant radius, and dashed lines contours of constant luminosity. Right: plot of ∇ as a function of stellar temperature (which is roughly monatonically decreasing throughout the star, so lower temperatures represent outer regions) for an M$_\odot$ and a 10 M$_\odot$ star. The dotted line represents $\nabla_{\rm ad}$, the dashed line $\nabla_{\rm rad}$ and the solid line ∇. Notice the dips in $\nabla_{\rm ad}$ in regions of partial ionization (labelled with their respective elements), and the deviation of ∇ from $\nabla_{\rm ad}$ in the outer layers of the M$_\odot$ star. From Kippenhahn & Weigert (1994), their Figs. 22.7 and 22.8.

4.8.1. *Derive the Ledoux Criterion.*

Suppose in a star a parcel of material has a thermal perturbation DT and responds adiabatically throughout its motion (because there is no time to transport heat out; this is a reasonable estimate for the deep interiors of stars). The local sound crossing time means that dynamical expansion in reponse to overpressure of this parcel is rapid, meaning that $DP = 0$. Consequently (from the ideal gas law) $D\rho < 0$, and so buoyant forces will cause the parcel to move upward by Δr. In this new environment, $D\rho = \left(\frac{d\rho}{dr}_e - \frac{d\rho}{dr}_s\right)\Delta r$, where e refers to the parcel ("element") of material and s the surroundings. From Archimedes' Principle the buoyant force is given by $-gD\rho$, and for this force to be positive, we require $\frac{d\rho}{dr}_e - \frac{d\rho}{dr}_s > 0$. Density is a state function, which allows us to write

$$\begin{aligned}\frac{d\rho}{dT} &= \frac{\partial \ln\rho}{\partial \ln P}\frac{dP}{P} + \frac{\partial \ln\rho}{\partial \ln T}\frac{dT}{T} + \frac{\partial \ln\rho}{\partial \ln \mu}\frac{d\mu}{\mu} \\ &= \alpha\frac{dP}{P} - \delta\frac{dT}{T} + \phi\frac{d\mu}{\mu}. \end{aligned} \tag{152}$$

For an ideal gas $\alpha = \delta = \phi = 1$. The parcel does not change its chemical composition, and the pressure gradients for both parcel and surrounding are identical, giving us

$$-\left(\delta\frac{dT}{Tdr}\right)_e + \left(\delta\frac{dT}{Tdr}\right)_s - \left(\phi\frac{d\mu}{\mu dr}\right)_s > 0 \tag{153}$$

We may multiply these terms by the presure scale height $H_P = -\frac{dr}{d\ln P} = -P\frac{dr}{dP}$. ($H_P = P/\rho g$ from hydrostatic equilibrium, and in the Solar photosphere, $H_P \approx 10^7$ cm, while in the centre of the Sun $H_P \to \infty$.) This nets us

$$\left(\frac{d\ln T}{d\ln P}\right)_s < \left(\frac{d\ln T}{d\ln P}\right)_e + \left(\frac{\phi}{\delta}\frac{d\ln\mu}{d\ln P}\right)_s \tag{154}$$

Lastly, we define $\nabla = \left(\frac{d\ln T}{d\ln P}\right)_s$, $\nabla_e = \left(\frac{d\ln T}{d\ln P}\right)_e$ and $\nabla_\mu = \left(\frac{d\ln\mu}{d\ln P}\right)_s$. This gives us $\nabla < \nabla_e + \frac{\psi}{\delta}\nabla_\mu$. Suppose we force

a star to purely be radiative. Then the surroundings will be $\nabla = \nabla_{\rm rad}$. The parcel of material has $\nabla = \nabla_{\rm ad}$ by our assumptions. We have therefore obtained the Ledoux criterion:

$$\nabla_{\rm rad} < \nabla_{\rm ad} + \frac{\psi}{\delta}\nabla_\mu. \tag{155}$$

Assuming a uniform composition throughout the system gives the Schwarzschild Criterion. Reversing the inequality gives the criterion for instability, $\nabla_{\rm rad} > \nabla_{\rm ad} + \frac{\psi}{\delta}\nabla_\mu$.

4.8.2. *Describe convection.*

The simplest model, which gives us most of what we want, is the "mixing length theory", which is modelled off of molecular transport (where convective cells are molecules and their distance by which they travel the mean free path).

During convection, a blob of material rises to the surface; the flux it carries is $F_{\rm conv} = \rho v c_P DT$. In this convection zone, let us assume that these blobs can go anywhere between zero to l_m distance, where l_m is the mixing length. DT/T is then given by $\frac{1}{H_P}(\nabla - \nabla_e)\frac{l_m}{2}$, naively assuming the average length travelled is $l_m/2$. $D\rho/\rho$ is simply $-\delta DT/T$, and $F_{\rm buoy} = -g\frac{D\rho}{\rho}$. if we assume half of the work ($1/2 F_{\rm buoy}$) goes into kinetic energy of the convecting blob (the other half shoves stuff aside), then we can find v, and from this find

$$F_{\rm conv} = \rho c_P T\sqrt{g\delta}\frac{l_m^2}{4\sqrt{2}}H_P^{-3/2}(\nabla - \nabla_e)^{3/2}. \tag{156}$$

Noting that $L = 4\pi r^2 F$, Carroll & Ostlie (2006) in Ch. 10.4 calculate, using a similar formula to Eqn. 156, that ∇ only has to be slightly superadiabatic for almost all the luminosity to be transported via convection.

The effectiveness of convection at energy transport is given by $\Gamma = \frac{\nabla - \nabla_e}{\nabla_e - \nabla_{\rm ad}} = 2\tau_{\rm adj}/\tau_l$ (τ_l is the lifetime of a parcel of convecting material). In the interior of a star, a good approximation is $\nabla = \nabla_{\rm ad}$, $\Gamma \to \infty$ and all of the energy is carried by convection. In the outer convective envelope, some of our assumptions break down (for example, that $\nabla_e = \nabla_{\rm ad}$) and a full mixing length theory calculation has be done to determine the true ∇, which will be somewhere in between $\nabla_{\rm rad}$ and $\nabla_{\rm ad}$. This is known as superadiabatic convection. In the photosphere of a star, $\nabla = \nabla_{\rm rad}$ and convection is ineffective (even if $\nabla_{\rm rad} > \nabla_{\rm ad}$; see below); $\Gamma \to 0$ and radiation transports all the energy.

4.8.3. *How does convective energy transport change radiative energy transport?*

As noted above, in the deep interior of stars, this change can be disregarded, since all the luminosity is carried by convection, but in the outer regions of the star, radiative and convective transport are much closer to being on par. The flux of energy due to radiation alone is given by

$$F_{\rm rad} = -\frac{4ac}{3}\frac{T^3}{\kappa\rho}\frac{\partial T}{\partial r}. \tag{157}$$

Noting that $\frac{\partial T}{\partial r} = \frac{\partial T}{\partial P}\frac{\partial P}{\partial r}$ (and $\frac{\partial P}{\partial r}$ is defined from hydrostatic equilibrium), we can rewrite this as:

$$F_{\rm rad} = \frac{4acG}{3}\frac{T^4 m}{\kappa P r^2}\nabla \tag{158}$$

where $\nabla = \frac{d\ln T}{d\ln P}$. The specific case where all outward energy transport is radiative requires $\nabla = \nabla_{\rm rad}$. When convection occurs, ∇ is shallower.

4.8.4. *What are the main sources of opacity in the stellar atmosphere?*

See Sec. 4.3.1.

4.8.5. *Why is convection ineffective near the photosphere?*

We define $U = \frac{3acT^3}{c_P\rho^2\kappa l^2}\sqrt{\frac{8H_P}{g\delta}}$, which is the ratio of "conductiveness" between radiative diffusion and convection. U is directly related to Γ, above, in that $U \to 0$ when $\Gamma \to \infty$, and vice versa. From the derivation in Ch. 7 of Kippenhahn & Weigert, we know that if $U \to \infty$ convection becomes highly inefficient. This occurs near the photosphere because the star is too underdense (Kippenhahn & Weigert 1994, Ch. 10.3.2).

4.8.6. *What is the Hayashi Line?*

The Hayashi line(s) (as there is a separate line for any given M and chemical composition) is the locus of fully convective stellar models (except for their outermost layers, which are superadiabatic). These lines are located on the far right of the HR diagram, typically with surface temperatures around 3000 - 5000 K (colder for more massive objects) and are nearly vertical. The Hayashi line is also a boundary between the "allowed" and "forbidden" region of the HR diagram: it can be shown (Kippenhahn & Weigert 1994, pg. 229 - 230) that a star left of its Hayashi line has $\bar{\nabla} < \nabla_{\rm ad}$, and any star right of its line has $\bar{\nabla} > \nabla_{\rm ad}$. The star on the right must then be superadiabatic,

and vigourous convection will bring the star back to $\bar{\nabla} < \nabla_{\rm ad}$ and onto the left of the Hayashi line on a convective adjustment timescale. Stars that are not in hydrostatic equilibrium can certainly evolve to the right of the Hayashi line, but they change on a dynamical time, making their time to the right of the line potentially even shorter.

See Fig. 84.

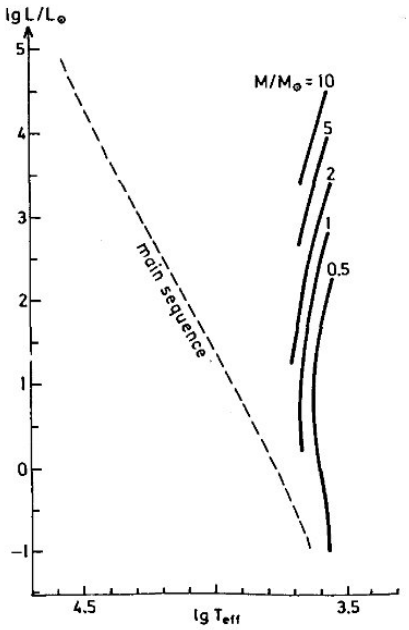

FIG. 84.— The position of the Hayashi Line for stars of masses between 0.5 and 10 $M_\odot$, composition $X = 0.739$, $Y = 0.24$ and $Z = 0.021$ and $l_m/H_P = 2$. From Kippenhahn & Weigert (1994), their Fig 24.4.

4.9. *Question 9*

QUESTION: Describe and compare the ages of the surfaces of Mercury, Venus, Earth and Mars.

Most of the information for this answer comes from Carroll & Ostlie (2006) Ch. 20. Age determination information comes from de Pater & Lissauer (2009).

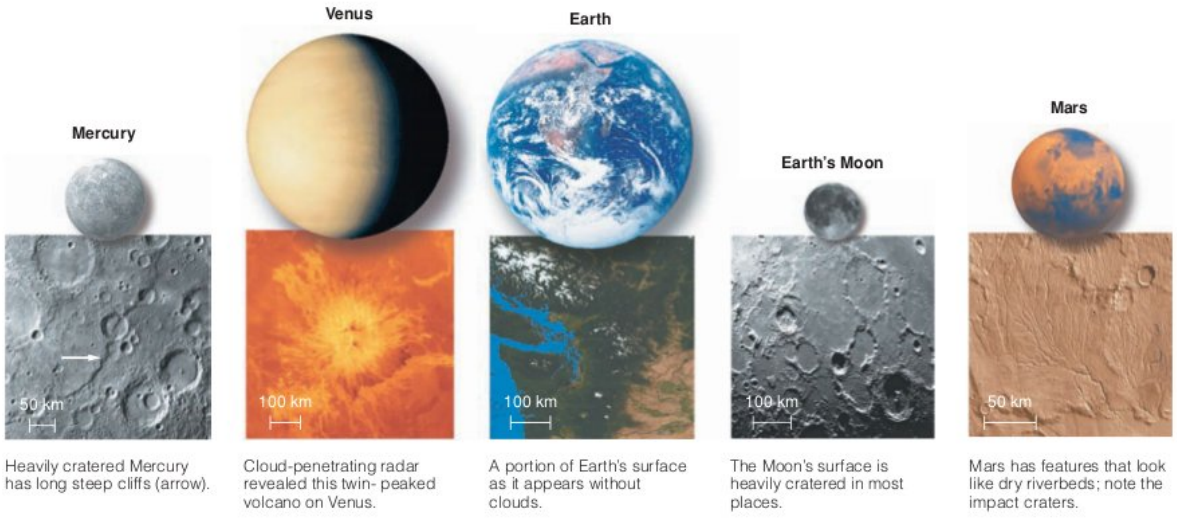

FIG. 85.— Size comparison of the four inner planets and the Moon that also gives an idea of the surface appearance of each world. Venus's clouds have been removed. From Bennett et al. (2007), their Fig 7.1.

The age of a rock is defined as the time since a rock has solidified. While there are a number of methods that can

be used to date rocks on the Earth (radioactive decay, for example; also used for the Moon), the primary means of determining the age of the surface of another planet is through the spatial density of its craters [36].

At its simplest, age can simply be determined by the number density of craters - radiometric dating of lunar rocks give a roughly linear relation between geological age and number of craters, up to 3 Gyr, and then an exponentially increasing relation. This is consistent with the early Solar System having many more planetismals, most of which eventually collided with a larger body (and formed craters) during the early bombardment era [37]. Age is derived from cratering history of Venus, Mars and Mercury by comparing these histories to that of the Moon's.

A large number of complications can arise. Cratering rates vary substantially both across and between bodies, due to synchronization (ex. with the Moon), gravitational focusing, and stray object density depending on distance from the Sun. Destruction of a crater by phenomena native to the planet (ex. rain, subduction) and saturation (when new impacts destroy the evidence of old impacts) can also affect results.

de Pater & Lissauer give the smooth plains of Mercury an age < 3.8 Gyr (since they are lightly cratered), and the cratered highlands > 4 Gyr, older than the early bombardment era. They give the age of Venus to be < 1 Gyr. They give the Martian southern hemisphere to be ~ 4.45 Gyr old, and the northern hemisphere to be 3 - 3.5 Gyr. Le Feuvre & Wieczorek give the average age of Venus to be 240 ± 13 Myr (compared to 650 - 750 Myr from Magellan cratering observations), and 3.73 Gyr for the Orientale and Caloris Basins on Mercury. Earth's ocean plates are < 200 Myr old, while some of its land plates are billions of years old (Wikipedia 2011b). From Carroll & Ostlie (2006), the oldest rocks ever found on Earth are 3.8 Gyr old, while 90% of the planet is < 600 Myr old.

The surfaces of the four terrestrial planets are as follows

- Mercury superficially resembles a Moon analogue, and is heavily cratered - the Caloris Basin (diameter 1550 km) is one of the largest impact basis in the solar system. Mercury's surface also has smooth plains relatively devoid of craters (likely formed volcanically), which therefore are younger than the late heavy bombardment (de Pater & Lissauer 2009). This is consistent with the fact that Mercury is larger, and closer to the Sun, making its cooling rate slower than that of the Moon. Mercury has a tenous atmosphere of captured solar wind and material liberated from the surface regolith (soil) (liberation can occur through solar wind stripping or micrometeorite impacts). Because of Mercury's tidal locking and lack of atmosphere, it is difficult for thermal energy from the lit regions of Mercury to migrate to the dark regions. Therefore, the day side temperature is ~ 800 K, while the night side is ~ 100 K, allowing certain shadowed regions of Mercury to play host to volatiles (such as water ice) that would otherwise melt (de Pater & Lissauer 2009).

- Venus is enshrouded by an atmosphere composed of about 97% CO_2, which acts as an extremely efficient insulator of infrared radiation - with an optical depth $\tau = 70$, the surface temperature of Venus, 740 K, is about three times what it would be if Venus had no surface. The pressure at the surface of Venus is 90 atm, indicating an atmosphere about two orders of magnitude more massive than that of Earth. This atmosphere likely comes from a combination of outgassing from volcanoes and deposited material from comets and meteorites. Due to these extreme surface conditions, space probes last only minutes once landed, and the most effective means of probing Venus's surface is through radar imaging.

  Venus's surface features, composed mainly of rock (and not soil), look much like those of Earth (lowland plains, upland hills, etc.), except that instead of a bimodal distribution of surface feature elevations like for Earth (one for oceans, one for continents), Venus has just one distribution, strongly indicating the absence of plate tectonics (de Pater & Lissauer 2009). Venus's surface is dotted with ~ 1000 volcanic structures, including shield volcanoes, weird pancake-like and spidery-looking domes, "coronae" and lava flows (de Pater & Lissauer 2009). It also has relatively few impact craters, which suggests the surface of Venus is only several hundred million years old, and that at some point massive geological activity resurfaced the entire planet. Changing atmospheric H_2SO_4 content observed since the late 1970s suggest major eruptions are still ongoing.

- Aside from the obvious presence of life and large bodies of liquid water, Earth's surface is relatively uncratered (due to erosion), and feature many active volcanoes. Carbon dioxide is readily dissolved into water, eventually becoming bound in carbonate rocks. This explains why the Earth's atmosphere is much more sparse than Venus's, and is mainly composed of nitrogen (equivalently, Venus's crust will contain insignificant amounts of carbonate rocks compared to the Earth). Earth appears unique among the inner planets in having plate tectonics: its lithosphere is fractured into crustal plates, which move ontop the convective upper mantle. As a result of the subduction of these plates into the mantle, and the creation of new plate from mantle, no sea plates are older than 200 Myr, while some regions of land are over 3 billion years old (Wikipedia 2011b).

- Mars's main global feature is the difference between the north and south hemispheres of the planet, known as a crustal dichotomy. The southern hemisphere is heavily cratered and elevated by several kilometres, while the northern hemisphere is smooth and lies several kilometres below. Because of the lower surface gravity Mars's features can be much taller than their equivalents on earth, and prominent on Mars's surface are enormous shield

[36] How craters have changed or been changed by geological processes can also be used to date certain geological features (ex. mountain ridges that crinkle a crater suggest the mountains formed after the crater did).

[37] Impact melts in rocks returned from the Moon by the Apollo missions suggest a second period of extreme bombardment, called the late heavy bombardment, around 3.8 - 3.9 Gyr.

volcanoes in the Tharsis region, and a large canyon system, Valles Marineris, that formed along with the Tharsis region volcanoes. These volcanoes were active until $\sim$ 2 Myr ago. It is most likely that the entire surface of Mars is more than 4 Gyr old, and the northern plains were resurfaced later. Two hypotheses for why the crustal dichotomy exists are large-scale mantle convection and a giant impact.

Mars' surface and atmosphere are both quite red, indicative of high concentrations of FeO_2. This is perhaps because the planet cooled faster than Earth, preventing Mars from undergoing the same degree of gravitational separation.

At 6 mbar and a temperature around 130 - 300 K, all water is either solid or vapour (indeed, a thin layer of frost forms on Mars's surface each night, to sublimate away at dawn). CO_2 in the atmosphere will also condense out in the winter, forming the dry ice polar caps (which, therefore, grow and recede every Martian year). Mars also has plenty of features that look like they were carved out by water; whether or not it still flows today is controversial. Regardless, Mars is very likely to have had large amounts of water in the past, and significant reservoirs of it exist today either frozen in the ice caps, or in the Martian permafrost in upper and mid-latitudes.

This section was from de Pater & Lissauer (2009) Ch 5.5.4.

4.9.1. *Compare the interiors of the inner planets.*

All four inner planets (Fig. 86) have stratified interiors (i.e. an iron core surrounded by a mantle and crust of lighter material) due to gravitational differentiation of heavy and light materials. Among the four, Mercury has the largest iron core, suggesting that much of the outer material was stripped away by a giant impact with an object a fifth its mass early in its history. Mars, being close to Jupiter, was sapped of planetismals due to the giant planet's gravitational cross-section, and is therefore smaller than the other terrestrial planets. This is insufficient to explain Mars' lack of iron, though (it has a lower density than Earth and Venus); the reason for this dearth of iron is not yet understood.

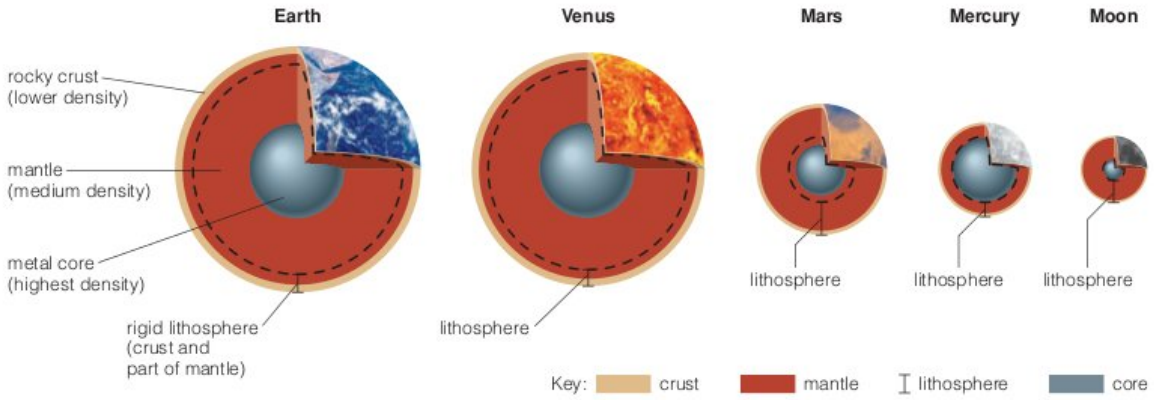

FIG. 86.— Comparison of the interiors of the inner planets and the Moon. From Bennett et al. (2007), their Fig 7.2.

4.9.2. *Why are there tectonic plates on Earth, but not other planets?*

The information in this section comes from Valencia et al. (2007).

The appearance of plate tectonics on terrestrial planets is related to planetary mass. In order for plates to move, convective stresses in the mantle have to be large enough to overcome lithospheric resistance to deformation, and the plate has to reach negative buoyancy to drive subduction. As planetary mass increases, the shear stress available to overcome resistance to plate motion increases while the plate thickness decreases, thereby enhancing plate weakness. A thinner crust also allows it to be negatively buoyant.

Earth-mass objects are at the limit at which plate tectonics will naturally occur, at which point other processes become important. On Venus, high surface temperatures may mean a layer of Venus's crust is highly deformable, reducing the chances for Venus's crust to fracture. On Earth, material permeated with water tends to be less tolerant of stresses. Objects with less than an Earth mass have crusts that can withstand mantle convection.

4.9.3. *Why has Mars lost its atmosphere while Venus has not?*

Atmospheres are replenished by outgassing (volcanic eruptions), evaporation/sublimation of solids and surface ejection of particles from cosmic ray and micrometeorite impacts (Bennett et al. 2007). Atmospheres are depeted by condensation, chemical reactions, solar wind stripping and thermal escape (where a small part of the tail of the atmosphere's Maxwell-Boltzmann distribution has the escape velocity) (Bennett et al. 2007).

Mars once was warm enough to have a convecting outer core, which, coupled with its rotation, produced a substantial magnetic field to ward off solar wind stripping. As Mars cooled at a much more substantial rate than Earth or Venus

(because it is smaller), the magnetosphere weakened, and outgassing was reduced. The atmosphere was therefore not being replenished, and what gases it did have were eventually stripped away by the solar wind, or, in the case of water, dissociated by UV radiation and then ejected via thermal escape. Some gas also condensed onto the surface due to the declining greenhouse effect from a thinning atmosphere.

Due to its slow rotation, Venus also does not have a strong magnetic field, and its upper atmosphere is bombarded by solar wind particles. It has, however, higher surface gravity than Mars, and its interior is still warm due to its larger mass. As a result, outgassing is still significant, and loss mechanisms from solar radiation are dampened (since escape velocity is higher). Because of high temperatures from the runaway greehouse effect, the atmosphere stays in gas form.

4.9.4. *Why does Venus have a much thicker atmosphere than Earth?*

Naively one would expect Venus and Earth to have the same atmospheres (with Venus being somewhat hotter). The proximity of Venus to the Sun, however, gave it a higher surface temperature, resulting in more water evaporation than on Earth (Carroll & Ostlie 2006). This lead to a runaway greenhouse effect whereby all liquid water on Venus' surface vapourized. Because water is needed for CO_2 sequestration into carbonate rocks, outgassed CO_2 was never reprocessed, and remained in the atmosphere (Bennett et al. 2007). Earth has 10^5 times more CO_2 dissolved in carbonate rock than in the atmosphere, explain why our atmosphere is much thinner than that of Venus (Bennett et al. 2007). The water, lighter than CO_2, floated to the top of the atmosphere, where it was photodissociated by solar UV radiation. The hydrogen (being lighter but having the same temperature as the rest of the atmosphere) was then peeled away through thermal escape, while the oxygen sequestered into rock (Bennett et al. 2007).

4.10. *Question 10*

QUESTION: What is the Eddington Luminosity? Give examples where it is important.

A system that emits copious amounts of radiation will naturally feel a radiation pressure - each electron will see the photons from its neighbours. Suppose the electron sees a flux S (we reserve F for force). Since each photon of energy E imparts momentum $p = E/c$, the pressure from this flux is S/c. If we assume all atoms are ionized, then the electron cross-section can be approximated as the Thomson scattering cross-section $\sigma_{\rm T} = 6.65 \times 10^{-25}$ cm^2. This gives us $F = \frac{\sigma_{\rm T}}{c} S$. If we now assume the system is spherically symmetric and optically thick and we are concerned mostly with the outer layers (photosphere if the system is optically thick) blowing off, we approximate S with a blackbody, and:

$$F_{\rm ph} = \frac{\sigma_{\rm T}}{c} \frac{L}{4\pi r^2}, \tag{159}$$

where r is the radius of the photosphere. Equating this to the force of gravity $F_{\rm g} = GMm_{\rm p}n_{\rm ne}/r^2$ (nuclei are electromagnetically coupled to electrons and tend to flow with them even if the system is ionized), where $n_{\rm ne}$ is the number of nucleons per electron (1 for hydrogen, 2 for carbon and oxygen):

$$L_{\rm Edd} = \frac{4n_{\rm ne}\pi GMm_p c}{\sigma_{\rm T}} \tag{160}$$

An equivalent method (Carroll & Ostlie 2006, pg. 341) of looking at this through bulk properties, rather than the force balance on a single electron. We recall $\frac{dP_{\rm rad}}{dr} = -\frac{\bar{\kappa}\rho}{c} F_{\rm rad}$ from Sec. 4.3. We again equate $F_{\rm rad} = L/4\pi r^2$. If the radiation pressure provides the exact pressure gradient to conteract gravity, then

$$-\frac{\bar{\kappa}\rho}{c} \frac{L}{4\pi r^2} = -G\frac{M\rho}{r^2} \tag{161}$$

From this we can derive

$$L_{\rm Edd} = \frac{4\pi Gc}{\bar{\kappa}} M \tag{162}$$

Note that $\bar{\kappa} = \sigma_T / n_{\rm ne} m_{\rm p}$, which simply indicates opacity is a cross-section per unit mass of material.

The Eddington Luminosity is an important feature of many astrophysical systems, since a process that produces a larger luminosity must then drive an outflow. AGB-tip stars, for example, may exceed the Eddington Luminosity and drive a superwind. Steady-state accretion onto objects (see below) is also limited by the Eddington Luminosity - if too much mass is dumped onto an object at once, the increase in the object's luminosity can drive an outflow, buoying the infalling material and driving accretion rates down to the Eddington accretion rate.

4.10.1. *What is the Eddington accretion rate?*

A related value is the Eddington accretion rate. Suppose accretion is approximated as spherical, and gravitational potential energy is turned into radiation with efficiency ϵ. The mass accretion rate corresponding to L_{Edd} is then:

$$\dot{m}_{Edd} = \frac{4n_{\text{ne}}\pi GMm_p c}{\sigma_T \epsilon \Delta\phi}. \tag{163}$$

$\Delta\phi$ is the amount of potential energy lost per unit mass accreted. In a stationary system (i.e. looks the same for all time) all the mass is moving closer to the accretor, and the net energy loss is equal to $\dot{m}GM/R$, where R is the radius of the accretor (visualize a long chain of particles moving radially toward the accretor to convince yourself of this), making $\Delta\phi = GM/R$. For disk accretion $\epsilon \sim 0.1$.

4.10.2. *Under what conditions does the Eddington luminosity not apply?*

Many astrophysical systems (in particular disk accretion) do not radiate isotropically. Moreover, this entire derivation implicitly assumes the system is stationary (since it is at equilibrium); astrophysical systems can easily exceed L_{Edd} temporarily, such as SN, where the photons drive a radiation pressure-dominated outflow.

4.11. *Question 11*

QUESTION: State the central temperature of the Sun.

The central temperature of the Sun is 1.57×10^7 K. The central pressure is 2.342×10^{16} N/m^2, and the central density is 1.527×10^5 kg/m^3 (Carroll & Ostlie 2006, pg. 351). (The hydrogen content is only 0.34, and the helium content is 0.64, due to 5 Gyr of nuclear processing.)

We can perform a very simple estimate of the central temperature, pressure and density of the Sun using the hydrostatic equilibrium equation and the ideal gas law. Assuming $dP/dm \to \Delta P/M$, $r \to R/2$, $m \to M/2$ (and $P_s = 0$) we obtain

$$P_c = \frac{2GM^2}{\pi R^4} \tag{164}$$

The average density of the star is $\bar{\rho} = \frac{3M}{4\pi R^3}$. The ideal gas law at the centre of the star is $P_c = \frac{\rho_c}{\mu m_H} k_B T_c$. Combining these things together, we obtain:

$$T_c = \frac{8\mu m_H GM}{3k_B R}\frac{\bar{\rho}}{\rho_c} \tag{165}$$

Assuming that $\rho_c \approx \bar{\rho}$, we obtain $T_c = 3 \times 10^7$ K, $P_c = 7 \times 10^{14}$ Pa and $\rho_c = 1.4 \times 10^3$ kg/m^3.

4.11.1. *What assumptions have you made, and how good are they?*

We assume that the average density is of the same order of magnitude as the central density, but this is a severe underestimate - more detailed models suggest by a factor of 10^2. As it turns out we are also underestimating the central pressure, which is much higher; this is because pressure is not a linear function inside a star. Our mean molecular weight was assumed to be 0.5; in fact it should be closer to 1, since the Sun's core is made of more He than H. We assume the Sun is not rotating - it in fact is, but at a rate about 0.5% its breakup velocity (a correction, known as the Chandrasekhar approximation, can be made by adding $\frac{2}{3}\rho\Omega^2 r^2$ to the hydrostatic equilibrium equation). The Sun has an outer convection zone, which ostensibly changes the central temperature, but since luminosity is governed by the much longer radiative diffusion time (through the radiative zone), and, by mass, convection only makes up a few percent of the Sun. Indeed, much better estimates of central values can be obtained just by assuming an appropriate polytrope and integrating.

Fig. 87 shows the temperature, pressure, density, mass, luminosity and inverse gradient structure of the Sun, as computed by modern models.

4.11.2. *Why are the centres of stars so hot?*

Because they are massive gravitating bodies, by the equation of hydrostatic balance $\frac{dP}{dr} = -\frac{Gm\rho}{r^2}$ there needs to be a pressure gradient within the star. This results in an extremely high central pressure, and since pressure in a main sequence star is provided by non-degenerate gas pressure, and per the ideal gas law, $P = \frac{\rho}{\mu m_H} k_B T$, a high thermal energy density $f\rho c_V T$ is needed. Stars radiate as blackbodies, and so over time the thermal energy needed to hold the star up is lost. To replenish it, a star's central temperature must be high enough to light nuclear fusion, which generates enough energy to maintain a high thermal energy density at the star's core. This, in essense, is why the centre of stars are so hot.

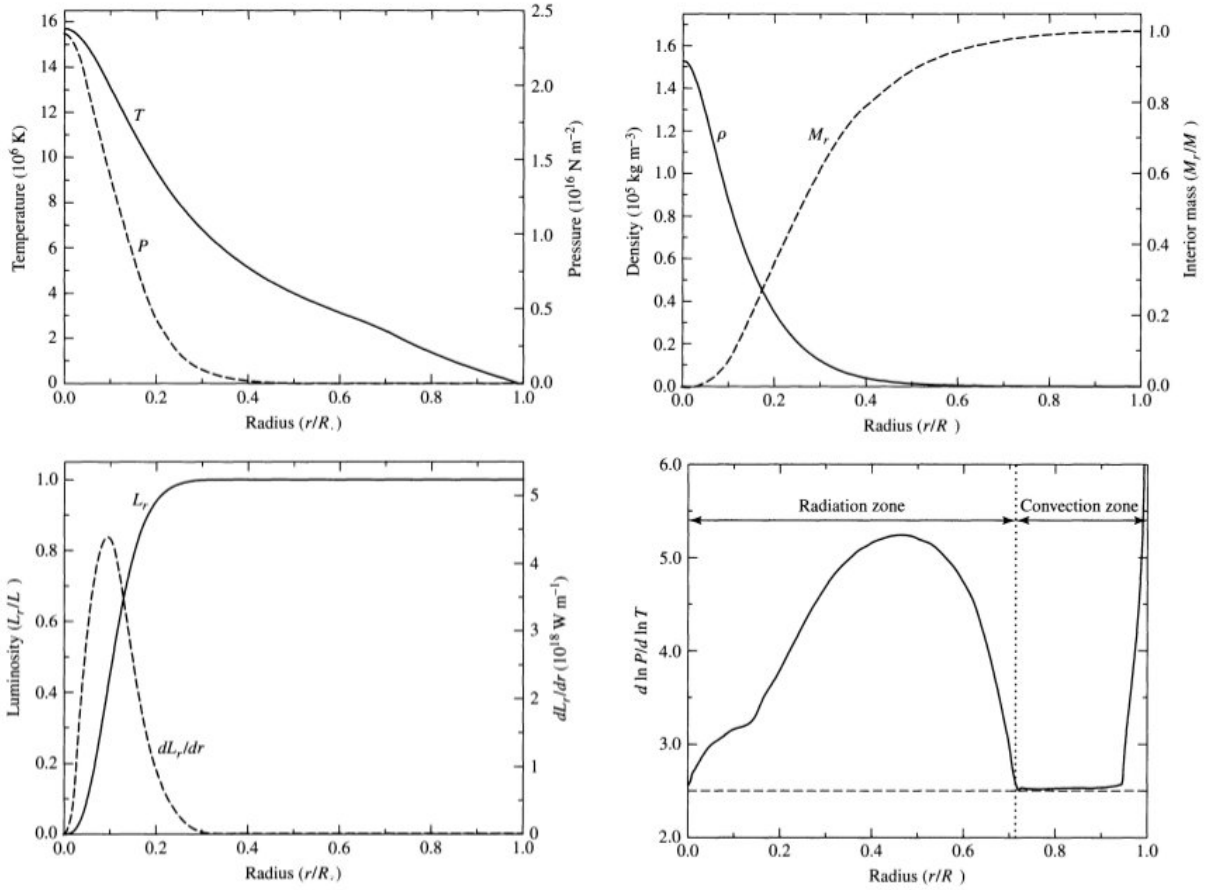

FIG. 87.— Various properties of the Sun as a function of normalized radius. dL_r/dr may seem strange at first, but recall this is a radial derivative of something that depends on volume. Also note that the gradient plotted is $d \ln P/d \ln T$, rather than the more common reverse. From Carroll & Ostlie (2006), their Figs 11.4 - 11.7.

4.12. *Question 12*

QUESTION: Which have higher central pressure, high-mass or low-mass main-sequence stars? Roughly, what is their mass-radius relation? Derive this.

Let us combine Array 128 with the ideal gas law, which says $P \propto \rho T \propto \frac{M}{R^3} T$. Combining $P \propto \frac{M^2}{R^4}$ with $P \propto \rho T \propto \frac{M}{R^3} T$ nets us

$$T \propto \frac{M}{R}. \tag{166}$$

Combining this with $T^4 \propto \frac{M}{R^4} L$ gives us

$$L \propto M^3 \tag{167}$$

Note that the empirical relation is $L \propto M^{3.2}$, reasonably close to this value (Kippenhahn & Weigert 1994, pg. 209). Now, recall from 4.21 that the p-p chain has energy generation per unit mass $\epsilon \propto T_c^4$ and the CNO cycle has energy generation per unit mass $\epsilon \propto T_c^{20}$, in the range of 10^7 K, which is the approximate central temperature of our Sun. $L \propto M\epsilon \propto M^3$ [38], which gives us $\epsilon \propto M^2$. For both the p-p chain ($T_c \propto M^{1/2}$) and CNO cycle ($T_c \propto M^{1/10}$), mass is a very shallow function of temperature.

The main sequence contains stars from the hydrogen burning limit (0.08 $M_\odot$) to around 100 $M_\odot$, and switches from being p-p chain to CNO cycle dominated at around 1 $M_\odot$ (Kippenhahn & Weigert 1994, pg. 212). Central temperature, then, changes by less than one order of magnitude, according to our scaling relations. We assume it is a constant.

[38] $L \propto M\epsilon$ suggests that the amount of material currently undergoing fusion is linearly proportional to the total mass of the system.

At this point, there is a bit of an ambiguity: does temperature in Eqn. 166 refer to central or surface temperature? Marten's solution is to assume that $T_c \propto \frac{M}{R}$, from which the mass-radius relation follows immediately:

$$M \propto R, \tag{168}$$

and from $P_c \propto \frac{M}{R^3} T_c$ we obtain

$$P_c \propto M^{-2}. \tag{169}$$

(The same result is obtained if we use $P \propto \frac{M^2}{R^4}$.)

Alternatively, we can assume Eqn. 166 refers to exterior temperature T. This, however, is wrong, since luminosity is related to the central temperature, while the photosphere temperature has to do with the "response" of the star to having a particular central temperature, density and fusion rate.

We obtain that radius either scales linearly or quite shallowly with mass, and pressure scales with the inverse of mass (more massive stars have less central pressure). The latter point should not be surprising - 100 $M_\odot$ stars like Eta Carinae can barely hold themselves together, while stars like the Sun have no such problem.

More advanced modelling techniques net the general formula

$$R \propto M^{\frac{1-n}{3-n}}, \tag{170}$$

noting that $\gamma = \frac{C_P}{C_V} = 1 + \frac{1}{n}$. While generally useful (ex. for a non-relativistic WD), this expression fails for the Sun, which can be approximated by an $n = 3$ polytrope. (In the Sun, if we assume $\kappa l/m$ is a constant (not a bad assumption), then radiation pressure ($P_{\rm rad} \propto T^4$) divided by gas pressure will also be a constant. As a result, $P \propto T^{1/4}$. Combining this with the ideal gas equation gives us $P \propto \rho^{4/3}$, an $n = 3$ polytrope.)

