# AN EDUCATIONAL NOTE ON THE SAHA EQUATION AND ITS SOLUTION FOR THE IONIZATION STATE OF A GAS

David J. Jeffery[1]

## ABSTRACT

The solution of the Saha equation (understood to be actually the solution of what we call in this note the Saha-CC equation) for the ionization state of a gas in local thermodynamic equilibrium (LTE) must in general be done numerically. There are some tricks needed to do the solution accurately and robustly. In this education note, we elucidate some aspects of the Saha equation and its solution not expanded on by Mihalas' book *Radiative Transfer* and try elucidate some of the tricks.

*Subject headings:* atomic processes — methods: numerical — radiative transfer — stars: atmospheres — supernovae: general

## 1. INTRODUCTION

The solution of the Saha equation (understood to be actually the solution of what we call in this note the Saha-CC equation) gives the ionization state (i.e., the ionization occupation numbers) for an ionized gas (i.e., a plasma) in the state of local thermodynamic equilibrium (LTE). Except for the hydrogenic case (neutral and singly-ionized for one species), there is no simple analytic solution. The numerical solution given density and composition is in principle straightforward, but tricks are needed to do the solution accurately and robustly. In this education note, we try elucidate the Saha equation and some of those tricks.

As indicated in the first paragraph the (formal) Saha equation itself is not what is solved for the ionization state. It is an equation combining the Saha equation for all the ionization stages of all the elements with charge conservation (i.e., overall charge neutrality) that is solved for the ionization state. This combined equation seems to have no common name, and so we call it the aforementioned Saha-CC equation since we need a name for it.

---

[1]Department of Physics & Astronomy, University of Las Vegas, Nevada, 4505 S. Maryland Parkway Las Vegas, Nevada 89154, U.S.A.

It should be pointed out that the assumption of LTE is extremely simplifying and when applicable extremely useful in obtaining the ionization state. One feature of LTE ionization states is that usually only 1 or 2 ionization stages of an element are abundant. We will show why this is so in § 5????.

Note that in many astrophysical environments non-LTE (NLTE) holds and the LTE ionization state may be very wrong. In particualar, there may be 3 or more abundant ionization states for an element.

Supernovae in the photospheric phase are peculiar case. Parameterized synthetic spectra using the the LTE ionization state can give qualitatively good fits to observed spectra (e.g., Jeffery & Branch 1990, esp. p.170, 214), suggesting the LTE ionization state is qualitatively a valid approximation for the photospheric phase. However, it is well known that supernova atmospheres are pretty much always NLTE for at least three reasons:

1. The density is too low to enforce thermodynamic equalilibrium via collisions.

2. The radiation is in general non-Planckian both in shape and scale everywhere in the atmosphere, and so cannot enforce LTE in the matter.

3. Radioactive species mixed far outward into the ejecta can lead to non-thermal ionization via ionizing radiations.

Thus, quantitative accuracy in the calculation of synthetic spectra for supernovae needs NLTE in general.

In § 2 of this note, we expand on a fine point of the derivation of the Saha equation. Section 3 presents the Saha equation and the inverse Saha equation. Section 4 presents the hydrogenic Saha-CC equation and its analytic solution. Analytic solutions of the Saha-CC equation are derived in § 6. Section 7 presents what we call the ionization function which is useful in understanding the numerical solution of the Saha-CC equation. Section 8 describes the numerical solution of the Saha-CC equation. Conclusions are in § 9. Appendix A presents numerically good solutions of the quadratic equation: these are needed in the numerical solution of the Saha-CC equation.

## 2.   THE STATISTICAL WEIGHT OF FREE ELECTRONS

One fine point of the derivation of the Saha equation not elucidated by Mihalas (1978, p. 112) is reason for assigning the volume element of the (differential) statistical weight for

free electrons to be $n_e^{-1}$, the inverse of the electron density of the gas which is just the volume per electron $V_e$. Here we give the elucidation.

In fact, statistical weights for different degrees of freedom are multiplied to get the total statistical weight and also, in fact, the statistical weights appear only in ratios in Saha equation Mihalas (1978, p. 112–113). There is one ratio for the atomic energy levels and another for the free electrons and these multiplied together. Now the statisical weights for an differentially narrow energy band (containing states $g_e$) for free electrons (or any fermions) of number $N_e$ is

$$w_e(N_e) = \frac{g_e!}{N_e!(g_e - N_e)!} \tag{1}$$

(e.g., Pointon 1978, p. 61). The ratio of statistical weights for free electrons for the case one ionization stage to the next higher one is then

$$\frac{w_e(N_e + 1)}{w_e(N_e)} = \frac{N_e!}{(N_e + 1)!} \frac{(g_e - N_e)!}{[g_e - (N_e + 1)]!} = \frac{g_e - N_e}{N_e + 1} \approx \frac{g_e}{N_e} , \tag{2}$$

where the approximation holds for $g_e >> N_e$ (i.e., a nondegenerate gas: Pointon e.g., 1978, p. 65) and large numbers of electrons. The approximate ratio holds for applications of the Saha equation. The number of states $g_e$ follows from box quantization

$$g_e = g \frac{d^3k}{(2\pi)^3} V , \tag{3}$$

where spin degeneracy $g = 2$ for electrons (and all fermions), $d^3k/(2\pi)^3$ is the differential wavenumber volume available for electrons with wavenumber $\vec{k} = (k_x, k_y, k_z)$, and $V$ is the volume of the box. The $g_e$ formula is easily derived a rectangular box of volume $V$. Although no proof ever seems to be given, the box quantization formula holds asymptotically for large irregular volume or locally uniform gases where there is no specified boundaries at all and volume is indeterminate, but not needed for calculations.

Given equation (3), the ratio of statistical weights for free electrons (for the nondegenerate gas case) is

$$\frac{w_e(N_e + 1)}{w_e(N_e)} = g \frac{d^3k}{(2\pi)^3} \frac{V}{N_e} = g \frac{d^3k}{(2\pi)^3} \frac{1}{n_e} \tag{4}$$

which combined with other relevant factors (including the appropriate Boltzmann factor) is used in the derivation of the Saha equation (Mihalas 1978, p. 112–113).

