National Aeronautics and Space Administration

Analytical Solutions and Approximations for the Equation ydy/dx = (ay+b) h(x) with Applications to Drift-Diffusion

Larry D. Edmonds
Jet Propulsion Laboratory
California Institute of Technology
Mail Stop 303-220
4800 Oak Grove Drive
Pasadena, California 91109-8099
Phone: (818) 354-2778
FAX: (818) 393-4559
Email: larry.d.edmonds@jpl.nasa.gov

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

JPL Publication 09-13

National Aeronautics and Space Administration

Analytical Solutions and Approximations for the Equation ydy/dx = (ay+b) h(x) with Applications to Drift-Diffusion

Larry D. Edmonds
Jet Propulsion Laboratory
California Institute of Technology
Mail Stop 303-220
4800 Oak Grove Drive
Pasadena, California 91109-8099
Phone: (818) 354-2778
FAX: (818) 393-4559
Email: larry.d.edmonds@jpl.nasa.gov

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

JPL Publication 09-13

The research in this paper was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

Contents

Abstract

The problem considered is

$$y(x)\frac{dy(x)}{dx} = \left(a\,y(x) + b\right)h(x) \quad for \quad x \in (x_1, x_2)$$

subject to a point-value condition of the type $y(x_C) = y_C$, where $y(x)$ is the function to be solved when all other parameters and functions are given. In principle, this problem is solvable using elementary textbook methods (separation of variables). In practice, two difficulties are encountered. The first difficulty is the complex coupling of qualifiers associated with existence and uniqueness of solutions and with the identification of the "required solution" (defined by the physical problem that the equation describes) when there are multiple solutions. In the language of a programmer, "qualifiers" means *if-then* statements and "complex coupling" means that these *if-then* statements produce a complicated logic flow. The second difficulty is that exact solutions to the differential equation are expressed as solutions to transcendental algebraic equations that require numerical root-finding algorithms. This paper simplifies the qualifiers and reduces the number of qualifiers that are encountered by confining attention to "uniform solutions," defined as solutions that do not change sign. The second problem is avoided by finding accurate, yet simple, approximations for the exact solutions. These approximations are derived for the physical application of charge-carrier drift-diffusion in a quasi-neutral semiconductor material. Exact results are also given for the problem of charge collection, showing that a sufficiently large carrier generation rate creates a sensitive volume in the quasi-neutral region. However, the sensitive volume is a symbolic model and has limited applicability. An alternate model that is a more literal description of charge-collection physics is ambipolar diffusion with a cutoff.

Key words: Ambipolar diffusion, charge collection, charge-collection efficiency, drift-diffusion, funnel, sensitive volume.

1. Introduction

An earlier paper [1] used an analytical (as opposed to numerical) method to calculate the I-V curve of a semiconductor diode under conditions general enough to range from low-injection to high-injection. Although the analysis was successful in the sense of producing correct predictions, the complexity obscured physical insight, making the solution nothing more than a number crunching routine that has no obvious advantage over a more traditional numerical method of solution. "Insight" is the ability to predict some qualitative properties from an inspection of the equations without having to actually calculate numbers. An analytical method that is too complicated to deliver insight does not meet expectations. The complexity of the diode analysis can be traced to a particular mathematical equation that is the first subject of this paper (the second subject, in Section 6, revisits the semiconductor problem). A better understanding of this mathematical problem will add insight into the physical problem of a diode, as well as other problems in science and engineering that encounter the same equation. It therefore seems appropriate to dedicate an in-depth study to this mathematical problem.

The mathematical problem is as follows: the user selects an open interval on the real axis (x_1, x_2) (with $x_2 > x_1$), two real constants a and b, and a real-valued function of a real variable $h(x)$. The objective is to solve for $y(x)$ (a real-valued function of a real variable) satisfying

$$y \text{ is defined and continuous on the closed interval } [x_1, x_2] \tag{1a}$$

$$\frac{dy(x)}{dx} \text{ is defined for all } x \in (x_1, x_2) \tag{1b}$$

$$y(x)\frac{dy(x)}{dx} = \big(a\, y(x) + b\big)h(x) \quad \textit{for} \quad x \in (x_1, x_2) \tag{1c}$$

subject to a point-value condition of the type

$$y(x_C) = y_C \tag{1d}$$

where x_C is a user-specified point in the closed interval [x_1, x_2] and y_C is a user-specified real number. To keep the analysis simple, but still general enough for practical applications, we will assume that $h(x)$ is defined and continuous at each point in the closed interval [x_1, x_2].