Assuming a stationary, radiation-dominated star with a uniform composition, and power-law representations for the density, power generated by fusion and opacity (Kippenhahn & Weigert (1994), Eqn. 20.9), we find that $R/R' = (M/M')^{z_1}$, where primes indicate a different star, and z_1 ranges from 0.4 to 0.8, depending on whether the p-p chain or CNO cycle are used. We also find $P/P' = \frac{M^2 R'^4}{M'^2 R^4}$ (since the solution is a homology it does not matter if we are describing central or average pressure), so very roughly $P \propto M^{-0.4}$, again giving us that pressure should decrease with mass. Homology relations also give us $T_c \propto M^{0.4}$ while $\rho_c \propto M^{-0.8}$ (the last can be derived from the ideal gas law and the T_c and P scalings).

Detailed numerical calculations of main sequence models allow us to obtain $R \propto M^{0.68}$ and $L \propto M^{3.2}$; combining this with $L \propto R^2 T^4$, where T is surface temperature, we can obtain the slope of the main sequence on a colour-magnitude diagram:

$$L \propto T^7 \tag{171}$$

4.12.1. *Why do we use the radiative diffusion equation rather than a convection term?*

The photon diffusion rate, which ultimately sets the luminosity (thermal energy divided by the diffusion timescale gives the luminosity of the star), is determined by the radiative zone(s) of the star. Even fully convective stars will have radiative zones in their very outermost layers, and because convective energy transport is so efficient, the timescale for photon escape from the star is determined by this thin radiative zone. If we also assume that these thin radiative envelopes scale in the same manner as their thicker counterparts, the scaling relation used in the calculation above is still valid.

4.12.2. *Where does this approximation fail?*

We have assumed above that the central temperature of the star is roughly constant throughout the main sequence due to the high temperature dependence of nuclear fusion. The homology relation ($T_c \propto M^{0.4}$, $\rho_c \propto M^{-0.8}$) is consistent with this. Detailed modelling, however, show that at around 1 $M_\odot$ the density stays roughly constant while temperature changes half an order of magnitude. This is related to the fact that at about 1 $M_\odot$ the star is almost completely radiative (Kippenhahn & Weigert, Fig. 22.7), while in the flat regions of the curve above and below, the star is convective to a significant degree. This region is also host to two other major changes (THOUGH IT'S NOT OBVIOUS IF THEY ARE A CAUSE OR EFFECT OF THE CONSTANT DENSITY REGIME): at the upper tip of the constant density regime, temperatures are high enough for the CNO cycle to begin dominating the pp chain as the star's source of energy, and degeneracy effects become important for stars below 0.5 $M_\odot$.

4.12.3. *What about extreme masses on the main sequence?*

Attempts at stellar modelling for very low masses is mired by having to treat extensive convection, and the opacity of complex molecules which form in cool stellar atmospheres. Moreover, the time for the star to reach equilibrium burning can be very long.

For very massive stars, the adiabatic gradient $\nabla_{\rm ad} \to 1/4$, which is quite shallow, and therefore the star is largely convective. The adiabatic structure requires constant specific entropy s across the star. Taking the specific energy

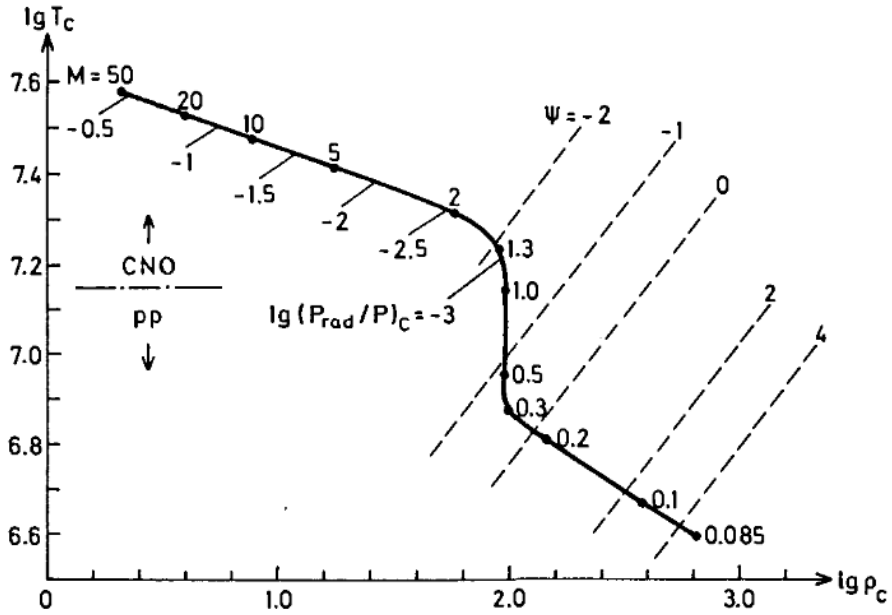

FIG. 88.— A $\rho - T$ diagram of central temperature and density for the main sequence. Different masses are labelled above the main sequence. Labels below the main sequence indicate the relative contribution of radiation to the pressure. The p-p/CNO horizontal line indicates which nuclear process should dominate. Dashed lines indicate lines of constant degeneracy. Note the two flat regions (qualtiatively similar to what is obtained by homology arguments) and the steep region around 1 $M_\odot$. How central pressure behaves can be inferred from this trend and the ideal gas law. From Kippenhahn & Weigert (1994), their Fig. 22.5.

density $u = aT^4/\rho$ and pressure $P = aT^4/3$, and using the first law of thermodynamics ($ds = \frac{1}{T}\left(du - \frac{P}{\rho^2}d\rho\right)$) we find $s = \frac{4aT^3}{3\rho}$, which nets us $\rho \propto T^3$. Since $P \propto T^4$, we obtain an $n = 3$ polytrope for a radiation-dominated star.

4.12.4. *What is the scaling relation when the pressure support of the star itself is radiation-dominated?*

For stars near the Eddington Luminosity, $P \propto T^4$ (from $P_{\rm rad} = \frac{1}{3}aT^4$). Combining this with $P \propto M^2/R^4$, we obtain $T \propto M^{1/2}/R$. Then combining this with $T^4 \propto \frac{M}{R^4}L$ gives us

$$L \propto M \tag{172}$$

We continue to assume T_c is approximately constant, for the same reasons as listed above. This immediately gives us

$$P = \text{constant} \tag{173}$$

and (from $T \propto M^{1/2}/R$)

$$M \propto R^2 \tag{174}$$

4.13. *Question 13*

QUESTION: Derive the scaling of luminosity with mass for a (mostly) radiative star. Do you need to know the source of energy for the star to derive this scaling?

We have, to good order, already completed this question in the last question, Sec. 4.12. Combining $P \propto \frac{M^2}{R^4}$ with $P \propto \rho T \propto \frac{M}{R^3}T$ nets us $T \propto \frac{M}{R}$. Combining this with $T^4 \propto \frac{M}{R^4}L$ gives us $L \propto M^3$. We do not need to know the source of energy for the star (we do need to know it to determine the mass-radius relation); this is because the luminosity is set not by the rate of nuclear fusion or gravitational contraction, but by the rate at which photons can diffuse out of the star. This sets an energy loss rate that the star has to compensate with nuclear reactions.

When radiation pressure begins to dominate, $L \propto M$, but by that point the star has become almost entirely convective.

4.13.1. *Does more detailed modelling give better results?*

This derivation comes from Hansen et al. (2004).

TABLE 4
Power-law Indices

Energy Generation Mode	λ	ν	Opacity Mode	n	s	Pressure Mode	χ_ρ	χ_T
pp-chain	1	4	Thomson Scattering	0	0	Ideal Gas	1	1
CNO cycle	1	20	Kramers' Law	1	3.5	Radiation	0	4
Triple-α Process	2	40						

We may once again derive scaling relations from Array 127 (though we keep opacity as a variable in the radiative transfer equation). Now, however, we also generate power laws (see Table 4) for energy generation, opacity and pressure (equation of state): $\epsilon \propto \rho^\lambda T^\nu$, $\kappa \propto \rho^n T^{-s}$, and $P \propto \rho^{\chi_\rho} T^{\chi_T}$. We can then write total logarithmic differentials for the hydrostatic equilibrium equation and the various power laws; i.e. $d\ln\epsilon = \lambda d\ln\rho + \nu d\ln T$. Lastly, we assume that R, ρ, T and L all have power law scaling relations with mass M. If we then write the scaling relations for $\frac{\partial m}{\partial r} = 4\pi r^2\rho$, $\frac{\partial P}{\partial r} = -\frac{Gm\rho}{r^2}$, $\frac{\partial l}{\partial r} = 4\pi r^2\rho\epsilon_{\rm nuc}$ and $l = -\frac{64\pi^2 acT^3 r^4}{3\kappa}\frac{\partial T}{\partial m}$, and insert the power laws in, we obtain the matrix equation 1.74 in Hansen et al.. The result can be found on pg. 27 of Hansen et al.. If convection is wholly used instead of radiation as the means of energy transport, the fact that $T \propto \rho^{\gamma-1}$ replaces the radiative diffusion equation.

The results: for electron scattering opacity, CNO cycle fusion and the ideal gas law (all of which are representative of more massive main sequence stars), $R \propto M^{0.78}$ and $L \propto M^3$. For pp chain, Kramers opacity, and the ideal gas law along with radiation, $L \propto M^{5.5}$, but this is not empirically reproduced. Using convection rather than radiation gives $L \propto M^{8.3}$, which is even more untenable. Evidently, something (perhaps increasing degeneracy pressure in the core, or a better treatment of convection) must be added.

Detailed modelling and empirical results can be found in Fig. 89. Over the entire range plotted, $M \propto L^{3.2}$, though the exponent is highest near 1 $M_\odot$, at 3.88.

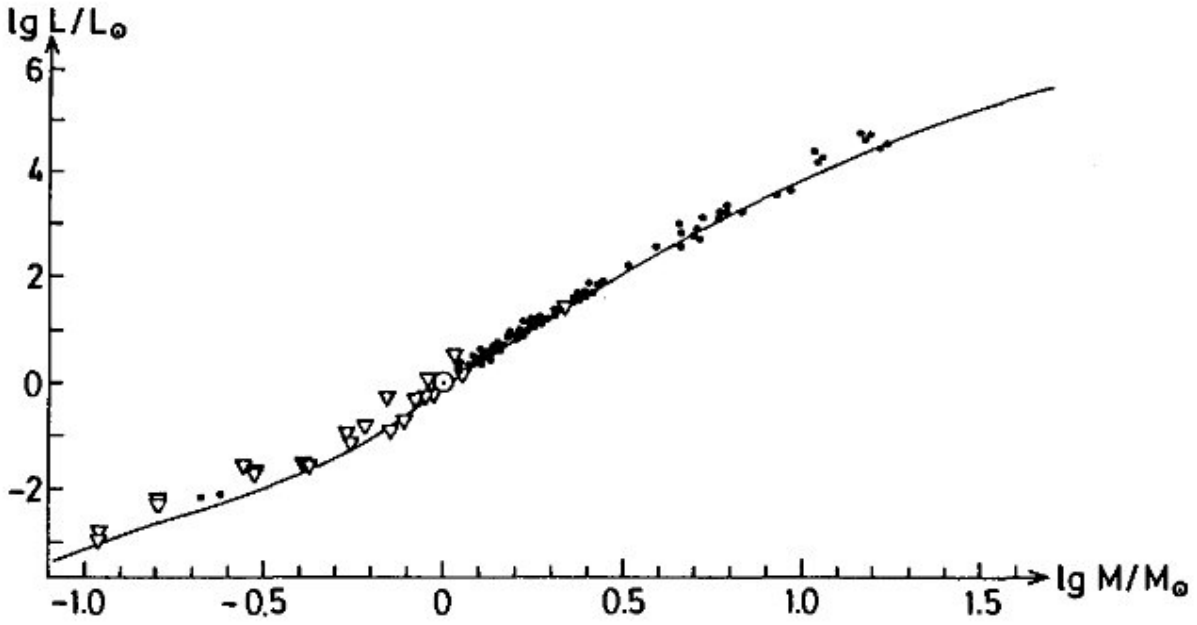

Fig. 89.— The mass-luminosity relationship for stars on the main sequence. The solid line gives the theoretical mass-luminosity relationship, while the points give observed masses derived from detatched (points) and visual (triangles) binary systems. The decreasing slope at higher masses is due to increasing radiation pressure. From Kippenhahn & Weigert (1994), their Fig 22.3.

4.14. *Question 14*

QUESTION: What source of pressure prevents gravitational collapse in a normal star, a white dwarf, and a neutron star? Why, physically, is there a maximum mass for the latter two? And why can we calculate this easily for a white dwarf, but not for a neutron star?

The source of pressure preventing gravitational collapse in a normal star is gas pressure, as given by the ideal gas law, $P = \frac{\rho}{\mu m_H} k_B T$. The source of pressure for a white dwarf is electron degeneracy pressure, and the source of pressure for a neutron star is neutron degeneracy pressure. Degeneracy pressure is covered in Sec. 5.12.

Degeneracy pressure is physically generated by forcing electrons into increased momentum states so as to satisfy the

Pauli exclusion principle - the quantization of phase space requires that electrons sharing the "same" position must have different momenta. As such, an increased compression generates a higher distribution of momentum states (i.e. the top of the Fermi sea is higher, i.e. E_F is larger). Eventually, E_F becomes infinite, but even then there are still a large percentage of particles at low energies, providing little momenta, and so the maximum pressure degeneracy can provide is finite. If the stellar structure requires a higher pressure, which would be the case past a critical mass, then collapse sets in. The critical mass is known as the Chandrasekhar Mass, $M_{\rm ch}$, in a WD.

$M_{\rm ch}$ is easy to calculate for a WD (an estimate is below). If neutrons behaved like electrons, the equivalent mass for neutron stars would also be easy to calculate, except that the radius of a neutron star is on order R_s, the Schwarzschild radius, and therefore a self-consistent general relativistic solution would be needed (Townsend 1997). Materials at neutron star densities, however, are poorly understood, as a complete theoretical description of the behaviour of a sea of free neutrons interacting via the strong nuclear force in the presence of electrons and protons does not yet exist, and there is little experimental data on the matter (Carroll & Ostlie 2006, pg. 581). Moreover, more massive baryon resonances, quark "bags", and bose condensates of mesons might also exist at such densities (Camenzind 2007).

4.14.1. *Estimate the Chandrasekhar Mass for a WD. Estimate the equivalent for an NS.*

The momentum state density (i.e. number of states from p to $p+dp$ in a volume dV) of a fermionic gas is $f(p) = \frac{8\pi p^2}{h^3}$, and in a degenerate gas the occupation number for all states less than momentum p_F is 1 (and the occupation number for all states above is zero). The integral of $f(p)$ from 0 to p_F is equal to the number density of fermions. As a result, $n = \frac{8\pi}{3h^3}p_F^3$, from which we can (through $E_F = p_F/2m$ for non-relativistic and $E_F = p_F c$ for relativistic) obtain E_F, which is

$$E_F = \frac{\hbar^2}{2m_e}(3\pi^2 n)^{2/3} \tag{175}$$

for a non-relativistic gas and

$$E_F = (3\pi^2 n)^{1/3}\hbar c \tag{176}$$

for a relativistig gas. We then can perform a similar calculation in both cases as we did in Sec. 3.13.2. Using $\xi E_i + E_g = 4\pi R^3 P_{\rm ext}$, in the non-relativistic case, we have

$$-f_g\frac{GM^2}{R} + f_F\frac{\hbar^2}{2m_e\mu^{5/3}m_p^{5/3}}(3\pi^2)^{2/3}\frac{M^{5/3}}{R^2} = 4\pi R^3 P_{\rm ext}. \tag{177}$$

In this case for $R \to 0$, $P_{\rm ext} \to \infty$, meaning the system is stable against completely collapse. (There is a turning point in $P_{\rm ext}(R)$, past which the object will collapse to the R of the turning point, but (again recalling Sec. 3.13.2) $P_{\rm ext} = 0$ is inside this turning point.) Assuming $P_{\rm ext} = 0$ gives us an $R \propto M^{-1/3}$ relationship.

In the relativistic case,

$$-f_g\frac{GM^2}{R} + f_{FR}\hbar c(3\pi^2)^{1/3}\frac{(M/\mu m_p)^{4/3}}{R} = 4\pi R^3 P_{\rm ext}. \tag{178}$$

This system is critically stable, and there is only one mass which corresponds to $P_{\rm ext} = 0$. Solving for this, we obtain:

$$M_{\rm ch} = (3\pi^2)^{1/2}\left(\frac{f_{FR}}{f_g}\right)^{3/2}\left(\frac{\hbar c}{G}\right)^{3/2}(\mu m_p)^{-2} \tag{179}$$

f_{FR}/f_g is of order unity, and if we ignore it, we obtain 2.5 $M_\odot$. Due to the M^2 dependence of gravitational potential energy vs. the $M^{4/3}$ dependence of Fermi energy, $M > M_{\rm ch}$ stars always have negative external pressures, making them unstable to collapse, while $M < M_{\rm ch}$ stars will expand until becoming less relativistic. We can estimate $R_{\rm ch}$ by taking the mass-radius relation for non-relativistic WD, and plugging $M_{\rm ch}$ in.

Eqn. 179 is independent of electron mass, and so applies equally well to neutron star, except for neutron stars $\mu \approx 1$ - the NS equivalent to the Chandrasekhar mass is therefore about a factor of four more massive. This turns out not to be the case for two reasons: general relativity increases the effective gravitational potential energy of the NS, and the neutron star equation of state is much more complicated than the white dwarf equation of state.

Detailed calculations give 1.4 $M_\odot$ for WDs, and 2.2 $M_\odot$ for NSs (Carroll & Ostlie 2006, pg. 583).

4.14.2. *Why is proton degeneracy unimportant in a WD?*

While Coulombic corrections between proton-electron pairs do affect the WD equation of state, to first order protons do not matter in a WD except as a source of gravity. This is because degeneracy pressure is inversely proportional to mass. Since electrons and protons have the same density at any given region in the neutron star (this is mediated by gravity), proton degeneracy pressure becomes unimportant. Proton degeneracy pressure *can* become important at neutron star densities, but by this point the star is unstable to electron capture onto protons.

4.14.3. *What is the structure of a neutron star?*

This information is from (Camenzind 2007, pg. 188) and (Carroll & Ostlie 2006, pg. 581 - 582)
From the outside in, a neutron star is composed of

- The cm-thick atmosphere.
- The outer crust, essentially relativistic electron degeneracy material.
- The inner crust, which extends from 4×10^14 kg/m^3 (when free neutrons begin to appear) to 2×10^{17} kg/m^3 (when nuclei dissolve), which contains free neutrons, relativistic electrons and neutron-rich nuclei. The material is superconducting. Neutron degeneracy dominates electron degeneracy past 4×10^{15} kg/m^3.
- The core, composed of a superconducting zero-viscosity fluid of free neutrons, protons, electrons and a possible assortment of hyperons (more massive baryon resonances), quarks and mesons.

See Fig. 90.

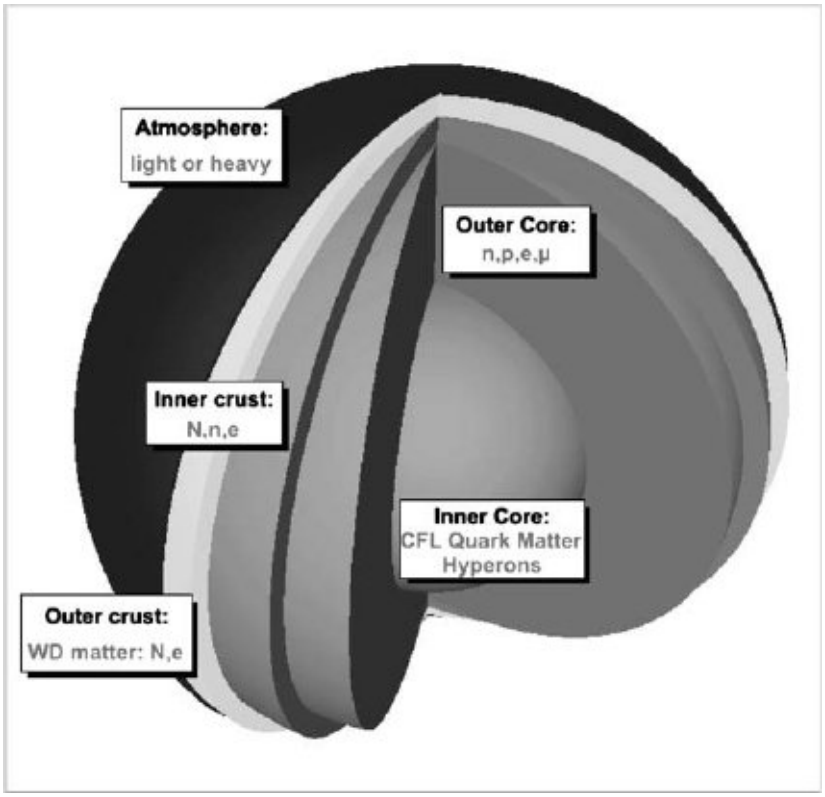

FIG. 90.— A schematic of the interior of a neutron star. From Camenzind (2007), their Fig 6.1.

4.14.4. *How can rotational support change the maximum mass?*

Since rotational energy can help support a star, a fast-spinning WD or NS could theoretically exceed their hydrostatic (non-rotating) mass limits.

A rigidly rotating WD can have a mass up to 1.48 $M_\odot$, but a differentially rotating WD (with an artificially constructed rotation profile) can achieve masses in excess of 4 $M_\odot$ (Yoon & Langer 2004). WDs this massive cannot be created by canonical stellar evolution, but they could possibly be spun up while accreting. The upper limit to a differentially rotating WD's mass, ~ 2 $M_\odot$, is given by the maximum amount of material it could acquire by mass accretion from a companion (Yoon & Langer 2004).

2.9 $M_\odot$ NSs are possible with significant rotational support (Carroll & Ostlie 2006, pg. 583).

4.14.5. *Could you heat a WD or NS to provide thermal support?*

Of course! Heat a WD enough, and in such a way as to not ignite nuclear burning, and you would have a hot cloud of H, He, C and O (and possibly Ne). Heat an NS enough, and the additional thermal pressure support will lift the NS degeneracy, and you would possibly get a cloud of hot deuterium.

4.15. *Question 15*

QUESTION: Sketch the SED of an O, A, G, M, and T star. Give defining spectral characteristics, and describe physically.

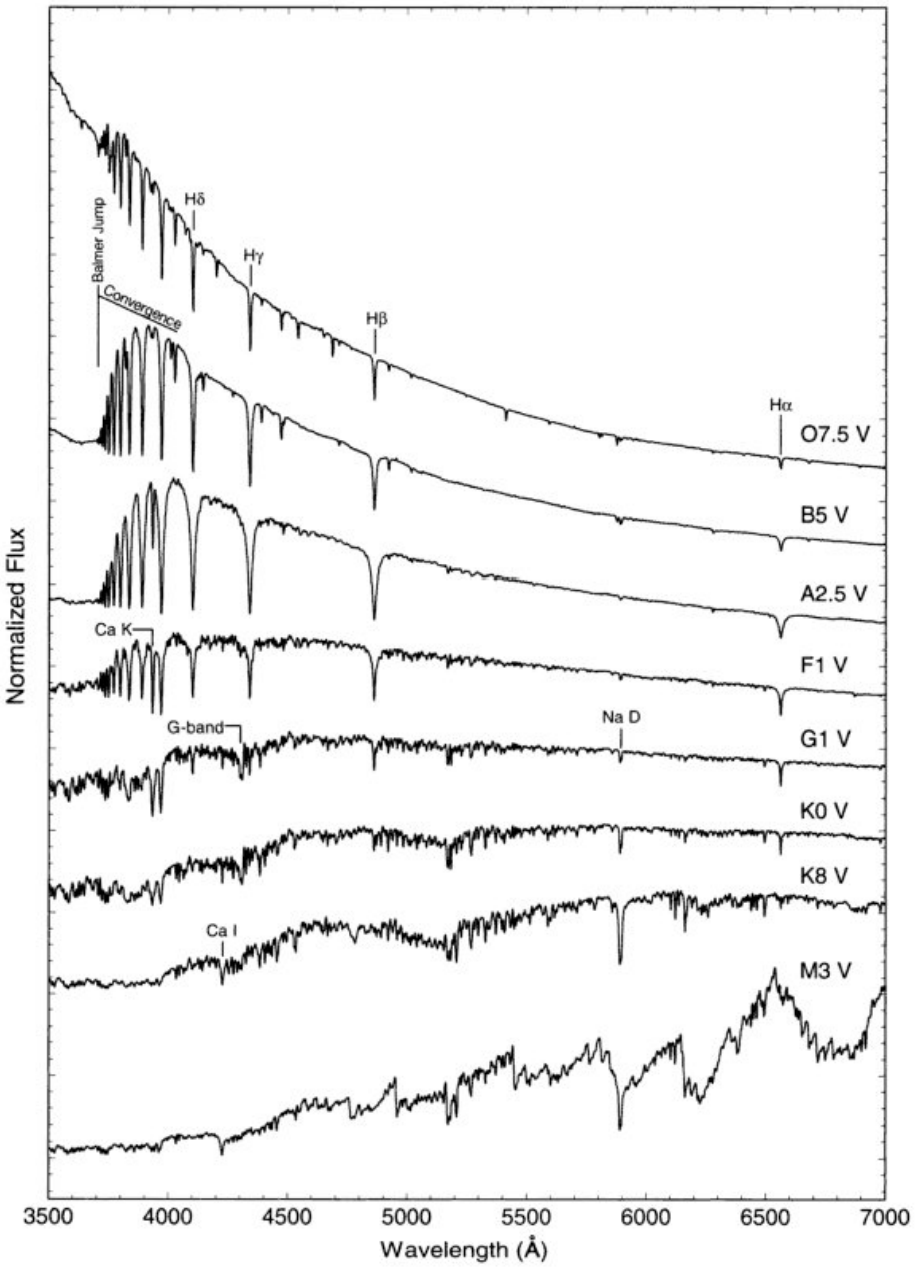

FIG. 91.— Representative spectra of main sequence, with prominent line features labelled. The spectra have been normalized at a common wavelength, and stacked on top of each other for the sake of presentation. From Gray & Corbally (2009), their Fig 2.1.

The information for this question comes from Carroll & Ostlie (2006), Ch. 8.1, and Gray & Corbally (2009), Ch. 2.

Figs. 91, 92 and 93 show the spectrum of main sequence stars from O to M from 1100 Åto 2 μm. Most prominent in all early-type (more massive than the Sun) spectra are the Balmer lines (Hα, 6562 Å; Hβ, 4861 Å, Hγ, 4340 Åand Hδ, 410.2 Å; ionization from the first excited state occurs past 3651 Å, the location of the "Balmer jump"), which increase in intensity from O to A2, and then decrease for later types, ceasing to be prominent by K type. He I follows a similar pattern, peaking in B2. He II lines are already decreasing for O stars, indicating their peak is in a range not normally reached by stars. Ca II lines (the strongest being the H and K lines at 3968 and 3933 Å) become stronger until they

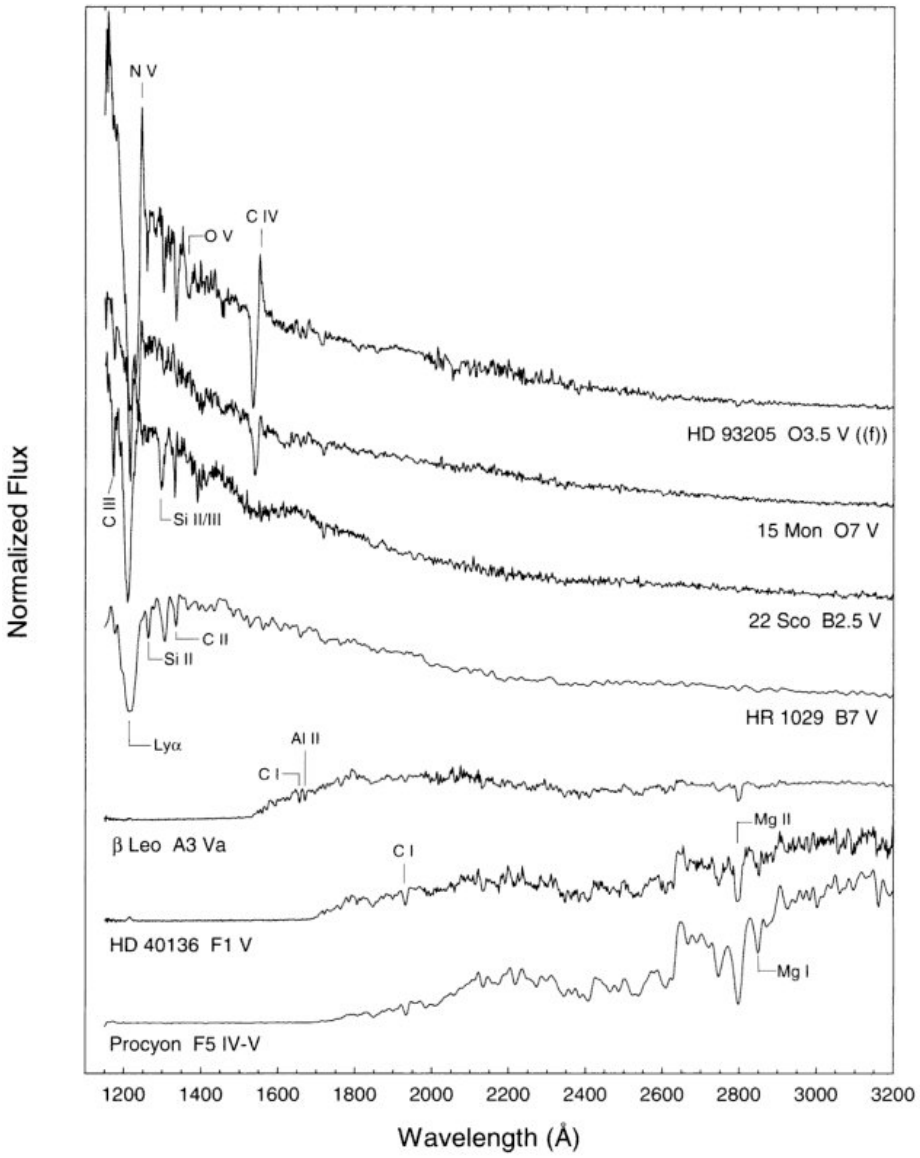

FIG. 92.— Representative spectra of main sequence in the UV, with prominent line features labelled. The spectra have been normalized at a common wavelength, and stacked on top of each other for the sake of presentation. The spectra of 15 Mon and HD 93205 have been de-reddened. From Gray & Corbally (2009), their Fig 2.6.

peak at K0. Neutral metal lines, particularly from Fe I and Cr I, begin to show in F, and dominate the spectrum of K stars before diminishing in M stars. Molecular absorption bands appear first in F stars (most noticably CH, also known as the "G-band") and become strong for M stars (mostly TiO and VO). These features are summarized in Table 2.

T stars (Fig. 94) contain strong CH_4 absorption bands in the NIR (this distinguishes T stars from L stars), complementing strong H_2O and NH_3 bands. Collisionally induced H_2 absorption (caused by collisionally induced quadrupolar moments in H_2) also factor. The only prominent atomic absorption is K I.

L stars, not shown, are intermediate sub-stellar objects between M and T stars, and have blackbodies that peak in the infrared and are littered with strong molecular absorption bands. S and C class objects are evolved stars that overlap with traditional K and M stars. C ("carbon") stars have an abundance of carbon-rich molecules in their atmospheres, while S types show ZrO lines rather than TiO lines, as is standard for M stars.

In the UV, early-type stars contain highly ionized metal absorption lines, some with P Cygni features indicating an outflow, while for later type stars than our Sun, photospheric emission is minimal in the UV and chromospheric and coronal emission lines dominate instead (their strengths are set by rotation and magnetic fields). The dearth of UV emission for late-type stars is a combination of continuum metal absorption and a cooler blackbody. In the IR, early type stars feature the Brackett series (H-transitions where $n_f = 4$), which disappears by G0 stars. Late type IR spectra are dominated first by metal lines (ex. Mg and Si) and then by molecular lines.

To summarize:

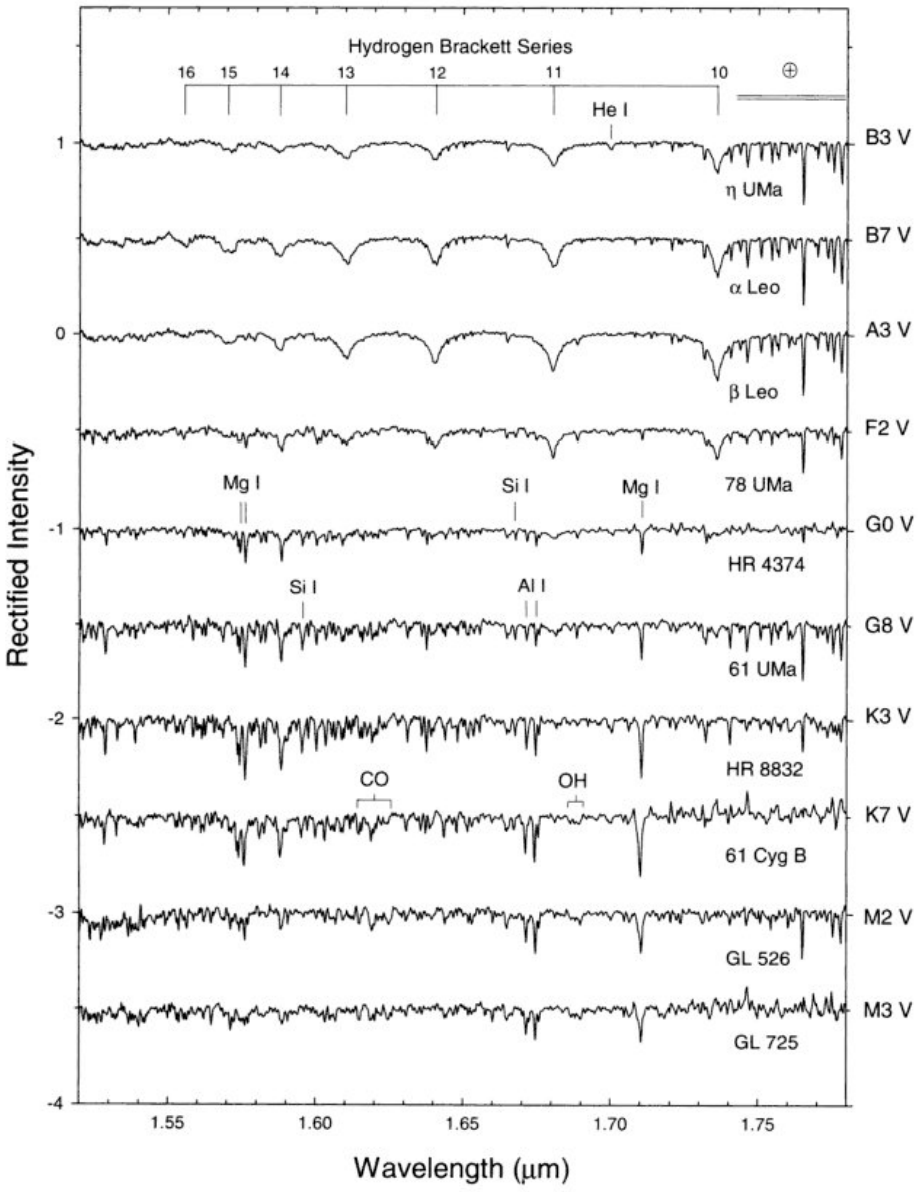

FIG. 93.— Representative spectra of main sequence in the IR, with prominent line features labelled. The spectra have been normalized at a common wavelength, and stacked on top of each other for the sake of presentation. From Gray & Corbally (2009), their Fig 2.7.

1. **O stars**: highly ionized metals, He II absorption, weak Balmer lines.

2. **A stars**: strongest Balmer lines, He I no longer visible, weak neutral metal lines and Ca II H and K.

3. **G stars**: strong Ca II doublet, stronger neutral metal lines, G-band.

4. **M stars**: spectra dominated by molecular absorption bands, especially TiO. Neutral metal absorption lines remain strong.

5. **T stars**: molecular absorption, mainly in CH_4 and H_2O.

As noted earlier, absorption lines are caused by wavelength-dependent opacity in the stellar photosphere (or, equivalently, scattering and absorption of a blackbody by colder, tenuous material). The amount of resonant scattering and absorption from a certain species i, j, where i is the degree of ionization and j is the degree of excitation, is dependent on amount of such species in the stellar photosphere, $N_{i,j}$.

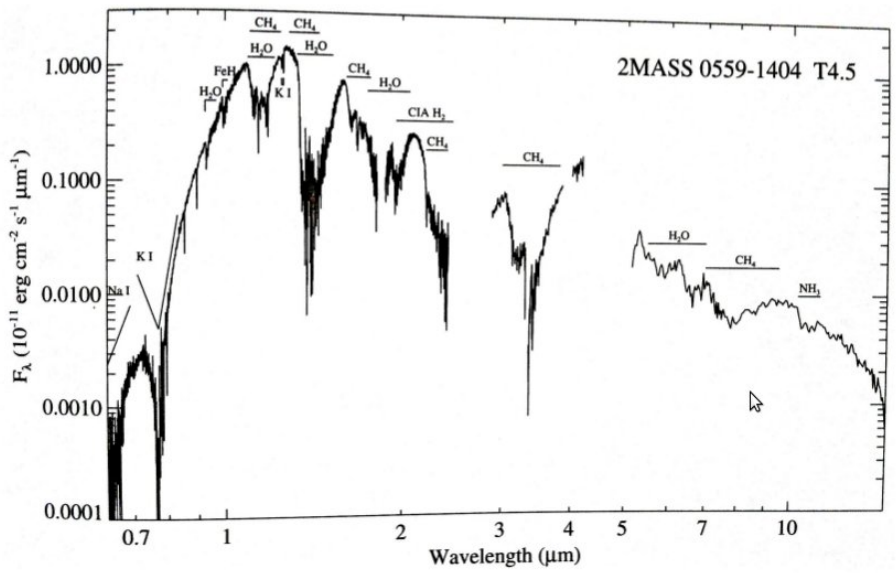

FIG. 94.— Observed spectrum of a T star from 2Mass. From Gray & Corbally (2009), Fig. number unknown.