## 3. THE SAHA EQUATION

The Saha equation is

$$\frac{N_{jk}}{N_{j+1,k}} = n_e \Phi_{jk}(T) = n_e C_I \frac{U_{jk}(T)}{U_{j+1,k}(T)} T^{-3/2} \exp\left(\frac{X_{jk}}{kT}\right) , \qquad (5)$$

where $k$ is atom/element (or the Boltzmann constant in a different meaning understood from context), $j$ is ionization stage, $T$ is temperature, $\Phi_{jk}(T)$ can be called the $\Phi$ function, $N_{jk}$ is number density of an ionization stage, $U_{jk}(T)$ is the partition function of an ionization stage, $X_{jk}$ is the ionization energy from stage $j$ to state $j+1$, and

$$C_I = \frac{1}{2}\left(\frac{h^2}{2\pi m_e k}\right)^{3/2} = (2.070\ 665\ 141\ 6)\times 10^{-16}\,\mathrm{cm^3\,K^{3/2}} \qquad (6)$$

can be called the Saha constant and has approximately 11 significant digits based on CO-DATA constants (NIST CODATA, 2020). The notation for the Saha equation follows Mihalas (1978, p. 112–113) with some obvious simplifications. The inverse Saha equation is

$$\frac{N_{j+1,k}}{N_{jk}} = \frac{1}{n_e}\Psi_{jk}(T) = \frac{1}{n_e C_I}\frac{U_{j+1,k}(T)}{U_{jk}(T)}T^{3/2}\exp\left(-\frac{X_{jk}}{kT}\right) , \qquad (7)$$

where $\Psi_{jk}(T)$ can be called the $\Psi$ function. The inverse Saha equation is useful in understanding the solution for ionization state and electron density of a gas.

It is enlightening to the write the Saha equation and inverse Saha equation in terms of fiducial quantities. We have chosen those quantities to be hydrogen in neutral and once ionized stages at temperature $T = 10^4$ which is approximately effective temperature of A0 stars (Wikipedia: A-type main-sequence star). The formulae are

in terms of natural units with $kT_{\mathrm{Rydberg}} = E_{\mathrm{Rydberg}} = (13.605\ 693\ 122\ 994)\,\mathrm{eV}$ (i.e., the Rydberg energy) and $T_{\mathrm{Rydberg}} = E_{\mathrm{Rydberg}}/k = 157887.5124\,K$ (NIST CODATA, 2020). One obtains

$$\frac{N_{j+1,k}}{N_{jk}} = \frac{1}{n_e}\Psi_{jk}(T) = \frac{U_{j+1,k}(T)}{U_{jk}(T)}\frac{n_{e,\mathrm{fid}}}{n_e}\left(\frac{kT}{E_{\mathrm{Rydberg}}}\right)^{3/2}\exp\left(-\frac{X_{jk}}{kT}\right) , \qquad (8)$$

where the fiducial Rydberg (electron density is

$$n_{e,\mathrm{fid}} = 2\left(\frac{2\pi m_e E_{\mathrm{Rydberg}}}{h^2}\right)^{3/2} = (3.029\ 784\ 927\ 541\ 21)\times 10^{23}\,\mathrm{cm^{-3}} \qquad (9)$$

which has approximately 14 significant digits based on CODATA constants (NIST CODATA, 2020). Obviously, the inverse Saha equation for hydrogen for ionization states 0 and 1 is

$$\frac{N_{1,1}}{N_{0,1}} = \frac{1}{n_e}\Psi_{0,1}(T) = \frac{U_{1,1}(T)}{U_{0,1}(T)}\frac{n_{e,\mathrm{fid}}}{n_e}\left(\frac{kT}{E_{\mathrm{Rydberg}}}\right)^{3/2}\exp\left(-\frac{E_{\mathrm{Rydberg,reduced}}}{kT}\right) , \qquad (10)$$

where $E_{\text{Rydberg,reduced}} = E_{\text{Rydberg}}(m_{e,\text{reduced}}/m_e) = 13.598\ 287\ 264\ 287\,\text{eV}$ is the reduced

In fact, $n_{e,\text{fid}}$ is about 10000 times higher than the density of air at standard temperature and pressure (Wikipedia: Number Density: Examples; Wikipedia: Density of Air: Dry Air) and is much higher than the particle number density (which is comparable to the electron density which is part of it) in most optically thin astrophysical plasmas. If temperatures in those plasmas are of order $T_{\text{Rydberg}}$ or greater, then they will be ionized or highly ionized.

Supernovae in the photospheric phase are peculiar case as mentioned in § 1. The Saha equation predicts their ionization state not so badly even though there are reasons why it should not: see § 1 for the reasons. Typically supernovae have singly or doubly ionized atoms as known from spectroscopy. The inverse Saha equation in terms of natural units does not, alas, allow a by inspection prediction of ionization for supernovae. Supernovae in the photospheric and nebular eras typically have particle densities many orders of magnitude less than $n_{e,\text{fid}}$, but temperatures of order a few a few thousand kelvins much less than $T_{\text{Rydberg}}$. One has to actually calculate the Saha equation ionization prediction numerically for particular cases.

## 4. THE HYDROGENIC SAHA-CC EQUATION

The Saha equation combined with charge conservation gives an equation can be used to calculate the electron density. We will call this equation the Saha-CC equation, where CC stands for charge conservation. When speaking one often just calls this equation the Saha equation too: context tells you which equation is meant.

For the hydrogenic case where one has only neutral and singly ionized (and is not considering the negative ion: e.g., the $H^-$ ion (i.e., hydride ion)), the Saha-CC equation is simple derive and can be solved analytically. The derivation is obviously

$$n_e = N_k \left( \frac{N_{1,k}}{N_{0,k} + N_{1,k}} \right) = N_k \left( \frac{1}{n_e \Phi_{0,k} + 1} \right) = N_k \left( \frac{1}{n_e/\Psi_{0,k} + 1} \right) , \qquad (11)$$

where here $k$ labels a general atom treated as hydrogenic. For the scaled solution, we define scaled electron density and atom density by, respectively,

$$x = n_e \Phi_{0,k} = \frac{n_e}{\Psi_{0,k}} \qquad \text{and} \qquad X = N_k \Phi_{0,k} = \frac{N_k}{\Psi_{0,k}} , \qquad (12)$$

both of which are small for high ionization ($\Psi_{0,k}$ large) and large for low ionizaion ($\Psi_{0,k}$ small). The scaled hydrogenic Saha-CC equation is

$$x = \frac{X}{x+1} , \qquad (13)$$

which is just a quadratic equation. The scaled solutions are

$$x = \begin{cases} \dfrac{1}{2}(\sqrt{1+4X}-1) & \text{numerically-poor exact solution;} \\[2mm] \dfrac{2X}{1+\sqrt{1+4X}} & \text{numerically-good exact solution;} \\[2mm] X - X^2 + 2X^3 + \ldots \approx X & \text{small } X \text{ (high ionization) series solution;} \\[2mm] -\dfrac{1}{2} + \sqrt{X}\Big(1 + \dfrac{1}{8}\dfrac{1}{X} - \dfrac{1}{128}\dfrac{1}{X^2} & \\[2mm] \qquad\qquad + \dfrac{1}{4096}\dfrac{1}{X^3} + \ldots\Big) \approx \sqrt{X} & \text{large } X \text{ (low ionization) series solution;} \\[2mm] \dfrac{\sqrt{5}-1}{2} = \phi - 1 = 0.6018\ldots & \text{for } X = 1, \text{ where } \phi \text{ is the golden ratio;} \\[2mm] 1 & \text{for } X = 2. \end{cases} \tag{14}$$

The numerically-poor exact solution is poor because for $X$ small one gets roundoff from the subtraction between nearly equal values. On the other hand, the numerically-good exact solution agrees with both small/large $X$ series solutions to within a relative error of $\lesssim 10^{-18}$ as $X$ goes to zero/infinity. So the numerically-good exact solution is good everywhere and should be used in numerical calculations.