Because of the deceptively simple appearance of (1), it might not be obvious that an analysis of (1) is actually quite complex. In particular, (1c) is separable in the sense that dividing both sides by $ay(x)+b$ makes each side an integrable combination — a seemingly simple procedure. One complication is that the solution obtained this way involves transcendental algebraic equations that require numerical root-finding algorithms. Another complication is associated with existence and uniqueness of solutions. Those of us that are not mathematicians rarely pay enough attention to existence or uniqueness theorems and often take it for granted (a seemingly small leap of faith) that we have found the most general solution to (1c) if we found a solution containing an arbitrary integration constant. All that remains to satisfy all equations in (1) is to select the integration constant to satisfy (1d). In reality, there can be very different functions of x, with each containing an integration constant and satisfying (1c). Furthermore, it might be

possible to select integration constants so that each of these very different solutions to (1c) also satisfy (1d); i.e., solutions to the set of equations in (1) are not unique.

An illustration of multiple solutions is provided by an example that was constructed for this purpose. This example is defined by

$$x_1 = -1, \quad x_2 = 1, \quad x_C = -1, \quad y_C = 1, \quad a = 1, \quad b = 0, \quad h(x) = 2x \quad \text{(example)}. \tag{2}$$

Two solutions for this example are

$$y(x) = x^2 \quad \textit{for all } x \in [-1,1] \quad \text{(one solution for the example)} \tag{3a}$$

$$y(x) = \begin{cases} x^2 & \textit{if} \quad -1 \le x \le 0 \\ 0 & \textit{if} \quad 0 < x \le 1 \end{cases} \quad \text{(another solution for the example)}. \tag{3b}$$

The two branches on the right side of (3b) join smoothly enough so that the derivative of y is defined and satisfies (1c) for all x, including the point $x = 0$; (3b) is, therefore, a valid solution, but is not the same as the solution (3a).

The fact that different solutions can sometimes be constructed by joining different branches, as in (3b), can lead to an assortment of solutions. The assortment becomes very complex for the general case in which a, b, and h are arbitrary, producing a tangled weave of *if-then* statements that we must sort through in order to ensure that all solutions have been found. Fortunately, science or engineering problems for which (1) must be solved often dictate that the physically meaningful solution has a particular sign. For example, the physical problem in Section 6 requires that carrier density be positive, while y is related to the carrier density in such a way so that a positive carrier density implies that y does not change sign. The required sign of y can be different for different examples, but in all cases the required sign of y is the same as the sign of the given point-value y_C. In other words, the required solution is a "uniform solution," defined here to be a solution that does not change sign and, therefore, has the same sign as y_C (a more precise definition is given below). Confining our attention to uniform solutions maintains enough generality to have applications in science and engineering, but greatly simplifies the bookkeeping because the next section will show that uniform solutions are unique when they exist.

In order for the definition of a uniform solution to make sense, we confine our attention to those cases in which y_C is not zero. The precise definition can then be stated as follows:

$$\begin{aligned} &\text{If } y_C > 0, \text{ then a solution } y(x) \text{ to (1) will be called} \\ &\text{uniform if it satisfies } y(x) > 0 \text{ for all } x \in [x_1, x_2]. \end{aligned} \tag{4a}$$

$$\begin{aligned} &\text{If } y_C < 0, \text{ then a solution } y(x) \text{ to (1) will be called} \\ &\text{uniform if it satisfies } y(x) < 0 \text{ for all } x \in [x_1, x_2]. \end{aligned} \tag{4b}$$

The next section demonstrates that uniform solutions are unique when they exist. However, strict inequalities in (4) are intentional. The motive for them is not merely to make the uniqueness proof in the next section easier; they are essential for uniformity to imply uniqueness. For example, if the strict inequality in (4a) were replaced by $y(x) \geq 0$ for all x, then both solutions in (3) would be classified as uniform; hence, uniformity would not imply uniqueness (even when $y_C \neq 0$) if the definition of uniformity were altered in this way. Using the definition as given by (4), neither solution in (3) is uniform. This leads to the question of whether a uniform solution exists. Sufficient conditions for existence are proven by constructing solutions in Section 3 (for one set of conditions) and in Section 4 (for another set of conditions).