For a system in local thermodynamic equilibrium, the Boltzmann equation gives the probability of being in state j as

$$p_j = \frac{g_j \exp^{-E_j/k_B T}}{Z} = \frac{g_j \exp^{-E_j/k_B T}}{\sum_j g_j \exp^{-E_j/k_B T}} \tag{180}$$

where Z is the partition function and g_j is the degeneracy of a particular energy state. Meanwhile, the Saha Equation, which can be derived from the grand partition function, gives

$$\frac{N_{i+1}}{N_i} = \frac{2Z_{i+1}}{n_e Z_i} \left(\frac{2\pi m_e k_B T}{h^2} \right)^{3/2} e^{-\chi_i/k_B T} \tag{181}$$

where i indicates the degree of ionization, χ_i is the ionization energy from degree i to degree $i+1$. The strength of a line is given by $N_{i,j}/N_{\rm total}$, which is given by a combination of the Boltzmann and Saha equations. Hydrogen is the easiest to calculate, since it only has one ionization state. The strength of Balmer absorption, for example, is determined by

$$\frac{N_{I,2}}{N_{\rm total}} = \frac{g_2 \exp^{-E_2/k_B T}}{Z_I} \frac{1}{1 + N_{II}/N_I} \tag{182}$$

where $N_{II}/N_I = \frac{2Z_{II}}{n_e Z_I} \left(\frac{2\pi m_e k_B T}{h^2} \right)^{3/2} e^{-\chi_I/k_B T}$. Eqn. 182 contains two competing temperature dependencies: $\exp^{-E_2/k_B T}/Z_I$ increases with temperature (at least, up to a certain point - atoms have infinite numbers of excited states, and at infinite temperature (and artificially suppressing ionization) each of these states would be equally populated), and so does N_{II}/N_I. Plotting $\frac{N_{I,2}}{N_{\rm total}}$ as a function of temperature, we obtain Fig. 95.

In general all line strengths follow the pattern described by Fig. 95. The exception are lines in the ground state neutral atoms, which should only increase in strength with decreasing temperature, and ions, which should also increase with decreasing temperature so long as temperatures are sufficiently high for their existence (ionization is a fairly sharp transition in many cases, which is what causes the sudden disappearance of these lines for sufficiently cool stars). Lyman-α (1216 Å), Ca II H and K, and Ca I (4226 Å) are examples of such "resonance lines".

4.15.1. *What are the primary sources of line broadening?*

Lines are not delta functions, and their width can be measured either by the equivalent width W (where $F_{\rm continuum} W = \int F_{\rm continuum} - F d\lambda$) or the full-width half-maximum, which is obtained by fitting a Gaussian.

Spectral lines can be broadened by

- **Natural broadening**: the time-energy uncertainty principle, $\Delta E \Delta t \approx \hbar$, states that for a system with rapidly changing observables (i.e. a short lifetime in a certain state), the energy cannot be entirely certain. Very roughly, ΔE of a transition is equal to $\Delta E_f + \Delta E_i$, which translates to $\delta\lambda = \frac{\lambda^2}{2\pi c} \left(\frac{1}{\Delta t_i} + \frac{1}{\Delta t_f} \right)$. A more involved calculation

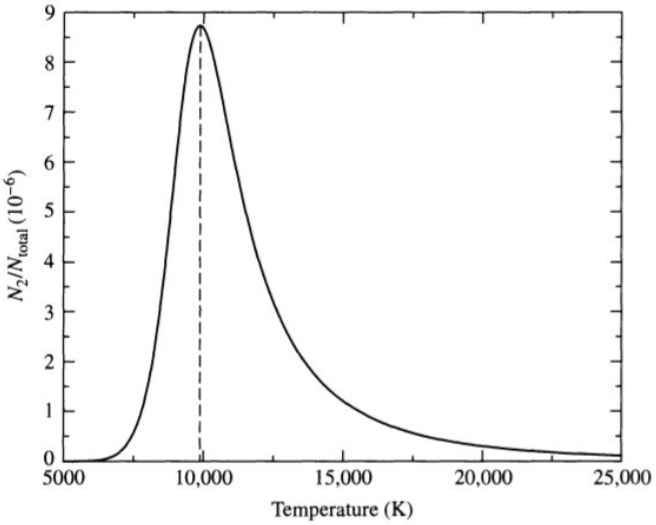

FIG. 95.— Estimate (assuming only ground and first excited state occupation for the Boltzmann equation) of the fraction of first excited state neutral hydrogen in a stellar atmosphere of varying temperature. n_e is estimated from $P_e = n_e k_B T = 20$ N/m^2. The peak occurs at 9900 K, corresponding to the temperature at which Balmer lines will be at their strongest. From Carroll & Ostlie (2006), their Fig. 8.9.

gives the FWHM of natural line broadening as $\frac{\lambda^2}{\pi c}\frac{1}{\Delta t}$, where Δt is the lifetime of a state before transition to another state. Natural broadening forms a Lorentz profile.

- **Doppler broadening**: this occurs because the atoms are Maxwell-Boltzmann distributed, with a most probable velocity $v_{\text{mp}} = \sqrt{\frac{2k_BT}{m}}$. $\Delta\lambda = \frac{2\lambda}{c}v_{\text{mp}}$ is a decent estimate. Integration over the distribution gives the FWHM as $\frac{2\lambda}{c}\sqrt{\ln 2\left(\frac{2k_BT}{m} + v_{\text{turb}}^2\right)}$, where v_{turb} is the most probable turbulent velocity (turbulent velocity is assumed to be Maxwell-Boltzmann distributed). Turbulence is of importance in giant and supergiant stars. Non-Maxwellian broadening, such as in fast-rotating, pulsating or superwind stars, also affect spectra.
- **Pressure (and collisional) broadening**: pressure broadening occurs when an atoms orbital structure is modified by a large number of closely passing ions. Close encounters with ions and collisions with neutral atoms also can disturb orbitals. We can estimate the effect of line broadening by setting the Δt in the natural broadening equation to $\Delta t = l/v = \frac{1}{n\sigma\sqrt{2k_BT/m}}$. Pressure/collision line broadening is therefore proportional to both temperature and density.

The combined line profile from these effects is known as the Voigt profile. The core of a Voight profile is dominated by Doppler broadening, while the wings by the Lorentz profile of natural and pressure broadening (this is because a Lorentz profile does not decay as quickly).

4.15.2. *What is the curve of growth?*

The curve of growth is a plot of W as a function of N_a, the abundance of a certain atom in the stellar atmosphere (N_a can be thought of as a column density). In the regime where the absorption line is optically thin (i.e. at its resonant wavelength, not all the light from the background is absorbed), W cales with N_a. Once the centre of the line becomes optically thick (and absorbs all incoming radiation at the resonant wavelength), the optically thin wings continue to grow, and W scales like $\sqrt{\ln N_a}$. If N_a continues to increase, it reaches a stage at which pressure line broadening exceeds natural line broadening, and W begins to scale like $N_a^{1/2}$.

4.15.3. *How is a P Cygni feature produced?*

A P Cygni profile consists of a blueshifted absorption line alongside a redshifted emission line (see Fig. 96), and is caused by an outflow. The blueshifted absorption is caused by continuum absorption due to material headed toward the observer, while the redshifted emission comes from re-radiation of absorbed photons (due to the same transition) from material headed away from the observer. The rest-frame resonant emission comes from material moving perpendicular to the observer. An inverted P Cygni profile indicates accretion.

4.15.4. *What are the main differences between supergiant and main sequence spectra, assuming the same spectral classification?*

Absorption lines in a supergiant are generally narrower than their dwarf star counterparts, as they have tenuous, extended (but still optically thick) atmospheres and therefore low surface densities at there photospheres. A reduced electron density also increases ionization, and therefore the Balmer lines (for example) will be suppressed at lower

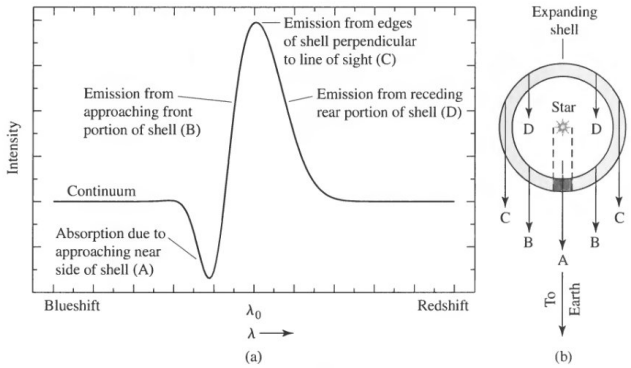

FIG. 96.— Schematic of a P Cygni profile, annotated to describe how each component of the feature is created. From Carroll & Ostlie (2006), their Fig. 12.17.

temperatures. These features are used to determine the Morgan-Keenan (MK) luminosity class, which can then be used to determine the (approximate) luminosity of the star.

4.16. *Question 16*

QUESTION: What physical and orbital parameters of an extra-solar planet can be determined a). from radial velocity (Doppler) measurements alone, b). from transit observations alone, and c). from the combination of both Doppler and transit observations?

Most of the information here comes from Santos (2008) and Seager (2010).

A planet-star system will orbit a common centre of mass. As the star orbits this centre of mass, its radial velocity will shift sinusoidally, with an amplitude that depends both on the star's orbital velocity v and the inclination of the system to us, i (where $i = 0$ means the system is edge-on to us). We may measure $v_s \sin i$ using $\frac{\Delta\lambda}{\lambda} = v_s \sin i/c$. While the typical spectrograph resolution (0.1 Å) is insufficient to detect $v_s \sin i$ (typically $\Delta\lambda = 10^{-4}$ Å) using a single line, fitting for the entire stellar spectrum can greatly increase accuracy - errors of 1 - 100 m/s are achievable.

Using the fact that $M_s/M_p = v_p/v_s = a_p/a_s$ (i.e. the CM has position 0 and does not move), $v = 2\pi a/P$ and Kepler's Third Law, we are able to find that $K_s = v_s \sin i$ has the form

$$K_s = 2\pi G \left(\frac{M_s}{P}\right)^{1/3} \frac{q}{(1+q)^{2/3}} \frac{\sin i}{\sqrt{1-e^2}}. \tag{183}$$

where $q = M_p/M_s$. Generally, the "mass function" $f(M) = \frac{PK_s^3}{2\pi G} = \frac{M_p \sin^3 i}{(1+q)^2}$ is reported. $1 + q \approx 1$ for planets, and all other quantities except for q and $\sin i$ can be measured: P is obtained from timing the period of radial oscillatory motion, and a_p can be found by noting that $M_s a_s \sin i = M_p a_p \sin i$ and noting we know M_s, $a_s \sin i$ and $M_p \sin i$. e can be determined by the shape of the binned velocity curve. It is not possible to obtain i, and therefore we can only find a minimum mass using radial velocities (statistically, there is an 87% chance $\sin i > 0.5$, as there is a bias against finding low-inclination orbits.

Transits occur when the star and planet orbit line up so that, from the perspective of an observer, the planet periodically passes in front of its parent star (Fig. 97). Full transit probability can be estimated by $p = R_s/a$, indicating that a close-in planet orbiting a giant star is much more likely to be spotted via transit assuming any transit that occurs can be detected. The dip in starlight due to a transiting planet can be estimated as

$$\frac{\Delta L}{L} \approx \left(\frac{R_P}{R_S}\right)^2. \tag{184}$$

For Jupiter transiting the Sun, a $\sim 1\%$ drop in the Sun's luminosity is expected. Since a large $\Delta L/L$ is easier to detect, giant planets and small stars are favoured.

A transit schematic is given in Fig. 97. As the planet moves in front of the star, there will be two periods of time (between I and II and between III and IV on the schematic) where only part of the planet is blocking any starlight; these periods are known as ingress (I - II) and egress (III - IV) (periods are denoted $\tau_{\rm ing}$ and $\tau_{\rm egr}$, and the two values are generally almost the same) (Seager 2010). These times, along with the total transit time, may be used to determine the impact parameter (b in Fig. 97) and R_s/a, which can then be used to determine the inclination angle i (see Seager 2010, pg. 60). In reality, the star is not a uniform disk, and a model of limb darkening must be used to accurately determine ingress and egress times.

If secondary eclipses are visible, it is actually possible to constrain the eccentricity and angle between transit and periastron (Seager 2010). The semi-major axis can be found if the star's mass is known. More complex features on

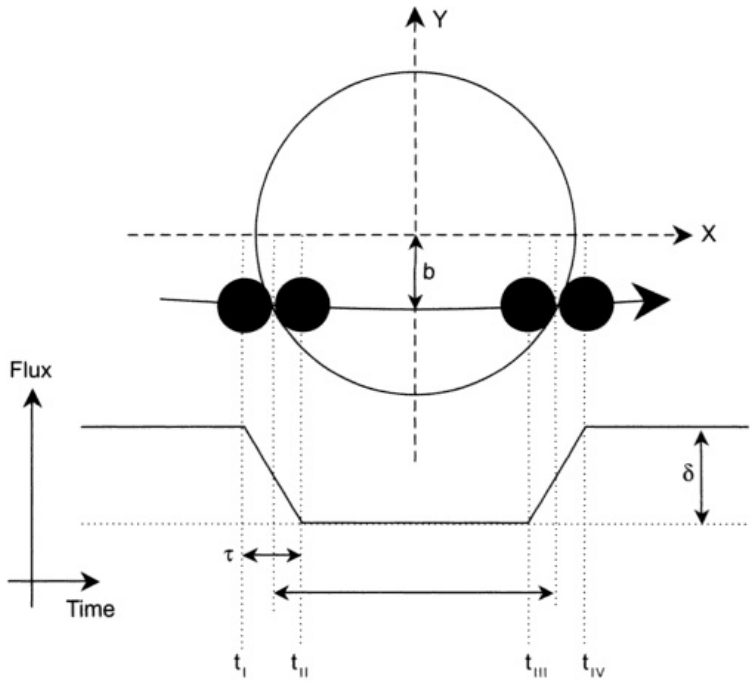

FIG. 97.— A schematic of a transient and its ideal light curve, assuming that the star is a uniform disk and the planet uniformly absorbs all light. The four times correspond to I). when the planet first begins to transit the star, II). when the entire cross-section of the planet is within the stellar disk, III). last point in the transient when the entire planet cross-section is within the stellar disk, and IV). end of transient. Changing the impact parameter b will change the ratio of the time period between I and II (or III and IV, which is almost identical) and I and IV. Limb darkening is not modelled here. From Seager (2010), their Ch. 2, Fig. 2.

both the star and the planet (such as sunspots or rings) can be determined through modelling of the transit light curve. The mass, however, cannot be determined.

From this, we can see that if we had only radial velocity measurements, we would be able to determine the semi-major axis (if we knew the mass of the star), period and orbital eccentricity. The mass will only be determined up to $M_p \sin i$. If we had only transits, we would know i, planetary radius (assuming we know the stellar radius)[39], semi-major axis (if the parent star mass is known), eccentricity (if secondary eclipses are visible) and period. If we knew both, we would have all orbital properties (semi-major axis, period, orbital eccentricity and inclination) and all planetary properties (mass, radius and average density). If no information about the star is included, planetary masses and radii would be known only up to scaling constants for the stellar radius and mass, though we would still be able to determine period and eccentricity (Seager 2010).

A number of curious properties can be determined without any knowledge of the parent star. So long as the period and R_S/a is known (both determined from transits), one can determine $\rho_S + \frac{R_P^3}{R_S^3}\rho_P$ (Seager 2010, pg 60). Since the latter term is tiny, this provides a means of constraining stellar modes through the mean stellar density ρ_S. With radial velocities and transits combined, one can determine the surface gravity of the planet using the star's radial velocity, the period, inclination and R_P/a (derivable from R_P/R_S and R_S/a) (Seager 2010).

Transits and radial velocities are binned in order to increase signal-to-noise. If, however, the orbit precessed, or if the planet has other planetary bodies exerting gravitational forces on it, then the time of the transient and the shape of the radial velocity curve will change. Practically, transit timing has much greater sensitivity, and is often used to infer the existence of additional, unseen planets in a stellar system (changes in orbit will also effect the shape of the transit, which could also be used) (Seager 2010). If transits are observed at multiple wavelengths, how the shapes of the transit ingress and egress change with wavelength relates to the optical properties of the planet's atmosphere (Seager 2010).

4.16.1. *What methods are there to detect planets?*

There are six methods by which exoplanet detection is made:

- **Radial velocity measurements**.
- **Transits**.

[39] Technically the mass can be determined here from the mass-radius relation, but as noted in Sec. 4.2, the mass-radius relation for giants is flat.

- **Astrometry** - measurement of the semimajor axis, a_s, of the orbit of a planet-hosting star, and the period P, allows us to measure the mass of the planet via $P^2 = 4\pi^2 a^3/G(M_p + M_s)$ and $a_s = M_p a/(M_s + M_p)$, if the mass of the star M_s is known. In practice the entire system can be inclined to us, but ellipse fitting and radial velocity measurements can help disentangle any degeneracies. Stellar activity may also shift the photocentre of a star, creating false astrometric movements. A more pressing concern is the fact that the expected astrometric proper motions of stars reacting even to Hot Jupiters is of order 1 μas at best, and as of 2008 no instrument could definitively obtain astrometric evidence of the existence of a planet without radial velocity or transit data to reinforce the case. Note that this same procedure can be used to determine the masses of unseen stellar-mass companions much more easily. If the companion is visible, a_{s1} and a_{s2} can both be measured, allowing us to solve for M_{s1} and M_{s2}.
- **Gravitational lensing** - not generally used.
- **Pulsar timing** - planets orbiting pulsars can shift the pulsar pulsation rate in a regular manner.
- **Direct imaging** - through stellar masking, faint emission/reflection from the neighbourhood of the star can be detected. This method can be used to observe dust disks as well as find planets.

4.16.2. *What are common issues when using radial velocities and transits?*

Intrinsic stellar features, like non-radial pulsation, inhomogeneous convection, or starspots may be responsible for the detection of radial-velocity variations. These hamper planet detections using radial velocities both by adding noise to signal, and by creating false positives. Starspots, eclipsing stellar binaries, dwarves eclipsing giants, and blended eclipsing binary systems may create false positives in transit searches. Many of these false positives can be weeded out by looking for suspicious patters in the transit photometry (ex. starspots do not form secondary eclipses, and are temporary, eclipsing stellar binary transits are dominated by ingress/egress).

4.16.3. *How do Hot Jupiters form?*

Gas giants need to form beyond the snow line, as water ice and other volatiles are needed to build cores of 10 Earth masses or greater, at which point gas accretion can begin in earnest. In particular, there simply was never enough material in the vicinity of Hot Jupiters to have constructed them. The solution to this issue is planetary migration.

Type I planetary migration involves interactions between the planet and density waves in the accretion disk, which are generated by non-axisymmetry perturbations of the disk. Density-wave planet interactions tend to transfer angular momentum outward and mass inward, and are proportional to mass, meaning that as planets develop the effect strengthens. It is possible for a planet that started forming beyond the snow line to collide with the star in $10^7 - 10^8$ years. Type I migration ends when the planet accretes enough (locally, since its movement through the accretion disk brings it fresh material) to form a gap in the accretion disk. Type II migration occurs once the gap is formed - maintenance of the gap results in the planet coupling to the viscous evolution of the accretion disk (which, of course, tends toward rigid rotation), generally driving the planet closer to the star. Type II migration is general slower than Type I migration. Gravitational scattering between planets and other planets or planetismals can drive inward or outward migration, depending on the nature of the scattering.

In our own Solar System, it is likely that Jupiter migrated about 0.5 AU inward from where it formed, while Saturn, Uranus and Neptune migrated outward. During this motion, the outer planets passed through certain resonances with each other and with minor bodies, displacing significant numbers of objects.

4.17. *Question 17*

QUESTION: What spectroscopic signatures help distinguish a young (pre-main sequence) star from a main sequence star of the same spectral type?

Most of this information is from Carroll & Ostlie (2006) Ch. 12.3, and cribbed from Emberson (2012). Followups are from the same sources.

A young stellar object (YSO) is a star that has yet to settle on the ZAMS and include both protostars and pre-main sequence stars (the dividing line between the two is the appearance of significant amounts of outflowing material, which ends the spherical accretion phase and begins the disk accretion phase). Most of the internal evolution of these stars, such as the switch between gravitational contraction and nuclear fusion, cannot be seen, and so instead YSOs are classified spectroscopically (see below) into Class 0, I, II and III. These classes trace out evolution over time, as by the time a YSO reaches Class II or III (which we will denote "late-type"), bipolar outflows have cleared out much of the material surrounding the pre-main sequence star, and we are left with a star, disk and substantial outflow material. A schematic of this process can be found in Fig. 98

Photometric and spectroscopic signatures of late-type YSOs (i.e. the ones that have ejected much of their surrounding material) include:

- **Emission lines** in the optical from regions of the hot disk and/or the outflow. These lines may include the Balmer series, the Ca II H and K lines, and forbidden lines [O I] and [S II], which are evidence of a very tenuous

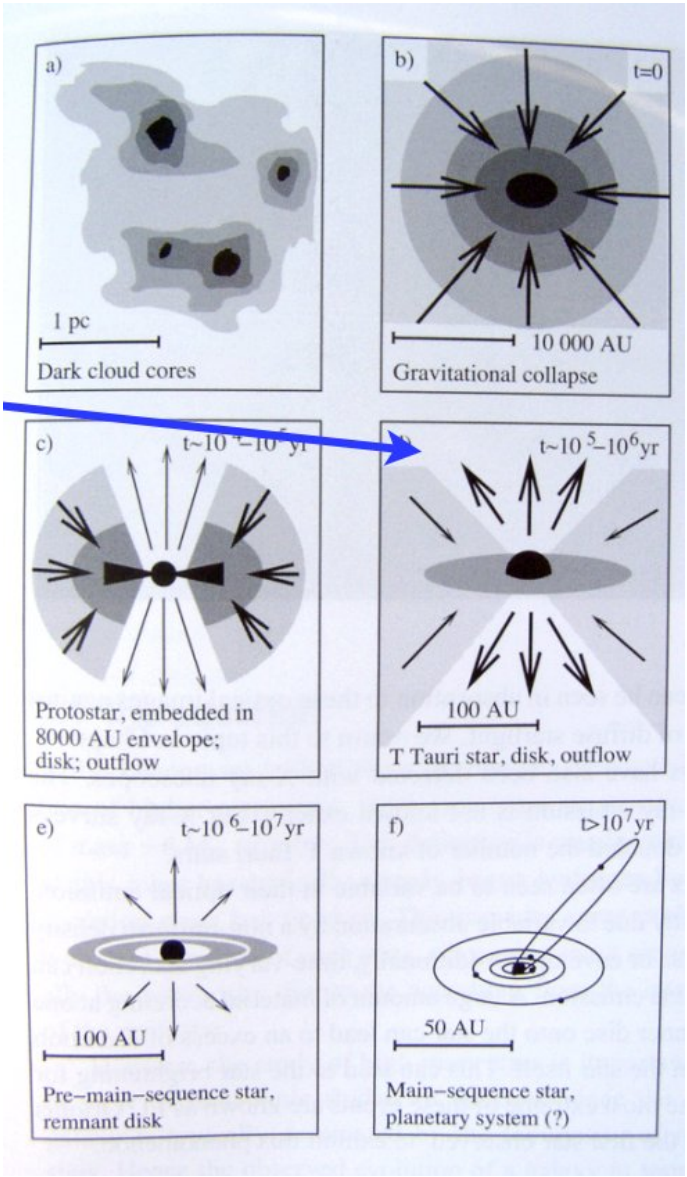

FIG. 98.— A schematic of low-mass stellar formation from 10 K clouds to the zero-age main sequence. Recovered from Stolte (2011), hence the blue arrow.

material (since forbidden lines are very easily collisionally destroyed). These lines tend to have a characteristic P Cygni feature (Sec. 4.6, though for late-type YSOs the emission is much stronger than the absorption), which is due to absorption of the line by the outflowing atmosphere in front of the YSO combined with the line emission by perpendicular and receding outflows. For T Tauri stars (Sec. 4.18) the difference in velocities between the blue and redshifted trough/peak is 80 km/s. Occasionally the P Cygni profile will reverse, signalling periods of significant mass accretion. G to M type absorption from the YSO's photosphere can also be observed. See Fig. 99.

- **Lithium absorption lines**. Since Li is generally destroyed in stars near when they begin nuclear fusion (Sec. 3.13), identifying Li absorption lines indicates the star has yet to achieve nuclear fusion.
- **High variability** (hours to days) in the optical, since material is still being accreted.
- **An extended emission bump in the infrared** due to dust absorption. This emission comes at the expense of optical emission, and so the YSO is shifted significantly to the right on the HR diagram compared to the MS star it will become (in fact depending on its age it can be near or past the Hayashi line).
- **A high level of activity** (flares, starspots, coronal emission), which can result in variability in x-ray. This is due to the fact that these stars are fast-rotators and are fully convective; these two facts combine to generate a strong poloidal field in the YSO, which drives magnetic field-related activity.

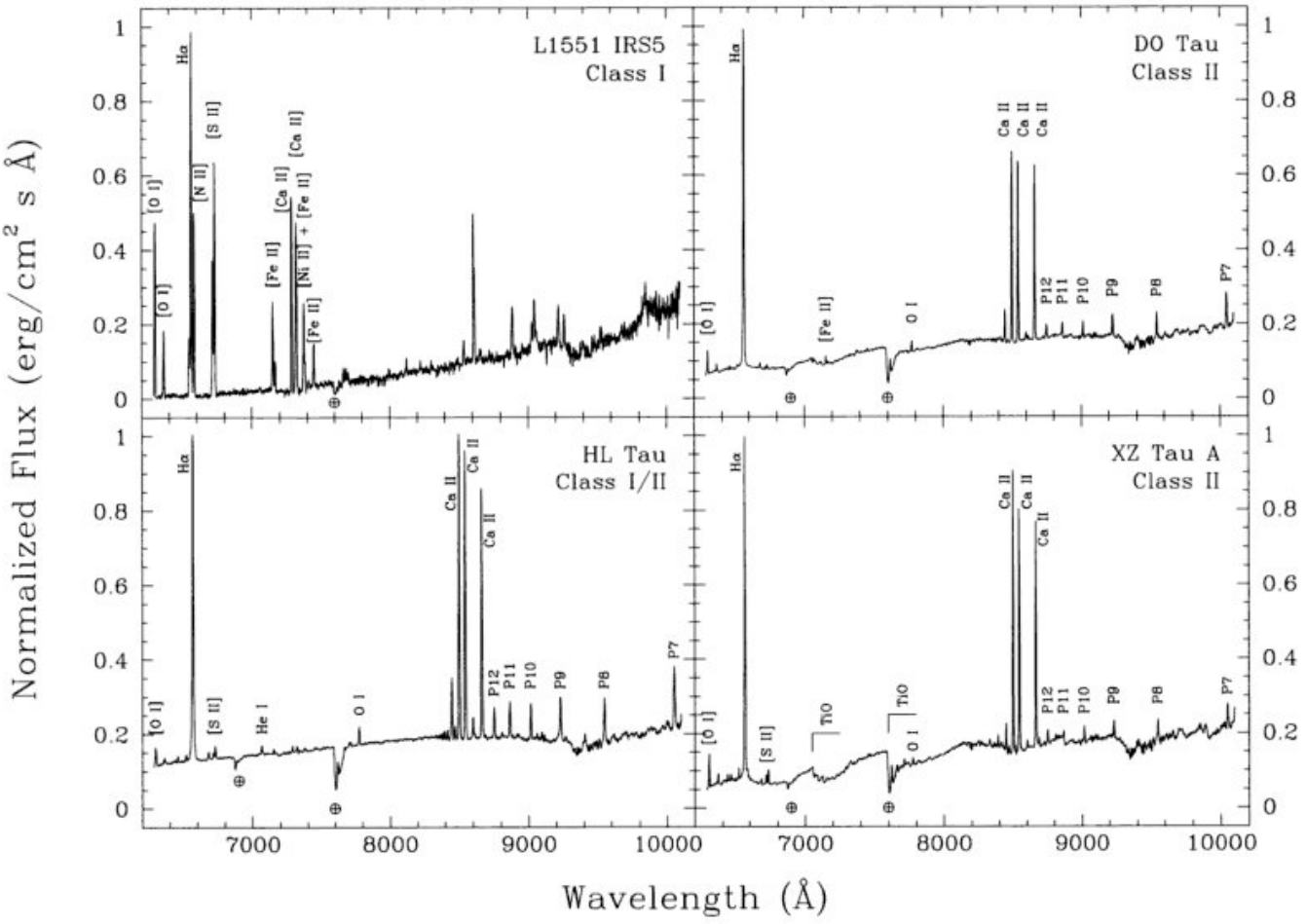

FIG. 99.— Optical spectra of various types of YSOs, normalized to unity at the peak of Hα emission. In the Class I and I/II sources all absorption is due to the Earth's atmosphere, and not intrinsic to the object. From Gray & Corbally (2009).

4.17.1. *How are YSOs classified spectroscopically?*

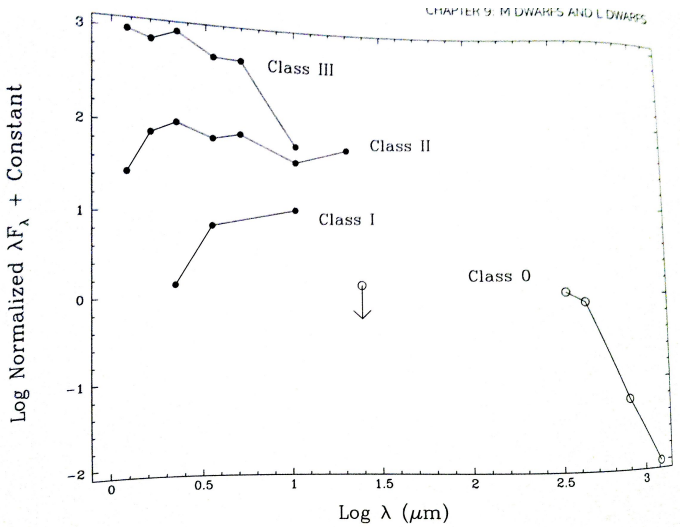

FIG. 100.— Continuum emission from different classes of YSOs. From Emberson (2012).

YSOs are intrinsically faint in the optical due to being surrounded by dust envelopes or disks. This is why their

classification uses the power law index

$$\alpha \equiv \frac{d\ln(\lambda F_\lambda)}{d\ln\lambda} \tag{185}$$

over the wavelength interval $2\,\mu\text{m} \lesssim \lambda \lesssim 25\,\mu\text{m}$. Class 0 sources have negiligible emission shortward of 10 μm, Class I have $0 < \alpha < 3$, Class II has $-2 < \alpha < 0$ and Class III has $-3\alpha - 2$, which more closely resembles ZAMS stars. See Fig. 100.

Alternatively, we can define classes using a bolometric temperature: Class 0 has T < 70 K, Class I has 75 K < T < 650 K, Class II has 650 K < T < 2880 K, and Class III has T > 2880 K.

Classes correspond to different stages of the lives of YSOs:

- Class 0 YSOs are cloud cores just beginning their collapse into YSOs.
- Class I YSOs lie in a cocoon of CSM that is assembling itself into a disk. Near the end of Class I an outflow breaks out from the star and ejects much of the CSM.
- Class II YSOs undergo disk accretion but are still surrounded by residual material.
- Class III YSOs have moved beyond the accretion stage but have yet to move onto the ZAMS.

See Fig. 101.

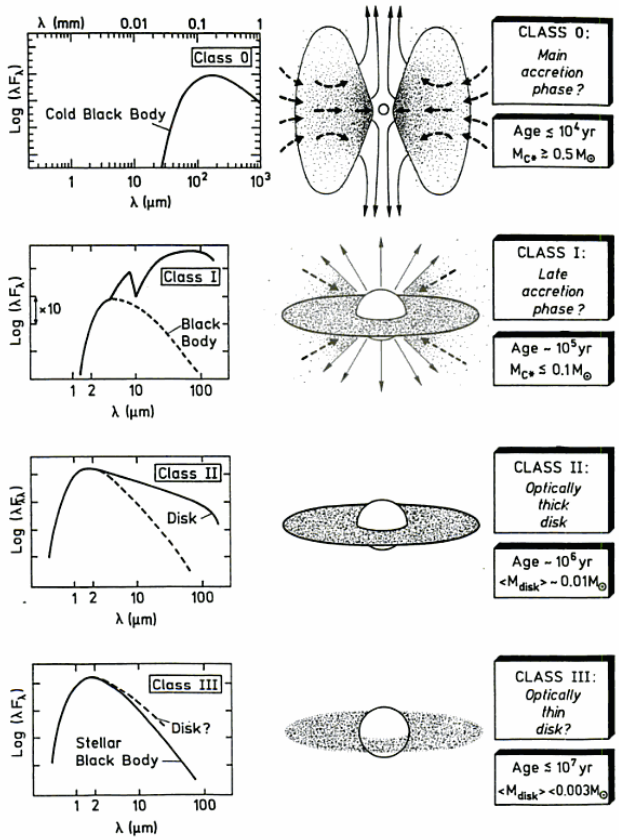

FIG. 101.— Schematic of the physical nature of various YSO types. From Emberson (2012).

4.17.2. *How do YSOs evolve on the HR diagram?*

The rough trajectory of YSO evolution is given by the classical model of star formation (Sec. 3.13, in particular Fig. 64). If accretion is slow, then the object falls down and leftward along the Hayashi track (it will deviate upward and to the left if accretion is increased) until the luminosity decreases enough so that the interior becomes radiative (this does not occur for very low-mass stars, which simply "drop" along the Hayashi line until they reach their ZAMS positions). The luminosity then begins to increase once more. For an ideal gas with no nuclear generation, $L \propto M^{4.5}T_{\text{eff}}^{0.8}$, which is followed until the the star reaches the MS.

One modification to the classical model of star formation comes from the shell of material around the YSO. Because this shell is not ejected until the star begins moving down the Hayashi line, the YSO actually cannot be seen until then (we would still see the extended IR emission of the envelope before). This is known as the "birth line".

4.17.3. *What classes of PMS stars are there?*

Pre-main sequence stars have burst out of their cocoons, and drive significant outflows. A number of classes exist:

1. **T Tauri stars** have masses from 0.5 - 2 $M_\odot$. They are characterized as having an envelope, disk and outflow, and have most of the photometric and spectroscopic features described above.
2. **FU Orionis stars** are T Tauri stars that are currently undergoing a period of extreme mass accretion ($\dot{M} \sim 10^{-4}$ $M_\odot$/yr), which also drives up luminosities by superheating the inner disk. High-velocity winds are also driven.
3. **Herbig Ae/Be stars** have masses from 2 - 10 $M_\odot$, which look like brighter (since their blackbody spectrum more closely resembles an A or B star with extended infrared emission) versions of T Tauri stars.
4. **Herbig-Haro objects** are pre-main sequence stars that fire jets as well as an expanding shell outflow. They can be associated with either T Tauri or Herbig Ae/Be stars.

4.17.4. *How is this picture changed for massive star formation?*

More massive stars will in general be more violent (see above), but the means of star formation for truly massive stars is relatively unknown and difficult to pin down observationally (because they are rare and evolve quickly).

4.18. *Question 18*

QUESTION: Sketch the spectral energy distribution (SED) of a T Tauri star surrounded by a protoplanetary disk. How would the SED change a). if the disk develops a large inner hole, b). if the dust grains in the disk grow in size?

This answer was cribbed from Emberson (2012) with a bit of help from Kissin (2012) and whomever answered this question in the 2011 qualifier.

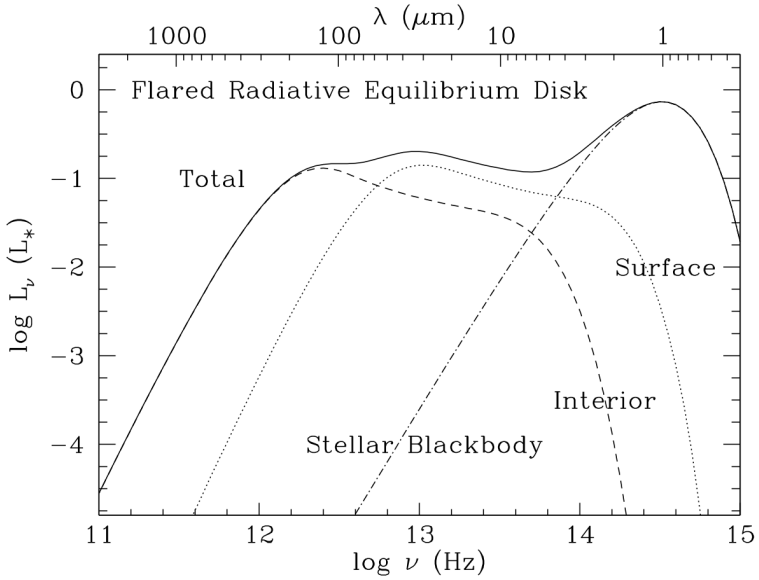

FIG. 102.— Schematic of the continuum SED of a T Tauri star. The contribution from each component of the system is noted. From Chiang & Goldreich (1997), their Fig. 6.

T Tauri stars (TTSs) were mentioned in Sec. 4.17 as being Class II YSOs that have an extended disk and an outflowing envelope, but insufficient optical depth to shield the central pre-main sequence star. TTSs have the same curious SED features listed in Sec. 4.17 that distinguish them from main sequence stars: emission lines (some forbidden) with P Cygni profiles, Li absorption and an extended IR emission tail due to a circumstellar disk of gas and dust (dust provides most of the absorption/re-emission). 10 μm SiO line emission from dust is also seen in the specetrum.