Note it is slightly paradoxical that what we call high/low ionization in equation (14) gives the lower/higher electron density. The resolution of the paradox, which holds for all plasma cases, is just that atoms are in a high/low ionization state for low/high atom density, and therefore low/high electron density. Atoms are just more recombined at high density. For LTE calculations, this has nothing do to with recombination rates which could be anything. It is because the allowed phase space for free electrons decreases as their density increases, forcing a larger number into the bound state. The allowed phase space is, of course, limited by the Pauli exclusion principle. The average ionization stage, which we call the ionization function, is electron density over atom density when charge conservation holds (i.e., the plasma is has overall neutrality) For the hydrogenic case, the ionization function is

$$\frac{x}{X} = \begin{cases} \dfrac{2}{1+\sqrt{1+4X}} & \text{exact formula;} \\[2mm] 1 - X + 2X^2 + \ldots & \text{small } X \text{ (high ionization) formula;} \\[2mm] -\dfrac{1}{2X} + \dfrac{1}{\sqrt{X}}\Big(1 + \dfrac{1}{8}\dfrac{1}{X} - \dfrac{1}{128}\dfrac{1}{X^2} & \\[2mm] \qquad\qquad + \dfrac{1}{4096}\dfrac{1}{X^3} + \ldots\Big) \approx \dfrac{1}{\sqrt{X}} & \text{large } X \text{ (low ionization) series solution.} \end{cases} \tag{15}$$

The formula shows clearly that hydrogenic ionization decreases with $X$ with the only stationary point at $X = \infty$.

The exact solution for the hydrogenic Saha-CC equation (eq. (14) above) can be used to calculate an approximate electron density in general (for LTE). Let us call this the 1st hydrogenic approximation. For this approximation, one simply treats all protons as if they were hydrogen atoms and $N_k$ in the formulae above is replaced by the number density of protons $N_{\mathrm{p}}$ for which the formula is

$$N_{\mathrm{p}} = \rho \sum_k \frac{X_k Z_k}{A_k m_{\mathrm{u}}} \ , \tag{16}$$

where $\rho$ is density, $X_k$ is mass fraction, $Z_k$ is atomic number, $A_k$ is atomic mass, and $m_{\mathrm{u}}$ is the atomic mass unit (amu or u).

The 1st hydrogenic approximation will obviously give an overestimate of the electron density almost always since beyond the first ionization energy most ionization energies are greater than the reduced Rydberg. However, the approximation gives the right answer in two important limits: 1) for pure hydrogen (neglecting the negative ion); 2) for completely ionized atoms. Also however, the overestimate is probably at worst a factor of $\sim 30$ since the iron-peak element ejecta of supernovae Ia (SNe Ia) is probably the highest metalicity ejecta in astrophysics. A factor $\sim 30$ is not bad if the calculated electron density is only used as initial value in numerical solution for electron density from the (general) Saha-CC equation (see § 5 below).

Note negative ions are probably always a trace abundance in astrophysical plasma, and so probably can always be neglected in calculating electron density. However, negative ions can have high opacity, and so be important in radiative transfer, and so can have a significant indirect effect on electron density. For important example, the negative hydrogen ion ($H^-$ or hydride) is an important opacity source in solar type stars and stars cooler than solar type stars (e.g., Mihalas 1978, p. 104). Note overall charge conservation requires that there must be positive ions if there are negative ions.

We can do somewhat better than 1st hydrogenic approximation with what we will call the 2nd hydrogenic approximation which handles the low and high ionization limits exactly and interpolates between them approximately with an overestimate usually.

For low ionization limit, we treat all atoms as have just the stages 0 and 1. The Saha-CC equation becomes for the low ionization limit is

$$n_e = \sum_k N_k \left( \frac{1}{n_e/\Psi_{0,k} + 1} \right) = \sum_k \frac{N_k \Psi_{0,k}}{n_e} \ , \tag{17}$$

where both equalities are in the limit of low ionization: i.e., all $\Psi_{0,k} << 1$ and all $\Psi_{j>0,k}$ considered negligible. By inspection, the low ionization limit solution for $n_e$ is

$$n_e = \sqrt{\sum_k N_k \Psi_{0,k}} = \sqrt{N_a \sum_k \frac{N_k}{N_a} \Psi_{0,k}} = \sqrt{N_a \sum_k f_k \Psi_{0,k}} = \sqrt{N_a \Psi_{0,a}} , \qquad (18)$$

where we have used following formulae

$$N_a = \rho \sum_k \frac{X_k}{A_k m_u} , \qquad f_k = \frac{N_k}{N_a} , \qquad \text{and} \qquad \Psi_{0,a} = \sum_k f_k \Psi_{0,k} , \qquad (19)$$

where $N_a$ is the number density of atoms, $f_k$ is the number fraction of atom $k$, and $\Psi_{0,a}$ is the weighted mean $\Psi$ function. The low ionization limit solution shows that just as for hydrogenic solution, the electron density for low ionization state goes as the square root of the number of atoms. We can also write a scaled low ionization limit solution:

$$x = \sqrt{X} , \qquad (20)$$

with definitions $x = n_e/\Psi_{0,a}$ and $X = N_a/\Psi_{0,a}$.

The 2nd hydrogenic approximation itself is the interpolation formula

$$n_e = \begin{cases} \dfrac{N_a}{N_a/N_p + \sqrt{N_a/\Psi_{0,a}}} & \text{in general;} \\[2ex] \sqrt{N_a \Psi_{0,a}} & \text{in the low ionization limit (i.e., small } \Psi_{0,a} \text{ limit);} \\[1ex] N_p & \text{in the high ionization limit (i.e., large } \Psi_{0,a} \text{ limit).} \end{cases} \qquad (21)$$

The 2nd hydrogenic approximation is exactly correct in the two limits as mentioned above neglecting, of course, negative ions and molecule formation. However, in between the two limits it will be usually be an overestimate as mentioned above since the $\Psi_{0,k}$ functions will become large sufficiently many atoms are nearly all singly ionized and this well before the high ionization limit. However, as argued above for the 1st hydrogenic approximation, the overestimate is probably at worst a factor of $\sim 30$ since the iron-peak element ejecta of supernovae Ia (SNe Ia) is probably the highest metalicity ejecta in astrophysics. A factor $\sim 30$ is not bad if the calculated electron density is only used as initial value in numerical solution for electron density from the general Saha-CC equation.