This paper considers only uniform solutions; therefore, throughout this paper we take it as given that

$$y_C \neq 0. \tag{5}$$

Also, the function H is defined here for later use throughout this paper by

$$H(x) \equiv \int_{x_C}^{x} h(\xi)\, d\xi \quad \textit{for} \quad x \in [x_1, x_2]. \tag{6}$$

2. Uniqueness of Uniform Solutions

This section shows that uniform solutions are unique when they exist. The question of existence is discussed in the next two sections. For the analysis in this section, existence is regarded as given.

Let $y(x)$ and $z(x)$ be two uniform solutions to (1). Neither function is zero anywhere in the interval considered, so their differential equations can be written as

$$\frac{dy(x)}{dx} = \left(a + \frac{b}{y(x)} \right) h(x) \quad for \quad x \in (x_1, x_2)$$

$$\frac{dz(x)}{dx} = \left(a + \frac{b}{z(x)} \right) h(x) \quad for \quad x \in (x_1, x_2) .$$

Subtracting equations gives

$$\frac{d}{dx}\big(y(x) - z(x)\big) + \big(y(x) - z(x)\big)\frac{b\,h(x)}{y(x)\,z(x)} = 0 \quad for \quad x \in (x_1, x_2). \tag{7}$$

Note that for any $x \in [x_1, x_2]$, $y(x)$ and $z(x)$ differ from zero and have the same sign as y_C and, therefore, have the same sign as each other. This gives $y(x)z(x) > 0$ for each $x \in [x_1, x_2]$. The product function is continuous, and the interval is a closed interval. The inequality, together with these two facts, implies that the product is bounded above zero; i.e., there is an $\varepsilon > 0$ such that $y(x)z(x) \geq \varepsilon > 0$ for each $x \in [x_1, x_2]$. [1] The implication is that the quantity $1/y(x)z(x)$ is integrable on any subinterval of $[x_1, x_2]$. Also, h is continuous on the closed interval (hence bounded), so the quantity $h(x)/y(x)z(x)$ is integrable on any subinterval of $[x_1, x_2]$. We can, therefore, define W by

$$W(x) = \int_{x_C}^{x} \frac{b\,h(\xi)}{y(\xi)\,z(\xi)}\,d\xi \quad for \quad x \in [x_1, x_2]. \tag{8}$$

The chain rule, together with (8), gives

$$\frac{d}{dx}\left\{\big(y(x) - z(x)\big)\,e^{W(x)}\right\} = \left\{\frac{d}{dx}\big(y(x) - z(x)\big) + \big(y(x) - z(x)\big)\frac{b\,h(x)}{y(x)\,z(x)}\right\} e^{W(x)} .$$

Combining this with (7) gives

$$\frac{d}{dx}\left\{\big(y(x) - z(x)\big)\,e^{W(x)}\right\} = 0 \quad for \quad x \in (x_1, x_2) .$$

[1] A simple proof notes that a continuous function maps a closed interval onto a closed interval, so the range of $y(x)z(x)$, regarded as a function of x with $x \in [x_1, x_2]$, is some closed interval $[a_1, a_2]$. All points in this closed interval are positive. This includes the lower endpoint, which serves as the ε used here.

That is, the curly bracket on the left is constant in x. However, the curly bracket is zero at x_C because $y(x_C) = y_C = z(x_C)$, so the curly bracket is zero at each point in the open interval. The curly bracket is also continuous on $[x_1, x_2]$, so it is zero at each point in the closed interval $[x_1, x_2]$. The exponential function is not zero, so $y(x) - z(x) = 0$ at each point in the closed interval $[x_1, x_2]$, which proves uniqueness.