The extended flared dust disk can outpower emission from the star itself. We can consider the system as a simple flared, passive (only radiates what it absorbs from the star) disk surrounding a central star. The disk is then separated into two regions: a surface layer containing grains that are directly exposed to light from the central star and the shielded interior. The surface layer receives light directly from the star, and, in thermal equilibrium, radiates one half of incoming radiation into space and the other half into the inner regions of the disk. The equilibrium temperature is

larger than it would be for a pure blackbody because the emissivity of the grains in the IR is less than their absorptivity in the visible Chiang & Goldreich (1997). The disk interior is more complicated: at small r_{xy}[40] it is optically thick to its own emission and to the incoming radiation from the superheated surface. Past a certain r_{xy} it becomes transparent to its own radiation **I suppose because of Rayleigh?** but not to that of the surface, and past another critical point it becomes transparent to surface radiation as well Chiang & Goldreich (1997). This means that a large portion of the disk is optically *thin* at large r_{xy}. Detailed calculations show both disk and interior temperatures drop with $r_{xy}^{-1/2}$; the disk is simply an order of magnitude colder (this is not true for a flat disk) Chiang & Goldreich (1997).

The total SED, then, is a combination of the star and a disk with a hot surface and a cold interior. Emission from both disks can be modelled as a modified blackbody (since there is a temperature gradient) peaking at around $\lambda \approx 20$ μm for the surface and $\lambda \approx 200$ μm for the interior.

There are substantial viewing angle dependencies in the system. Since the disk is flared, if we were to view the TTS edge on, we would only see the outermost (i.e. large r_{xy}) reaches of the cold interior disk and parts of the superheated surface. As we moved toward viewing the system face-on, the inner (i.e. small r_{xy}) parts of the disk surface and interior, as well as the star, become visible. We would also see dust absorption (lines associated with H_2O and Si molecules) edge on, but dust emission (since the surface is superheated) face on. Since optical emission is damped for edge-on systems, they would be classified as earlier-type YSOs. See Fig. 103.

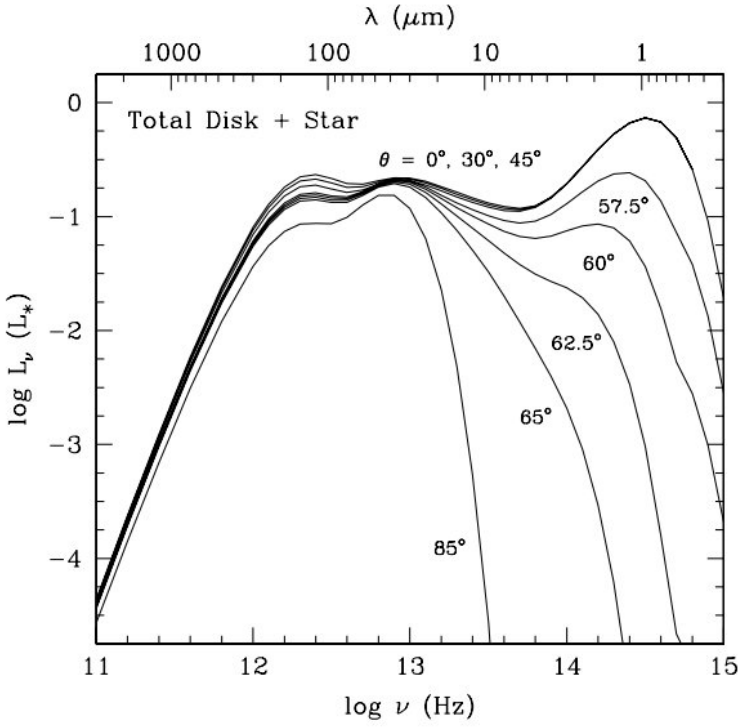

FIG. 103.— Schematic of the continuum SED of a T Tauri star at different viewing angles. From **?**, their Fig. 4.

If the disk were to develope a large inner hole, there would be a reduction in radiation from small r_{xy} regions of the disk (since they no longer exist). The outer surface layers of the disk, however, are largely unaffected by this change, and receive about the same amount of (slightly more) incoming flux. The interior disk, which does not directly receive flux from the star, is also unaffected. We therefore simply see a drop in emission from the inner disk.

If the grain size were increased, these larger grains yield a larger millimetre emissivity. At the same time, the number of small grains is reduced to account for the mass locked up in large grains, and this decreases the optical and near-infrared opacity. This results in less absorption of stellar ux, lower disk temperatures, and a decreased IR emission (D'Alessio et al. 2006). The combination of the two gives overall less disk emission, but relatively greater disk emission in the mm and sub-mm. See Fig. 104.

Note that while the figures presented above assume a passive disk, there should be some heat generation from viscous dissipation in the disk.

4.18.1. *Why does the disk have to be flared?*

This information is from Chiang & Goldreich (1997) and D'Alessio et al. (1999).

Originally, the disk around a TTS was assumed to be flat. This turned out to result in too little radiation coming from the disk, i.e. the outer regions of the disk were too cold to fit with observations. A number of alternatives (an

[40] We assume x and y are oriented onto the equatorial plane of the system.

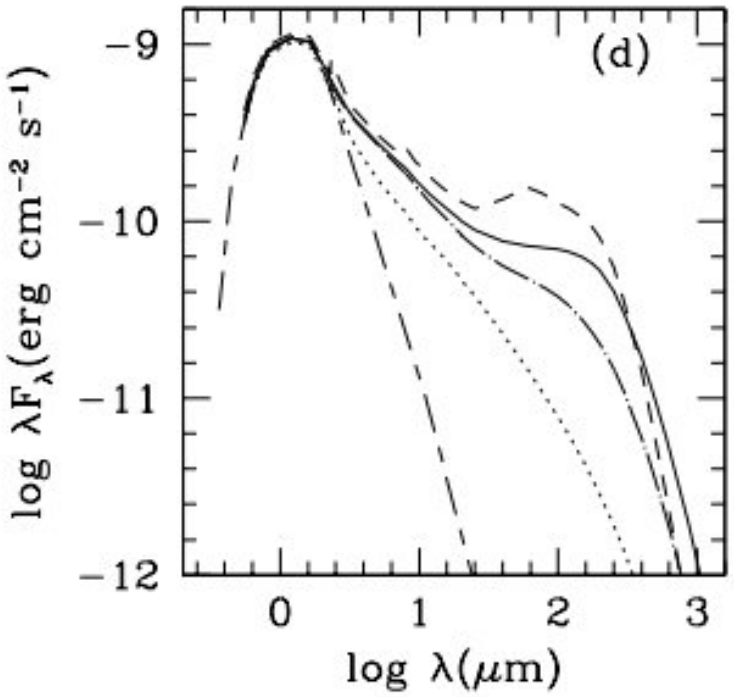

FIG. 104.— The continuum SED of a T Tauri star, given a grain distribution $n(a) \propto a^{-2.5}$ (a is grain size) and different maximum grain sizes. $a_{\max} = 10$ μm (dashed line), 1 mm (solid line), 1 cm (dot-dashed line) and 10 cm (dotted line). The system is assumed to be face-on. Differences are less substantial with a greater exponent. From D'Alessio et al. (2001), their Fig. 5.

active disk or an extended envelope to re-radiate some emission back onto the disk) were suggested, but the most robust solution was the hydrostatic equilibrium flared disk, which collects more light from the star.

4.19. *Question 19*

QUESTION: What are the primary origins of the heat lost to space by Jupiter, Earth, and Io?

The information for this question comes from Hartmann (2005), pg. 205 - 209, and Cole & Woolfson (2002) (various sections that have "heating" in their titles).

A planet's internal heat comes from several sources:

- **Gravitational contraction** - the planets formed out of material that had to collapse from a much larger size, releasing significant amounts of gravitational potential energy as heat.
- **Planetismal impact** - the impact of minor bodies onto planets is highly inelastic, and therefore adds heat to planets.
- **Radioactive decay** - radioactive isotopes within planets eject gamma radiation and particles, and the scattering of these radioactive products off other atoms produces heat.
- **Differentiation** - within a planet, lighter material tends to rise, and heavier material tends to sink, which releases gravitational potential energy through friction (the total energy released is the total potential energy of the system before and after differentiation). This process is based on the same principle as gravitational contraction, but is not identical, since the radius of the entire planet does not decrease.
- **Rotational spin-down** - tides between a planet and a large moon can spin the planet down (if the planet's spin angular velocity exceeds the orbital angular velocity of the moon).
- **Solar radiation** - photons are absorbed by the surfaces of planets, adding to their thermal energy budget. Solar wind may also contribute through electromagnetic induction heating (a process much more important in asteroids than planets, and during the early Solar System, with its significantly stronger T Tauri wind).

Heat produced in a planet's interior is transported outward through convection in molten regions of the interior, and through conduction in solid regions. At the surface, heat is radiated back into space.

Early in the life of the solar system, gravitational contraction, accretion of planetismals and rapid radioactive decay of isotopes such as ^{26}Al and ^{129}I created molten interiors inside terrestrial planets, which enables differentiation to occur (Hartmann 2005, pg. 207). Over time, terrestrial planets release their heat via radiation until they come into radiative balance with the solar radiation impinging on their surfaces - the value for Earth is 0.078 W/m^2 (Carroll & Ostlie 2006, pg. 752). Giant planets obtain much of their initial heat from gravitational contraction, and likewise slowly radiate it away over time.

The solar flux on Earth is 1365 W/m^2, meaning the heat radiated from the Earth's interior is 10^{-4} the reflected/absorbed-then-re-emitted light from the Sun. The Earth's albedo is 0.3, meaning that ~ 1000 W/m^2 of sunlight is absorbed. Thus, most of the heat lost to space by Earth is actually absorbed sunlight. The majority of internal heat lost to space comes from a combination of remnant internal heat from accretion/gravitational contraction/early radioactive decay (20% of thermal energy lost), and modern radioactive decay of heavy elements like ^{238}U, ^{235}U, ^{232}Th and ^{40}K, all with half-lives in the Gyr range (80% of thermal energy lost) Wikipedia (2012c). (Radioactive elements heat about 10^{-11} W/kg of mantle (which makes up the majority of mass on Earth), which gives a total energy generation rate of $\sim 4 \times 10^{13}$ W, translating to ~ 0.1 W/m^2 (Wikipedia 2012c, supported by Cole & Woolfson 2002, pg. 46 - 48).)

Jupiter actual emits 2.5 times more heat than it absorbs from the Sun (Jupiter's albedo is 0.343) the source of this energy is unknown, but is hypothesized to be one of two possibilities (Cole & Woolfson 2002, pg. 70). Gravitational contraction of Jupiter could still be occuring, with a rate of 0.05 cm per century being enough to explain the excess heating (contraction of 0.05 cm/century over the lifetime of Jupiter is only 3% its current radius). Alternatively (or complementary to contraction), separation and settling of He in Jupiter's atmosphere could provide a source of heat - the amount of material that would need to be differentiated is a fraction of the total He content of Jupiter.

Saturn radiates 2.3 times the heat it receives from the Sun, likely due to differentiation. Cole & Woolfson (2002), pg. 78, suggests Uranus actually may be emitting less energy than it receives from the Sun, but Hartmann (2005) suggests this is not the case, and Uranus radiates about the same heat it receives from the Sun. Neptune also radiates more heat than it receives from the Sun.

Io receives about 50 W/m^2 on its surface, and has a radius of about 1800 km, meaning that it receives about 5×10^{14} W of Solar energy. Its main source of internal heat is tidal stress from Jupiter (caused by the fact that Io is in resonance with the other Galilean moons, giving it a slight eccentricity), which generates some 10^{13} W. While most of Io's heat radiated into space comes from the Sun, Io's internal heat keeps it volcanically active. Io is also heated by Jupiter's magnetic field through Joule heating.

4.19.1. *Can you determine an order of magnitude estimate for Io's heating?*

This comes from (Cole & Woolfson 2002, pg. 402 - 404), except with a as the semimajor axis and R as Io's radius.

Suppose we divide Io into two hemispheres, one centred at $a - \frac{1}{2}R$, and the other at $a + \frac{1}{2}R$. In the CM frame, the near side has an acceleration

$$A = \frac{GM_J}{(a - \frac{1}{2}R)^2} - \frac{GM_J}{R^2} - (a - \frac{1}{2}R)\omega^2 + R\omega^2 \tag{186}$$

toward Jupiter, which can be simplified to

$$A = \frac{3GM_J a}{2R^3} \tag{187}$$

Giving a force on $\frac{1}{2}M_I$ of $F = \frac{3}{4}\frac{GM_I M_J a}{R^3}$. Io stretches WHY NOT MULTIPLY FORCE BY TWO? by an extension $\epsilon = \frac{2FR}{4R^2 Y} = \frac{F}{2RY}$, where Y is the Young's modulus (a measure of deformation given a force). The energy associated with this stretching is the work it takes to stretch it to ϵ, i.e. WHY 1/2??? $\Phi = \frac{1}{2}F\epsilon = \frac{1}{4}\frac{F^2}{RY}$. As Io moves about its slightly eccentric orbit, its orbital separation from Jupiter changes by a factor $2ae$ (from $a(1+e)$ to $a(1-e)$). The change in stretching energy is then $\Delta\Phi = \frac{d\Phi}{da}2ae$. If Io were completely elastic, this change in work would not heat it, but some energy is dissipated into Io as heat. The attenuation of seismic waves into thermal energy is given by $W = \frac{2\pi E}{Q}$, where Q, the "quality factor" is -2π times the energy in a seismic wave over the energy lost to dissipation. Io gains this much energy per period P. The heating rate of Io is then given by

$$W/P = (GM_I M_J)^2 \frac{27\pi Re}{8a^6 QPY} \tag{188}$$

Plugging in realistic numbers gives us about 3×10^{13} W.

4.20. *Question 20*

QUESTION: Consider a small test-mass orbiting around a compact object. What are the essential differences in the properties of its orbits between Newtonian gravity and General Relativity?

Most of this information comes from Charles Dyer's notes, and Carroll (2011). While the mathematics to derive the orbit of a test mass around a compact object is reasonable, it is not possible to perform such derivations in the time alloted for a qual question. We will therefore answer this question qualitatively.

Suppose the compact object is not rotating, and can be approximated as having spherical symmetry. Newtonian dynamics would have the object moving in an elliptical orbit. Far (defined below) from the object, GR effects make it so that the gravitational force is effectively no longer GMm/r^2; this deviation can be derived from the Schwarzschild line element,

$$ds^2 = \left(1 - \frac{2m}{r}\right) dt^2 - \left(1 - \frac{2m}{r}\right)^{-1} dr^2 - r^2(d\theta^2 + sin^2\theta d\phi^2) \tag{189}$$

where $m = GM/c^2$. Deriving the Lagrangian, and then the equations of motion, we obtain

$$u'' + u = \frac{m}{B^2} + 3mu^2 \tag{190}$$

where $u \equiv 1/r$ and primes indicate differentiation with respect to ϕ. Contrasting this equation with the Newtonian $u_c'' + u_c = \frac{GM}{L^2}$, we note the extra $3mu^2$, and look for first order deviations to the Newtonian solution. The solution to the Newtonian equation is $u_c = \frac{m}{B^2}(1 + \epsilon\cos\phi)$; this gives us an ellipse with semi-major axis $a = \frac{B^2}{m}\frac{1}{1-\epsilon^2}$. We can plug this solution back into Eqn. 190 in order to eventually obtain

$$u = \frac{m}{B^2}\left(1 + \epsilon\cos\left(\phi\left(1 - 3\frac{m^2}{B^2}\right) + ...\right)\right), \tag{191}$$

which is nearly identical to a Keplerian orbit, but precesses such that the major orbit moves by $\Delta\phi = 6\pi\frac{m^2}{B^2} = \frac{6\pi m}{a(1-\epsilon^2)}$. If we were to plug in the values for Mercury, we would obtain 0.1" per orbit, translating to the observed ∼43" per century. This expression holds when $m^2/B^2 << 1$, equivalent to $m/r_c = GM/c^2 r_c << 1$, which sets a scale for how far the test mass must be away from the compact object.

If the compact object is not "very" compact (i.e. it is a white dwarf), then these modifications to Newtonian gravity are the extent of the changes, but if the compact object is a neutron star or black hole, then greater deviations from Newtonian gravity exist. Noting that by Birkhoff's Theorem all of these objects, if spherically symmetric, can be described by the Schwarzschild Solution, we can assume the compact object is a black hole (trajectories around a neutron star will be identical up to the surface of the star).

At $r = 3m$, we reach the "photon sphere", at which photons are able to circle the black hole indefinitely. To obtain the stability criterion of particle orbits, we convert Eqn. 189 into a Lagrangian, and rewrite the result in the form $\frac{1}{2}\dot{r}^2 + V(r) = \frac{1}{2}A^2$, where $V(r) = \frac{1}{2}\mathbb{L} - \mathbb{L}\frac{m}{r} + \frac{B^2}{2r} - \frac{mB^2}{r^3}$, where $\mathbb{L}$ is zero for null and 1 for timelike geodesics. Attempting to find circular orbits by determining when $dV/dr = 0$ for varying L, we find a set of stable circular orbits for $r = 6m$ and above, and a set of unstable circular orbits from the photon sphere to $r = 6m$ (objects on unstable orbits may, with a slight perturbation, be pushed into the black hole or ejected out to infinity). Below a critical $L = \sqrt{12}GM$, no stable orbits exist. Precessing elliptical orbits can dip down to $r = 4m$ (only unbound trajectories can get lower). Unlike for Newtonian effective potentials which always go to ∞ as $r \to 0$ (due to the L^2/mr^2 term, which becomes infinite at $r = 0$ for any non-zero L), for the Schwarzschild metric all $V(r)$ go to $-\infty$ (for both photons and particles) due to the additional r^{-3} term. This is why there are minimum stable test particle orbits in GR, but not in Newtonian dynamics.

From Eqn. 189 and the Euler-Lagrange equations we can obtain for radially infalling particles

$$\frac{dr}{d\tau} = \pm\sqrt{A^2 - (1 - 2m/r)}$$
$$\frac{dt}{d\tau} = \frac{A}{1 - 2m/r}, \tag{192}$$

and for radially infalling photonss

$$\frac{dr}{dt} = \pm\left(1 - \frac{2m}{r}\right). \tag{193}$$

As it turns out, the integral of $\frac{dt}{dr}$ is divergent if $r = 2m$ is included in the integration bounds. At the same time, however, the integral of $\frac{d\tau}{dr}$ is finite even if the integration bounds extend to $r = 0$. This indicates that while the proper time of an infalling object is finite when the object reaches the singularity, an observer at infinity at rest with

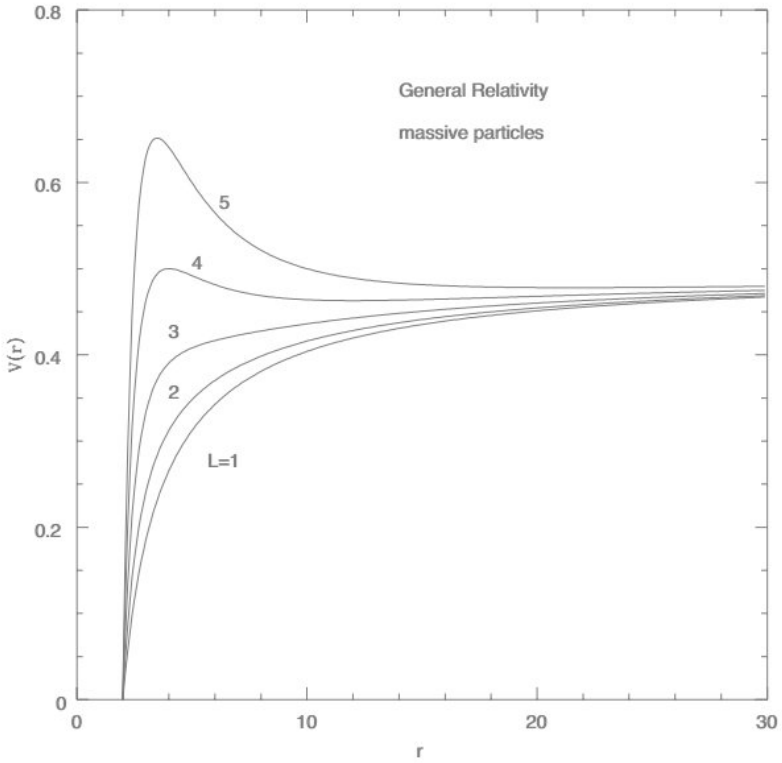

FIG. 105.— A plot of various effective potentials for a test particle in a Schwarzschild metric. Note that below a critical L, there is no maximum in the function at low r. From Carroll (2011), pg. 178.

respect to the black hole would have to wait an infinite amount of time to see the object cross the event horizon. As space and time are not coupled in Newtonian mechanics, a Newtonian "black hole" would not have these features.

To summarize, in Newtonian mechanics any bound particle with non-zero L orbiting a compact object will trace out an elliptical orbit. In GR, if the particle's distance $r >> GM/c^2$, it will trace out an elliptical that precesses. If the particle is close to the object, and the object is a black hole (and therefore small), the particle will only be able to occupy a stable circular orbit if its radius $r \geq 6m$, and a stable closed orbit if its radius $r \geq 4m$. If its energy is high enough to cross the potential barrier at $rsim3m$, or if it has $L < \sqrt{12}GM$, then the particle cannot maintain a stable orbit and can (will for $L < \sqrt{12}GM$ or $\dot{r} < 0$) fall into the black hole. A photon sphere exists at $r = 3m$. An event horizon exists at $r = 2m$, and infalling objects will be seen by distant observers to take an infinite amount of time to cross it.

4.20.1. *What would change if the object were rotating?*

If we drop the assumption of non-rotation and spherical symmetry for our black hole, but assume a stationary state, we obtain the Kerr solution, which, at large distances, resembles the Schwarzschild metric, and our post-Newtonian approximations for how orbits are deflected are reasonably accurate for the Kerr metric. The (outer) event horizon of the Kerr metric is larger than that of the Schwarzschild metric. Moreover, analysis of the system's Killing vectors gives that infalling geodesics will tend to move along spiral trajectories rather than radial ones as in the Schwarzschild metric, an effect known as "frame dragging". To see the metric as stationary, observers would have to move against the direction of metric rotation. This effect becomes more extreme with decreasing distance, and there is a oblate spheroidal region around the outer event horizon known as the "ergosphere" where no time-like observer would see the metric as stationary.

If an object attempts to orbit the black hole with an angular momentum not aligned with the black hole's axis of rotation, frame dragging results in a precession to this orbit which, in the weak field limit, is known as the Lense-Thirring effect. Lense and Thirring showed that for a circular orbit in a Kerr metric, the nodes of the orbit will be rotated about the spin axis of the black hole by an angle $\delta\Omega$ each orbit (this is different than the precession described above, since that effect cannot change the orbital plane). $\delta\Omega$ is dependent on spin plane distance from the singularity, and therefore the Lense-Thirring effect has an effect on accretion disks: if the disk plane and black hole spin plane are misaligned, the inner disk will warp. This will often be the case for disks from tidal disruptions of close-passing stars, since there is no reason these stars should have any particular alignment to the spin of the black hole.

4.20.2. *Charged Compact Objects*

Charged objects can be represented by the Reissner-Nordstrøm metric. Adding charge effectively changes the locations of the event horizon, adds an inner null horizon past which the radius r switches back to being a spacelike

coordinate (in a Schwarzschild metric, lines of constant radius are timelike outside the event horizon and spacelike inside, indicating that inside the horizon no timelike observer can stay at constant radius), and, in the case of this metric, changes the singularity at $r = 0$ to a timelike singularity. The outer event horizon is at a smaller r than for a Schwarzschild metric with the same M, and additional r^{-2} and r^{-4} terms enter the effective potential. The second horizon and a timelike singularity means that travellers can move freely inside the inner event horizon, and can actually fly back out and *exit* the black hole (see Carroll pg.41 for details). Charged particles orbiting such a metric will be affected by a Lorentz force from the black hole's electromagnetic field.

The most general metric incorporates both charge and rotation, and is known as the Kerr-Newman metric.

4.20.3. *What would happen if you fell into a black hole?*

While at $r = 2m$ Eqn. 189 appears to be undefined, we can transform from our standard coordinates to Kruskal-Szekeres coordinates, which leads to the Kruskal diagram, Fig. 106, which summarizes many of the properties of trajectories that wander too close to a Schwarzschild black hole. For example, an object orbiting close to the black hole experiences time dilation compared to observers at infinity, an effect that can be seen by picking a radius in Fig. 106 and drawing light signals being emitted at regular intervals from this radius out to much larger radii. The receiver at much larger radii will find these intervals to be longer than what the emitter sees. (This effect explains gravitational redshift.) The light signal from an object moving past the event horizon takes an "infinite" amount of time to reach an observer at infinity, meaning that observers at infinity will never see an object cross the horizon, in accordance with our calculation from earlier.

If a person were launched into the black hole, their head and feet could be represented by two timeline trajectories. Assuming the person is launched feet-first, over time the free-fall trajectory of the feet would carry them away from the head if no restoring forces existed for the human body. This is known colloquially as "spaghettification", and has to do with the extreme tidal forces near the event horizon. Passing the event horizon would be uneventful, and, if spaghettification has not already killed the person, s/he will eventually reach the singularity, which is space-like. Because light signals cannot propagate to larger r from the singularity, the person will not feel his/her feet "hit" the singularity before his/her head does.

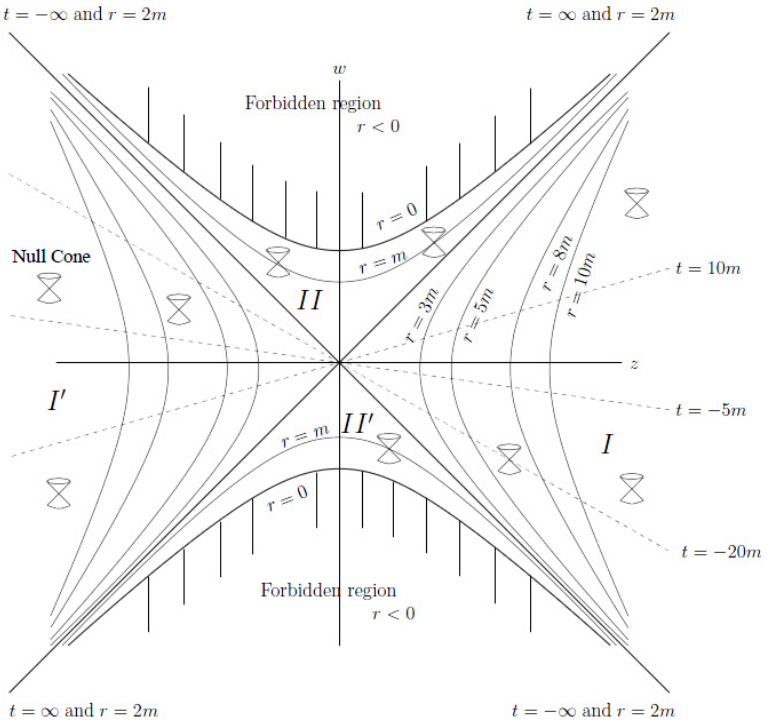

FIG. 106.— The Kruskal diagram. Null trajectories are lines with slope 1. Lines of constant time follow lines of constant w/z, and lines of constant r are hyperbolae. Note that $r = 2m$ and $t = \infty$ are identical lines that cross the origin of the diagram. From Dyer (2010b).

4.21. *Question 21*

QUESTION: Write out the p-p chain. Summarize the CNO cycle.

This material was taken from Carroll & Ostlie (2006) and Kippenhahn & Weigert (1994).

In any nuclear reaction process, baryon number, lepton number, and charge must all be conserved (note that antiparticles subtract from these numbers, so that particle-particle pair results in zero change in any of these numbers, allowing this process to occur /textitex nihilo). The energy liberated or captured in a nuclear reaction can be determined by the difference in net rest mass between the reactants and products ($\Delta E = \Delta mc^2$) - this energy difference is the difference in nuclear strong force binding energy between the two sets of particles.

The net reaction for hydrogen fusion inside stars is $4\,{}^1_1\mathrm{H} \rightarrow {}^4_2\mathrm{He} + 2e^+ + 2\nu_e + 2\gamma$ [41]. The net energy gain in this reaction is 26.731 MeV, ten times the energy liberated from any other fusion process (2 - 30% of this energy is carried away in neutrinos). The rate at which 4 H atoms can come together and tunnel to form an ^{4}He atom is tiny compared to the rate of a more energetically favourable chain known as the proton-proton chain.

The first proton-proton chain (PPI) is

$$\begin{aligned} {}^1_1\mathrm{H} + {}^1_1\mathrm{H} &\rightarrow {}^2_1\mathrm{H} + e^+ + \nu_e \\ {}^2_1\mathrm{H} + {}^1_1\mathrm{H} &\rightarrow {}^3_2\mathrm{He} + \gamma \\ {}^3_2\mathrm{He} + {}^3_2\mathrm{He} &\rightarrow {}^4_2\mathrm{He} + 2\,{}^1_1\mathrm{H}. \end{aligned} \tag{194}$$

(Note that lepton number is conserved in these reactions by the inclusion of neutrinos. Also, to be exacting, two ${}^1_1$H that entered the chain return, making them catalysts, meaning that PPI could be called a p-p cycle.) The first step is rate limiting, since it requires proton decay. The production of Helium-3 in the third step of PPI allows for direct interaction with Helium-4 through PPII:

$$\begin{aligned} {}^3_2\mathrm{He} + {}^4_2\mathrm{He} &\rightarrow {}^7_4\mathrm{Be} + \gamma \\ {}^7_4\mathrm{Be} + e^- &\rightarrow {}^7_3\mathrm{Li} + \nu_e \\ {}^7_3\mathrm{Li} + {}^1_1\mathrm{H} &\rightarrow 2\,{}^4_2\mathrm{He}. \end{aligned} \tag{195}$$

Beryllium-7 can also capture an electron, leading to the PPIII chain

$$\begin{aligned} {}^7_4\mathrm{Be} + {}^1_1\mathrm{H} &\rightarrow {}^8_5\mathrm{B} + \gamma \\ {}^8_5\mathrm{B} &\rightarrow {}^8_4\mathrm{Be} + e^+ + \nu_e \\ {}^8_4\mathrm{Be} &\rightarrow 2\,{}^4_2\mathrm{He}. \end{aligned} \tag{196}$$

The branching ratios of the three different chains (i.e. the percentage of p-p reactions that utilize each chain) depends on the exact temperature, density and chemical composition of the material undergoing fusion. Fig. 107 gives values for the interior of the Sun.

The nuclear energy generation rate is given by the rate-limiting first step of the chain. This rate can be approximated by a power law near $T = 1.5 \times 10^7$ K,

$$\epsilon_{pp} \approx 1.08 \times 10^{-12} \rho X^2 f_{pp} \psi_{pp} C_{pp} T_6^4 \mathrm{Wkg}^{-1}, \tag{197}$$

or $\epsilon_{pp} \propto T^4$.

4.21.1. *The CNO Cycle*

In the presence of carbon, another means of producing helium from hydrogen becomes available, the CNO cycle:

$$\begin{aligned} {}^{12}_6\mathrm{C} + {}^1_1\mathrm{H} &\rightarrow {}^{13}_7\mathrm{N} + \gamma \\ {}^{13}_7\mathrm{N} &\rightarrow {}^{13}_6\mathrm{C} + e^+ + \nu_e \\ {}^{13}_6\mathrm{C} + {}^1_1\mathrm{H} &\rightarrow {}^{14}_7\mathrm{N} + \gamma \\ {}^{14}_7\mathrm{N} + {}^1_1\mathrm{H} &\rightarrow {}^{15}_8\mathrm{O} + \gamma \\ {}^{15}_8\mathrm{O} &\rightarrow {}^{15}_7\mathrm{N} + e^+ + \nu_e \\ {}^{15}_7\mathrm{N} + {}^1_1\mathrm{H} &\rightarrow {}^{12}_6\mathrm{C} + {}^4_2\mathrm{He}. \end{aligned} \tag{198}$$

$$\tag{199}$$

Approximately 0.04% of the time, ${}^{15}_7\mathrm{N} + {}^1_1\mathrm{H} \rightarrow {}^{16}_8\mathrm{O} + \gamma$, leading to a second chain in the CNO cycle

[41] Subscript e is used to denote electron and positron neutrinos (ν_e and $\bar{\nu}_e$, respectively). Other flavours of neutrinos exist, including muon neutrinos and tauon neutrinos, neither of which are relevant for the reactions mentioned here.

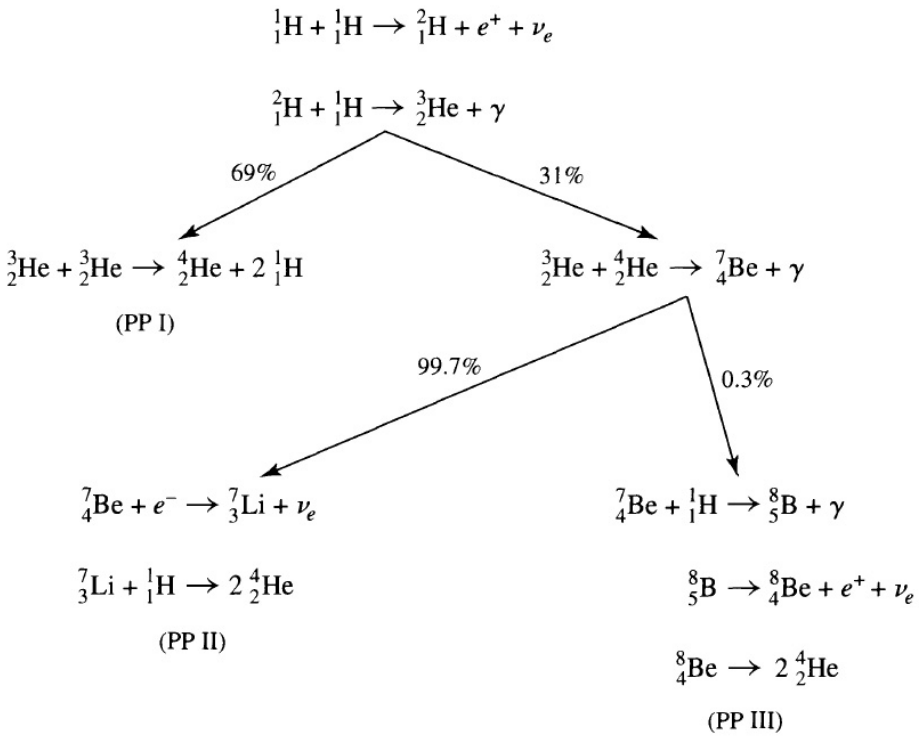

FIG. 107.— The p-p chain in schematic form, with branching ratios appropriate for the interior of the Sun. From Carroll & Ostlie (2006), their Fig. 10.8.

$$
\begin{aligned}
{}^{15}_{7}\mathrm{N} + {}^{1}_{1}\mathrm{H} &\rightarrow {}^{16}_{8}\mathrm{O} + \gamma \\
{}^{16}_{8}\mathrm{O} + {}^{1}_{1}\mathrm{H} &\rightarrow {}^{17}_{9}\mathrm{F} + \gamma \qquad (200)\\
{}^{17}_{9}\mathrm{F} &\rightarrow {}^{17}_{8}\mathrm{O} + e^+ + \nu_e \\
{}^{17}_{8}\mathrm{O} + {}^{1}_{1}\mathrm{H} &\rightarrow {}^{14}_{7}\mathrm{N} + {}^{4}_{2}\mathrm{He}.
\end{aligned}
$$

(201)

This cycle can be summarized as ${}^{12}_{6}\mathrm{C} + 4{}^{1}_{1}\mathrm{H} \rightarrow {}^{12}_{6}\mathrm{C} + {}^{4}_{2}\mathrm{He} + 2e^+ + 2\nu_e + 3\gamma$ and ${}^{12}_{6}\mathrm{C} + 6{}^{1}_{1}\mathrm{H} \rightarrow {}^{14}_{7}\mathrm{N} + {}^{4}_{2}\mathrm{He} + 3e^+ + 3\nu_e + 5\gamma$, where the catalysts have been left in the reactions (note that 3γ is shorthand for three photon releases, not three of the same photon). During the reaction cycle heavy elements capture protons, and then beta decay, increasing their total mass until a tightly-bound alpha particle buds off. The rate limiting step is ${}^{14}_{7}\mathrm{N} + {}^{1}_{1}\mathrm{H} \rightarrow {}^{15}_{8}\mathrm{O} + \gamma$, making it so that most of the CNO in the star will be in the form of ${}^{14}\mathrm{N}$.

The energy generation rate is given by

$$\epsilon_{CNO} \approx 8.24 \times 10^{-31} \rho X X_{CNO} T_6^{19.9} \mathrm{Wkg}^{-1}, \qquad (202)$$

in the range of $T = 1.5 \times 10^7$ K ($\epsilon_{cno} \propto T^{19.9}$). Since Eqn. 202 has a much greater temperature dependence than Eqn. 197, we expect low-mass stars to be p-p chain dominated, while higher mass stars are CNO dominated (the transition occurs near the mass of the Sun).

4.21.2. *What about fusion of heavier nuclei?*

Useful primarily for post-merger evolution of stars, these reactions turn atoms with $Z > 1$ into heavier masses. The triple-alpha (3α) process goes like

$$
\begin{aligned}
{}^{4}_{2}\mathrm{He} + {}^{4}_{2}\mathrm{He} &\leftrightarrow {}^{8}_{4}\mathrm{Be} \qquad (203)\\
{}^{8}_{4}\mathrm{Be} + {}^{4}_{2}\mathrm{He} &\rightarrow {}^{12}_{6}\mathrm{C} + \gamma. \qquad (204)\\
& \qquad (205)
\end{aligned}
$$

The first step of the process is highly endothermic (${}^{8}_{4}\mathrm{Be}$ is 100 keV more massive than two ${}^{4}_{2}\mathrm{He}$), and therefore ${}^{8}_{4}\mathrm{Be}$ needs to be struck by another alpha particle before it decays back by the reverse reaction. As a result, this process

is approximately a three-body interaction. At the relevant temperature of 10^8 K (this process becomes prominent at this step), $\epsilon_{3\alpha} \propto T_8^{41.0}$, a dramatic temperature dependence.