The advantage of the 2nd hydrogenic approximation over the 1st hydrogenic approximation is that it is exact for general compositions in the low ionization state limit (i.e., small $\Psi_{0,a}$ limit)

## 5.   THE SAHA-CC EQUATION

The Saha-CC equation (i.e., the general Saha-CC equation) is

$$n_e = N_{\mathrm{a}} \sum_k f_k \sum_{j=L_k}^{J_k} j f_{jk} \; , \tag{22}$$

where recall $N_{\mathrm{a}}$ is the number density of atoms, $f_k$ is the number fraction for atom $k$, and $f_{jk}$ is the number fraction for ionization stage $j$ of atom $k$.

The formula for $f_{jk}$ is

$$f_{jk} = \frac{N_{jk}}{N_k} = \frac{N_{jk}}{\sum_{m=L_k}^{J_k} N_{mk}} = \frac{N_{jk}/N_{M_k k}}{\sum_{m=L_k}^{J_k} N_{mk}/N_{M_k k}} \; , \tag{23}$$

where $L_k$ is the lowest ionization stage considered (usually zero for the neutral atom or $-1$ for singly-charge negative atom: e.g., the $\mathrm{H}^-$ ion (i.e., hydride ion)), $J_k$ is the highest ionization stage considered (the highest possible being the atomic number), and $M_k$ is a fiducial ionization stage. Note that $L_k$, $J_k$, and $M_k$ are all functions of $k$ (i.e., of the atom under consideration). Since, in fact, the ionization state of an atom is usually completely dominated by one or at most two ionization stages, it is sometimes mentally clarifying and probably numerically best to choose $M$ to be the ionization stage of maximum occupation number. Note that Mihalas (1978, p. 114) sets $M_k$ to $J_k$ in all cases which choice has the advantage that the formulae for the $f_{jk}$ have only positive powers of $n_e$ in their numerators and denominators and look simpler than for other choices of $M_k$. We make use of this advantage below in § 6.

Now what can be called the generalized Saha equation is

$$
\frac{N_{jk}}{N_{M_k k}} =
\begin{cases}
\dfrac{N_{M_k+1,k}}{N_{M_k k}} \cdots \dfrac{N_{jk}}{N_{j-1,k}} = \displaystyle\prod_{\ell=M_k}^{j-1} (n_e \Phi_{\ell k})^{-1} & \\[2ex]
\qquad\qquad\qquad = n_e^{(M_k-j)} \displaystyle\prod_{\ell=M_k}^{j-1} \Psi_{\ell k} & \text{for } j > M_k; \\[3ex]
\dfrac{N_{jk}}{N_{j+1,k}} \cdots \dfrac{N_{M_k-1,k}}{N_{M_k k}} = \displaystyle\prod_{\ell=j}^{M_k-1} (n_e \Phi_{\ell k}) & \\[2ex]
\qquad\qquad\qquad = n_e^{(M_k-j)} \displaystyle\prod_{\ell=j}^{M_k-1} \Phi_{\ell k} & \text{for } j < M_k; \\[3ex]
\displaystyle\prod_{\ell=M_k\pm1}^{j} \left( n_e \Phi_{\left\{\substack{\ell-1\\\ell}\right\},k} \right)^{\mp1} = n_e^{(M_k-j)} \left( \Phi_{\left\{\substack{\ell-1\\\ell}\right\},k} \right)^{\mp1} & \text{where the upper case} \\
& \text{is for } j > M_k \\
& \text{and the lower case} \\
& \text{is for } j < M_k.
\end{cases}
\tag{24}
$$

Note that the product symbol expressions for $j = M_k$ is set to 1 in all cases in equation (24). Also note the denominator terms in equation (23) are, of course, given by equation (24) with the $j$ replaced by the dummy index $m$.

If we choose $M_k = J_k$ (as per Mihalas 1978, p. 114), the Saha-CC equation is

$$
n_e = N_{\mathrm{a}} \sum_k f_k \frac{\sum_{j=L_k}^{J_k} j\, n_e^{J_k-j} \prod_{\ell=j}^{J_k-1} \Phi_{\ell k}}{\sum_{j=L_k}^{J_k} n_e^{J_k-j} \prod_{\ell=j}^{J_k-1} \Phi_{\ell k}} \, ,
\tag{25}
$$

where the product symbol expressions for $j = J_k$ are set to 1 in all cases and for cases where $j = 0$ there is a zero term in the numerators. Equation (25) explicitly shows the powers of $n_e$ dependence of Saha-CC equation and, as aforesaid, has only positive powers of $n_e$ in the numerators and denominators, and so looks simpler than for other choices of $M_k$.


## 6.   ANALYTIC SOLUTIONS OF THE SAHA-CC EQUATION

The Saha-CC equation cannot be solved analytically in general. In the hydrogenic case, it can easily be solved analytically as we have already shown in § 4 since it is just a quadratic equation. Analytic solutions are also possible as long as there are only powers of $n_e$ less than or equal to $n_e^4$ occur when the Saha-CC equation is rearranged as a polynomial (Wikipedia: Quartic function: History) and in at least one other case as we show below.

To elucidate possible analytic solutions, let's consider a gas consisting of hydrogen ($k = 1$), helium ($k = 2$), and lithium ($k = 3$) which is of interest since an example actually occurs: i.e., the primordial cosmic composition given by Big Bang nucleosynthesis (Wikipedia: Big Bang nucleosynthesis). In this case (allowing for $H^-$ but not other negative ions), equation (25) from § 5 specializes to

$$
\begin{aligned}
n_e \;=\; & N_1 \left( \frac{-n_e^2 \Phi_{-1,1} \Phi_{0,1} + 1}{n_e^2 \Phi_{-1,1} \Phi_{0,1} + n_e \Phi_{0,1} + 1} \right) + N_2 \left( \frac{n_e \Phi_{1,2} + 2}{n_e^2 \Phi_{0,2} \Phi_{1,2} + n_e \Phi_{1,2} + 1} \right) \\
& + N_3 \left( \frac{n_e^2 \Phi_{1,3} \Phi_{2,3} + 2 n_e \Phi_{2,3} + 3}{n_e^3 \Phi_{0,3} \Phi_{1,3} \Phi_{2,3} + n_e^2 \Phi_{1,3} \Phi_{2,3} + n_e \Phi_{2,3} + 1} \right) \; .
\end{aligned}
\tag{26}
$$

We see that cases for hydrogen, helium, and lithium alone all have analytic solutions since they give, respectively, a cubic, a cubic, and a quartic. However, no combination of the atoms gives an analytic solution since one gets quintic or higher order polynomial equations. But if we neglect the negative ion state for hydrogen, the combination of hydrogen and helium gives a quartic, and so does have analytic solution.

In fact, general cubic and quartic analytic solutions are complicated, and so not useful for simple understanding. Also one has to identify the physically correct root and perhaps deal with numerical accuracy problems in some limits. The upshot is that fast, accurate numerical solutions to the cubic and quartic Saha-CC equations is probably preferrable, unless one has done a careful analysis to show that an analytic solution is better.