3. Explicit Solutions When *a* = 0 or *b* = 0

The more general treatment (given later) that allows *a* and *b* to both differ from zero will also insist that they must differ from zero (to avoid undefined terms); therefore, the cases that will be excluded there are treated separately here for completeness, even though these cases are trivial. Recall that uniform solutions are unique when they exist. If we construct "a" uniform solution, then we have found "the" uniform solution. By constructing a uniform solution, we also have verified the existence of a uniform solution. The conclusions below are obvious enough to be stated without proof.

Case 1: a = 0

If $a = 0$, then a uniform solution to (1) exists if

$$2b\,H(x) + {y_C}^2 > 0 \quad \textit{for all } x \in [x_1, x_2]. \tag{9}$$

If this condition is satisfied, the uniform solution is

$$y(x) = y_C \sqrt{1 + \frac{2b\,H(x)}{{y_C}^2}} = \begin{cases} \sqrt{2b\,H(x) + {y_C}^2} & \textit{if} \quad y_C > 0 \\ -\sqrt{2b\,H(x) + {y_C}^2} & \textit{if} \quad y_C < 0. \end{cases} \tag{10}$$

Case 2: b = 0

If $b = 0$, then a uniform solution to (1) exists if

$$a\,y_C\,H(x) + {y_C}^2 > 0 \quad \textit{for all } x \in [x_1, x_2]. \tag{11}$$

If this condition is satisfied, the uniform solution is

$$y(x) = a\,H(x) + y_C\,. \tag{12}$$

4. Explicit Solutions When $a \neq 0$ and $b \neq 0$

Throughout this section we assume that

$$a \neq 0 \quad and \quad b \neq 0. \tag{13}$$

It is convenient to change variables in (1) by defining

$$v(x) \equiv a\, y(x)/b, \qquad g(x) = a^2 h(x)/b, \qquad A \equiv a\, y_C / b\,. \tag{14}$$

For later use, we also define

$$G(x) \equiv \int_{x_C}^{x} g(\xi)\, d\xi = a^2 H(x)/b. \tag{15}$$

Using (14), we can write (1) as

$$v(x)\frac{dv(x)}{dx} = \big(v(x)+1\big)g(x) \quad for \quad x \in (x_1, x_2) \tag{16a}$$

$$v(x_C) = A. \tag{16b}$$

Solutions to (16) will be expressed in terms of three special functions, so we digress long enough to briefly review these functions, denoted E_1, E_2, and E_3. In-depth discussions of these functions, including limits and asymptotic behaviors, are given in Appendix A. Plots of these functions are shown in Figures A2, A3, and A4 in Appendix A. Several properties of these functions that are derived in Appendix A are repeated below:

$$E_1(x) \text{ is defined when } x \geq 0, \quad \frac{dE_1(x)}{dx} \text{ is defined when } x > 0 \tag{17a}$$

$$E_2(x) \text{ is defined when } x \geq 0, \quad \frac{dE_2(x)}{dx} \text{ is defined when } x > 0 \tag{17b}$$

$$E_3(x) \text{ is defined for all } x, \quad \frac{dE_3(x)}{dx} \text{ is defined for all } x \tag{17c}$$

$$E_1(0) = 0 \text{ and } E_1(x) > 0 \text{ for } x > 0 \tag{18a}$$

$$E_2(0) = 0 \text{ and } -1 < E_2(x) < 0 \text{ for } x > 0 \tag{18b}$$

$$E_3(x) < -1 \text{ for all } x \tag{18c}$$

$$\xi = E_1\big(\xi - \ln\big(\xi + 1\big)\big) \quad for \quad \xi \geq 0 \tag{19a}$$

$$\xi = E_2\big(\xi - \ln(\xi+1)\big) \quad for \quad -1 < \xi \le 0 \tag{19b}$$

$$\xi = E_3\big(\xi - \ln(-\xi-1)\big) \quad for \quad \xi < -1 \tag{19c}$$

$$E_1(x)\frac{dE_1(x)}{dx} = E_1(x)+1 \quad for \quad x > 0 \tag{20a}$$

$$E_2(x)\frac{dE_2(x)}{dx} = E_2(x)+1 \quad for \quad x > 0 \tag{20b}$$

$$E_3(x)\frac{dE_3(x)}{dx} = E_3(x)+1 \quad for \quad -\infty < x < +\infty. \tag{20c}$$