If both carbon-12 and oxygen-16 exist alongside alpha particles, the reactions ${}^{12}_{6}\mathrm{C}+{}^{4}_{2}\mathrm{He} \rightarrow {}^{16}_{8}\mathrm{O}+\gamma$ and ${}^{16}_{8}\mathrm{O}+{}^{4}_{2}\mathrm{He} \rightarrow {}^{20}_{10}\mathrm{Ne}+\gamma$. At still higher temperatures, two ${}^{12}_{6}\mathrm{C}$ atoms can fuse together to form ${}^{16}_{8}\mathrm{O}$, ${}^{20}_{10}\mathrm{Ne}$, ${}^{23}_{11}\mathrm{Na}$, ${}^{23}_{12}\mathrm{Mg}$ and ${}^{24}_{12}\mathrm{Mg}$, while two ${}^{16}_{8}\mathrm{O}$ atoms can fuse together to form ${}^{24}_{12}\mathrm{Mg}$, ${}^{28}_{14}\mathrm{Si}$, ${}^{31}_{15}\mathrm{P}$, ${}^{31}_{16}\mathrm{S}$ and ${}^{32}_{16}\mathrm{S}$, among other things. Some of these reactions are endothermic, and therefore have to be powered by gravitational collapse or other exothermic reactions. Fusion alone may continue to be energetically favourable until ${}^{56}\mathrm{Fe}$ is fused (see Fig. 108). Past this point, fusion is energetically unfavourable, and therefore must come at the expense of loss of gravitational potential energy - in stars this is what eventually causes supernovae.

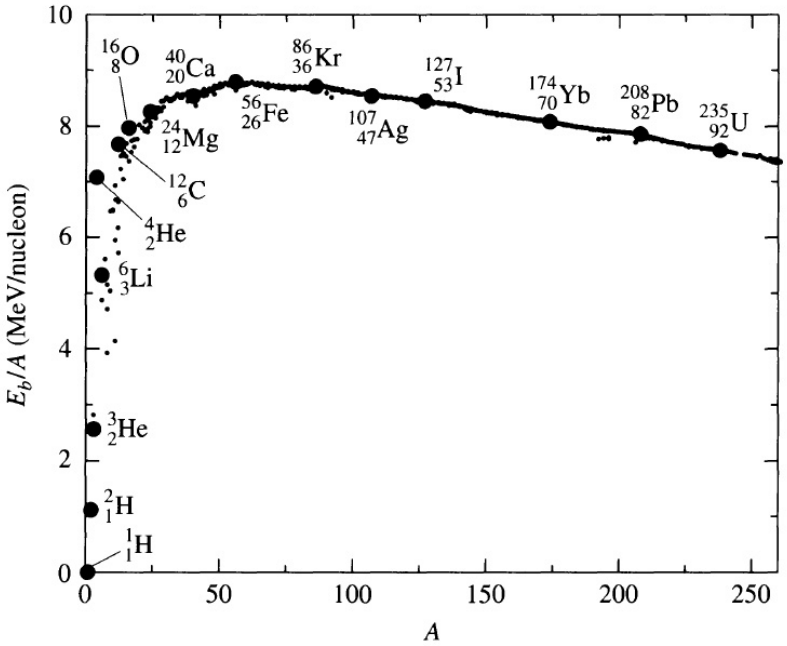

FIG. 108.— The binding energy per nucleon (total binding energy divided by the number of nucleons) for various isotopes arranged by mass number A. Note the broad peak with a maximum at ${}^{56}\mathrm{Fe}$. From Carroll & Ostlie (2006), their Fig. 10.9.

4.21.3. *Why is quantum tunnelling important to nuclear fusion in stars?0*

The potential well of a nucleus-nucleus pair (in the CM frame using reduced masses, of course) consists of an electromagnetic repulsion term and a strong nuclear force binding term. Classically, for two nuclei to bind requires a kinetic (i.e. thermal) energy equal to the potential $U(r)$ at the EM-nuclear turnover - this gives the temperature required as $T_{classical} = Z_1 Z_2 e^2/(6\pi\epsilon_0 k_B r)$. Assuming this radius is on the order of 1 fm, and the nuclei are both protons, this gives 10^{10} K , three orders of magnitude higher than the temperature at the centre of the Sun. Clearly, quantum mechanical tunneling is required. A first order approximation for the temperature at which quantum tunneling becomes prominent can be made by assuming that $r = \lambda$, the de Broglie wavelength, and $K = p^2/2\mu = h^2/(2\mu\lambda^2)$ (where μ is the reduced mass), and then performing the same estimate as above for the classical temperature. This gives $T_{quantum} = Z_1^2 Z_2^2 e^4 \mu/(12\pi^2\epsilon_0^2 h^2 k_B)$, which corresponds to 10^7 K for hydrogen fusion, the correct order of magnitude for the central temperature of the Sun.

4.21.4. *What is the Gamow peak?*

Let us consider a Maxwellian gas where the average kinetic energy is non-relativistic. The number of incident particles that will cause fusion on a target nucleus over an interval dt is given by [42]

$$dN_E = \sigma(E) v(E) n_{iE} dE dt, \tag{206}$$

where $v(E) = \sqrt{2E/\mu}$ and $n_{iE}dE$ is the number density of incident particles with energy E to $E + dE$. Note that $n_{iE} = \frac{n_i}{n} n_E dE$ if we assume that whether or not a particle is incident is independent of its energy ($n_i/n < 1$ if the

[42] We will subsume the probability that an incident particle fuses onto a target into $\sigma(E)$.

incident and target particles are not of the same species). Integrated over all energies, and assuming a density n_x of targets per unit volume, we obtain rate r_{ix}:

$$r_{ix} = \int_0^\infty n_x n_i \sigma(E) v(E) \frac{n_E}{n} dE. \tag{207}$$

To obtain what $\sigma(E)$ is, we note that it must at least be proportional to the size of the target nucleus, and the tunneling probability through the potential well. The size should be proportional to the square of the de Broglie wavelength, so $\sigma(E) \propto \lambda^2 \propto 1/E$ (for non-relativistic incident nuclei). The tunneling probability, on the other hand, can be calculated using the WKB approximation and an incident plane wave substituting a particle with constant energy. This gives $\sigma \propto e^{-bE^{-1/2}}$, where $b = \frac{Z_1 Z_2 e^2}{4\pi\epsilon_0} \frac{\pi\sqrt{2\mu}}{h}$. Combining our two results, we obtain $\sigma(E) = S(E) e^{-bE^{-1/2}}/E$, where $S(E)$ can be assumed a constant if it is slowly varying in the range of energies we consider. Putting everything together, and assuming a Maxwellian gas, we obtain

$$r_{ix} = \left(\frac{2}{k_B T}\right)^{3/2} \frac{n_i n_x}{(\mu_m \pi)^{3/2}} \int_0^\infty S(E) e^{-bE^{-1/2}} e^{-E/k_B T} dE. \tag{208}$$

The $e^{-bE^{-1/2}} e^{-E/k_B T}$ in Eqn. 208 gives the solid line in Fig. 109; the peak of that curve indicates there is an ideal energy at which fusion is most efficient, due to a combination of higher energy incident particles being more effective at fusion, and these particles also being more rare. This peak is known as the Gamow Peak, and has value $E_0 = (bkT/2)^{2/3}$. Since most fusion occurs for inbound particles with energies near the Gamow Peak, we generally assume that $S(E) \approx S(E_0)$.

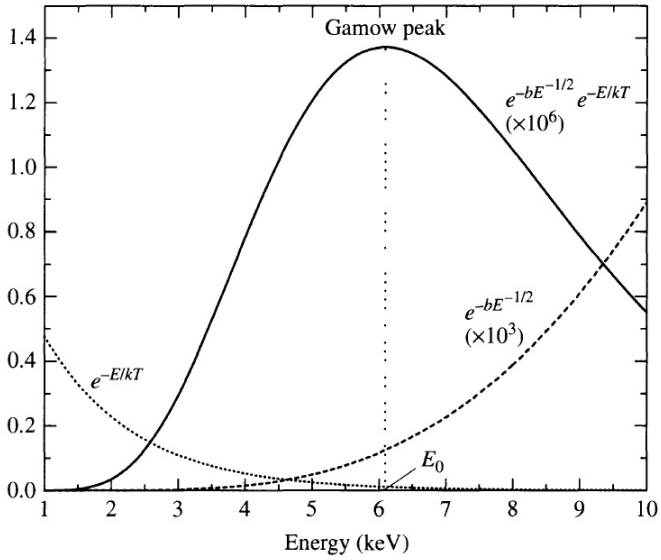

FIG. 109.— The curve $e^{-bE^{-1/2}} e^{-E/k_B T}$, plotted alongside $e^{-E/k_B T}$ and $e^{-bE^{-1/2}}$. The Gamow Peak is indicated. The $e^{-bE^{-1/2}} e^{-E/k_B T}$ and $e^{-bE^{-1/2}|}$ have been multiplied by 10^3 and 10^6, respectively, to help illustrate their functional forms. From Carroll & Ostlie (2006), their Fig. 10.6.

Two complicating factors will affect fusion rates. Firstly $S(E)$ is usually slowly varying, but does at certain energy levels contain resonances. Additionally, the potential well must be corrected for electron screening, which reduces the effective charge of the nuclei.

In short, the Gamow Peak is a peak in the distribution of energies of fusing nuclei within a star, and most nuclei that fuse have energies within an order of magnitude of the Gamow Peak. The peak exists because higher energy nuclei are more likely to tunnel ($e^{-bE^{-1/2}}$), but fewer of them exist ($e^{-E/k_B T}$).

5. MATH AND GENERAL PHYSICS (INCLUDES RADIATION PROCESSES, RELATIVITY, STATISTICS)

5.1. *Question 1*

QUESTION: Draw the geometry of gravitational microlensing of one star by another, and estimate the angular displacement of the background star's image.

This is from Carroll & Ostlie (2006) pg. 1130 - 1133, Dyer (2010c) and Schneider (2006), Sec. 2.5.1.

Since we are contemplating stars, we may use the Swarzschild metric with impunity (since a spherically symmetric, static solution (meaning the solution is independent of time) is a good approximation in this case). The Schwarzschild metric is

$$ds^2 = (1 - \frac{2m}{r})dt^2 - (1 - \frac{2m}{r})^{-1})dr^2 - r^2(d\theta^2 + \sin^2\theta d\phi^2). \tag{209}$$

Since $ds^2/d\tau^2 = L$, where L is the Lagrangian and over-dots mean differentiation with respect to τ, an affine parameter (we may use the proper time for simplicity, since all affine parameters scale with each other). We may now convert the above equation as a Lagrangian. We note that due to symmetry of the metric we can without loss of generality define $\theta = \pi/2$ (for all time) for our geodesic. This means that lensing for a point mass is completely fixed on a plane, and what Schneider (2006) does in Sec. 2.5.1 is superfulous. Following a lengthy derivation from the Euler-Lagrange equations of motion (see my other qual notes, or Dyer (2010c)), we obtain the deflection angle of a photon due to a Schwarzschild metric in the small deflection limit

$$\phi_d \approx \frac{4m}{a}, \tag{210}$$

or $\phi_d \approx 4GM/ac^2$ if we rewrite Eqn. 210 in SI units. a here is the impact parameter.

We will now switch to the notation of Carroll & Ostlie to determine the angular displacement of a star being lensed by another star (so that what is written matches with Fig. 110). The lensing geometry is shown in Fig. 110, where the line connecting O to L is known as the "optical axis". The deflection angle $\phi \approx 4GM/r_0c^2$. Suppose we also knew d_S and d_L from, say, the luminosity distances of the two stars[43]. Also assume we use the small angle approximation throughout, meaning that $\tan x = \sin x = x$. The distance between the source S's real position and the optical axis is βd_S, while the apparent distance is θd_S. Their difference is approximately equal to $\phi(d_S - d_L)$, which allows us to equate:

$$(\theta - \beta)d_S = \frac{4GM}{r_0c^2}(d_S - d_L) \tag{211}$$

we eliminate r_0 by noting $r_0 = \theta d_L$. This gives us

$$\theta^2 - \beta\theta = \frac{4GM}{c^2}\frac{d_S - d_L}{d_S d_L}. \tag{212}$$

Lastly, from the quadratic equation, we find two solutions for θ

$$\theta_{1,2} = \frac{1}{2}\left(\beta \pm \sqrt{\beta^2 - \frac{16GM}{c^2}\left(\frac{d_S - d_L}{d_S d_L}\right)}\right) \tag{213}$$

There are then always two images for one source (one on each side of the lens). Note that $\beta = \theta_1 + \theta_2$. Conversely, if we observationally knew θ_1 and θ_2, we can obtain

$$M = -\frac{\theta_1\theta_2 c^2}{4G}\left(\frac{d_S d_L}{d_S - d_L}\right) \tag{214}$$

In general, only one plane can be drawn through three points, so there are only two images (because light being emitted in other directions will end up converging at the location of other observers). There is one case in which this is not true, however: when $\beta = 0$. In this case, we observe an Einstein ring, with

$$\theta_E = \sqrt{\frac{4GM}{c^2}\frac{d_S - d_L}{d_S d_L}} \tag{215}$$

[43] L, S and O are not precisely aligned, but in the small angle limit, the distances to them are approximately the projected distances to them along the optical axis (i.e. D_S, the true distance to S, is $d_S/\cos\beta \approx d_S$).

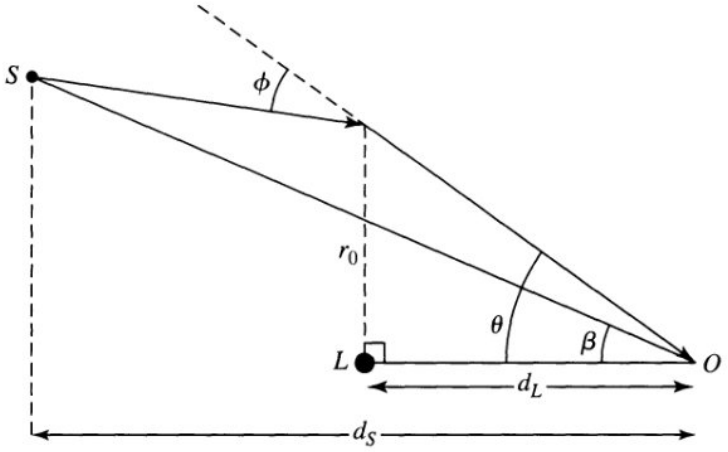

FIG. 110.— Geometry for a gravitational lens. From Carroll & Ostlie (2006), their Fig. 28.35.

5.1.1. *What happens if lensing is not done by a spherically symmetric object on a point source?*

If the background object is not a point source, we obtain both distortion and magnification of the image due to differential lensing. See Fig. **??**. If the lens is also not spherically symmetric, it can be approximated as a distribution of point masses, each with a different distance to the lens. The resulting image is highly dependent on both geometry and the structure of the source and lens.

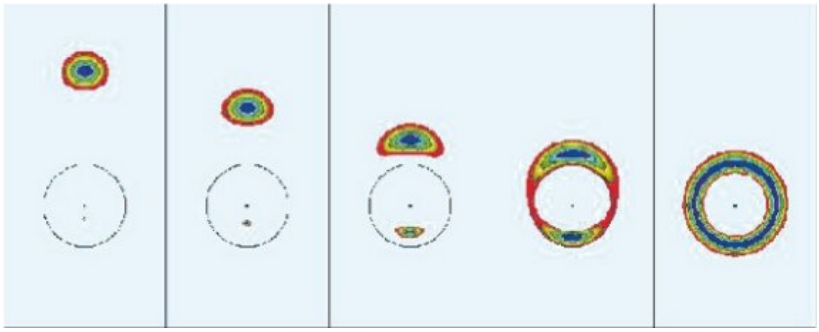

FIG. 111.— Image of a circular source with a radial brightness profile (blue indicates brighter) being lensed by a spherically symmetric point source. Note that if the source is distant from the lens, the two images have different sizes. If the source is directly behind the lens, an Einstein ring is formed. From Schneider (2006), his Fig. 2.25.

5.1.2. *What is lensing used for?*

Lensing provides a useful test for general relativity. Even if we assume that photons can be deflected by Newtonian gravity (treat the photon as a light-speed particle with a small mass) the resulting deflection from classical mechanics is $2m/a$, not $4m/a$. Lensing is also the only means by which we can find dark matter in the absence of nearby gravitationally-affected luminous matter (ex. a completely star-less dark matter halo could only be detected by lensing). In cases where dark matter mass can also be detected kinematically by looking at luminous matter, lensing serves as a good consistency check. Historically, lensing has been used to attempt to determine the nature of dark matter (revealing that MACHOs do not make up the majority of dark matter), and determine the masses of galaxy clusters and therefore probe the matter power spectrum.

5.2. *Question 2*

QUESTION: A two-element interferometer consists of two telescopes whose light is combined and interfered. Sketch the response of such an interferometer to a nearby red giant star, as a function of the (projected) separation between the two telescopes. The red giant subtends one-fiftieth of an arc second on the sky, and the telescope operates at a wavelength of 2 microns.

This information is from AstroBaki (2010) and Jackson (2008).

An interferometer is a set of two or more detectors collectively used to measure the interference pattern produced by the angular brightness distribution of a very distant source at a given frequency range. They are commonly used in the optical and radio.

The operational principle behind an interferometer is the Young's double slit experiment (see Jackson (2008) for details). A point source creates a fringe pattern with spacing λ/d, where d is the slit separation. If a number of point sources were attached together, we would obtain the sum of a number of fringe patterns with the same frequency and a slight phase offset (due to the difference in position of each point). These fringe patterns can destructively interfere if the sources were sufficiently offset. Lastly, decreasing the slit separation decreases the frequency λ/d. This thought experiment tells us, essentially, that the interference pattern is a Fourier transform of the source: a point source (delta function) gives us the same amplitude in the fringe pattern no matter what d (analogous to spatial frequency) value is used, while an extended source gives us a sharp peak.

Let us switch gears slightly (and actually answer the question). Consider a two-element interferometer, as seen in Fig. 112. The telescopes are aimed along unit vector $\vec{s}$ at some target. In general, the target will have some spatial extent σ (measured in radians), which we will parameterize using $\vec{\sigma}$, which describes the location, from the part of the object $\vec{s}$ points to, of the piece of the sky we are considering. In the limit in which the sky is small, the direction of incoming plane waves from the source will be $\vec{s} + \vec{sigma}$.

The response, i.e. the specific flux received as a function of viewing angle, baseline, etc., of a two-element interferometer is

$$R = \int I_\lambda \exp(ik\vec{B} \cdot (\vec{s} + \vec{\sigma}) d\sigma \tag{216}$$

where $k = \frac{2\pi}{\lambda}$. We note that the projection of $\vec{B}$, $\vec{b}$, sits on a plane that is parallel to the plane of $\vec{\sigma}$, and so

$$R = \exp(ik\vec{B} \cdot \vec{s}) \int I_\lambda \exp(ik\vec{b} \cdot \vec{\sigma}) d\sigma. \tag{217}$$

The simplest thing we can pass through is a plane wave from a point source, in which case $R = e^{ik\vec{B}\cdot(\vec{s}+\vec{\sigma_0})} I_{\nu,0}$, where $\vec{B} \cdot (\vec{s} + \vec{\sigma_0})$ is just the dot product between the baseline and the direction to the point source.

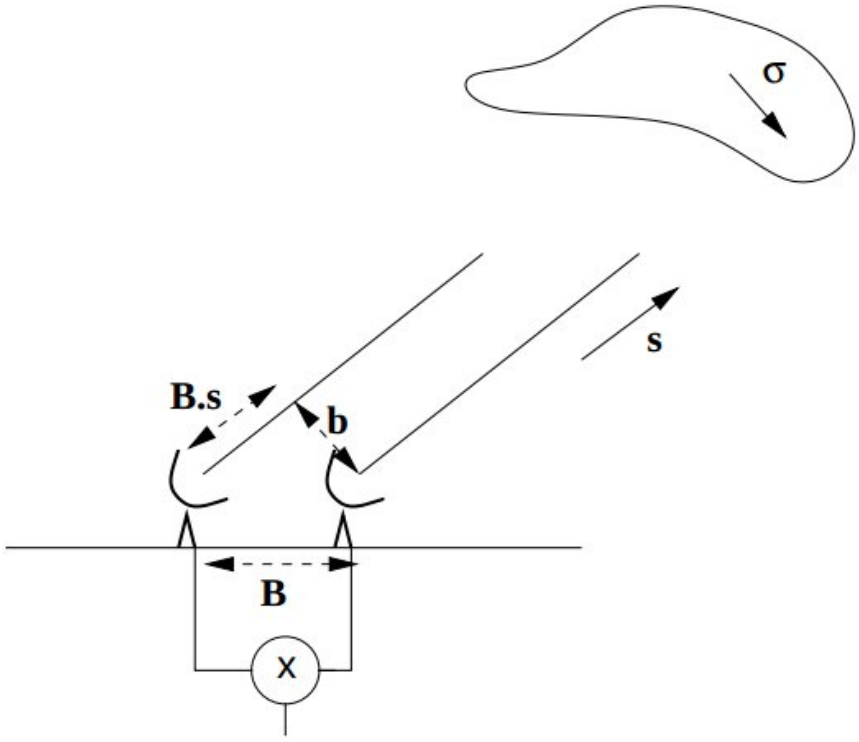

FIG. 112.— Schematic representation of an interferometer. Note that in reality the two-element interferometer is in 3D, which means $\vec{b}$ is not necessarily aligned with $\vec{\sigma}$. From Jackson (2008).

We can now make one more shift, from physical to (u, v) space. Let us assume that $\vec{b} = u\hat{x} + v\hat{y}$, and $\sigma = \sigma_x\hat{x} + \sigma_y\hat{y}$. Then $\vec{b} \cdot \vec{\sigma} = u\sigma_x + v\sigma_y$ and Eqn. 217 becomes

$$R(u, v) = \exp(ik\vec{B} \cdot \vec{s}) \int \int I_\lambda \exp(ik(u\sigma_x + v\sigma_y)) d\sigma_x d\sigma_y. \tag{218}$$

This equation is a bit different than the equation on page 6 of Jackson, because $\vec{b}$ remains in physical units. Generally, we would switch to units of "wavelengths", $u = B/\lambda$. So long as $\sigma << 1$ radian, then, we are simply picking out a particular value (u, v) in the 2D Fourier transform of the sky (if $\sigma \sim 1$ we would have to perform a 3D Fourier transform). This is a variant of the Van Cittert-Zernicke theorem.

Take $v = 0$ (without loss of generality, since the star is a circle), $u = b = B \sin\theta$ ($\sin\theta = \vec{B} \times \vec{b}$ and is a constant) and assume $\vec{s}$ is straight up (otherwise $\exp(ik\vec{B} \cdot \vec{s})$ remains as a multiplicative constant), which also gives us $B = b$. $I_\lambda(\sigma_x, \sigma_y)$ is a Gaussian, $I_0 \exp\left(\frac{-\sigma_x^2 - \sigma_y^2}{\sigma_c^2}\right)$, where $\sigma_c = 4.848 \times 10^{-8}$ radians. $\lambda = 2 \times 10^{-6}$ m. Our integral then becomes (neglecting constants because we only desire the functional form)

$$\begin{aligned} R(B) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I_0 \exp\left(\frac{-\sigma_x^2 - \sigma_y^2}{\sigma_c^2}\right) \exp(\frac{2\pi i}{\lambda}(B\sigma_x)) d\sigma_x d\sigma_y. \\ &\propto \int_{-\infty}^{\infty} \exp\left(\frac{-\sigma_x^2}{\sigma_c^2}\right) \exp(\frac{2\pi i}{\lambda}(B\sigma_x)) d\sigma_x \\ &\propto \int_{-\infty}^{\infty} \exp\left(-\frac{1}{\sigma_c^2}\left[(\sigma_x - i\pi\sigma_c^2 B/\lambda)^2 + \frac{\pi^2}{\lambda^2}\sigma_c^4 B^2\right]\right) d\sigma_x \\ &\propto \exp\left(-\frac{\pi^2}{\lambda^2}\sigma_c^2 B^2\right) \end{aligned} \tag{219}$$

We have used the fact that the integral of a Gaussian is a constant, and in step three did some algebraic gymnastics using $(\alpha - \beta)^2 = \alpha^2 - 2\alpha\beta + \beta^2$. By varying B, we are tracing the 2D Fourier transform of the star (the FT of a Gaussian is a Gaussian!). See Fig. 113.

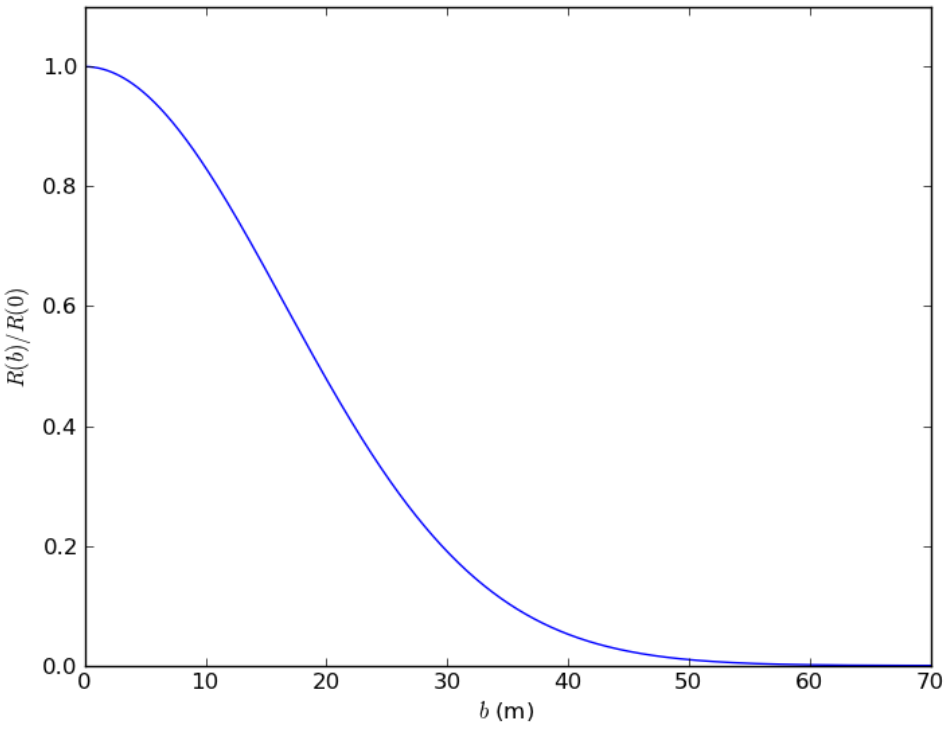

FIG. 113.— Plot of Eqn. 219.

It should be noted that the physical situation described above has actually been done before: Burns et al. (1997) performed optical aperture synthesis observations of Betelgeuse in 1995.

5.2.1. *Complications?*

One implicit assumption we have made is that the telescopes have infinite primary diameters. If the diameter were finite, like with real-life telescopes, we would have to multiply our FT with the Airy disk of a single telescope. All our

interferometry information still exists, however, and our analysis is still accurate. We would simply have an additional source of noise.

Another issue is that for large delays ($\vec{B} \cdot \vec{s}$ in Fig. 112), the phase shift from different wavelengths is different. The largest delay occurs at largest σ, which translates to a narrowing of the field of view. This issue is resolved by using narrow bandwidths, and using multiple bands to generate a large effective $\Delta\lambda$ if we wish to increase S/N (letting in a larger $\Delta\lambda$ increases the photon count).

5.2.2. *What is aperture synthesis?*

Using Eqn. 218 we could determine that two point sources with a separation x radians will generate a series of stripes separated $1/x$ wavelengths (see the definition of wavelengths, earlier) apart on the uv plane. Now, suppose we could determine all baselines (i.e. all values of u from 0 to λK). We would then have information about the 2D Fourier transform up to wavelength K, which, as we just noted, corresponds to an angular size $1/K = \lambda/B$ on the sky. This means means that for interometers the effective resolution limit is

$$\theta \approx \frac{\lambda}{B}. \tag{220}$$

By an identical argument, the minimum separation is equivalent to the field of view. We now see that the various baselines are simply sampling points (u, v) for reconstructing the 2D Fourier transform of the image on the sky. We therefore want to sample many points.

Practically, telescope design prevents us from simply putting down an enormous number of telescopes to produce a huge variation of (u, v) (though see the Allen array). The Very Large Array (VLA), for instance, has a y-shaped structure along which telescopes can be moved to generate multiple baselines. If the Earth did not rotate, we would only be able to obtain a y-shape in uv space. The Earth does rotate, however, which rotates the image of the object in the sky, and as a result we can actually sample multiple tracks in uv space (Fig. 114). This technique is known as aperture synthesis, the building of an effective aperture thorough the repositioning of multiple apertures.

5.2.3. *What is interferometry used for?*

Interferometers can be used to obtain data of objects at very small angular separations, and therefore generate extremely high-resolution images in the sky. This is particularly useful in situations where the beam of the telescope is enormous (ex. in the radio). Interferometers can also be used to provide high-precision absolute positions of objects in the sky (Kitchin 2009, pg. 454 - 455).

5.2.4. *Name some interferometers.*

In the radio, there is VLA, Giant Metre-wave Radio Telescope (GMRT), Low Frequency Array (LOFAR), and the Atacama Large Millimetre Array (ALMA) (the Square-Kilometre Array (SKA) will be built in the next few decades). In the optical, there are the CHARA (Center for High Angular Resolution Astronomy) Array and the COAST (Cambridge Optical Aperture Synthesis Telescope) array, which also works in the near-IR (Kitchin 2009, pg. 455, Burns et al. 1997).

5.3. *Question 3*

QUESTION: Define and describe the 'diffraction limit' of a telescope. List at least three scientifically important telescopes which operate at the diffraction limit, and at least three which do not. For the ones which do not, explain why they do not. In both categories include at least one telescope not operating in optical/near IR wavelengths.

This is cribbed from Emberson (2012) and Kitchin (2009), unless otherwise cited.

Consider a circular aperture of radius R with a lens attached to it (a focusing mirror will have an identical effect). Light emitted from distant objects can be approximated by plane waves impinging on the aperture. When the plane wave crosses through the aperture, the plane wave is "cut" so that only a section of plane wave, in the shape of the aperture, is allowed through (imagine a two-dimensional Heaviside function where the amplitude is unity when $r < R$). We are in the Fraunhofer limit since we are focusing our image with a lens (we would also be if $R^2/L\lambda << 1$, where L is the distance between the aperture and the image). Therefore, at the focal plane, Fourier optics demands that the image be the Fourier transform of the image at the aperture. The FT of the aperture is a diffraction pattern known as an Airy disk

$$I(\theta) = \frac{\pi^2 R^4}{m^2}[J_1(2m)]^2 \tag{221}$$

where $m = \frac{\pi R}{\lambda}\sin\theta$, J_1 is a Bessel function of the first kind of order unity, and θ is the angle from the normal to the aperture (see Fig. 1.27 of Kitchin), but is equivalently the angle on the sky from the centre of the image. The shape

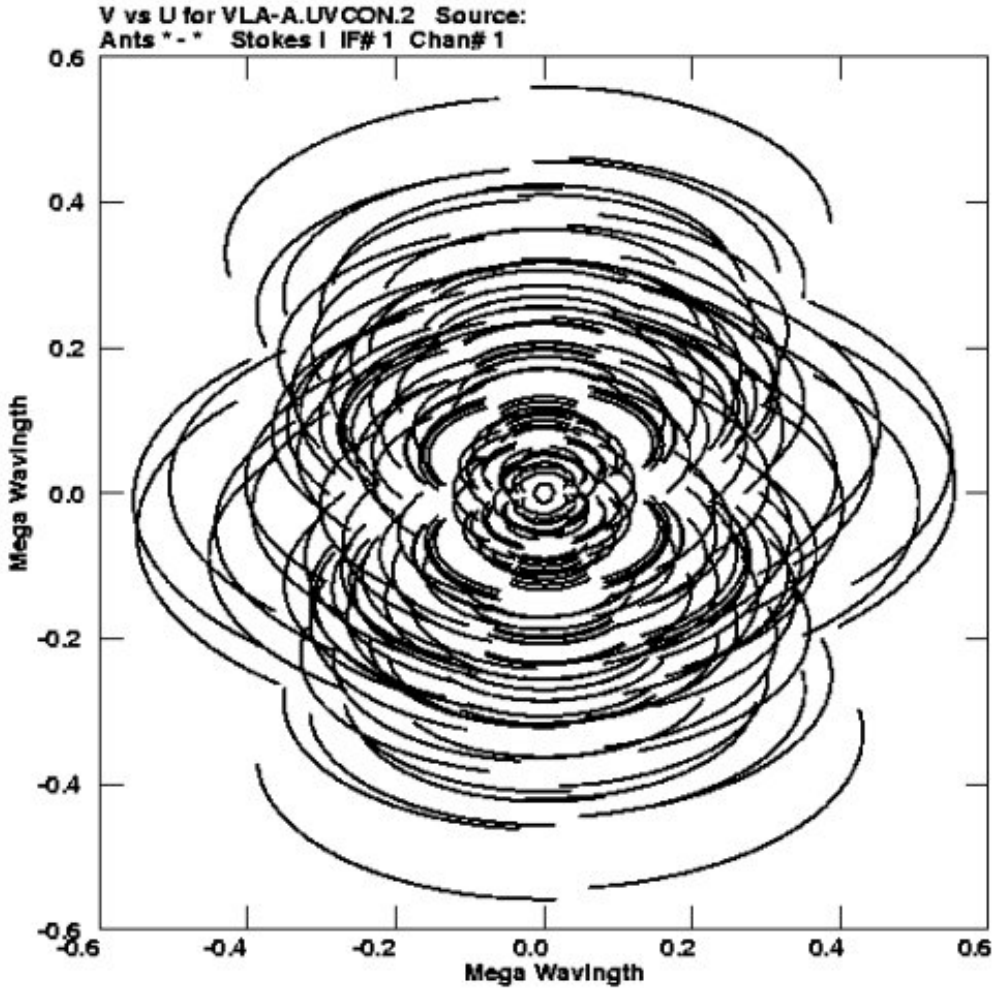

FIG. 114.— Example of aperture synthesis for the VLA. Each arc in uv space is due to a single telescope separation combined with the rotation of the Earth to produce different (u, v) values. In this way, coverage of uv space is greatly improved. From Jackson (2008), his Fig. 7.

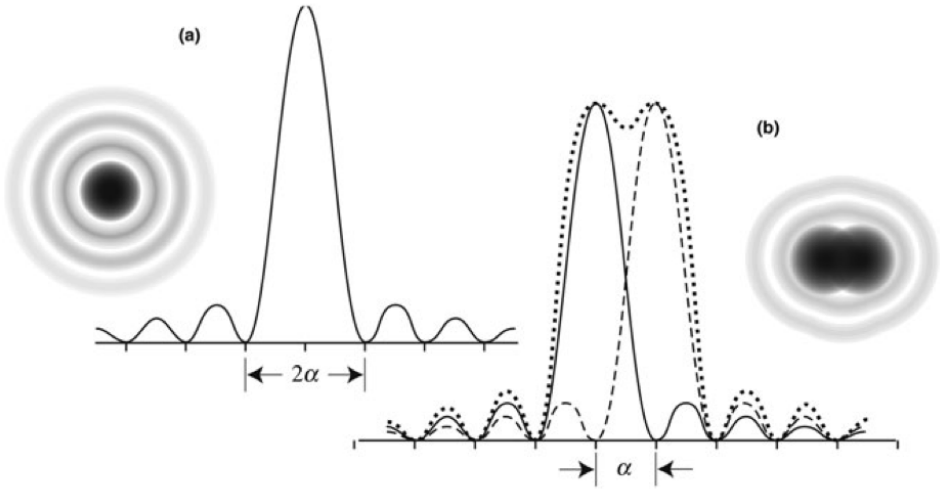

FIG. 115.— Left: a single Airy disk. Right: two Airy disks superimposed, with the centre of one at the first minimum of the other. From Emberson (2012).

of this function can be seen in Fig. 115. The first minimum of this function occurs at

$$\alpha \approx \sin(\alpha) = \frac{1.220\lambda}{2R} = \frac{1.220\lambda}{D} \tag{222}$$

We note that the plane wave is the FT of a distant point source - if we had an infinite sized aperture and an infinite lens, our focal plane image would be the FT of an FT (i.e. roughly equivalent to an FT and then an inverse FT), convolved with the FT of the aperture. If two objects were to have an angular separation less than α, their Airy disks would begin to merge (Fig. 115). This blend is still distinguishable for $\theta < \alpha$, but this becomes very difficult, especially if one source is much less bright than the other. This is why we consider Eqn. 222 the diffraction limit (i.e. a resolution limit) of a telescope (also known as the Rayleigh criterion).

Earthbound telescopes often have a much more stringent limit set on it by turbulence in the atmosphere. Temperature gradients in the Earths atmosphere caused by turbulence create compressions and rarefactions in the air and correspondingly higher and lower indices of refraction. These regions in the air retard some sections of a plane wavefront while allowing other sections to progress forward. For telescopes operating past the Fried coherence length (the distance over which the phase difference is one radian) $r_0 \approx 0.114 \left(\frac{\lambda \cos z}{550}\right)^{0.6}$ m (where z is the angle to the zenith), the atmosphere sets up an seeing limit (Hickson 2002)

$$\alpha_{\text{eff}} \approx \frac{\lambda}{r_0} \tag{223}$$

which limits the resolution of the telescope (For telescopes smaller than this limit, atmospheric turbulence creates scintillation, or twinkling, hence why the stars twinkle to the naked eye.)

There can also be instrumental reasons for why a telescope is not operating at its diffraction limit. These include mirror/lens imperfections and inherent design features, such as the ones on the Chandra space telescope.

Three telescopes that are operating at the diffraction limit are:

- **Giant Meterwave Radio Telescope** (GMRT), which has an effective mirror diameter (since GMRT is an array) of 25 km, operates on the GHz frequency and has a ~ 2" resolution.
- The **Hubble Space Telescope** (HST), which has an aperture diameter of 2.4 m, works in the optical (~ 500 nm) and has a resolution limit of about 0.052".
- The **Spitzer Space Telescope**, which has an aperture diameter of 0.85 m, works in the near-IR (8 μm), and has a resolution limit of about 2".