The aforementioned "at least one other" analytic solution can be obtained for the unrealistic case where all the $\Phi/\Psi$ functions are equal and we have ionization stages zero to infinity. Let's call this case the infinite-ionization-stage case. To start on getting the solution, we define the scaled inverse electron density by

$$
y = x^{-1} = (n_e \Phi)^{-1} = \frac{\Psi}{n_e} < 1 \; ,
\tag{27}
$$

where $\Phi$ and $\Psi$ are the common equal $\Phi$ and $\Psi$ functions, and $x$ is the scaled electron density as in § 4, eq. (12). For the infinite-ionization-stage case, the Saha-CC equation (eq. 22) with $M_k = 0$ and $X = N_a/\Psi$ (as in § 4, eq. (12)) specializes to

$$
x = y^{-1} = X \frac{\sum_{j=0}^{\infty} j y^j}{\sum_{j=0}^{\infty} y^j} = \frac{X y (\partial/\partial y)[\sum_{j=0}^{\infty} y^j]}{\sum_{j=0}^{\infty} y^j} = \frac{X y (1-y)^{-2}}{(1-y)^{-1}} = \frac{X y}{1 - y} \; ,
\tag{28}
$$

where we have used the infinite geometric series formula which converges for $y < 1$ (e.g.,

Arfken 1985, p. 279). Equation (28) is a quadratic equation for $y$ with solutions

$$x = y^{-1} = \begin{cases} \dfrac{2X}{\sqrt{1+4X}-1} & \text{numerically-poor exact solution;} \\[2ex] \dfrac{1}{2}(\sqrt{1+4X}+1) & \text{numerically-good exact solution;} \\[2ex] 1 + X - X^2 + 2X^3 + \ldots \approx 1 + X & \text{small } X \text{ (high ionization) series solution;} \\[2ex] \dfrac{1}{2} + \sqrt{X}\Big(1 + \dfrac{1}{8}\dfrac{1}{X} - \dfrac{1}{128}\dfrac{1}{X^2} \\[2ex] \qquad\quad + \dfrac{1}{4096}\dfrac{1}{X^3} + \ldots\Big) \approx \sqrt{X} & \text{large } X \text{ (low ionization) series solution;} \\[2ex] \dfrac{\sqrt{5}+1}{2} = \phi = 1.6018\ldots & \text{for } X = 1, \text{ where } \phi \text{ is the golden ratio;} \\[2ex] 2 & \text{for } X = 2. \end{cases}$$
$$(29)$$

At this point, one should recall the discussion in § 4 explicating why small/large $X$ gives high/low ionization.

The infinite-ionization-stage case scaled solutions are similar to the scaled hydrogenic Saha equation solutions and the two solutions are the same asymptotically in the limit that $X$ becomes large: they both have $x \to \sqrt{X}$ as $X$ becomes large. This is understandable since as $X$ becomes large, the zeroth and first ionization stages of infinite-ionization-stage atom become asymptotically the only non-neglegible ones: i.e., the infinite-ionization-stage atom becomes asymptotically the hydrogenic atom. On the other hand, as $X$ becomes small there is a profound difference between the hydrogenic case and the infinite-ionization-stage case. In the former, asymptotically $x = X$ (electron density scales with atom density) and in the latter $x = 1 + X$ (electron density approaches a constant value). The constant electron density (scaled value 1) as $X$ approaches zero is the limit electron density for our formalism for infinitely many ionization stages with atom density going to zero.

Note that $x = 1$ (implying $y = 1$) violates our assumption in deriving the infinite-ionization-stage case scaled solutions that $y < 1$. However, $x = 1$ as a limit is still valid.

As pointed out above, the infinite-ionization-stage case is unrealistic, but it is interesting in that it does have an analytic solution.

## 7. THE IONIZATION FUNCTION

What we called the ionization function $F_{\rm I}$ in § 4 is just the average ionization stage. In § 4, we gave the ionization function for the hydrogenic atom. For the general ionization

function, obtained using the right-hand side of the Saha-CC function in the form of equation (25) (see § 5), we find

$$F_\mathrm{I} = \frac{n_{e,\mathrm{out}}}{N_\mathrm{a}} = \sum_k f_k \frac{\sum_{j=L_k}^{J_k} j z^{J_k - j} \zeta_{j J_k}}{\sum_{j=L_k}^{J_k} z^{J_k - j} \zeta_{j J_k}} \; , \tag{30}$$

where we have defined the scaled electron density and scaled product of $\Phi$ functions by

$$z = \frac{n_e}{N_\mathrm{a}} \qquad \text{and} \qquad \zeta_{j J_k} = \prod_{m=j}^{J_k - 1} (N_\mathrm{a} \Phi_{mk}) \; , \tag{31}$$

and where $n_{e,\mathrm{out}}$ is an electron-density-like parameter which is the electron density when charge conservation (i.e., overall neutrality of the plasma) is imposed which we have done hitherto in this note. In nature, overall neutrality usually holds although special conditions may give a net charge in relatively small regions. We are not interested in such cases, but rather in the use of ionization function in numerical solutions for $n_e$ with overall neutrality: i.e.., numerical solutions of the Saha-CC equation. What we want to prove about the ionization function in this section is that it is a strictly decreasing function of $n_e$ (except for a minium at $n_e = \infty$) since that feature greatly aids in obtaining numerical solutions as we discuss in § 8.

In fact, in § 4, we have already given an argument that shows why $F_\mathrm{I}$ should decrease with increasing electron density. The argument actually applies whether charge conservation holds or not. However, here we give a mathematical proof that verifies definitively that argument and shows that $F_\mathrm{I}$ is strictly decreasing (except for a minium at $n_e = \infty$).

Now we only need to consider the single-element ionization function $F_{\mathrm{I},k}$ since all single-element ionization functions have the same behavior. Also to avoid finicky, unimportant cases we will assume $J_k > \max(L_k, 0)$. The single-element ionization function (now using