Note that although these three functions are solutions to a common differential equation (20), they have very different behaviors, as indicated by the plots in Figures A2, A3, and A4. Horizontal translations of these functions are also solutions to the differential equation. For example, if c is any constant, then $E_1(x+c)$ is a solution to (20) at any x satisfying $x+c > 0$. Analogous considerations apply to E_2 and E_3; consequently we have found three families of solutions to the differential equation, with each containing an arbitrary integration constant to be selected to satisfy a given point-value condition. The choice depends on the given point value. If the given point value is positive, the family $E_1(x+c)$ is the only candidate. If the given point value is between -1 and 0, the family $E_2(x+c)$ is the only candidate. If the given point value is less than -1, the family $E_3(x+c)$ is the only candidate.

We now return to (16). Uniqueness of uniform solutions was already established; therefore, if we can find "a" uniform solution by constructing it, then we have found "the" uniform solution. Furthermore, by constructing a solution, we have also answered the question of existence. The approach used here is to first propose the solution to (16) and then verify that the proposed solution really is a solution. The proposed solution will make sense if

$$A \ne 0 \quad and \quad (A+1)\exp\big(-A-G(x)\big) < 1 \ for\ all\ x \in [x_1, x_2]. \tag{21}$$

Therefore, we assume that this condition is satisfied. It will be seen below that this is a sufficient condition for the existence of a uniform solution. The proposed solution to (16) is

$$v(x) = \begin{cases} E_1\big(A - \ln(A+1) + G(x)\big) & if \quad A > 0 \\ E_2\big(A - \ln(A+1) + G(x)\big) & if \quad -1 < A < 0 \\ -1 \quad if \quad A = -1 & \\ E_3\big(A - \ln(-A-1) + G(x)\big) & if \quad A < -1. \end{cases} \tag{22}$$

To verify that the proposed solution (22) really is a uniform solution to (16), first consider the case where $A > 0$. The assertion to be proven is that the upper expression on the right side of (22) is a uniform solution to (16). Note that the condition $A > 0$ together with (21) implies that the argument to E_1 is defined and positive for all $x \in [x_1, x_2]$. One implication, from (18a), is that the upper expression is a uniform solution if it is a solution. Another implication, from (17a), is that

the derivative of the upper expression is defined for all $x \in [x_1, x_2]$. Using the chain rule together with (15) and (20a) shows that the upper expression is a solution to the differential equation (16a) for all $x \in [x_1, x_2]$. Finally, the fact that $G(x_C) = 0$ (implied by (15)) together with $A > 0$ and (19a) implies that the upper expression in (22) satisfies the point-value condition (16b). This verifies that (22) correctly gives the uniform solution when $A > 0$. Similar arguments for the remaining nontrivial cases listed in (22) verify that (22) correctly gives the uniform solution for all cases, given that (21) is satisfied. Therefore, (21) is a sufficient condition for the existence of uniform solutions. Note that if $A < -1$, then (21) imposes no restrictions on x, but no restrictions are needed because E_3 is defined at all arguments. A condition that is equivalent to (21) is

$$\begin{aligned} &\textit{Either} \quad A \le -1 \quad \textit{or} \\ &A \ne 0 \; \textit{and} \; A - \ln(A+1) + G(x) > 0 \; \textit{for all} \; x \in [x_1, x_2]. \end{aligned} \tag{23}$$

Finally, the results are written in the original notation appearing in (1). Using (14) to change notation, the sufficient condition (21) for the existence of a uniform solution is written as

$$y_C \ne 0 \quad \text{and} \quad \left(\frac{a\, y_C}{b} + 1 \right) \exp\left(-\frac{a\, y_C}{b} - \frac{a^2}{b} H(x) \right) < 1 \;\; \text{for all } x \in [x_1, x_2]. \tag{24}$$