In general, space telescopes and radio telescopes operate at the diffraction limit. Three telescopes that do not operate at the diffraction limit are:

- **Chandra X-ray Observatory**, which has an aperture diameter of 1.2 m, looks at x-rays with wavelength 1 nm and has an angular resolution limit of 0.5". This is due to imperfections in Chandra's concentric mirror shells.
- **W.M. Keck Observatory**, which has a diameter of 10 m, looks in the optical ($\sim$500 nm), but is seeing limited to $\sim$1". Adaptive optics decreases this limit almost to the diffraction limit.
- **XMM Newton**, with an effective diameter of 0.74 m, works in the x-ray, and has a resolution limit of about 6".

In general, high-energy telescopes and telescopes on the ground do not operate at the diffraction limit.

5.3.1. *What if the lens or mirror is not perfect?*

This information is from Hickson (2002) and Kitchin (2009), Ch. 1.1.17.2.

A number of optical imperfections can distort images:

- **Spherical aberration**, where different areas of the spherical mirror/lens have different focal points.
- **Chromatic abberation**, where different wavelength light have different focal points.
- **Coma**, where a paraboloidal mirror/lens generates comet-like distortions.
- **Astigmatism**, where bar-like distortions are generated.
- **Distortion**, a variation of magnification across the image.

5.3.2. *How does adaptive optics work? Does it truly create diffraction-limited images?*

Adaptive optics (AO) can correct wavefront distortions that arise from atmospheric turbulence and sharpen blurred images on long exposures. A guide star or laser artificial star is used as a reference to determine atmospheric distortions. A set of small, deformable mirrors or a lenslet array is then employed to correct the atmospheric distortions generated from turbulence. This process is somewhat easier in the near-infrared because of the longer wavelengths involved. As a result, AO have been successful in providing near-diffraction-limited images in the near IR.

Since AO is a complicated system, it generates its own intrinsic errors. Moreover, atmospheric turbulence cannot completely correct for atmospheric turbulence, since there is a delay in image correction, as well as a directional dependence on turbulence. Solutions include more efficient computational algorithms and adding additional guide stars (ex. via multiple lasers).

5.4. *Question 4*

QUESTION: What's the minimum mass of a black hole you could survive a fall through the event horizon without being ripped to shreds? Why would you be ripped to shreds for smaller black holes?

This is from Kissin (2012), with a little help from last year's qualifier notes.

We can approximate the black hole with a Schwarzschild metric. The event horizon is then

$$r_{BH} = \frac{2GM}{c^2} \tag{224}$$

The Newtonian approximation for the tidal force per unit mass is

$$f_t \approx \frac{dg}{dr} L \tag{225}$$

which is a fine approximation so long as $\frac{dg}{dr} >> \frac{d^2g}{dr^2} L$. Calculating this nets us $f_t \approx L\frac{2GM}{r^3}$. If we are at the event horizon of the black hole, then

$$f_t \approx \frac{c^6 L}{4G^2 M^2}. \tag{226}$$

Solving for M, we obtain

$$M = \sqrt{\frac{c^6 L}{4G^2 f_t}} \tag{227}$$

Traditional forms of execution include dismembering an individual through the tidal force of horses pulling in different directions. A very ballpark estimate suggests 100 g's would tear a human limb from limb (horses pulling cannot be more than 10^4 N, which translates, for a 100 kg person, into 100 g's). If we also assume $L \approx 2$ m, then we obtain a mass $M \approx 10^{34}$ kg, or 5×10^3 $M_\odot$.

From Eqn. 226, it is obvious that as mass decreases, the tidal force *increases*. This has to do with the Schwarzschild radius being dependent on mass, and the tidal force being dependent on M^{-3}.

5.4.1. *How are supermassive black holes (SMBHs) formed?*

This is from Wikipedia (2012g) and Carroll & Ostlie (2006), pg. 639.

This is still an open question. The most obvious hypothesis is that black holes of tens or perhaps hundreds of solar masses that are left behind by the explosions of massive stars grow by accretion of matter. It is possible SMBHs could have formed from the collisions of galaxies. They could also have formed out of pre-galactic halos through various instabilities, culminating in the formation of a ~ 10 $M_\odot$ black hole, which then accretes at super-Eddingthon rates until it reaches $\sim 10^5$ $M_\odot$. Black holes could also have been formed from the Big Bang itself.

The primary unsolved issue today is how to feed these black holes until they reach SMBH status. The existence of high-redshift quasars requires that SMBHs be formed soon after the first stars. No method is known to gorge primordial BHs to stimulate such rapid growth.

5.5. *Question 5*

QUESTION: Let's say the LHC produces microscopic black holes in their high energy proton-anti-proton collisions? What will happen to them? Will they destroy the Earth?

This information comes from my own notes, buoyed with wikipedia (2012) and last year's notes on the problem.

If the LHC produces microscopic black holes, they will likely almost instantly decay through Hawking Radiation. Black hole thermodynamics demands that a Schwarzschild black hole radiate like a blackbody with temperature

$$T = \frac{\hbar c^3}{8\pi k_B G M}. \tag{228}$$

This will be stated without proof, but a useful motivation is below. The total radiation from the black hole is then $L = 4\pi R_s^2 \sigma_{SB} T^4$, giving us $L = \frac{2\sigma\hbar^4 c^{12}}{(8\pi)^3 k_B^4 G^2 M^2}$. Since only the rest mass of the BH can provide energy, $L = -c^2 \frac{dM}{dt}$. We can then equate the two, integrating from the time $t = 0$ when the BH had $M = M_0$ to time $t = t_{ev}$ when $M = 0$:

$$\int_{M_0}^{0} M^2 dM = -\int_0^{t_{ev}} dt \tag{229}$$

where we have thrown a lot of constants into K. Integrating this, solving for t_{ev} and plugging in the numbers of constants, we obtain:

$$t_{ev} = 8.4 \times 10^{-17} \text{s}(M/kg)^3 \tag{230}$$

For a 1 TeV black hole, we would obtain $t_{ev} = 5 \times 10^{-88}$ s!

We can make this a bit more difficult by assuming that the black hole can swallow nearby masses to stave off its own death. The accretion rate is then $dM/dt = f\pi R_s^2 \rho v$ (from $\Gamma = n\sigma v$, where Γ is the number of collisions per second), where f is of order unity. We then require, to stave off death:

$$L/c^2 = \frac{2\sigma\hbar^4 c^{10}}{(8\pi)^3 k_B^4 G^2 M^2} = f\pi R_s^2 \rho v. \tag{231}$$

If we then assume $v \approx c$ to minimize ρ, we calculate $\rho = \frac{4\sigma\hbar^4 c^{13}}{(8\pi)^4 k_B^4 G^3 M^3} = 1.8 \times 10^{155}$ kg/m^3 (i.e. much greater than black hole densities)! In order to survive by accreting particles while travelling at the speed of light (any less and the ρ goes up) through pure iron, the black hole would need an rest energy of $\sim 10^{42}$ GeV.

Even if we suppose all our theory above is wrong, and black holes can be produced by the LHC and survive indefinitely, it is still unlikely they would destroy the Earth. The LHC collision of two 7 TeV proton streams is equivalent to a stationary Earth atom being struck by a 10^8 GeV cosmic ray, and the cutoff for cosmic rays is around 10^{11} GeV. We would then expect to see microscopic black holes being produced all the time in our atmosphere. These black holes would be fast-moving, but as they accreted charge drag from electrostatic effects would keep them trapped in massive objects such as the Sun. Unless these black holes could rapidly neutralize (the only known mechanisms for doing this would also cause Hawking radiation), they would have destroyed the Sun and the Earth long ago.

Similar arguments can be made to eliminate the possibility of strangelets or magnetic monopoles destroying the world.

5.5.1. *Can the LHC produce microscopic black holes?*

Not according to GR. The LHC produces an energy of 1 TeV within a length of 10^{-19} m. A black hole with a Schwarzschild radius 10^{-19} m would have a rest energy of 10^{35} GeV, far more than the LHC could produce.

5.5.2. *What if there were more dimensions than 4 to the universe?*

If there were extra spatial dimensions, the number of which is given by d (i.e. $d = 0$ for 4 dimensions), then $T \propto M^{-1/(1+d)}$. The temperature dependence on mass would therefore be shallower and it would take longer for the BH to evaporate. Current experiments with modified gravity at very small scales have shown $d \leq 2$, and even with $d = 2$, our analysis above does not change substantially. The minimum rest energy a black hole must have to stave off thermodynamic death by accreting iron while travelling at the speed of light is $\sim 10^{24}$ GeV.

5.5.3. *Can you motivate why temperature is inversely propotional to mass for black hole thermodynamics?*

This information comes from Wayne Ngan's notes on the problem.

Suppose we have a black hole made of photons, and each photon has wavelength $\lambda = R_s$. The number of photons within the black hole is then

$$N = Mc^2\lambda/hc = \frac{2GM^2}{hc} \tag{232}$$

The entropy of the system is $S = k_B \ln\Omega \approx k_B \ln(N!)$, and if N is large, $\ln(N!) \approx N\ln(N) - N$. For a very rough order of magnitude, say that $\ln(N) - 1 \sim 1$ (we will perhaps be an order of magnitude off). Then $S \sim E/T \sim k_B N$, and $E = Mc^2$, which gives us

$$T \sim \frac{hc^3}{k_B G M} = \frac{\hbar c^3}{8\pi k_B G M} \tag{233}$$

5.6. *Question 6*

QUESTION: How is synchrotron radiation generated? In outline, how can it be removed from Cosmic Microwave Background maps?

This information is from Zhu (2010).

Synchrotron radiation is generated by electrons looping inside magnetic fields. Suppose, without loss of generality, that $\vec{B}$ is pointed along the z-axis. Then any electron caught in the field will travel along a trajectory $\vec{r}(t) = \hat{z}v_z t + \frac{v_{xy}}{\omega_B}(\hat{x}\cos(\omega t) + \hat{y}\sin(\omega t))$, with $\omega_B = \frac{eB}{\gamma mc}$. As this electron loops inside the magnetic field at relativistic speeds, it tends to generate highly beamed emission. This emission comes at the frequency at which the electron orbits along the magnetic field, $\omega = \frac{3}{2}\gamma^3\omega_B \sin\alpha = \frac{3}{2}\gamma^2\omega_L \sin\alpha$, where ω_L is the Larmor frequency, and α is the pitch angle, between the projection of the vector $\vec{v}$ onto the xz or yz plane, and the magnetic field. If we Fourier transform the square of the electric field, we get the emission spectrum. If we then assume a power law distribution with index p (i.e. $n(\gamma) = n_0\gamma^{-p}d\gamma$), the total emission becomes a power law,

$$I_\nu \propto \nu^{-(p-1)/2}. \tag{234}$$

If the region is sufficiently dense with electrons, synchrotron self-absorption sets in, and the low ν end of the spectrum is given a power law $\nu^{5/2}$ (the frequency at which the spectrum transitions back to a power law is dependent on the magnetic field). In galactic emission p is often about 8, giving $I_\nu \propto \nu^{3.5}$. See Fig. 117.

Note that synchrotron is highly polarized. This is because emission is always orthogonal to the direction of the field. This means a measure of synchrotron polarization allows us to measure the direction of the magnetic field (projected onto the sky).

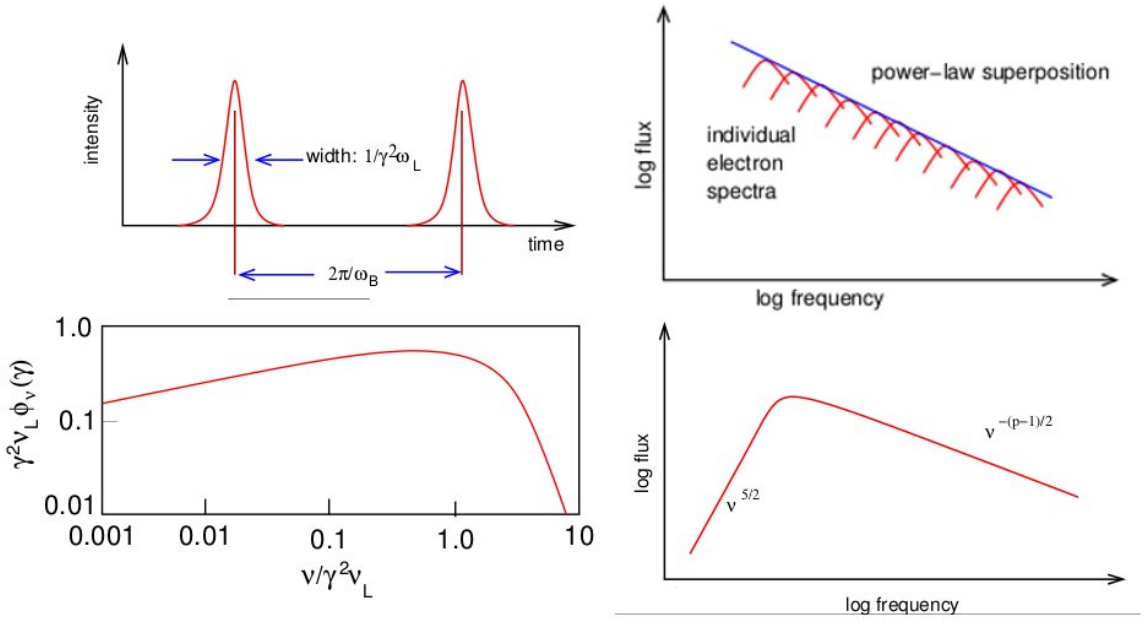

FIG. 116.— Schematic of how synchrotron radiation is derived. Upper left: electric field from a single electron in a magnetic field. Bottom left: Fourier transform of E^2 from a single electron. Upper right: emission from a distribution of electrons (all FT'ed). Bottom right: the full spectrum, taking into account self-absorption at low frequency. From Marleau (2010), Lecture 6.

See Sec. 1.13 for information on how synchrotron radiation is removed from the CMB.

5.6.1. *What is cyclotron radiation?*

This information is from Rybicki & Lightman (1979), pg. 182

The non-relativistic limit of synchrotron radiation is cyclotron radiation, which occurs when the electrons are not

moving relativistically. The emission spectrum from a single electron is then a delta function given by the frequency

$$f = \frac{qB}{mc}. \tag{235}$$

This means that, even moreso than synchrotron radiation, cyclotron radiation traces the distribution of electron energies in the gasy. Cyclotron is polarized.

5.6.2. *What are common sources of synchrotron radiation?*

Essentially anything with a strong magnetic field and free electrons nearby will radiate synchrotron emission. Sources include AGN jets, supernova remnants, pulsar wind nebulae.

5.6.3. *What do the spectra of other major radiation sources look like?*

- **Blackbody radiation**
- **Bremsstrahlung**
- **Thomson scattering**
- **Compton scattering**

5.7. *Question 7*

QUESTION: What are "forbidden lines" of atomic spectra? In what conditions are they observationally important?

This comes from Draine (2011), pg. 60 - 62, and Griffiths (2005) Ch. 9.2.

Spontaneous emission comes in two camps: allowed and forbidden, which simply indicates the strength of the transition. The strength can be calculated using time-dependent perturbation theory (Griffiths (2005), Ch. 9.2). We first determine stimulated emission and absorption by looking at a two-level system being perturbed by the Hamiltonian

$$H' = -qE_0 z\cos(\omega t), \tag{236}$$

recalling that corresponds classically to $V = -qEz$, i.e. if the electric field is uniform, V is a linear function of position, and we orient the z-axis to the movement of the electron. As it turns out, this means that the probability of transitioning between the states, $P_{ab} = P_{ba} \propto \langle\psi_a|\mathbf{r}|\psi_b\rangle$. These are *stimulated* transition probabilities, but by the nature of Einstein's coefficients, it means that the *spontaneous* transition probabilities also depend on $\langle\psi_a|\mathbf{r}|\psi_b\rangle$. Because $\langle\psi_a|\mathbf{r}|\psi_b\rangle$ looks a little like a dipole moment, this is known as electric dipole radiation.

As it turns out, there are large numbers of different states a and b for which $\langle\psi_a|\mathbf{r}|\psi_b\rangle$ is zero. We have made an approximation in Eqn. 236, which is that the electric field is uniform over space. In general $\mathbf{E}(\mathbf{r},t) = \mathbf{E_0}\cos(\mathbf{k}\cdot\mathbf{r} - \omega t)$, and we can expand this out into $\mathbf{E}(\mathbf{r},t) \approx \mathbf{E_0}(\cos(\omega t) + \mathbf{k}\cdot\mathbf{r}\sin(\omega t))$. The first term give electric dipole transitions, while the second gives electric quadrupole transitions and magnetic dipole transitions. These transition probabilities are substantially weaker, and therefore spontaneous emission due to electric dipole, or allowed, transitions, occurs at a much more prolific rate than spontaneous emission due to quadrupole or magnetic dipole, or forbidden, transitions.

Allowed transitions must follow the following rules:

1. Parity must change. Parity is defined as $\sum_i l_i$ (i.e. adding the angular momentum quantum number of all electrons) and can either be even or odd.
2. Change in total angular momentum $\Delta L = 0, \pm 1$.
3. Change in total angular plus spin momentum $\Delta J = 0, \pm 1$, but $J = 0 \to 0$ is forbidden.
4. Only one single electron wavefunction nl changes, with $\Delta l = \pm 1$.
5. Total spin $\Delta S = 0$ (spin cannot change).

Allowed transitions are denoted without square brackets. There are transitions that follow all these rules except the last, and are known as semiforbidden, intercombination, or intersystem transitions. These are denoted with square brackets only on the right. Forbidden transitions break any rules from 1 to 4, and are denoted with square brackets. Roughly speaking, semiforbidden lines are 10^6 times weaker than allowed transitions, and forbidden lines are $10^2 - 10^6$ times weaker than semiforbidden lines. I believe magnetic dipole transitions are stronger than electric quadrupole transitions.

Because they are so much weaker, forbidden lines are easy to collisionally wash out at high densities. In the ISM and around certain stars, however, it is possible to have extremely tenuous gas, and in these gases forbidden transitions become important. Indeed, forbidden line emission that have different critical densities $n_{\rm crit}$ at which they begin to be suppressed can be used as density diagnostics. (Forbidden lines can also be used as temperature diagnostics, which means that density effects must be accounted for when performing calculations. See Sec. 3.17.)

5.7.1. *Derive Einstein's coefficients.*

This comes from Hickson (2002).

Suppose there are two states 1 and 2 in a system at thermodynamic equilibrium; this means that

$$n_1 B_{12} \bar{J} = n_2 (A_{21} + B_{21} \bar{J}), \tag{237}$$

where B_{ab} represents stimulated transition between a and b, A_{ab} represents spontaneous emission, and $\bar{J}$ represents the intensity of photons at the correct frequencies to drive transitions. Since the system is in thermodynamic equilibrium,

$$\frac{n_2}{n_1} = \frac{g_2}{g_1} e^{-h\nu_0/k_B T}, \tag{238}$$

where ν_0 is the wavelength difference between states 1 and 2. Moreover, a system in thermodynamic equilibrium radiates as a blackbody, so $\bar{J} \approx B_\nu(\nu_0)$ (the $\approx$ is because there is some variation in ν over a line profile, but this variation is small). Combining Eqns. 237 and 238 gives us

$$B_\nu(\nu_0) = \frac{A_{21}/B_{21}}{\frac{g_2 B_{21}}{g_1 B_{12}} e^{h\nu_0/k_B T} - 1}, \tag{239}$$

and comparing this to the Planck spectrum gives us:

$$\begin{aligned} g_1 B_{12} &= g_2 B_{21} \\ A_{21} &= \frac{2h\nu^3}{c^2} B_{21}. \end{aligned} \tag{240}$$

B can be calculated from time-dependent perturbation theory.

5.7.2. *How do atoms end up in states where they can only transition back out through a forbidden line?*

Atoms can either be collisionally exited into the state, absorb a photon with the right frequency (both collision and absorption could happen multiple times before the atom returns to the ground state), or cascade down from a higher energy state and become "trapped".

5.8. *Question 8*

QUESTION: What is a polytropic equation of state? Give examples of objects for which this is a very good approximation.

Most of this answer is from Kissin (2012), with help from last year's qualifier notes and Kippenhahn & Weigert (1994).

In general, the equation of state for a gas has the form $P = f(\rho, T)$, i.e. pressure is a function of both density and temperature. This is analytically difficult to work with, since there are essentially two free parameters and setting P cannot uniquely determine both ρ and T. In many situations, however, T is a function of ρ, and in these situations $P = f(\rho)$, known as a barytropic equation of state. A subset of these equations of state, where

$$P = K\rho^\gamma = K\rho^{1+1/n} \tag{241}$$

are known as polytropic equations of state. γ is known as the polytropic exponent, K the polytropic constant and n the polytropic index. K can either be fixed (as we will see below) or remain as a free-parameter that varies from star to star.

Of course, in many situations polytropes are not realistic equations of state. Examples of where they are include:

1. **Radiation-dominated stars**: consider the total pressure in a region of a star to be the sum of the gas kinetic pressure and the radiation pressure, i.e.

$$P = \frac{k_B}{\mu m_H}\rho T + \frac{a}{3}T^4 = \frac{k_B}{\mu m_H \beta}\rho T \tag{242}$$

where $\beta = P_{\text{ideal}}/P$. Then $1 - \beta = \frac{P_{\text{rad}}}{P} = \frac{aT^4}{3P}$. We can then replace temperature with a function of pressure to get

$$P = K\rho^{4/3} = \left(\frac{3k_B^4}{a\mu^4 m_H^4}]\right)^{1/3} \left(\frac{1-\beta}{\beta^4}\right)^{1/3} \rho^{4/3}, \tag{243}$$

which is an $n = 3$ polytrope, if β is constant. Extremely massive stars are nearly completely radiation-dominated, completely convective (i.e. roughly isentropic), which for radiation means $\rho \propto T^3$; combining this with $P \propto T^4$ nets us $\gamma = 4/3$, which is an $n = 3$ polytrope and consistent with our prior derivation. The Sun can be approximated by an $n = 3$ polytrope, though the match would not be perfect since the Sun has a convective envelope.

2. **Zero-temperature degenerate stars**: these are considered in detail in Sec. 5.12. Suffice it to say, a non-relativistic completely degenerate gas has $P \propto \rho^{5/3}$ and a relativistic completely degenerate gas has $P \propto \rho^{4/3}$. An $n = 3$ polytrope is inherently unstable, and that manifests itself here in a maximum mass for degenerate stars. The non-relativistic equation works well for cold white dwarfs, brown dwarfs, and gas giants.

3. **Adiabatic stars**: consider an ideal gas with 2α degrees of freedom. Then $U = \alpha N k_B T = \alpha PV$. The first law of thermodynamics then says $dU = -PdV$. Substituting $dU = \alpha PdV + \alpha V dP$, we obtain $\frac{dP}{P} = -\frac{\alpha+1}{\alpha}\frac{dV}{V}$, which nets us

$$P \propto \rho^{(\alpha+1)/\alpha}. \tag{244}$$

   For a monotonic gas, $\alpha = 3/2$, giving $\gamma = 5/3$, or $n = 3/2$.

4. **Isothermal stars**: this is obvious: $P \propto \rho$, with $K = \frac{k_B T}{\mu m_H}$. True isothermal stars do not exist, but this can give the structure of the inert core of a shell-burning star.

5.8.1. *Why are polytropes useful?*

The advantage of using a polytropic equation of state is that it immediately simplifies the equation of hydrostatic balance $\frac{dP}{dr} = -\frac{d\Phi}{dr}\rho$ into

$$\frac{d\Phi}{dr} = -\gamma K \rho^{\gamma-2}\frac{d\rho}{dr}. \tag{245}$$

This can be integrated to obtain $\rho = \left(\frac{-\Phi}{(n+1)K}\right)^n$ if $\gamma \neq 1$ and we have set $\Phi = 0$ when $\rho = 0$. The Poisson equation assuming spherical symmetry reduces to

$$\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\Phi}{\partial r}\right) = 4\pi G\rho. \tag{246}$$

We can then substitute in ρ. We now define $z = Ar$, where $A^2 = \frac{4\pi G}{(n+1)K}\rho_c^{1-1/n}$ and $w = \frac{\Phi}{\Phi_c} = \left(\frac{\rho}{\rho_c}\right)^{1/n}$, where subscript c refers to central values. We may then rewrite Eqn. 246 as

$$\frac{d^2w}{dz^2} + \frac{2}{z}\frac{dw}{dz} + w^n = 0. \tag{247}$$

This is known as the Lane-Emden equation. Since it is equivalent to $\frac{1}{z^2}\frac{d}{dz}\left(z^2\frac{dw}{dz}\right) + w^n = 0$, to prevent $w(z=0) = \infty$, we set $\frac{dw}{dz}(z=0) = 0$.

While there are only a few analytical solutions, this is a relatively easy integral to solve (to remove the regular singularity at zero we may expand $w(z)$ into a power series). We are then able to determine the density ρ as a function of the central density as a function of radius.

To scale this solution to a star, we use the relation

$$M = 4\pi\rho_c R^3\left(-\frac{1}{z}\frac{dw}{dz}\right)_{z=z_n} \tag{248}$$

If we do not fix K, then all we need to specify are n and two out three variables, M, R, and ρ_c. If we do fix K, then all we need to specify are n and one out of M, R, and ρ_c. Fixing K naturally gives the mass-radius relationship $R \propto M^{(1-n)/(3-n)}$.

5.8.2. *What numerical methods are used to for calculating polytropes?*

A brief description of shooting goes here.

5.8.3. *How is the Chandrasekhar Mass estimated using polytropes?*

With electricity and meat.

n	z_n	$\left(-z^2\frac{dw}{dz}\right)_{z=z_n}$	$\varrho_c/\bar{\varrho}$
0	2.4494	4.8988	1.0000
1	3.14159	3.14159	3.28987
1.5	3.65375	2.71406	5.99071
2	4.35287	2.41105	11.40254
3	6.89685	2.01824	54.1825
4	14.97155	1.79723	622.408
4.5	31.8365	1.73780	6189.47
5	∞	1.73205	∞

FIG. 117.— Results from integrating the Lane-Emden equation for a given polytrope. Integration ends when $z = z_n$, at which point $w(z) = 0$. From Kippenhahn & Weigert (1994), their Table 19.1.

5.9. *Question 9*

QUESTION: What was the solar neutrino problem, and how was it resolved?

This answer is from Carroll & Ostlie (2006), pg. 356 - 360.

The solar neutrino problem was first discovered by Davis et al. in 1970. Davis et al.'s neutrino detector at Homestake Gold Mine in South Dakota contained approximately a thousand metric tonnes of cleaning fluid, C_2Cl_4, in order to observe the weak reaction

$$^{37}_{17}\mathrm{Cl} + \nu_e \leftrightharpoons ^{37}_{18}\mathrm{Ar} + e^-. \tag{249}$$

$^{37}_{18}$Ar has a half-life of 35 days. The threshold energy for this reaction is less than the energies of the neutrinos produced in every step of the pp-chain save for the first ($^1_1\mathrm{H} + ^1_1\mathrm{H} \rightarrow ^2_1\mathrm{H} + e^+ + \nu_e$), though most of the neutrinos detected should be from

$$^8_5\mathrm{B} \rightarrow ^8_4\mathrm{Be} + e^+ + \nu_e. \tag{250}$$

Careful calculations of solar neutrino emission and reaction with C_2Cl_4 suggested that 7.9 Solar Neutrino Units (SNU, defined as 10^{-36} reactions per target atom per second) should be recorded in the Homestake experiment. The actual data gave 2.56 ± 0.16 SNU. This was corroborated by other detectors, such as Japan's Super-Kamiokande (which uses pure H_2O surrounded by detectors; neutrinos scattering off electrons produce Cherenkov radiation, which can then be detected) and the Soviet-American Gallium Experiment and GALLEX in Gan Sasso, Italy (which use the reaction $\nu_e + ^{71}_{31}\mathrm{Ga} \rightarrow ^{71}_{32}\mathrm{Ge} + e^-$).

The resolution to this deficit of neutrinos, known as the solar neutrino problem, is known as the Mikheyev-Smirnov-Wolfenstein (MSW) effect, which involves the transformation of neutrinos from one flavour to another. All neutrinos produced by the Sun are electron neutrinos (ν_e), but during their journey from the Sun to Earth they oscillate between electron, muon (ν_μ) and tau (ν_τ) neutrinos. All experiments reporting a neutrino deficit were sensitive only to ν_e.

In 1998 Super-Kamiokande was used to detect atmospheric neutrinos produced when high-energy cosmic rays collide with Earth's upper atmosphere, which is capable of creating electron and muon neutrinos, but not tau neutrinos. They found that the muon neutrino flux coming from underneath the detector was substantially less than the flux coming from overhead. This can be interpreted as muon neutrinos generated on the other side of the planet oscillating into tau neutrinos on their way to the detector, while the muon neutrinos generated on the near side of the planet have not had the time to do so. The Sudbury Neutrino Observatory has been able to detect all three flavours of neutrino.

5.9.1. *Why do we believe neutrinos have masses?*

This answer is from Carroll & Ostlie (2006), pg. 359, and Schneider (2006), pg. 351 - 352.

One testable consequence of the MSW effect was that flavour oscillation was only allowed by grand unified theories if neutrinos had mass (electroweak theory does not predict neutrinos to have mass). The current established upper limit for neutrino masses is around 0.3 eV, which is still greater than the mass needed for the MSW effect to operate.

Experimentally, neutrino mass can be determined via particle accelerator experiments, which have set a lower limit for the total neutrino mass of all three flavours to be ~ 0.05 eV. Since neutrinos are considered hot dark matter, they suppress both the CMB power spectrum and the matter power spectrum at small scales by free-streaming. An $\Omega_\nu h^2 < 0.0076$ implies that the maximum neutrino mass is < 0.23 eV.

5.10. *Question 10*

QUESTION: Why is nuclear fusion stable inside a main-sequence star? Under what conditions is nuclear fusion unstable? Give examples of actual objects.

This answer is from Kippenhahn & Weigert (1994), Sec. 20.3 and 25.3.4, as well as van Kerkwijk (2011).

Nuclear fusion inside main sequence stars is stable due to the "solar thermostat". Suppose fusion inside a star increases on a local thermal adjustment timescale. The thermal energy generated by the increases the temperature, which, by the ideal gas law, generates an overpressure. The system is no longer in hydrostatic equilbrium and the star expands. Since work is done during this expansion, the star must then cool, which quenches any increase in fusion rate due to increase in temperature. The star then relaxes back into its previous equilibrium.

This is actualy something that can be determined from the virial theorem, which holds if the star is in hydrostatic equilibrium. $2E_i = -E_g$ requires that if the star expands, and therefore its in

The following two equations are derived in more detail in Sec. 20.3 in Kippenhahn & Weigert (1994), but here we perform an order of magnitude scaling. Density for a star is given by $\rho \propto M/R^3$, and mass is fixed, so $\dot{\rho}/\rho = -3\dot{R}/R$. For the same reason ($P \propto M^2/R^4$) we obtain $\dot{P}/P = -4\dot{R}/R$. If our system expands and contracts homologously, the same equations hold for any radius r (and so ex. $\dot{P}/P = -4\dot{r}/r$). Lastly, we assume an equation of state $\rho \propto P^\alpha T^{-\delta}$; taking the total differential gives us $\dot{\rho}/\rho = \alpha\dot{P}/P - \delta\dot{T}/T$. Since $\dot{\rho}/\rho = -3\dot{R}/R$, we obtain $\dot{T}/T = -((4\alpha - 3)/\delta)\dot{r}/r$. To summarize:

$$\begin{aligned} \frac{\dot{\rho}}{\rho} &= -3\frac{\dot{r}}{r} \\ \frac{\dot{P}}{P} &= -4\frac{\dot{r}}{r} \\ \rho &\propto P^\alpha T^{-\delta} \\ \frac{\dot{T}}{T} &= -\frac{4\alpha - 3}{\delta}\frac{\dot{r}}{r} \end{aligned} \tag{251}$$

For an ideal gas $\alpha = \delta = 1$, so we immediately see

$$\frac{\dot{T}}{T} = -\frac{\dot{r}}{r} \tag{252}$$

i.e. an increase in temperature drives an outward, homologous expansion, which by this equation requires a decrease in temperature. Indeed, we can also derive $\dot{P}/P =$ if we use the first law of thermodynamics ($dQ = dU + PdV$) and adjust it to our system ($dQ = c_P dT - \frac{\delta}{\rho}dP$, $\nabla_{\rm ad} = \frac{P\delta}{T\rho c_P}$; see Kippenhahn & Weigert (1994), Eqns. 4.18 and 4.21), we can derive the heat capacity of a star

$$c^* = c_P(1 - \nabla_{\rm ad})\frac{4\delta}{4\alpha - 3}, \tag{253}$$

which is negative for an ideal gas. This is equivalent to our derivation from before (though perhaps less, since it states that an increase in fusion rates results in a decrease in temperature. Note that Kippenhahn & Weigert (1994) Sec. 25.3.4 adds c ("central") to all its variables; this is to denote that we are supposing homologous expansion properly describes changes in central values.

Examples where nuclear fusion is unstable:

- **Thin shell source**: when the centre of a star becomes too cold for nuclear fusion, fusion continues in a shell around an inert core. Since the temperature on the inside of the shell is hotter than on the outside, which affects fusion rate, the shell thins over time. A thin shell homology a distance r_0 from the centre of the star has mass within the burning shell as $m \approx \rho r_0^2 D$; assuming mass is fixed, this translates to $\dot{\rho}/\rho \propto -\dot{D}/D = -(r/D)\dot{r}/r$ (since $r = r_0 + D$). Performing the same calculation as before using $\dot{P}/P = -4\dot{r}/r$ and the equation of state, we obtain $\frac{\dot{T}}{T} = \left(\frac{r}{D} - 4\right)\frac{\dot{r}}{r}$. If $D < r/4$, an expansion actually increases temperature. This leads to thermal pulses on the asymptotic giant branch.
- **Central fusion in a degenerate equation of state**: for a non-relativistic degenerate object, $\delta = 0$ and $\alpha = -3.5$, so from Eqn. 253 the heat capacity is positive. Equivalently, Array 251 shows that pressure is independent of temperature. This means in a degenerate object pressure does not increase with temperature. Since there is no cooling mechanism, and temperature increases fusion rates, this is a nuclear runaway. This mechanism destroys white dwarfs in SNe Ia.

- **Faster than dynamical nuclear fusion**: a star can be unstable in any situation where the nuclear burning rate becomes faster than the dynamical time, since the star no longer has the ability to readjust its structure to compensate for the increasing energy. An example of this is a pair-instability supernova: stars that produce oxygen cores of $\gtrsim 40$ $M_\odot$ (these stars are > 100 $M_\odot$ and could be prototypical Pop III stars) will, on their way to producing iron, begin producing electron-positron pairs. This creates a $\gamma < 4/3$ equation of state inside the star (energy is being siphoned off to produce pairs), and the star collapses until carbon fusion is ignited on a timescale less than the dynamical time, and destroys the star.

5.10.1. *Why does the star expand homologously?*

It actually does not, but homologous expansion is a useful approximation to make the derivation of fusion stability and negative heat capacity tractable.

5.10.2. *What if you increased fusion tremendously inside a star? Would that not destroy it?*

It would. Another factor to why a star is stable to nuclear fusion is that changes to fusion occur on a thermal adjustment timescale in main-sequence and giant stars (for the most part), and therefore hydrostatic equilibrium is maintained. An example of when it is not is listed above: the pair-instability supernova.

5.11. *Question 11*

QUESTION: Why do neutrons inside a neutron star not decay into protons and electrons?

This information is from Townsend (1997).

Degeneracy pressure is physically generated by forcing electrons into increased momentum states so as to satisfy the Pauli exclusion principle - the quantization of phase space requires that electrons sharing the "same" position must have different momenta. As such, an increased compression generates a higher distribution of momentum states (i.e. the top of the Fermi sea, given by energy E_F, is higher).

The mass of a neutron is 1.001378 the mass of a proton, and the mass of an electron is 5.44617×10^{-4} the mass of a proton (Wolfram 2012). Electron capture,

$$e^- + p^+ \rightarrow n + \nu_e \tag{254}$$

therefore requires about $8 \times 10^{-4} m_p c^2$ worth of kinetic energy in the reactants, approximate the same value as the rest energy of the electron. When the neutrino energy is accounted for, approximately 3 times the rest energy of the electron is needed.

For ordinary (even relativistic) WDs, $E_F \lesssim m_e c^2$, so WDs are stable to electron capture. A collapsing WD, however, will increase its E_F to infinity (at infinite E_F there is still insufficient electron degeneracy pressure to halt collapse, since there are still tons of low-energy electrons), making electron capture possible.

Once electron capture occurs, it cannot come into equilibrium with the reverse reaction ("assisted beta decay", $n + \nu_e \rightarrow e^- + p^+$) because the neutrino escapes the WD. Beta decay ($n \rightarrow e^- + p^+ + \bar{\nu}_e$) *can* can occur, however, and do, to the point at which E_F filled to $\sim 3 m_e c^2$ when inverse beta decay becomes ENTROPICALLY? preferred. This makes it so that in the regime of nuclear saturation density (i.e. the nucleons "touch"; $\sim 3 \times 10^{14}$ g/cm^3) electrons and protons still make up $\sim$12% of the material (van Kerkwijk 2012). Neutrons, however, dominate, since E_F is rapidly filled at neutron star densities. The limiting value, set by the balance between beta decay and electron capture and mediated by filling the electron Fermi sea, is 8 neutrons to 1 proton and 1 electron (Carroll & Ostlie 2006, pg. 581)

It is therefore not entirely true that neutrons do not decay into protons and electrons, but a neutron star is stable to beta decay.

5.12. *Question 12*

QUESTION: Give examples of degenerate matter.

This is from my own notes and Kippenhahn & Weigert (1994), Ch. 15.

Degenerate matter is matter in which pressure due to the Pauli Exclusion Principle dominates over pressure calculated from the ideal gas. (Partly degenerate matter is when the two are on the same order of magnitude, indicating thermal pressure support still plays a significant role.) This pressure comes from two factors. The first is that in quantum mechanics the number of possible states in phase space is discretized; this "density of states" can be calculated by, for example, determining the discretization of eigenstates in an infinite square well, and then taking the boundaries of this well to infinity (Sturge 2003). The second is the Pauli Exclusion Principle, which prevents the occupation of a single phase space state by particles with the same spin. For a gas of electrons and neutrons, for example, this results in

$$f(p) dp dV \leq 8\pi p^2 dp dV / h^3 \tag{255}$$

where $f(p)$ is the "occupation number" of a state p, and runs from 0 to 1. As lower p values become full ($f(p) = 1$), electrons and neutrons are forced to occupy higher p states - this is what generates degeneracy pressure. In a completely degenerate gas, the highest p state is known as the Fermi momentum, and its associated energy the Fermi energy. Since the momentum (equivalently the energy) is nonzero, the Fermi momentum and energy are associated with a Fermi pressure. In a partly degenerate gas, there is no critical value of p, though the fact that $f(p) \leq 1$ can still have an effect on the overall pressure.