distinct dummy indices) is

$$
F_{\mathrm{I},k} = \begin{cases}
\dfrac{\sum_{j=L_k}^{J_k} j z^{J_k-j}\zeta_{jJ_k}}{\sum_{m=L_k}^{J_k} z^{J_k-m}\zeta_{mJ_k}} & \text{in general;} \\[2ex]
\dfrac{J_k + (J_k-1)z\zeta_{J_k-1,J_k} + (J_k-2)z^2\zeta_{J_k-2,J_k}}{1 + z\zeta_{J_k-1,J_k} + z^2\zeta_{J_k-2,J_k}} + \ldots & \\[2ex]
\quad = J_k\left[1 - \dfrac{z}{J_k}\zeta_{J_k-1,J_k} + z^2\left(\dfrac{1}{J_k}\zeta_{J_k-1,J_k}^2 - \dfrac{2}{J_k}\zeta_{J_k-2,J_k}\right)\right] + \ldots & \\[2ex]
\quad = J_k - z\zeta_{J_k-1,J_k} + z^2\left(\zeta_{J_k-1,J_k}^2 - 2\zeta_{J_k-2,J_k}\right) + \ldots & \text{for } z \text{ small;} \\[2ex]
\dfrac{L_k z^{J_k-L_k}\zeta_{L_kJ_k} + (L_k+1)z^{J_k-L_k-1}\zeta_{L_k+1,J_k}}{z^{J_k-L_k}\zeta_{L_kJ_k} + z^{J_k-L_k-1}\zeta_{L_k+1,J_k}} + \ldots & \\[2ex]
\quad = L_k + z^{-1}(N_{\mathrm{a}}\Phi_{L_k k})^{-1} + \ldots & \text{for } z \text{ large;} \\[2ex]
\dfrac{1}{zN_a\Phi \pm 1} & \text{with upper/lower case for the} \\
& \text{hydrogenic/infinite-ionization-stage case,} \\
& \text{where we have written the hydrogenic } \Phi_{0,k} \\
& \text{as just } \Phi \text{ for simplicity.}
\end{cases}
$$

$$(32)$$

We note that at $z = 0$ the ionization function has value $J_k$: i.e., $z = 0$ gives complete ionization. On the other, at $z = \infty$ the ionization function has value $L_k$ which is zero ionization in the case that $L_k = 0$. Note that if charge conservation (i.e., overall neutrality) is imposed, ionization function less than zero is not possible since then it equals $n_e/N_{\mathrm{a}}$ where $n_e$ is the number density of free electrons which can never be negative.

The derivative of the single-element ionization function is

$$
\frac{dF_{\mathrm{I},k}}{dz} = \begin{cases}
-\dfrac{1}{2}\left[\dfrac{\sum_{j=L_k}^{J_k}\sum_{m=L_k}^{J_k}(j-m)^2 z^{2J_k-j-m-1}\zeta_{jJ_k}\zeta_{mJ_k}}{\left(\sum_{m=L_k}^{J_k} z^{J_k-m}\zeta_{mJ_k}\right)^2}\right] & \text{in general;} \\[3ex]
-\zeta_{J_k-1,J_k} + 2z\left(\zeta_{J_k-1,J_k}^2 - 2\zeta_{J_k-2,J_k}\right) + \ldots & \text{for } z \text{ small;} \\[1ex]
-z^{-2}(N_{\mathrm{a}}\Phi_{L_k k})^{-1} + \ldots & \text{for } z \text{ large;} \\[2ex]
-\dfrac{N_a\Phi}{(zN_a\Phi \pm 1)^2} & \text{with upper/lower case for the} \\
& \text{hydrogenic/infinite-ionization-stage case.}
\end{cases}
$$

$$(33)$$

The general derivative is clearly always negative or possibly zero. However, since we assumed $J_k > L_k$ always and the $\zeta_{j,J_k}$ functions are always greater than zero, the numerator of the

general formula clearly has non-zero terms. Thus the derivative is always negative, except possibly at $z = 0$ and $z = \infty$. The small-$z$ derivative shows that the general derivative is not zero at $z = 0$. The large-$z$ derivative shows that the general derivative is zero at $z = \infty$. Thus, we have verified that the ionization function is strictly decreasing except that it has a minimum at $z = \infty$.

Note from equation (33) that since $\zeta_{J_k-1,J_k}$ can less than or greater than 1, the slope of the ionization function at $z = 0$ can be less than or greater than $-1$. Thus in general the slope of the ionization function can be less than or greater than $-1$. This result has an important consequence in the numerical solution of the Saha-CC equation as we elucidate in § 8.

The 2nd derivative of the single-element ionization function allows us to see if the ionization function curves up (positive 2nd derivative) and/or down (negative 2nd derivative) as $z$ increases from zero. The 2nd derivative of the single-element ionization function for the small and large $z$ cases is

$$\frac{d^2 F_{\mathrm{I},k}}{dz^2} = \begin{cases} 2\left(\zeta_{J_k-1,J_k}^2 - 2\zeta_{J_k-2,J_k}\right) + \ldots \\ \qquad = 2N_\mathrm{a}\Phi_{J_k-1,k}\left(N_\mathrm{a}\Phi_{J_k-1,k} - 2N_\mathrm{a}\Phi_{J_k-2,k}\right) + \ldots \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for } z \text{ small;} \\ 2z^{-3}(N_\mathrm{a}\Phi_{L_k k})^{-1} + \ldots \qquad\qquad\qquad\qquad\quad \text{for } z \text{ large;} \\ \dfrac{2(N_a\Phi)^2}{(zN_a\Phi \pm 1)^3} \quad \text{with upper/lower case for the} \\ \qquad\qquad\qquad \text{hydrogenic/infinite-ionization-stage case.} \end{cases} \tag{34}$$

The large-$z$ 2nd derivative is positive, and so the ionization function always curves up asymptotically as $z \to \infty$. The small-$z$ 2nd derivative will be positive almost always and maybe always. This is because $\Phi_{J_k-1,k} > \Phi_{J_k-2,k}$ is generally true since $\Phi_{j,k}$ depends on the an exponential of the ionization energy $X_{jk}$ (see the Saha equation (eq. 5) in § 3) and $X_{jk}$ increases with $j$ (i.e., ionization stage) as general rule. However, all though we know of no exceptions to this general rule, but there might be some odd case. Also however, $\Phi_{j,k}$ does depend on the ratio of partition functions $U_{jk}/U_{j+1,k}$ and this ratio is generally not monotonic with $j$. Also however again, there is that factor of 2 multiplying $N_\mathrm{a}\Phi_{J_k-2,k}$ small-$z$ 2nd derivative. The three "howevers" mean we cannot guarantee that $(N_\mathrm{a}\Phi_{J_k-1,k} - 2N_\mathrm{a}\Phi_{J_k-2,k})$ in the small-$z$ 2nd derivative will always be positive. This means that we cannot guarantee that the 2nd derivative in general will always be positive. There might be some small negative regions sometimes.

## 8.   THE NUMERICAL SOLUTION OF THE SAHA-CC EQUATION

The most direct numerical solution of the Saha-CC equation is by an iteration equation. To be general for a moment, consider the iteration equation

$$x = f(x) , \tag{35}$$

where the satisfying $x$ is the solution. One solves for $x$ by estimating somehow an initial value $x_0$ and substituting it into $f(x)$ as input to obtain $x_1$ as an output, and then using $x_1$ as input to obtain output $x_2$, and so on until the inputs and outputs converge to the same numerical value within some criterion. However, convergence is not obtained in general: there can be divergence or some kind of oscillation. There also can be multiple solutions, and so one has somehow select the solution one wants.