The equivalent statement (23) of this same condition is written as

$$\begin{aligned} &\text{Either} \quad a\, y_C / b \le -1 \quad \text{or} \\ &y_C \ne 0 \;\; \text{and} \;\; \frac{a\, y_C}{b} - \ln\left(\frac{a\, y_C}{b} + 1 \right) + \frac{a^2}{b} H(x) > 0 \;\; \text{for all } x \in [x_1, x_2]. \end{aligned} \tag{25}$$

The uniform solution (22) is written as

$$y(x) = \begin{cases} \dfrac{b}{a} E_1\left(\dfrac{a}{b} y_C - \ln\left(\dfrac{a}{b} y_C + 1 \right) + \dfrac{a^2}{b} H(x) \right) & \textit{if} \quad \dfrac{a}{b} y_C > 0 \\ \dfrac{b}{a} E_2\left(\dfrac{a}{b} y_C - \ln\left(\dfrac{a}{b} y_C + 1 \right) + \dfrac{a^2}{b} H(x) \right) & \textit{if} \quad -1 < \dfrac{a}{b} y_C < 0 \\ -\dfrac{b}{a} & \textit{if} \quad \dfrac{a}{b} y_C = -1 \\ \dfrac{b}{a} E_3\left(\dfrac{a}{b} y_C - \ln\left(-\dfrac{a}{b} y_C - 1 \right) + \dfrac{a^2}{b} H(x) \right) & \textit{if} \quad \dfrac{a}{b} y_C < -1. \end{cases} \tag{26}$$

A simple inspection of the differential equation in (1) suggests that the solution for the $a = 0$ case (Case 1 in the previous section) can be obtained by taking a suitable limit of the case considered here. This assertion is correct, but might not be obvious from a casual inspection of the explicit solutions given by (10) for one case and by (26) for the other. A formal proof derived from the explicit solutions can be given, but only an outline of the proof is given here because the details are cumbersome and the proof merely confirms the obvious conclusion that the $a = 0$

case is a limiting case of (26). The basic idea starts with the qualifier (9) for the $a = 0$ case and uses this, together with some elementary inequalities, to show that there is a sufficiently small but nonzero $|a|$ for which the qualifier (25) is satisfied for any smaller $|a|$; hence, (26) applies for sufficiently small $|a|$. Furthermore, one of the two upper expressions on the right side of (26) applies for sufficiently small $|a|$. Therefore, when taking the limit as $a \to 0$, we can use (26) and confine our attention to the upper two expressions. Then, two limits (A14) and (A15) derived in Appendix A, together with some elementary limits involving the logarithm function, can be used to show that the limit of the solution (26) is (10). Similarly, given the qualifier (11), the solution for the $b = 0$ case (Case 2 in the previous section) can also be obtained by taking a suitable limit of (26). However, there is an interesting point to be made for this case. If the qualifier (11) is not satisfied, it is still possible to take a one-sided limit of (26) in which $b \to 0$ in such a way so that $ay_C/b \to -\infty$. If we take this limit when (11) is violated, the limit of the solution will be $y(x) = 0$, which is not a uniform solution. Therefore, the qualifier (11) is essential, even when taking this one-sided limit. Finally, it is easy to show that the third solution in (26) can be obtained either by taking the limit of the second solution as $ay_C/b \to -1$ from above, or by taking the limit of the fourth solution as $ay_C/b \to -1$ from below.

5. Visual Aids

Previous sections have simplified and organized the qualifiers so that the appropriate solution can be immediately selected from the available choices. The available choices consist of (10), (12), and the four choices contained within (26). For those cases in which an E-function must be evaluated, a numerical evaluation can be done by using a root-finding algorithm with either (19) in the main text or with (A4) in Appendix A. What is still missing is intuitive insight. What does the solution look like? Under what conditions are various approximations useful? The first question is answered in this section. The next section investigates approximations useful for selected applications.