Examples of degenerate matter include:

- **Electron degenerate matter** - in white dwarfs, massive gas giants and brown dwarfs, as well as within post-main sequence stars.
- **Neutron degenerate matter** - in neutron stars.

More theoretical forms of degenerate matter include:

- **Proton degenerate matter** - is physically possible, but practically it is always overshadowed by electron degeneracy due to its much larger Fermi sea - the only way for proton degeneracy to become prominent is to have an object completely made of protons.
- **Quark degenerate matter (QCD matter)** - at T $\sim 10^{12}$ K or densities where the average inter-quark separation is less than 1 fm (or a combination of both), hadrons will tend to decompose themselves into quarks. This matter will be Fermi degenerate, but will also exhibit properties such as color superconductivity.
- **Preon degenerate matter** - if quarks and leptons themselves are comprised of sub-components, we would expect them to break up at extremely high temperatures and pressures. Under these conditions degeneracy pressure of the sub-components (called preons) could hold the system against gravitational collapse.

5.12.1. *How can you tell if something is degenerate?*

For a Fermi gas, the more general equation for the occupation of states is

$$f(p)dpdV = \frac{8\pi p^2 dpdV}{h^3} \frac{1}{1 + e^{E/k_B T - \psi}} \tag{256}$$

where $\psi = \infty$ at infinite density or zero temperature, and $\psi = -\infty$ at infinite temperature or zero density. From the argument in pg. 124 of Kippenhahn & Weigert (1994), for a non-relativistic gas,

$$\psi = \ln\left(\frac{h^3 n}{2(2\pi m_e k_B T)^{3/2}}\right) \tag{257}$$

and the degeneracy/non-degeneracy limit is approximately when $\psi = 1$. An easier way of determining degeneracy lines on a $\rho - T$ diagram is to use the degeneracy parameter:

$$\frac{E_F}{k_B T}. \tag{258}$$

This value is $\propto n^{2/3}/T$ for non-relativistic gases, and $\propto n^{1/3}/T$ for a relativistic gas.

5.12.2. *Why are white dwarfs composed of an electron gas (rather than, say, a solid)? Are WDs pure Fermi degenerate gases?*

In WDs, the gas is of sufficiently high density that it is nearly fully pressure-ionized.

WDs are not perfect zero-temperature Fermi gases. The first correction that must be made, especially for hotter $> 10^8$ K WDs is that $f(p)$ does not have to be either 0 or 1, as is the case for a zero-temperature Fermi gas. This adds a ($\lesssim 10\%$) thermal pressure component to the gas. The second correction that needs to be made is the ions in the system. From above, if $\psi_e \approx 1$ for electrons in the system, $\psi >> 1$ for the ions, and as a result the ions are not degenerate even when the electrons are, and contribute a negligible amount of pressure (this, by the way, is true even if the ions are degenerate; see Sec. 4.14). At high densities and particularly low temperatures, ions have a tendency to form crystal lattices; this occurs when $(Ze)^2/r_{\rm ion} k_B T \approx 1$, which means that it is proportional to $n_{\rm ion}^{1/3}/T$. This has no real effect on the overall pressure of the system, which is still dominated by degeneracy pressure, but it is an additional source of energy for cooling WDs.

5.12.3. *Sketch the derivation of the Fermi pressure.*

Recall from Sec. 4.14 that we get p_F for zero temperature perfect Fermi gas from:

$$n = \int_0^{p_F} f(p) dp \tag{259}$$

where $f(p) = \frac{8\pi p^2}{h^3}$ - the reason this is true is because $f(p)$ describes the number of filled states per momentum range dp that fit in volume dV. (Note we can derive E_F from p_F through $E_F = p_F/2m$ for non-relativistic and $E_F = p_F c$ for relativistic.) The energy density is then

$$u = \int_0^{p_F} E(p) f(p) dp \tag{260}$$

where $E = p/2m$ if the gas is not relativistic, $E = pc$ if the gas is completely relativistic, and $E = mc^2\sqrt{1 + \frac{p^2}{m^2c^2}}$ in general. The pressure is similarly derived: the pressure is simply the net momentum flux through an area $d\sigma$ (see pg. 120 of Kippenhahn & Weigert (1994)) and therefore

$$P = \frac{1}{4\pi}\int_{2\pi}\int_0^{\infty} f(p)v(p)p\cos^2\theta dp d\Omega = \frac{1}{3}\int_0^{p_F} f(p)v(p)pdp, \tag{261}$$

where we have used the fact that the flux should be isotropic. We note $v(p) = \frac{p/m}{\sqrt{1+p^2/(mc)^2}}$ in general, and that $v = p/m$ in the non-relativistic limit, and $v = c$ in the relativistic limit. Our results, then, are:

$$P = \frac{1}{20}\left(\frac{3}{\pi}\right)^{2/3}\frac{h^2}{m}n^{5/3}, \tag{262}$$

for a non-relativistic gas, and

$$P = \left(\frac{3}{\pi}\right)^{1/3}\frac{hc}{8}n^{4/3}. \tag{263}$$

for an relativistic gas. We can add subscript-e to mass and number density for electrons, or m for neutrons.

5.13. *Question 13*

QUESTION: What is the typical temperature of matter accreting on a star, a white dwarf, a neutron star, a stellar mass black hole, and a supermassive black hole? In what wavelength range would one best find examples of such sources?

This question is a streamlined version of pgs. 662 - 665 of Carroll & Ostlie (2006), with some help from last year's qualifier notes.

The question is not very specific about whether or not matter is being accreted spherically or onto a disk. We will assume it is a disk in our calculations, since in most physical circumstances the infalling material will retain angular momentum.

Assuming virialization, the energy lost due to a particle being accreted is (assuming it started very far away)

$$E = -G\frac{Mm}{2r} \tag{264}$$

We may obtain dE/dr by differentiating. The amount of energy lost when a particle moves a radial distance dr is then $G\frac{M\dot{m}}{2r^2}dr$, where $\dot{m}$ is the accretion rate through the slice dr. Because an accretion disk accretes through viscous evolution, the energy is turned into heat (we have already assumed virialization in Eqn. 264, so there is no need to consider rotational energy; also note material, as it streams down the accretion disk, is virial at all times with this formulation). Now we shall assume steady state. Then for a slice of material dr, the difference in energy that passes through its outer and inner boundaries per unit time must equal to its luminosity (otherwise energy would pile up in dr[44]). Then,

$$dL_{\rm ring} = G\frac{M\dot{M}}{2r^2}dr = 4\pi r\sigma_{SB}T^4 dr, \tag{265}$$

noting that with steady state $\dot{m} = \dot{M}$ for all dr. This gives us

$$T = \left(\frac{GM\dot{M}}{8\pi\sigma_{SB}R^3}\right)^{1/4}\left(\frac{R}{r}\right)^{3/4}, \tag{266}$$

[44] You may protest that dE/dr could simply go into heating the infalling material instead of being radiated, but recall that we require the disk to be completely virial at all times, requiring the total energy to be $E_g/2$. dE/dr is then the *excess* energy generated during the infall.

noting that R cancels in the equation. For our question, we will set $r = R$. We may now perform some calculations.

- **A white dwarf** recurrent nova has a mass 0.85 $M_\odot$, radius 0.0095 $R_\odot$ and accretion rate of perhaps 10^{-9} $M_\odot$/yr. This gives a temperature in the inner disk of 6.5×10^4 K, and a peak wavelength of 45 nm, which is in the hard UV. Accreting WDs will emit in the ultraviolet (actually steadily accreting WDs with $\dot{M} = 10^{-7.5}$ emit in the soft x-rays).
- **A neutron star** with a disk has a mass 1.4 $M_\odot$, radius 1.5×10^{-5} $R_\odot$ (10 km) and the same accretion rate. This gives us a temperature in the inner disk of 9.3×10^6 K, and a peak wavelength of 0.31 nm, which is in the x-rays.
- **A stellar mass black hole** with a disk has a mass 5 $M_\odot$, radius 2.1×10^{-5} $R_\odot$ ($\sim$10 km) and the same accretion rate. This gives us a temperature in the inner disk of 9.9×10^6 K, and a peak wavelength of 0.29 nm, which is in the x-rays.
- **A supermassive black hole** with a disk has a mass 10^8 $M_\odot$, radius 430 $R_\odot$ and an accretion rate of 3 $M_\odot$/yr (Carroll & Ostlie 2006, pg. 1112). This gives us 5.1×10^5 K, and a peak wavelength of 5.7 nm, which is in the soft x-rays.

5.13.1. *How will your results change if accretion was spherical?*

This is something I thought up myself, with support from Spruit (2010), so read with a grain of salt.

If we assume the accretion is spherical, then it is also infalling with spherical symmetry. The accreting material is therefore in free fall until it strikes the surface, and becomes shock heated. For there to be equilibrium, the temperature must be below the effective temperature at the Eddington luminosity, Eqn. 160. If this is the case, then

$$\frac{GM\dot{M}}{R} = 4\pi R^2 \sigma_{SB} T^4. \tag{267}$$

We can then solve for T. This is *not* the temperature gained by a single particle being shock heated after gaining GMm/R worth of gravitational energy! Except for one particular $\dot{M}$, the temperature T will be lower or higher than $\frac{2GM\mu m_H}{3Rk_B}$. This is because while instantaneously infalling particles do get shock heated, the actual shock heating energy gets spread across the surface, and at equilibrium (when energy into the system is equal to energy out of it), the total energy piled up on the surface creates a temperature such that the luminosity balances the incoming energy from falling material. A reasonable assumption is that $\dot{M}$ is indeed the Eddington accretion rate, in which case

$$T = \left(\frac{L_{\rm Edd}}{4\pi R^2 \sigma_{SB}}\right)^{1/4} = \left(\frac{n_{\rm ne} GMm_p c}{R^2 \sigma_T \sigma_{SB}}\right)^{1/4} \tag{268}$$

5.13.2. *What is Bondi accretion?*

This information comes from Wikipedia (2012b).

An object plowing through the ISM will accrete at a rate $\dot{M} = \pi R^2 \rho v$, where ρ is the ambient density and v is either the speed of the object (if subsonic) or the soundspeed c_s if supersonic. R is not the radius of the object, but the radius at which $\sqrt{\frac{2GM}{R}} = c_s$. This accretion, known as Bondi accretion, would then occur at

$$\dot{M} = \frac{4\pi\rho G^2 M^2 v}{c_s^4} \tag{269}$$

5.14. *Question 14*

QUESTION: The weak equivalence principle for gravity corresponds to being able to find a coordinate system for a region of spacetime. What kind of coordinate system is this, and why?

This is cribbed off of Kissin (2012), with assistance from Dyer (2010a).

The weak equivalence principle states that all laws of motion for freely falling particles are the same as in an unaccelerated coordinate system or reference frame[45].

In general relativity, the motion of objects is given by the geodesic equation,

$$\ddot{x}^e + \Gamma^e_{mb} \dot{x}^m \dot{x}^b = 0, \tag{270}$$

[45] This is sometimes described as "the world line of a freely falling test body is independent of its composition or structure". If a series of test bodies comprised various tests of the laws of motion were crammed into a small enough area, they would all follow (almost) the same world line, and therefore all tests would be consistent with an unaccelerated reference frame.

where x is the 4-vector of an object, and overdots represent derivation with respect to an affine parameterization (equivalent to the proper time of the object). If we wish to find an unaccelerated reference frame, we equivalently want to find a coordinate system in which

$$\ddot{x}^e = 0 \tag{271}$$

i.e. the rate of change of coordinate time and space with respect to the proper time of the object is zero (i.e. no accelerations). It can be shown (Dyer 2010a) that at point P a coordinate system can always be constructed such that Γ^e_{mb} is zero in the coordinate system. The coordinate curves are themselves locally geodesics (i.e. "in free fall"), and an observer fixed in these coordinates has no forces acting on them from gravity (in the Newtonian sense). Any experiments performed would find the same laws of motion as an unaccelerated reference frame, or inertial coordinate system.

Therefore, the weak equivalence principle corresponds to being able to find the locally geodesic coordinate system at point P for which $\Gamma^e_{mb} = 0$, because $\Gamma^e_{mb} = 0$ translates to $\ddot{x}^e = 0$, which translates to an local motions of objects following the same trajectories as if they were in an unaccelerated coordinate system.

In special relativity, it was possible to single out a special set of frames (inertial frames), and define mechanical and electromagnetic forces as causing acclerations in these frames. With gravity, however, it is always possible to find some locally inertial frame in which any "force" due to gravity is non-existent. This motivates the concept that gravity is a geometric effect, rather than a force in the same sense as an electromagnetic or mechanical force.

5.14.1. *What is the significance of only being able to find this coordinate system at P?*

Because the coordinate system is only geodesic in a small non-empty neighbourhood around P, as we move away from P, the coordinate curves begin to diverge from the geodesics. Indeed, the only place where $\ddot{x}^e$ is zero is exactly at P! This is why even in the free-falling frame only experiments that are local (within the small, non-empty neighbourhood) are negligibly affected by the curvature of spacetime.

5.14.2. *What is the strong equivalence principle?*

The strong equivalence principle states that all laws of physics (rather than just the equations of motion that we dealt with above) that hold in Special Relativity hold exactly the same in a local inertial frame. This is equivalent to stating that a marriage of GR and electrodynamics, quantum and particle physics must reduce to their SR forms in the free-falling frame, i.e. the influence of gravity on any physical process can locally be transformed away.

5.15. *Question 15*

QUESTION: What are the typical detectors used in gamma-ray, X-ray, UV, visible, infrared, sub-mm, and radio observations?

This information comes from Karttunen et al. (2007), Kitchin (2009) and last year's qualifier notes.
Commonly used detectors include:

- **Charge-coupled device** (CCD): a surface made of light-sensitive silicon (semiconductor) diodes, arranged in a rectangular array of image elements or pixels. A photon hitting the detector ejects an electron into a conduction band (and produces an electron "hole" in the valence band). This electron is then trapped in the diode's potential well until it comes time for readout. Readout occurs in the manner suggested by Fig. 118. The quantum efficiency of CCDs is in the 70 - 90% range (highest in the infrared). The range of CCDs excends from the near-IR to the UV, though leftward of 500 nm the sensitivity drops due to absorption of UV radiation by silicon. To correct this the chip is made thinner to reduce absorption, or a coating is applied to down-scatter UV photons to the optical. In the x-rays it is also possible to use CCDs (the absorption issue being specific to UV), though because of the high energies involved multiple electrons are ejected per photon. To minimize dark current (thermal noise), CCDs are cooled. CCDs have a maximum storage capacity, past which it becomes saturated

- **Photomultiplier**: a series of charged plates. Radiation hits the cathode, and releases electrons via the photo-electric effect. The electrons are then accelerated by an electric field until they strike a dynode, which releases additional electrons. This process continues until a mountain of electrons strike the anode, and are read out using a curret meter. A two-plate photomultipler is known as a photocathode. A variant of the photomultiplier, the microchannel plate, is a cathode with holes carved in it, with a long tube extending downward from each hole. Electrons being photomultiplied travel down the tubes. This gives fine spatial resolution.

- **Scintillator**: a block of material connected to a photodetector (ex. a photomultiplier tube). Gamma radiation and hard x-rays interacts with the material (ex. NaI, $Bi_4Ge_3O_{12}$) by ejecting low-energy electrons. Since these materials have lots of higher energy electrons, and a "hole" in low energy, a cascade of spontaneous emission occurs after a gamma ray or x-ray strike.

- **Cerenkov detector**: extremely high energy gamma radiation (GeV to TeV) produce Cerenkov radiation when they impinge on the upper atmosphere. The radiation is mainly in the UV and has a distinctive shock-cone. UV photomultipliers can translate this to a readable signal.
- **Proportional counters**: a variant of the Geiger counter, which works by holding a capacitor with a medium between the plates at almost exactly the discharge potential difference. Ionizing radition then causes the capacitor to discharge in a cascade of electrons. The proportional counter holds the capacitor at a lower potential difference, which reduces the number of electrons discharged, but also makes the number more proportional to the energy of the ionizing photon.
- **Photoconductive cell**: these semiconductor cells change their conductivity when subjected to illumination, due to photons elevating valence band electrons to conducting levels. Very similar to CCDs (in that cells can be linked together into arrays), except instead of silicon they are made of materials such as indium antimodide, mercury-cadmium telluride (both used for near-IR) and germanium doped with gallium (mid-IR).
- **Bolometer**: material that changes its electrical resistivity in response to heating by illumination. At its simplest, it can just be considered a set of light-sensitive resistors.
- **Dipole antenna**: a simple set of conducting strips connected to electrical cable; turns the photons into an electrical impulse, which is then amplified, detected and integrated by a receiver, the simplest being a heterodyne system similar to that for a transistor radio. An oscilloscope (or more likely nowadays, a computer) acts as the analog-to-digital converter. Used, generally in antenna farms, in the MHz radio.
- **Horn antenna**: An antenna with a frontal waveguide shaped like an announcement horn. Used for GHz radio.
- **Superconductor-insulator-superconductor** (SIS): a superconductor, an insulating layer and another superconductor. Incoming photons give electrons in one superconducting layer enough energy to cross the insulating gap and enter the other superconducting layer.

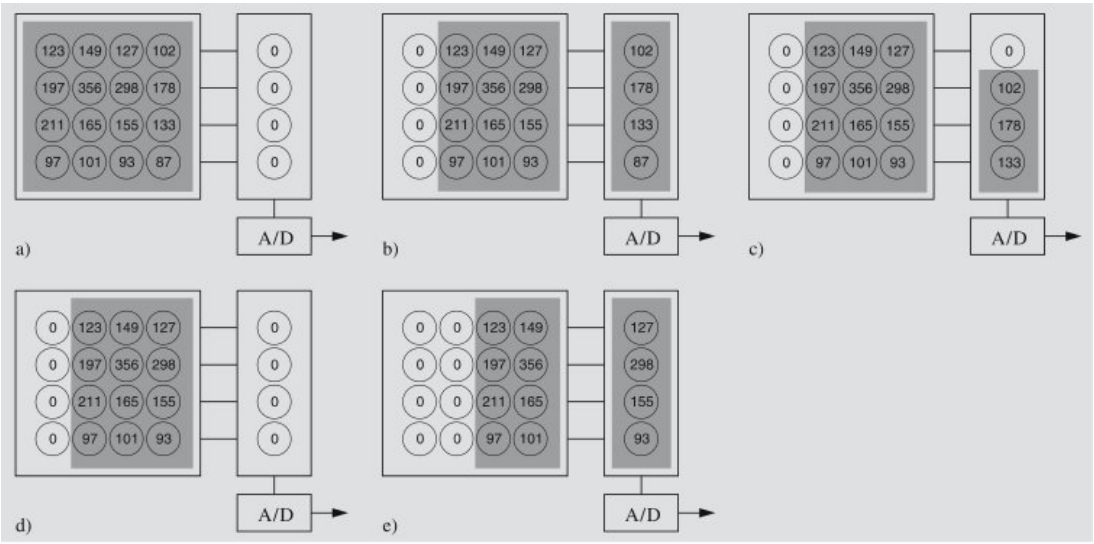

FIG. 118.— Schematic of CCD readout procedure. A staggered series of potential differences move columns of electrons to the right, where the rightmost is passed onto a readout row. The readout row is then read value by value. The process repeats until the entire CCD is expunged. From Karttunen et al. (2007), their Fig. 3.22a-e.

In terms of wavelength, the different detectors used are:

- **Gamma**: ($10^5 - 10^{14}$ eV (10^5 eV = 10^{-11} m)) scintillators, Cerenkov detectors
- **X-rays**: ($10^2 - 10^5$ eV, (0.01 - 10 nm)) CCDs (Chandra), photomultipliers, proportional (Geiger) counters, scintillators
- **Ultraviolet**: (10 - 400 nm) CCDs, photomultipliers, microchannel plates for hard UV
- **Optical**: (400 - 800 nm) CCDs
- **Infrared**: (Near-IR is ∼0.8 - 5 μm, mid-IR is 5 - ∼35 μm, and far-IR is ∼35 - ∼ 100 μm (IPAC 2012; Karttunen et al. 2007)) CCDs in the near-IR (but only out to about 1.1 μm), photoconductive cells in the near and mid-IR (note that Spitzer and IRAS use this technology), and photoconductive cells and bolometers in the far-IR

- **Sub-millimetre**: ($\sim$ 100 μm - 1 mm; this regime is sometimes included in the far-IR) bolometers (BLAST)
- **Radio**: ($>$ 1 mm) dipoles in MHz (GMRT), horn antennas in the GHz, SISs in the hundreds of GHz (mm).

5.15.1. *Which of these devices are photon counters?*

CCDs, possibly proportional counters and photoconductive cells?

5.16. *Question 16*

QUESTION: You don't usually need to cool down the detectors for short wavelength (e.g., X-ray) observations, but it's critical to cool down the detectors in long wavelength (e.g., far-IR) observations. Why is this?

This comes from Emberson (2012), Kitchin (2009) and last year's qual notes.

A number of issues due to the telescope's intrinsic temperature can affect observations. These include supplying thermal photons to the detector, and adding thermal electron noise.

Detectors in long-wavelength observations must be cooled sufficiently so that their blackbody radiation peak $\lambda_{\max} = \frac{0.0029}{T}$ is sufficiently below the range of wavelengths being observed. Otherwise, the telescope's own emission will contribute to the background emission. A hot telescope also excites electrons within the detector. If the detector is a semiconductor, and the energy gap between the valence and conduction bands is sufficiently small, the thermal tail of the electron distribution can promote electrons into the conduction bands. If the detector is a bolometer or antenna receiver, then thermal motions of electrons (known as thermal, Johnson or Nyquist noise) within the electrical components adds spurious signal.

Infrared detectors on the ground have to be cooled because of both issues. Room temperature (300 K) creates a blackbody peaking at about 10 μm, and so any measurements in the mid-IR would have to cool their instruments or contend with additional noise. Dark current becomes prohibitively large for infrared detectors due to the small band gap of the semiconducting materials used, and bolometers have to contend with thermal noise. As an example, Spitzer was cooled to $\sim$ 3 K in order to perform mid and far-IR observations. Once its liquid helium coolant ran out, it warmed to 30 K, and continued operation only in the near-IR. Bolometers are cooled to below 1 K.

While radio antennas themselves are generally not cooled, radio receivers are also often cooled to minimize thermal noise (Karttunen et al. 2007, pg. 70, and Kitchin 2009, pg. 135).

On the other hand, ultraviolet and x-ray detectors do not suffer from either problem: they are sensitive to photons far more energetic than the blackbody radiation of the telescopes they are housed in, and if they are semiconducting devices, their band gap is too large for thermal electrons to be an issue.

5.17. *Question 17*

QUESTION: Compare the S/N ratios between the following two cases where photon noise is dominant (assume an unresolved point source): [A] 1-minute exposure with a 10-m telescope; [B] 10-minute exposure with a 1-m telescope.

This information is from Mieda (2012).

Suppose the source has a flux of F impingent on the telescope. The number of photons per metre squared is then $\frac{F}{h\nu} = n$. Let us assume the noise is source-dominated. Photon noise is Poisson distributed, so the noise of a flux of N photons has error $\sqrt{N}$. The total number of photons received is

$$N = n\pi(D/2)^2\Delta t \tag{272}$$

If we suppose the sky is negligible, then the signal to noise ratio is

$$S/N = \frac{N}{\sqrt{N}} = \sqrt{N} \tag{273}$$

The ratio of S/N between case A). and case B). we shall call the preference factor P:

$$P = \sqrt{N_A}/\sqrt{N_B} = \frac{D_A\sqrt{\Delta t_A}}{D_B\sqrt{\Delta t_B}} \tag{274}$$

Plugging in values gives us $P = \sqrt{10}$. We prefer a larger diameter telescope.

5.17.1. *What if the sky were non-negligible? What if dark current and readout noise were non-negligible?*

The most general signal to noise equation is

$$S/N = \frac{N_*}{\sqrt{N_* + N_{\rm pix}\left(1 + \frac{N_{\rm pix}}{N_{\rm sky}}\right)(n_{\rm sky} + n_{\rm dark} + n_{\rm read}^2)}} \tag{275}$$

where[46]

- $N_{\rm pix}$ is the **number of pixels the object subtends**.
- $N_{\rm sky}$ is the **number of pixels over which the sky background was measured**.
- $n_{\rm sky}$ is the **sky background** per pixel. Abstractly this can be determined by taking a picture of any part of the sky and dividing it by the number of pixels on the CCD, but realistically this is determined by swinging the telescope slightly off target and taking an image of no object in particular, or by using an annulus in the same image around the target object. This is because different parts of the sky will have different photometric and spectroscopic properties.
- $n_{\rm dark}$ is the **dark current** per pixel. Dark current comes from thermal electrons in the CCD itself (Karttunen et al. 2007).
- $n_{\rm read}$ is the **readout noise** per pixel. This is noise inherent in the readout operation of the CCD (which turns electrons into a digital signal). Readout noise is not Poisson distributed, which is why it is $n_{\rm read}^2$.

This equation takes into account the fact that you have performed subtraction of all sources of signal that are not the source.

Suppose the source were quite faint. In this case, the measurement will be sky dominated. Option b). would still have a higher signal-to-noise: first, there would be more photons (both sky and source) collected, which would increase N_*, we would have a larger area over which to subtract the sky, which would increase $N_{\rm sky}$, and the PSF of the source would be smaller on the focal plane, which would reduce $N_{\rm pix}$ (assuming we had an AO system; otherwise the PSF would be atmosphere limited).

5.17.2. *What if you were doing this in the radio?*

No idea, though I believe this question implicitly assumes an optical telescope. No professional radio telescopes have dish sizes of 1 - 10 m.

5.18. *Question 18*

QUESTION: Describe the difference between linear and circular polarizations.

This is simplified from my own qual notes, and from last year's notes.

The polarization of monochromatic light at a position $\mathbf{x}$ is determined by the time evolution of the electric field vector $\epsilon(\mathbf{x}, t)$, which lies orthogonal to the wavevector $\mathbf{k}$ for most media. The electric field vector can generally be described by two orthogonal vectors, and the components of the field along these directions vary sinusoidally with time. Since each component's amplitude and phase will be different than the other, the most general shape $\epsilon(\mathbf{x}, \mathbf{t})$ traces is an ellipse, different at each point $\mathbf{x}$. In the paradial limit, where light propagates along directions that lie within a narrow cone centred about the optical axis, we can approximate light as a plane wave $\mathbf{E}(\mathbf{x}, t)$ with only one wave-vector:

$$\mathbf{E} = \mathbf{E}_0 e^{-i\omega t} = e^{-i\omega t}(\hat{\mathbf{x}} E_1 e^{i\phi_1} + \hat{\mathbf{y}} E_2 e^{i\phi_2}) \tag{276}$$

where we have chosen $\mathbf{x} = 0$ without loss of generality. Taking the real component of this complex expression[47], we obtain

$$\begin{aligned} E_x &= E_1 \cos(\omega t - \phi_1) \\ E_y &= E_2 \cos(\omega t - \phi_2) \end{aligned} \tag{277}$$

Plotting the vector $\mathbf{E} = \hat{\mathbf{x}} E_x + \hat{\mathbf{y}} E_y$, we find that in general the vector traces out an ellipse in xy space. Suppose now we force $\phi_1 = \phi_2$. $\mathbf{E}$ would then be a line in xy space (first row in Fig. 119); this is known as linear polarization. If

[46] James Graham disagrees with me here; he thinks it should be $\left(1 + \frac{1}{N_{\rm sky}}\right)$.

[47] Recall that Euler's Equation states $e^{ikx} = \cos(kx) + i\sin(kx)$, and that $Re(x + iy) = x$

we instead force $\phi_1 = \phi_2 + \pi/2$, and $E_1 = E_2$, then when E_x is at its maximum, E_y will equal zero, and vice versa. As a result, **E** will trace a circle in xy space, and this is known as circular polarization (second and third rows in Fig. 119). If we make no restrictions on ϕ or E, then **E** generates an ellipse, seen in Fig. 120.

To define left and right-handed circular polarization, we can switch to x' and y' coordinates, where we force x' to be along the major axis of the ellipse. Then,

$$E_{x'} = E_0 \cos\beta \cos(\omega t)$$
$$E_{y'} = -E_0 \sin\beta \sin(\omega t) \tag{278}$$

where $-\pi/2 \leq \beta \leq \pi/2$. When $\beta < 0$, the ellipse is traced counterclockwise, and when $\beta > 0$, the ellipse is trace clockwise. If we assume that the wave is propagating "out of the page" (along the positive z direction), the former is called right-handed polarization (or negative helicity), while the latter is called left-handed polarization (positive helicity) (second and third rows in Fig. 119).

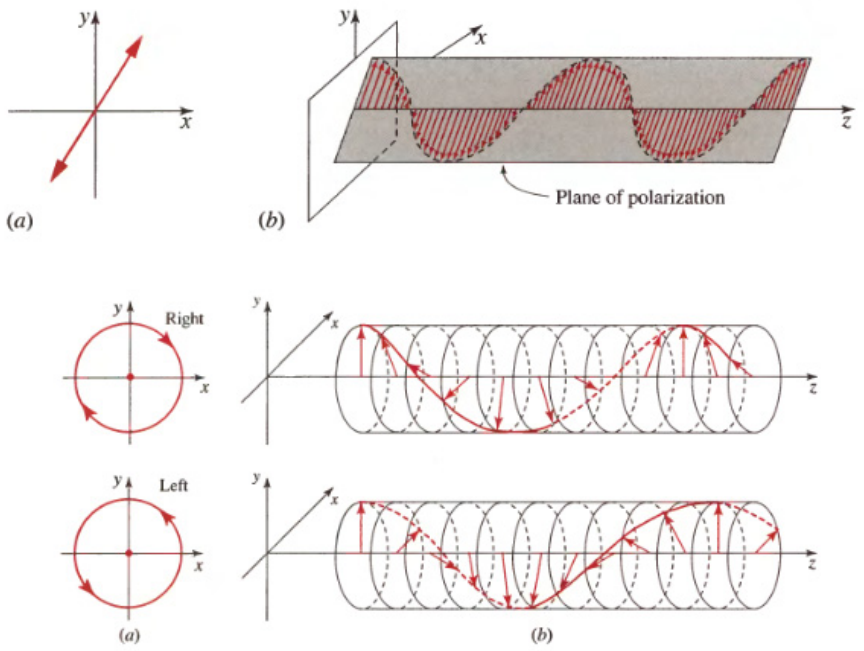

FIG. 119.— Graphical depiction of linear and circular polarization. a). represents the electric field vector in the plane normal to the wavevector **k**, while b). is a snapshot at fixed time, and represents how this vector changes in space. From Saleh & Teich (2007), their Figs. 6.1-3 and 6.1-4.

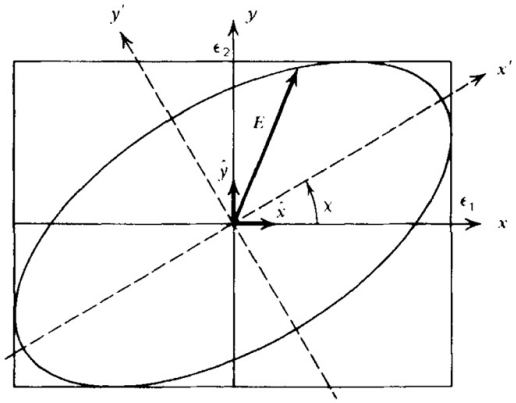

FIG. 120.— Generalized polarization ellipse (and one representative vector **E**), with the x-y and x'-y' coordinate systems overlaid. The angle between x and x' is denoted χ. From Rybicki & Lightman (1979), their Fig. 2.4.

Note that while we have chosen to write circular polarization as out-of-phase orthogonal linear polarization, it is completely possible to write linear polarization as out-of-phase counterpropagating circular polarization. This

formulation is useful when calculating the rotation of a linearly polarized wave due to ISM plasma in a magnetic field (Faraday rotation).

5.18.1. *What are Stokes' parameters?*

We may write E_1 and E_2 in terms of β, and the angle between the x and x' axes χ: $E_1 \cos\phi_1 = E_0 \cos\beta \cos\chi$, $E_1 \sin\phi_1 = E_0 \sin\beta \sin\chi$, $E_2 \cos\phi_2 = E_0 \cos\beta \sin\chi$, and $E_2 \sin\phi_2 = -E_0 \sin\beta \cos\chi$ (this can be derived from the rotation matrices). We can solve for E_0 β and χ given E_1, ϕ_1, E_2 and ϕ_2 using the Stokes parameters for monochromatic waves:

$$I = E_1^2 + E_2^2 = E_0^2 = |A_x|^2 + |A_y|^2 \tag{279}$$

$$Q = E_1^2 - E_2^2 = E_0^2 \cos(2\beta)\cos(2\chi) = |A_x|^2 - |A_y|^2 \tag{280}$$

$$U = 2E_1E_2\cos(\phi_1 - \phi_2) = E_0^2 \cos(2\beta)\sin(2\chi) = 2Re(A_x * A_y) \tag{281}$$

$$V = 2E_1E_2\sin(\phi_1 - \phi_2) = E_0^2 \sin(2\beta) = 2Im(A_x * A_y) \tag{282}$$

where $A_x = E_1 e^{i\phi_1}$ and $A_y = E_2 e^{i\phi_2}$. We can determine E_0, β and χ through

$$E_0 = \sqrt{I} \tag{283}$$

$$\sin(2\beta) = V/I \tag{284}$$

$$\tan(2\chi) = U/Q \tag{285}$$

The Stokes parameters do have physical meanings. I is (proportional to) the intensity, V is the circularity parameter that measures the ratio between major and minor axes (the wave is RH or LH polarized when V is positive or negative, respectively), and Q and U measure the orientation of the ellipse relative to the x-axis. Note that there are four Stokes parameters and three independent variables defining general elliptical polarization - thus there is a relation between Stokes parameters, which turns out to be $I^2 = Q^2 + U^2 + V^2$.

5.18.2. *Why is polarization useful in optics?*

Polarization plays an important role in optics (and therefore instrument design), as the interaction between light and certain materials will vary due to polarization. For example, the amount of light reflected at the boundary between two materials is polarization dependent, and the refractive index of anisotropic materials depends on polarization. These phenomena allow for the design of polarization-imparting and polarization-discriminating optical devices.

5.18.3. *Give some examples of linear and circular polarization in nature.*

Linear polarization can be seen in Thomson scattering of the CMB, as well as starlight scattered off of elongated dust particles in the ISM. When viewed from certain angles, synchrotron radiation is linearly polarized.

Synchrotron is circularly polarized when viewed from certain other angles.

5.19. *Question 19*

QUESTION: What's the field of view of a 2K x 2K CCD camera on a 5-m telescope with f/16 focal ratio. The pixel size is 20 micron. If you bring this to a 10-m telescope with the same focal ratio, what will be the field of view? Give your answer using the Étendue conservation rule.

This answer comes from last year's qual notes, with a bit from Wikipedia (2012d) and Marleau (2010), Ch. 1.

In an optical system where light can only propagate and suffer perfect reflections or refractions, the quantity

$$\epsilon = \int_A \int_\Omega \cos\theta d\Omega dA \tag{286}$$

known as the Étendue conservation rule. In telescopes, because everything is aligned on one axis, we can reduce this to the fact that $A\Omega = \epsilon$. We now use this rule to determine the fields of view of telescopes.

By similar triangles, $\Omega_{\rm sky} = \Omega_{\rm fp-p}$, and $A_{\rm fp}\Omega_{\rm p-fp} = A_{\rm p}\Omega_{\rm fp-p}$, where subscript p is primary, fp is focal plane, p − fp is primary as seen from focal plane, and fp − p is focal plane as seen from primary. The area of the primary is $\pi(D/2)^2$, and the distance f between the primary and focal plane is

$$f = f^{\#} D \tag{287}$$

Therefore, $\Omega_{\rm p-fp} = \frac{\pi(D/2)^2}{f^2} = \frac{\pi(D/2)^2}{(f^{\#}D)^2}$. The area of the focal plane is given by N^2x^2, where N is the number of pixels per side. From this, we obtain

$$\pi(D/2)^2\Omega_{\rm sky} = N^2x^2\frac{\pi(D/2)^2}{(f^{\#}D)^2}$$
$$\Omega_{\rm sky} = \frac{N^2x^2}{(f^{\#})^2D^2} \tag{288}$$

Plugging in $N = 2000$, $x = 2\times10^{-5}$ m, $f^{\#} = 16$ and $D = 5$ m, we obtain $\Omega_{\rm sky}$, or 1.06×10^4 square arcseconds, or 103" × 103". Plugging in $D = 10$ m, we obtain 2.6×10^3 square arcseconds, or 52" × 52".

5.19.1. *Derive the Étendue conservation rule.*

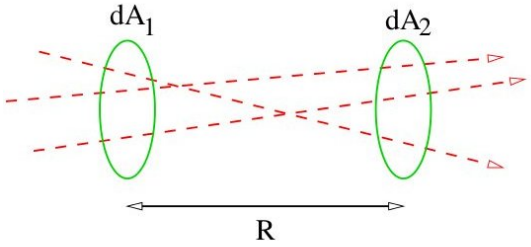

FIG. 121.— Schematic diagram of intensity conservation over free space. From Marleau (2010), Ch. 1.