However, if there is only one solution, there is a simple criterion guaranteeing convergence. To elucidate, say we have iteration equation (35) and, for simplicity and without loss of generality, we make the $x = 0$ the exact solution. There is no loss of generality because if $x_\infty \neq 0$ is the single solution, then we could define $x_{\text{new}} = x - x_\infty$ and $f_{\text{new}}(x_{\text{new}}) = f(x_{\text{new}} + x_\infty) - x_\infty$ and then use iteration equation $x_{\text{new}} = f(x_{\text{new}})$ just suppressing the subscripts "new".

We can graphically represent the solution of $x = f(x)$ (being $x = 0$ as just discussed without loss of generality) as the intersection of the curves $y = f(x)$ and $y = x$ in the $x$-$y$ plane. Now we divide the $x$-$y$ plane into 4 quadrants delimited by lines $y = x$ and $y = -x$. We label the quadrants 1, 2, 3, 4 going counterclockwise from quandrant 1 which is bisected by the positive $x$ axis. The convergence criterion is $f(x)$ has a single solution (only $x = 0$ satisfies $x = f(x)$) and $f(x)$ is entirely contained within quadrants 1 and 3 (without touching the boundaries of the quadrants, except at $x = 0$ itself). Note $f(x)$ does not have to be monotonic.

For a proof of the convergence criterion, say the $(i - 1)$th iterate is $x_{i-1}$. Then the $i$th iterate is $x_i = f(x_{i-1})$. Given the criterion, we know that

$$-|x_{i-1}| < x_i < |x_{i-1}| \quad \text{implying} \quad |x_{i-1}| > -x_i > -|x_{i-1}| . \tag{36}$$

Since the absolute value $|x_i|$ must be one of $x_i$ and $-x_i$, we find for all $i$ that

$$|x_i| < |x_{i-1}| . \tag{37}$$

Thus, the absolute value of the iterate $|x_i|$ decreases as $i$ increases. Since there is only a single solution $x = 0$, the iterate $x_i$ cannot go to any value other than 0 even as $i \to \infty$. Since equation (37) is a strict inequality, we cannot have $x_i = 0$ for finite $i$ since that would

imply $x_{i+1}$ equal to $x_i$. So convergence to $x_i = 0$ only occurs asymptotically as $i \to \infty$. That completes the proof of the convergence criterion along with the extra result that convergence only happens asymptotically as the iteration goes infinity.

Before leaving our general considerations, there is another useful result to prove. Say $f(x)$ (not necessarily satisfying the convergence criterion) is monotonic with single solution $x_\infty$. If $f(x)$ is monotonic increasing, then

$$x_i = f(x_{i-1} > x_\infty) > f(x_\infty) = x_\infty \qquad \text{and} \qquad x_i = f(x_{i-1} < x_\infty) < f(x_\infty) = x_\infty \qquad (38)$$

which implies that an iteration starting at greater/less than $x_\infty$ is always greater/less than $x_\infty$. On the other hand, if $f(x)$ is monotonic decreasing, then

$$x_i = f(x_{i-1} > x_\infty) < f(x_\infty) = x_\infty \qquad \text{and} \qquad x_i = f(x_{i-1} < x_\infty) > f(x_\infty) = x_\infty \qquad (39)$$

which implies that the iteration oscillates about the solution.

Now Saha-CC equation, formulated iteration equation in terms of the scaled variables of § 7, is

$$z = F_{\mathrm{I}}(z) \, , \qquad (40)$$

where recall $F_{\mathrm{I}}(z)$ is the ionization function. Unfortunately, the Saha-CC equation does not in general satisfy the simple convergence criterion given above since as discussed in § 7 (just after the derivative equation eq. (33)) since we cannot guarantee that the slope of the ionization function function (which is monotonic decreasing) is less than $-1$: thus, the ionization function can go outside quandrants 1 and 3. However, since the ionization function must be in the range from $\sum_k L_k \geq 0$ to $\sum_k J_k$, we can code the iteration never to diverge. Since $F_{\mathrm{I}}(z)$ is monotonic decreasing, the iteration will oscillate about the solution and we can code the allowed range to decrease between minimum and maximum values for $z$: i.e., $z_{\min}$ and $z_{\max}$. If the minimum and maximum values stop changing, we can get the iteration unstuck by choosing the next input value for the iteration to be

$$z = \frac{1}{2}(z_{\min} + z_{\max}) \, . \qquad (41)$$

The just described iteration procedure will converge albeit probably rather slowly. We call this the unaccelerated iteration or the 0th accelerated iteration.

As the last paragraph suggests, there are accelerated iterations. What we will call the 1st accelerated iteration is that we choose the accelerated input $z_{i-1}$ for the $i$ iteration step to be

$$z_{i-1} = \frac{1}{2}(z_{\min,i-1} + z_{\max,i-1}) \, , \qquad (42)$$

where $(i - 1) \geq 1$: i.e, you have to have done iterates 0 and 1 to start the acceleration. In fact, the 1st accelerated iteration is not very good. Numerically experiments show it can be slightly faster or slightly slower than the unaccelerated iteration. In some special cases, it might be much faster or slowler.

## 9.  CONCLUSIONS

The conclusions are essentially given in the abstract and the Introductuion (§ 1).

## A.  NUMERICALLY GOOD SOLUTIONS OF THE QUADRATIC EQUATION

Numerically good solutions of the quadratic equation (i.e., numerically good quadratic formulae) are needed for the numerical solution of the Saha-CC equation (see § 8). For reference, we derive them here.

The quadratic equation in conventional form is

$$ax^2 + bx + c = 0 . \tag{A1}$$

The quadratic equation is solved by completing the square:

$$
\begin{aligned}
0 &= ax^2 + bx + c \\
0 &= x^2 + \frac{b}{a}x + \frac{c}{a} \\
0 &= x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} - \frac{b^2}{4a^2} + \frac{c}{a} \\
0 &= \left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a^2} + \frac{c}{a} \\
x + \frac{b}{2a} &= \pm\mathrm{sgn}(a)\sqrt{\frac{b^2}{4a^2} - \frac{c}{a}} = \pm\frac{\sqrt{b^2 - 4ac}}{2a} \\
x_{a,\pm} &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} ,
\end{aligned}
\tag{A2}
$$

where there are real solutions only for discriminant $(b^2 - 4ac) \geq 0$, there two distinct solutions for $(b^2 - 4ac) > 0$, there is only one solution for $(b^2 - 4ac) = 0$, and where function $\mathrm{sgn}(x)$ returns $+1$ for $x \geq 0$ and $-1$ for $x < 0$.

Now equation (A2) is the conventional form of the solution to the quadratic equation: i.e., it is the conventional form of the quadratic formula. However, obviously it will be numerically poor for one or other of the upper and lower case solutions since there will be significant roundoff error when $|4ac/b^2| \ll 1$. However, which case is poor depends on the sign of $b$ which makes things tricky.

The path to get two numerically good solutions is to introduce a factor of $-\mathrm{sgn}(b)$ into the second to last equation in the derivation of equation (A2). We can do this since nothing forbids us: the new second to last equation is still a solution of the third to last equation in the derivation of equation (A2). We now solve to get

$$x + \frac{b}{2a} = \mp\mathrm{sgn}(b)\mathrm{sgn}(a)\sqrt{\frac{b^2}{4a^2} - \frac{c}{a}} = \mp\mathrm{sgn}(b)\frac{\sqrt{b^2 - 4ac}}{2a}$$

$$x_{a,\pm} = -\frac{1}{2a}\left[b \pm \mathrm{sgn}(b)\sqrt{b^2 - 4ac}\right] , \tag{A3}$$

where we have introduced the subscript "$a, \pm$" for clarity. It is now clear that the upper/lower case of solution $x_{a,\pm}$ is numerically good/poor since $b$ and $\mathrm{sgn}(b)$ always have the same sign.

How does one get a numerically good version of the lower case solution of equation (A3)? We define $x_{c,\pm} = x_{a,\mp}$ and obtain an expression for $x_{c,\pm}$ using difference of squares:

$$\begin{aligned} x_{c,\pm} = x_{a,\mp} &= = -\frac{1}{2a}\left[b \mp \mathrm{sgn}(b)\sqrt{b^2 - 4ac}\right] \\ &= -\frac{1}{2a}\left[\frac{-4ac}{-b \mp \mathrm{sgn}(b)\sqrt{b^2 - 4ac}}\right] \\ &= -2c\left[\frac{1}{b \pm \mathrm{sgn}(b)\sqrt{b^2 - 4ac}}\right] . \end{aligned} \tag{A4}$$

Note that the upper/lower case of $x_{c,\pm}$ is numerically good/poor for the same reason the upper/lower case of $x_{a,\pm}$ is numerically good/poor: $b$ and $\mathrm{sgn}(b)$ always have the same sign.

Point to emphasize, there are only two mathematical solutions, unless the discriminant $(b^2 - 4ac) = 0$. As the derivation of equation (A4) shows, the upper/lower case of $x_{c,\pm}$ (which numerically good/poor) is mathematically equal to the lower/upper case of $x_{a,\pm}$ (which is numerically poor/good).

We now define
$$q_\pm = -\frac{1}{2}\left[b \pm \mathrm{sgn}(b)\sqrt{b^2 - 4ac}\right] \tag{A5}$$
which is always numerically good and write the solutions for $x$ by

$$x_{a,\pm} = \frac{q_\pm}{a} \quad \text{and} \quad x_{c,\pm} = \frac{c}{q_\pm} , \tag{A6}$$

where the upper/lower case solutions are the numerically good/poor solutions.

Point to emphasize, the upper and lower cases of $x_{a,\pm}$ and of $x_{c,\pm}$ are distinct solutions, unless the discriminant $(b^2 - 4ac)$ in equation (A5) is zero. Just to verify explicitly:

$$\frac{x_{a,\pm}}{x_{a,\mp}} = \frac{x_{c,\mp}}{x_{c,\pm}} = \frac{q_\pm}{q_\mp} = \frac{q_\pm^2}{q_\mp q_\pm} = \frac{q_\pm^2}{ac} = \begin{cases} \text{real} \quad q_\pm^2/(ac) \neq 1 & \text{for } (b^2 - 4ac) > 0; \\ (b^2/4)/(4ac) = 1 & \text{for } (b^2 - 4ac) = 0; \\ \text{complex number} & \text{for } (b^2 - 4ac) < 0. \end{cases} \tag{A7}$$

Similarly, the numerically good/poor solutions are the distinct solutions (unless the discriminant $(b^2 - 4ac)$ in equation (A5) is zero) since

$$\frac{x_{a,\pm}}{x_{c,\pm}} = \frac{q_\pm^2}{ac} . \tag{A8}$$

Note the solutions $x_{a,\pm}$ and $x_{c,\mp}$ are identical solutions. Just to verify explicitly:

$$\frac{x_{a,\pm}}{x_{c,\mp}} = \frac{x_{a,\mp}}{x_{c,\pm}} = \frac{q_\pm q_\mp}{ac} = \frac{1}{4ac}\left(b^2 - b^2 + 4ac\right) = 1 \qquad \text{always.} \tag{A9}$$

In conclusion, to compactly get the 2 numerically-good mathematically-distinct solutions to a quadratic equation evaluate $q_+$ using equation (A5) and then evaluate $x_{a,+}$ and $x_{c,+}$ using equation (A6).

Note also the following special cases for the solutions in equation (A6):

1. If $a = b = c = 0$, there is no solution and $x$ is indeterminate.

2. If $a = b = 0$ and $c \neq 0$, there is no solution, $x$ is indeterminate, and equation is inconsistent.

3. If $a = 0$, $b \neq 0$, and $c = 0$, there is only one solution $x_{c,+} = 0$.

4. If $a \neq 0$ and $b = c = 0$, there is only one solution $x_{a,\pm} = 0$.

5. If $a \neq 0$, $b \neq 0$, and $c = 0$, there are solutions $x_{a,+} = -b/a$ and $x_{a,-} = x_{c,+} = 0$.

6. If $a \neq 0$, $b = 0$, and $c \neq 0$, there are solutions $x_{a,\pm} = \mp\sqrt{-c/a}$ and $x_{c,\pm} = \mp\sqrt{-c/a}$.

7. If $a = 0$, $b \neq 0$ and $c$ general, the only solution is $x_{c,+} = -c/b$ (i.e., the linear equation solution).

8. If the discriminant $(b^2 - 4ac) < 0$, there are no real solutions.

9. If the discriminant $(b^2 - 4ac) = 0$, $x_{a,\pm} = -b/(2a)$ and $x_{c,\pm} = -2c/b = -b/(2a) = x_{a,\pm}$. So there is only a single solution.

10. If the discriminant $(b^2 - 4ac) > 0$, there are two real solutions.

# REFERENCES

Arfken, G. 1970, Mathematical Methods for Physicists, 3rd edition (New York: Academic Press, Inc.)

Bevington, P. R. 1969, Data Reduction and Error Analysis for the Physical Sciences (New York: McGraw-Hill Book Company)

Jeffery, D. J., & Branch, D. 1990, in Jerusalem Winter School for Theoretical Physics, Vol. 6, Supernovae, ed. J. C. Wheeler, T. Piran, & S. Weinberg (Singapore: World Scientific), 149

Mihalas, D. 1978, Stellar Atmospheres, (San Francisco: W. H. Freeman and Company)

Pointon, A. J. 1978, Introduction to Statistical Physics (London: Longman Group Ltd.)

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. 1992, Numerical Recipes in Fortran (Cambridge: Cambridge University Press)