Consider the leftside panel of Fig. **??**. Flux passing through dA_1 that eventually goes through dA_2 is given by $dP_1 = I_1dA_1d\cos\theta_1d\Omega_{2f1}$ ($d\Omega_{2f1}$ is the solid angle of dA_2's projection as seen from dA_1, and the angle between dA_1's normal and R (not shown) is θ_1). Flux being received at dA_2 is given by $dP_2 = I_2dA_2d\cos\theta_2d\Omega_{1f2}$. We then note $d\Omega_{2f1} = dA_2\cos\theta_2/R^2$ and $d\Omega_{1f2} = dA_1\cos\theta_1/R^2$. Moreover, $dP_1 = dP_2$. As a result,

$$I_2 = I_1 \tag{289}$$

In free space, intensity is the same between the emitter and the receiver. At the same time, however, we have shown that $dA_1d\cos\theta_1\Omega_{2f1} = dA_2d\cos\theta_2\Omega_{1f2}$. If, then, no light is lost from the system between the two surfaces, the integral of $dA\cos\theta d\Omega$ over the entire area and all applicable solid angles is fixed for both systems (since each piece $dA\cos\theta d\Omega$ on one surface corresponds to $dA\cos\theta d\Omega$ on the other) We then have that

$$\epsilon = \int_A\int_\Omega \cos\theta d\Omega dA \tag{290}$$

which is known as the Étendue conservation rule (here denoted ϵ). It can be shown that in general, ϵ cannot decrease, and can increase only when reflections and refractions are not perfect. In good optical systems, then, ϵ is conserved.

5.20. *Question 20*

QUESTION: Sketch and give the equations for each of the following distributions: 1. Gaussian (Normal distribution); 2. Poisson distribution; 3. Log-normal distribution. Give two examples from astrophysics where each of these distributions apply.

This information is from Emberson (2012) and last year's qualifier notes.

The Gaussian (normal) distribution is drawn in the left of Fig. 122. Its functional form, the probability distribution of a random variable X with value x, is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-(x-\mu)^2/2\sigma^2}. \tag{291}$$

The mean is μ and the standard deviation σ. This distribution is called the normal distribution when $\mu = 0$, $\sigma = 1$. The Gaussian is especially important due to the Central Limit Theorem, which states that, given certain conditions, the mean of a sufficiently large number of independent random variables, each with finite mean and variance, will

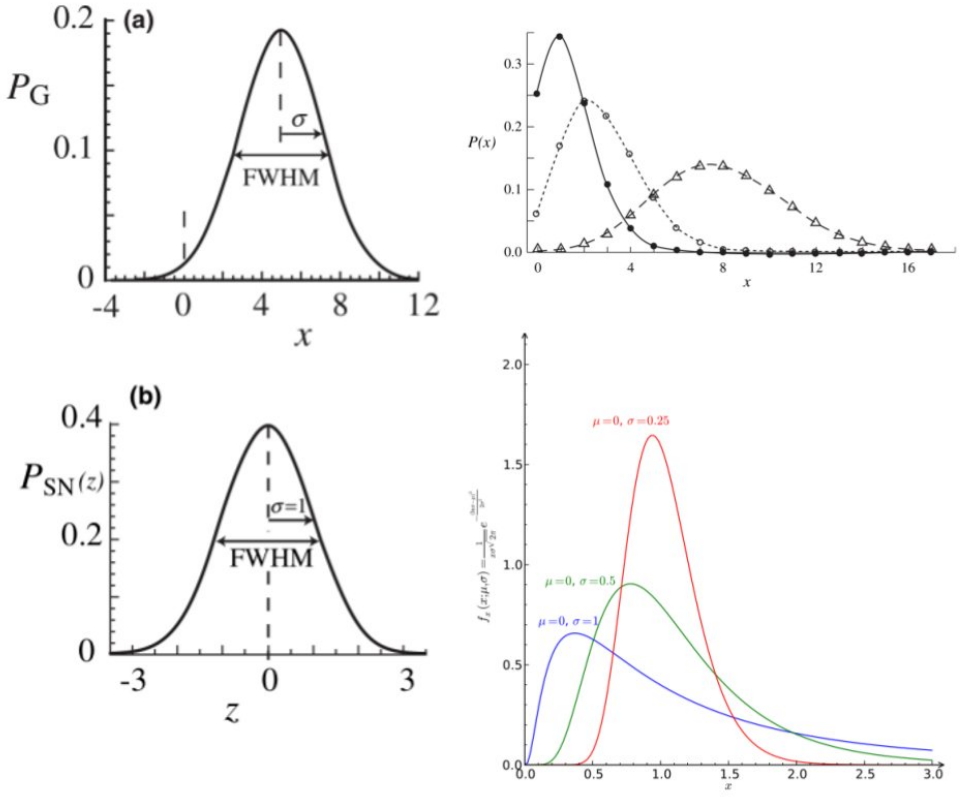

FIG. 122.— Left: two Gaussian distributions, one with $\mu = 5$, $\sigma = 2.1$, and the other, a normal distribution, with $\mu = 0$ and $\sigma = 1$. Upper right: Poisson distributions with $\mu = 1.4$, $\mu = 2.8$ and $\mu = 8$. Lower right: lognormal distributions (plotted with linear abscissa x). From Emberson (2012).

be approximately normally distributed. A weaker version of the theorem states that so long as the independent random variables are identically distributed, the distribution of the random variable $\frac{X_1+X_2+...+X_n-n\mu}{\sigma\sqrt{n}}$ tends toward the standard normal distribution. The summation of two Gaussians is also a Gaussian, with a new mean $\mu_1 + \mu_2$, and a new standard deviation $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$; this is why propagation of error requires that we add in quadrature.

Examples of Gaussians in nature include Doppler broadening of spectral lines ($\psi(\nu) = \frac{1}{\Delta\nu_D\sqrt{\pi}} e^{-(\nu-\nu_0)^2/\Delta\nu_D^2}$ ($\Delta\nu_D = \frac{\nu_0}{c}\sqrt{\frac{2k_BT}{m}}$), and radio signal random noise (due to the summation of many random oscillators, the voltage amplitude $v(t)$ has noise $P(v(t)) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-v^2/2\sigma}$). Gaussians are also used in many approximations, such as the shape of the PSF.

The Poisson distribution is drawn in the upper right of Fig. 122, and has the functional form

$$f(i) = e^{-\mu}\frac{\mu^i}{i!} \tag{292}$$

where $i = 0, 1....$ The Poisson random variable (RV) is discrete, in that i can only be an integer. The expected value of $f(i)$ (mean) is μ, while the variance $\sigma^2 = \mu$.

The advantage of the Poisson distribution is that it is the natural distribution for counting events that have an average rate λ, but where individual events occur randomly and are independent from all other events. If we consider a time-period t, then the probability we will observe n events is given by $f(n) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}$. This situation describes photon counting, which is why if we detect $N = \lambda t$ photons, we obtain a variance $\sigma^2 = N$. Other phenomena that follow Poisson statistics are the number of supernovae that will occur in a galaxy over a certain time interval, and the monthly mean number of sunspots.

The lognormal distribution is drawn in the lower right of Fig. 122, and has the functional form

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right] \tag{293}$$

This distribution is such that if Y is a random variable, $X = \exp(Y)$ is a lognormal random variable. The lognormal is used in the Chabrier IMF for stars with $M < 1$ $M_\odot$ (and a power law for stars with $M > 1$ $M_\odot$). Another example is x-ray flux variations in certain Seyfert 1 galaxies. In general, if the fluctuations on the quantity are proportional to the quantity itself, the distribution will be lognormal.

5.21. *Question 21*

QUESTION: You are trying to determine a flux from a CCD image using aperture photometry, measuring source(+sky) within a 5-pixel radius, and sky within a 20–25 pixel annulus. Assume you find 10000 electrons inside the aperture and 8100 electrons in the sky region, and that the flux calibration is good to 1%. What is the fractional precision of your measurement? (Ignore read noise.) More generally, describe how you propagate uncertainties, what assumptions you implicitly make, and how you might estimate errors if these assumptions do not hold?

This information comes mainly from Graham (2012) (though I disagree with him), Wall & Jenkins (2003) Ch. 3.3.2 and Ross (2007), Ch. 2.5.

The fractional precision of the measurement is another way of asking for the noise-to-signal ratio. We are therefore tasked with finding both the signal and its error. Since we are ignoring read noise, all other standard sources of error (Sec. 5.17) are Poisson distributed, and therefore have errors of $\sqrt{N}$. We will use the same notation as we did in Sec. 5.17.

We denote all non-source photons (i.e. we lump sky and dark current together) with subscript n. We wish to calculate n_n, the number of such photons per pixel. To do this, we say

$$n_n = N_n / N_{\rm sky} \tag{294}$$

where $N_n = 8100$ electrons. We now introduce the general means by which uncertainties are propagated, if our measurements are all independent of each other:

$$\delta f^2 = \sum_i \left(\frac{df}{dx_i}(\bar{x_i})\right)^2 \delta x_i^2 \tag{295}$$

and in this case we have $\delta n_n^2 = \delta N_n^2 / N_{\rm sky}^2$, so $\delta n_n = \frac{1}{N_{\rm sky}}\sqrt{N_n} = \sqrt{\frac{n_n}{N_{\rm sky}}}$. We now perform our subtraction to get N_*:

$$N_* = N_s - N_{\rm pix} n_n = N_s - \frac{N_{\rm pix}}{N_{\rm sky}} N_n \tag{296}$$

where $N_s = 10000$ electrons. The error on this measurement is $\sqrt{N_s + N_{\rm pix}^2 \delta n_n^2} = \sqrt{N_s + \frac{N_{\rm pix}^2}{N_{\rm sky}^2} N_n}$. (As a slight detour, we note we can also write the error as $\sqrt{N_s + N_{\rm pix}^2 \frac{n_n}{N_{\rm sky}}}$, and since by definition $N_{\rm obs} = N_* + N_{\rm pix} n_n$ we have reproduced Eqn. 275.) We now write

$$S/N = \frac{N_*}{\sqrt{N_s + \frac{N_{\rm pix}^2}{N_{\rm sky}^2} N_n}} \tag{297}$$

Lastly, we use the fact that the number of pixels in an area is proportional to the area (the constant is divided out in our expression) to get

$$S/N = \frac{N_*}{\sqrt{N_s + \frac{A_{\rm pix}^2}{A_{\rm sky}^2} N_n}} \tag{298}$$

$A_{\rm pix} = 25\pi$ and $A_{\rm sky} = (25^2 - 20^2)\pi = 225\pi$. Plugging in numbers gives us an S/N of 100, or a noise-to-signal ratio of 1%. Note that if our flux calibration is off, our S/N would be modified by either $\sqrt{1.01}$ or $\sqrt{0.99}$, since the calibration would affect all our photon counts. Our noise-to-signal ratio would then vary from 0.995% to 1.005%, a negligible change.

In general, we may propagate errors using Eqn. 295. If, however, our measurements are actually correlated, we need to add the covariance. The covariance between two sets of sample measurements is given by the sample covariance. For example, the sample covariance between two variables X_j and X_k that are sampled an equal number of times is given by the covariance matrix (Wikipedia 2012d)

$$q_{jk} = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \tag{299}$$

5.21.1. *Derive the propagation of uncertainty formula, Eqn. 295.*

We have a function f dependent on many random variables X_i, and therefore it itself is a random variable F. If we assume f varies slowly with x_i (the values of X_i), we can calculate how changes in x_i affect f:

$$df = \sum \frac{df}{dx_i}dx_i \tag{300}$$

We now have a function f dependent on many random variables, and therefore is itself random variable δF, where $\delta F = F - \bar{F}$. We can also define $\delta X_i = X_i - \bar{X}_i$. We can then calculate the variance of δF:

$$\mathrm{Var}(\delta F) = \mathrm{Var}\left(\sum \frac{df}{dx_i}(\bar{X}_i)\delta X_i\right) \tag{301}$$

Now, because variance is defined as $E\left[(X-\bar{X})^2\right]$ (E is expectation, i.e. either $\sum_i x_i p(x_i)$ or $\int xp(x)dx$) we have $\mathrm{Var}\left(\frac{df}{dx_i}(\bar{X}_i)\delta X_i\right) = \left(\frac{df}{dx_i}(\bar{X}_i)\right)^2 \mathrm{Var}(\delta X_i)$. Moreover, (Eqn. 2.16 of Ross)

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n}\mathrm{Var}(X_i) + 2\sum_{i=1}^{n}\sum_{j<i}\mathrm{Cov}(X_i, X_j) \tag{302}$$

If all our variables are independently distributed, then their covariances (defined as $\mathrm{Cov}(X,Y) = E[(X-\bar{X})(Y-\bar{Y})] = E[XY] - E[X]E[Y]$) will be zero. These two facts allow us to simplify Eqn. 301 to

$$\sigma_F^2 = \sum_i \left(\frac{df}{dx_i}(\bar{X}_i)\delta X_i\right)^2 \sigma_{X_i} \tag{303}$$

i.e. Eqn. 295 (where we have taken the usual definition of standard deviation as the root of variance).

Note it is not always true that f varies slowly with x_i. Higher order terms may be added in these cases.

5.21.2. *What observational and instrumental simplifications have you made to answer this question?*

We have assumed that pixels are infinitely small; in reality pixels are square, and not all pixels will be fully illuminated. Detector sensitivity variation has been neglected - in reality we need to perform what is called a "flat field" first. Bad pixels, cosmic ray hits, multiple electrons generated per photon (which occurs occassionally) and saturation are also not accounted for.

5.22. *Question 22*

QUESTION: Suppose you measure the brightness of a star ten times (in a regime where source-noise dominates; you will be given 10 values for the total number of counts): (1) How do you calculate the mean, median, and mode and standard deviation? (2) How can you tell if any points are outliers? Say some points are outliers, what do you do now (ie. how does this impact the calculation of the quantities in part 1)?

This answer comes from Press et al. (1992), Ch. 14 and 15.7, Wall & Jenkins (2003), Ch. 3, and NIST (2012).

The sample mean and standard deviation have regular formulae:

$$\bar{x} = \frac{1}{N}\sum_{j=1}^{N} x_j \tag{304}$$

$$\sigma^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \bar{x})^2 \tag{305}$$

The reason there is $N-1$ can be looked at in several ways (ex. pg. 42 of Wall & Jenkins notes that using $\bar{x}$ tends to generate minimum values of σ^2), but the easiest is in terms of degrees of freedom - if only one data point were given, there should be no constraint on the variance, and therefore it should not be zero[48]. There are also higher order moments (skew and kurtosis) but they are best not used.

The median is determined by first reordering all data points from minimum to maximum, and then picking

$$x_{\rm med} = \begin{cases} x_{(N+1)/2} & \text{if } N \text{ is odd} \\ \frac{1}{2}(x_{(N+1)/2} + x_{N/2+1} & \text{if } N \text{ is even} \end{cases}, \tag{306}$$

The mode is defined as the value of x where the probability distribution $p(x)$ is at its maximum. There is, consequently, no guarantee of a unique definition for mode.

Traditionally, the sample mean plus/minus the sample standard deviation is given as a measure of where the true mean is (i.e. the 1-sigma error bars around a value indicate a 68% chance that the true value is within the bounds of the bars). This is often not the case, however, for two reasons. The first is that convergence of a probability distribution to that of a Gaussian due to the central limit theorem varies both as a whole, and over different regions of the probability distribution. For example, as the mean number of counts μ of a Poisson-distributed RV goes up, the probability distribution converges toward a Gaussian, but the fractional difference between the Poisson distribution and the Gaussian is low near the peak of the distribution, and high in the wings. The second is that uncontrolled random events (such as instrument glitches and human error) will at times generate random outliers. Since these outliers are not properly sampling the RV of the quantity being measured, they tend to distort statistical measurements. Outliers may actually be scientifically interesting (ex. transients).

To test for outliers, we must first check to see if our RV can indeed be described, roughly, as normally distributed. This can be done by plotting a histogram, box plot, or (most effectively) a normal probability plot (see links in NIST (2012)). Once this is done (note that we can transform the data if it is lognormal distributed), one of several tests can be done. Here are a few examples, all of which assume Gaussianity:

- **The Z-score test**: we measure, for each data point we have taken,

$$z_i = \frac{x_i - \bar{x}}{\sigma} \tag{307}$$

  We then cut those values that are too large - NIST recommends $z_i = 3.5$. Since the maximum Z-score of a datapoint will be $(n-1)/\sqrt{n}$ (because the mean was calculated using the same data), this is not a useful test with very small datasets.

- **The Grubbs' test**: if we believe there is a single outlier in the system, we determine the maximum Z-score, and remove that. See links in NIST for statistical details.

- **The Tietjen-Moore test**: if we believe there are k outliers in the system, we calculate the sum of the variances of the k data points with the largest residual $r_i = |x_i - \bar{x}|$, and then reorder the data. Letting y denote this reordered data set, we calculate

$$E_k = \frac{\sum_{i=1}^{n-k}(y_i - \bar{y}_k)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{308}$$

  with $\bar{y}$ being the sample mean, $\bar{y}_k$ being the sample mean with the largest residual points removed, and $\sum_{i=1}^{n-k}$ being the sum of all data points except for the k removed due to having large residuals. Our hypothesis that there are k outliers is validated if E_k is much less than 1.

- **The Generalized (Extreme Studentized Deviate) test**: if we suspect there may be up to r outliers in the data set, we may use this test. It is, roughly speaking, the Grubbs' test iterated multiple times. See links in NIST for details.

In general, therefore, we would test for outliers (either by eye or through the methods described above), subtract our outliers, and obtain a new mean, standard deviation, mode and median using the new, reduced set of data.

In situations where Gaussianity cannot be assumed (ex. we have a lopsided RV, too few data points, etc.), then the above tests do not fare well. In these situations, we may also perform high deviation cuts, but would likely use another measure of the "centre" of the distribution. For example, if the distribution has a strong central tendency (i.e. it has a single peak), then the median (and mode) is an estimator of the central value, and would be much more robust than the mean. This is because a single, enormously deviated outlier would greatly shift the mean, but would not greatly shift the median. If outliers are much more likely to appear on one side of the real data, the mode may

[48] Some distributions actually have infinite variance; in those cases, $A = \frac{1}{N}\sum_{j=1}^{N}|x_j - \bar{x}|$.

be the most appropriate statistic, since it is insensitive to the tails of a distribution. The median and mode belong to a class of "robust" statistical estimators, where robust means that the estimator is insensitive to small departures from idealized assumptions for which the estimator is optimized. Robust estimators do well in both cases where many points depart somewhat from ideal assumptions (i.e. non-Gaussianity) and where a few points depart significantly from ideal assumptions (i.e. outliers). See Ch. 15.7 of Press et al. (1992) for details.

5.23. *Question 23*

QUESTION: Suppose you do an imaging search for binaries for a sample of 50 stars, and that you find companions in 20 cases. What binary fraction do you infer? Suppose a binary-star fraction of 50% had been found previously for another sample (which was much larger, so you can ignore its uncertainty). Determine the likelihood that your result is consistent with that fraction.

This information comes mostly from myself, with RV information from Ross (2007), my Math 318 notes, Wikipedia (2012e) and Press et al. (1992), Ch. 14.

The binary fraction inferred is 0.4. (This is obvious, since what else could be inferred without more information?).

We note that randomly sampling stars for the binary fraction is sampling a Bernoulli RV X

$$p(0) = P(X = 0) = 1 - pp(1) \qquad = P(X = 1) = p \tag{309}$$

multiple times (i.e. a weighted coin has probability p of attaining heads (1)). The outcome of n independent Bernoulli trials is distributed according to the Binomial RV:

$$p(i) = \left(\frac{n!}{(n-i)!i!}\right) p^i (1-p)^{n-i} \tag{310}$$

We have information from a much larger sample that $p = 0.5$. We will now perform a hypothesis test, that our binary search and that binary search are sampled from different distributions. Our null hypothesis is that they come from the same distribution, and if this were true, then the probability of getting a binary fraction i/n is given by $p(i)$, which is distributed according to Eqn. 310. Let us set our confidence interval at 2 σ, i.e. 95.45%[49], distributed on both sides of the mean (which is 0.5). This means that if in our system $i < i_c$ (this is often referred to as i being in the "critical region"), where i_c is defined as $p(i \leq i_c) < 0.0227$, we discard the null hypothesis and accept that our two samples come from different distributions (our confidence interval is 95.45% above *and* below the mean which means the regions excluded from the interval are beyond 97.72% above and 97.72% below the mean). Now,

$$p(i \leq i_c) = \sum_{i=0}^{i_c} p(i) = \sum_{i=0}^{i_c} \left(\frac{n!}{(n-i)!i!}\right) p^i (1-p)^{n-i}. \tag{311}$$

This should really be done on computer. $p(i \leq 18) = 0.0324$ and $p(i \leq 17) = 0.0164$, so $i_c \approx 18$. Since $20 > 18$, the two distributions could be the same, and our result of 0.4 is therefore consistent with the other study, with their 0.5[50].

5.23.1. *What is hypothesis testing?*

This information comes from Wikipedia (2012e), Press et al. (1992) Ch. 14.3, Wall & Jenkins (2003) Ch. 5.3 and my Math 318 notes.

A statistical hypothesis test is a method of making decisions using data, whether from a controlled experiment or an observational study (not controlled). In statistics, a result is called statistically significant if it is unlikely to have occurred by chance alone, according to a pre-determined threshold probability, the significance level. Our derivation above is an example of a hypothesis test: we wished to see if the two binary fraction studies were drawing from different distributions. We ended up supporting the null hypothesis, that they were in fact from the same distribution, because we showed that obtaining our result was not so improbable given the variance of the distribution governing the binary fraction.

The gamut of statistical tests to determine whether data is statistically significant is daunting (see the table in Wikipedia (2012e)). We present here three commonly used ones that may be asked about on the qualifier:

- **Student's t-test**[51] suppose you have a set of numbers sampled from a Gaussian random variable (or the number of samplings you perform is greater than 30) with unknown mean μ and unknown standard deviation σ. Let $\bar{x}$ be the statistical mean and s be the statistical standard deviation; then

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{312}$$

[49] From Wikipedia (2012c) 1 σ is 68.27%, 2 σ is 95.45%, 3 is 99.73%, 4 is 99.994% and 5 is 99.999%.
[50] We could alternatively have found $p(i \leq 20) = 0.1013$, and noted this value is above our cutoff.
[51] "Student" was the pen-name of one William Seeley Gossett.

is distributed according to the Student's t distribution. To determine error bars for the sample mean (in order to reject hypotheses like we did earlier), we note that $P(-\alpha \leq (\bar{x} - \mu) \leq \alpha) = P(\frac{-\alpha}{s/\sqrt{n}} \leq T \leq \frac{\alpha}{s/\sqrt{n}})$. We can then determine, given a probability (say, 95.45%), the value of α.

If the true σ is known, then the z-test (Wikipedia 2012e) should be used instead of the t-test.

- **Chi-square test**: consider observational data which can be binned into data points o_i, and a model/hypothesis which predicts the population of each bin e_i. The chi-square statistic describes the goodness-of-fit of the data to the model. It is defined as

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} \tag{313}$$

If χ^2 is particularly high (see pg. 87 and Table A2.6 in Wall & Jenkins (2003) for details), then it is likely the data are sampled from a distribution other than the one defined by e_i. Technically this test assumes the RVs are Gaussian, but this is a good approximation for a large dataset. In general, o_i are considered to be sampled from a different source than e_i when χ^2 exceeds $2N_{\rm bin} - 2$ (pg. 89 of Wall & Jenkins). Datasets with less than 5 bins cannot be analyzed using the Chi-square test. Minimizing the Chi-square is often used for modelling purposes.

- **Kolmogorov-Smirnov (K-S) test**: the K-S test is applied to continuous data, with no binning required. It is a four-step process: a). calculate the cumulative frequency distribution $S_e(x)$ of your predicted distribution (can be theoretical or experimental), b). calculate the cumulative frequency distribution $S_o(x)$ of your empirical data (this can be done by determining the fraction of points that have values smaller than x), c). finding

$$D = \max|S_e(x) - S_o(x)| \tag{314}$$

and d). looking D up in a table (A2.7 of Wall & Jenkins). If D is larger than a critical value, then the observed data was not sampled from the distribution defined by $S_o(x)$. The K-S test tends to excel relative to other tests for very small data sets. Sampling must be of continuous distribution functions. The K-S test cannot easily be used for modelling.

5.23.2. *In this problem, what could observationally go wrong?*

In general, astronomical surveys are biased toward brighter objects in the sky, making surveys both luminosity and distance-limited (the two are, of course, degenerate, since flux is dependent on both) (Wikipedia 2012b). The resolving power of a telescope sets a limit for observation of extended objects. Multi-wavelength surveys may be biased by different instruments having different brightness limits and resolutions. Temporal surveys are biased by the cadence (period between observations) used.

REFERENCES

Abraham, R. 2011a, Active Galactic Nuclei - AST 2040 Notes
—. 2011b, Gravitational Collapse - AST 2040 Notes
—. 2011c, The World of Galaxies - AST 2040 Notes
Arnett, D. 1996, Supernovae and Nucleosynthesis
AstroBaki. 2010, Interferometry I, https://casper.berkeley.edu/astrobaki/index.php/Interferometry_I:_2-element_interferometer%3b_interferometer_response
Baldry, I. 2008, Note on the use of stellar IMFs in cosmology, http://www.astro.ljmu.ac.uk/∼ikb/research/imf-use-in-cosmology.html
Bassett, B., & Hlozek, R. 2010, Baryon acoustic oscillations, ed. P. Ruiz-Lapuente, 246
Belczynski, K., Wiktorowicz, G., Fryer, C., Holz, D., & Kalogera, V. 2011, ArXiv e-prints
Bennett, J. O., Donahue, M., Schneider, N., & Voit, M. 2007, The Cosmic Perspective, Media Update, 5/E, ed. Bennett, J. O., Donahue, M., Schneider, N., & Voit, M.
Binney, J., & Merrifield, M. 1998, Galactic Astronomy, ed. Binney, J. & Merrifield, M.
Binney, J., & Tremaine, S. 2008, Galactic Dynamics: Second Edition, ed. Binney, J. & Tremaine, S. (Princeton University Press)
Brooks, A. M., Governato, F., Quinn, T., Brook, C. B., & Wadsley, J. 2009, ApJ, 694, 396
Bruzual A., G., & Charlot, S. 1993, ApJ, 405, 538
Burns, D., et al. 1997, MNRAS, 290, L11
Camenzind, M. 2007, Compact objects in astrophysics : white dwarfs, neutron stars, and black holes
Carroll, B. W., & Ostlie, D. A. 2006, An Introduction to Modern Astrophysics, 2nd edn. (Pearson Addison-Wesley)
Carroll, S. 2011, The Schwarzschild Solution and Black Holes, http://preposterousuniverse.com/grnotes/grnotes-seven.pdf
Chabrier, G. 2003, PASP, 115, 763
Chabrier, G., Baraffe, I., Leconte, J., Gallardo, J., & Barman, T. 2009, in American Institute of Physics Conference Series, Vol. 1094, American Institute of Physics Conference Series, ed. E. Stempels, 102–111
Chiang, E. I., & Goldreich, P. 1997, ApJ, 490, 368
Cole, G. H. A., & Woolfson, M. M. 2002, Planetary science : the science of planets around stars
Cucciati, O., et al. 2012, A&A, 539, A31
D'Alessio, P., Calvet, N., & Hartmann, L. 2001, ApJ, 553, 321
D'Alessio, P., Calvet, N., Hartmann, L., Franco-Hernández, R., & Servín, H. 2006, ApJ, 638, 314
D'Alessio, P., Calvet, N., Hartmann, L., Lizano, S., & Cantó, J. 1999, ApJ, 527, 893
Davis, T. M., & Lineweaver, C. H. 2004, PASA, 21, 97
De Angeli, F., Piotto, G., Cassisi, S., Busso, G., Recio-Blanco, A., Salaris, M., Aparicio, A., & Rosenberg, A. 2005, AJ, 130, 116
de Pater, I., & Lissauer, J. 2009, Planetary Sciences, ed. de Pater, I. & Lissauer, J.
Dodelson, S. 2003, Modern cosmology
Draine, B. T. 2011, Physics of the Interstellar and Intergalactic Medium, ed. Draine, B. T.
Dyer, C. 2010a, Christoffel Symbols and Special Coordinate Systems
—. 2010b, Just How Singular is r = 2m
—. 2010c, Orbits and Paths in Schwarzschild Spacetime
Dyson, J. E., & Williams, D. A. 1997, The physics of the interstellar medium, ed. Dyson, J. E. & Williams, D. A.
Eisenstein, D. J. 2005, NAR, 49, 360
Eisenstein, D. J., Seo, H.-J., & White, M. 2007, ApJ, 664, 660
Emberson, J. 2012, Answers to the Astronomy & Astrophysics Qualification Examination, jD's general qual notes (both files)
Faucher-Giguere, C.-A., & Quataert, E. 2012, ArXiv e-prints
Ferrière, K. M. 2001, Reviews of Modern Physics, 73, 1031
Filippenko, A. V. 1997, ARA&A, 35, 309
Fitzpatrick, E. L., & Massa, D. 2007, ApJ, 663, 320
Forbes, D. A., & Kroupa, P. 2011, PASA, 28, 77
Fortney, J. J., Marley, M. S., & Barnes, J. W. 2007, ApJ, 659, 1661
Furlanetto, S. R., & Oh, S. P. 2008, ApJ, 681, 1
Glover, S. C. O., & Clark, P. C. 2012, MNRAS, 421, 9
Graham, J. 2012, Modelling Noise in Photometry, http://www.astro.utoronto.ca/∼astrolab/Z-files/PhotNoiseModel.pdf
Gray, R. O., & Corbally, J., C. 2009, Stellar Spectral Classification
Griffiths, D. J. 2005, Introduction to Quantum Mechanics, 2nd edn. (Upper Saddle River, New Jersey: Pearson Prentice Hall)
Hahn, D. 2009, Light Scattering Theory, http://plaza.ufl.edu/dwhahn/Rayleigh%20and%20Mie%20Light%20Scattering.pdf
Hansen, C. J., Kawaler, S. D., & Trimble, V. 2004, Stellar interiors : physical principles, structure, and evolution
Hartmann, W. K. 2005, Moons & planets
Hickson, P. 2002, Astronomical and Astrophysical Measurements, from Ingrid Stairs' website, http://www.astro.ubc.ca/people/stairs/teaching/p688.pdf
—. 2010, AST402 Notes
House, E. L., et al. 2011, MNRAS, 415, 2652
Hu, W., & Dodelson, S. 2002, ARA&A, 40, 171
IPAC. 2012, Near, Mid & Far Infrared, http://www.ipac.caltech.edu/outreach/Edu/Regions/irregions.html
Jackson, N. 2008, Principles of interferometry, http://www.jb.man.ac.uk/∼njj/int.pdf
Jarosik, N., et al. 2011, ApJS, 192, 14
Kaler, J. 2006, Variable Stars on the HR Diagram, http://stars.astro.illinois.edu/sow/hrd.html
Karttunen, H., Krüger, P., Oja, H., Poutanen, M., & Donner, K. J., eds. 2007, Fundamental Astronomy
Kasen, D., & Woosley, S. E. 2009, ApJ, 703, 2205
Keel, W. 2002, Quasars and Active Galactic Nuclei, http://www.astr.ua.edu/keel/agn/
Kippenhahn, R., & Weigert, A. 1994, Stellar Structure and Evolution, ed. Kippenhahn, R. & Weigert, A.
Kissin, Y. 2012, General qualifier exam solutions, yevgeni's handwritten qual notes
Kitchin, C. R. 2009, Astrophysical Techniques, Fifth Edition (CRC Press)
Komatsu, E., et al. 2009, ApJS, 180, 330
—. 2011, ApJS, 192, 18
Koval', V. V., Marsakov, V. A., & Borkova, T. V. 2009, Astronomy Reports, 53, 785
Kroupa, P., Weidner, C., Pflamm-Altenburg, J., Thies, I., Dabringhausen, J., Marks, M., & Maschberger, T. 2011, ArXiv e-prints
Le Feuvre, M., & Wieczorek, M. A. 2011, Icarus, 214, 1
Longair, M. S. 2008, Galaxy Formation, ed. Longair, M. S.
Marleau, F. 2010, AST1440 Notes
Marsakov, V. A., Koval', V. V., Borkova, T. V., & Shapovalov, M. V. 2011, Astronomy Reports, 55, 667
Mészáros, P. 2005, Mészáros Effect, http://www2.astro.psu.edu/users/nnp/Meszaros_effect.pdf
Mieda, E. 2012, General qualifier exam solutions, etsuko's various qual notes
Navarro, J. F., Frenk, C. S., & White, S. D. M. 1996, ApJ, 462, 563
Nelson, L. A., Rappaport, S., & Chiang, E. 1993, ApJ, 413, 364
Nguyen, L. 2012, Star Formation 101, part of 2012 DI Summer Student Lecture
NIST. 2012, Detection of Outliers, http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm
Nitschelm, C. 2011, Hertzsprung-Russell Diagram, http://www.astrosurf.com/nitschelm/HR_diagram.jpg
Page, D. N. 2011, ArXiv e-prints
Page, M. J., et al. 2012, Nature, 485, 213
Pastorello, A. 2012, Memorie della Societa Astronomica Italiana Supplementi, 19, 24
Pogge, R. 2011, Introduction to the Interstellar Medium, www.astronomy.ohio-state.edu/∼pogge/Ast871/Notes
Polletta, M. 2006, SWIRE Template Library, http://www.iasf-milano.inaf.it/∼polletta/templates/swire_templates.html
Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical recipes in C. The art of scientific computing
Raskin, C., Timmes, F. X., Scannapieco, E., Diehl, S., & Fryer, C. 2009, MNRAS, 399, L156

Renzini, A. 2006, ARA&A, 44, 141
Riess, A. G., et al. 2004, ApJ, 607, 665
Robbins, S., & McDonald, M. 2006, A Brief History of the Universe, http://burro.astr.cwru.edu/stu/advanced/cosmos_history.html
Roberts, M. S., & Haynes, M. P. 1994, ARA&A, 32, 115
Ross, S. M. 2007, Introduction to Probability Models
Roy, N., Chengalur, J. N., & Srianand, R. 2006, MNRAS, 365, L1
Rybicki, G. B., & Lightman, A. P. 1979, Radiative processes in astrophysics, ed. Rybicki, G. B. & Lightman, A. P.
Ryden, B. 2003, Introduction to cosmology, ed. Ryden, B.
Ryter, C. E. 1996, Ap&SS, 236, 285
Saleh, B. E. A., & Teich, M. 2007, Fundamentals of Photonics, ed. Saleh, B.E.A.
Santos, N. C. 2008, NewAR, 52, 154
Schneider, P. 2006, Extragalactic Astronomy and Cosmology, ed. Schneider, P.
SDSS. 2012, Algorithms: Target Selection, http://www.sdss.org/dr5/algorithms/target.html
Seabroke, G. M., & Gilmore, G. 2007, MNRAS, 380, 1348
Seager, S. 2010, Exoplanets, ed. Seager, S.
Shaw, D. J., & Barrow, J. D. 2011, Phys. Rev. D, 83, 043518
Shim, H., Colbert, J., Teplitz, H., Henry, A., Malkan, M., McCarthy, P., & Yan, L. 2009, ApJ, 696, 785
Sijacki, D., Springel, V., Di Matteo, T., & Hernquist, L. 2007, MNRAS, 380, 877
Smith, B., Sigurdsson, S., & Abel, T. 2008, MNRAS, 385, 1443
Spruit, H. C. 2010, ArXiv e-prints
Stolte, A. 2011, Pre-main sequence evolution, http://www.astro.uni-bonn.de/∼astolte/StarFormation/Lecture2011_PMS.pdf
Sturge, M. 2003, Statistical and Thermal Physics, 1st edn. (A.K. Peters/CRC Press)
Townsend, P. K. 1997, ArXiv General Relativity and Quantum Cosmology e-prints
Valencia, D., O'Connell, R. J., & Sasselov, D. D. 2007, ApJ, 670, L45
van den Bergh, S. 2008, MNRAS, 385, L20
van Kerkwijk, M. 2011, AST3010 Transients Notes, http://www.astro.utoronto.ca/ mhvk/AST3010/
—. 2012, Private communication, oral
Vardanyan, M., Trotta, R., & Silk, J. 2009, MNRAS, 397, 431
Wall, J. V., & Jenkins, C. R. 2003, Practical Statistics for Astronomers, ed. R. Ellis, J. Huchra, S. Kahn, G. Rieke, & P. B. Stetson
Weijmans, A.-M., et al. 2009, MNRAS, 398, 561
Wikipedia. 2011a, Cosmic Distance Ladder, http://en.wikipedia.org/wiki/Cosmic_distance_ladder
—. 2011b, Geology of Solar Terrestrial Planets, http://en.wikipedia.org/wiki/Geology_of_solar_terrestrial_planets
—. 2011c, Hertzsprung-Russell Diagram, http://en.wikipedia.org/wiki/Hertzsprung%E2%80%93Russell_diagram
—. 2011d, Timeline of the Big Bang, http://en.wikipedia.org/wiki/Timeline_of_the_Big_Bang
—. 2012a, Anthropic principle, http://en.wikipedia.org/wiki/Anthropic_principle
—. 2012b, Bondi accretion
—. 2012c, Earth, http://en.wikipedia.org/wiki/Earth
—. 2012d, Etendue, http://en.wikipedia.org/wiki/Etendue
—. 2012e, Galaxy cluster, http://en.wikipedia.org/wiki/Galaxy_cluster
wikipedia. 2012, Hawking radiation, http://en.wikipedia.org/wiki/Hawking_radiation
Wikipedia. 2012a, Inflation (cosmology), http://en.wikipedia.org/wiki/Inflation_(cosmology)
—. 2012b, Malmquist bias, http://en.wikipedia.org/wiki/Malmquist_bias
—. 2012c, Normal distribution, http://en.wikipedia.org/wiki/Normal_distribution
—. 2012d, Sample mean and sample covariance, http://en.wikipedia.org/wiki/Sample_mean_and_sample_covariance
—. 2012e, Statistical hypothesis testing, http://en.wikipedia.org/wiki/Hypothesis_test
—. 2012f, Sunyaev-Zel'dovich effect, http://en.wikipedia.org/wiki/Sunyaev%E2%80%93Zel'dovich_effect
—. 2012g, Supermassive black hole, http://en.wikipedia.org/wiki/Supermassive_black_hole
Wolfram. 2012, Wolfram Alpha Search Engine
Wood, K. 2011, Nebulae, http://www-star.st-and.ac.uk/∼kw25/teaching/nebulae/nebulae.html
Yoon, S.-C., & Langer, N. 2004, A&A, 419, 623
Zhu, C. 2010, AST1440 Notes
—. 2011, Transients Resulting From the Merger of Degenerate Objects, aST3010 Report