As pointed out at the end of Section 4, the solutions given by (10), (12), or the third expression in (26) are limits of the more generic cases consisting of the first, second, and fourth solutions in (26), so the analysis is sufficiently general if we confine our attention to the latter three cases. Also note that if the $h(x)$ appearing in (1) is not a constant, we can construct a new independent variable, a generalized coordinate, in such a way so that h is replaced by a constant when (1) is expressed in terms of the new independent variable. The question of how the solution depends on the new independent variable can be answered by considering the dependence on x when h is a constant. Therefore, we confine our attention to the case in which h is a constant, so $H(x)$ is proportional to x-x_C. However, it is evident from (1) that proportionality constants can be absorbed in the a and b parameters, so the case considered is equivalent to the case in which $h(x) = 1$ and

$$H(x) = x - x_C \qquad \text{(the case considered)}. \tag{27}$$

We now answer the question of what the solution looks like. First, consider the case in which $ay_C/b > 0$. In this case the first solution in (26) applies for suitably restricted x. Using (27), this becomes

$$y(x) = \frac{b}{a} E_1\left(\frac{a}{b} y_C - \ln\left(\frac{a}{b} y_C + 1 \right) + \frac{a^2}{b}(x - x_C) \right) \qquad (\text{when } ay_C/b > 0). \tag{28}$$

Note that a plot of $y(x)$ versus x-x_C is a re-scaled and translated plot of the E_1 function. The coefficient b/a outside the function is a scale factor for the vertical coordinate, the coefficient a^2/b multiplying x is a scale factor for the horizontal coordinate, and the sum of the two terms containing y_C controls the horizontal translation of the plot. In fact, any solution of the type in which $ay_C/b > 0$ can be represented by one universal curve by simply changing the axis labels in a plot of E_1 (see Figure A2 in Appendix A) to obtain the plot shown in Figure 1. Similarly, any solution of the type in which $-1 < ay_C/b < 0$ is represented by Figure 2, while any solution of the type in which $ay_C/b < -1$ is represented by Figure 3.

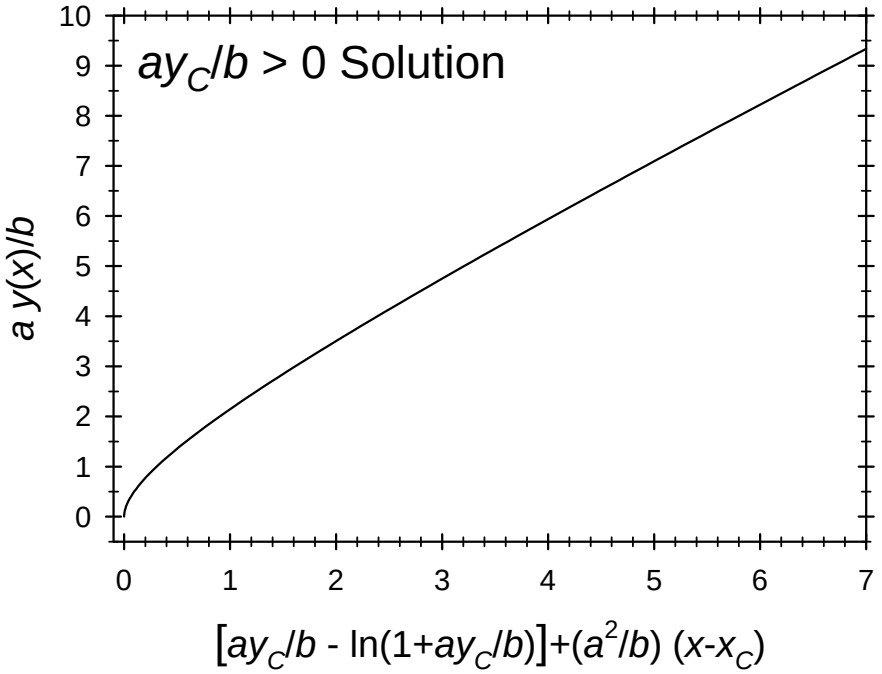

Figure 1. The $ay/b > 0$ Solution. This applies when $ay_C/b > 0$, in which case the square bracket in the horizontal axis label is positive. The $y(x)$ implied by the vertical coordinate is a uniform solution on any x-interval such that the horizontal coordinate is positive at each x within that interval. Only a finite region of the plot is shown but the solution domain includes all positive values of the horizontal coordinate.

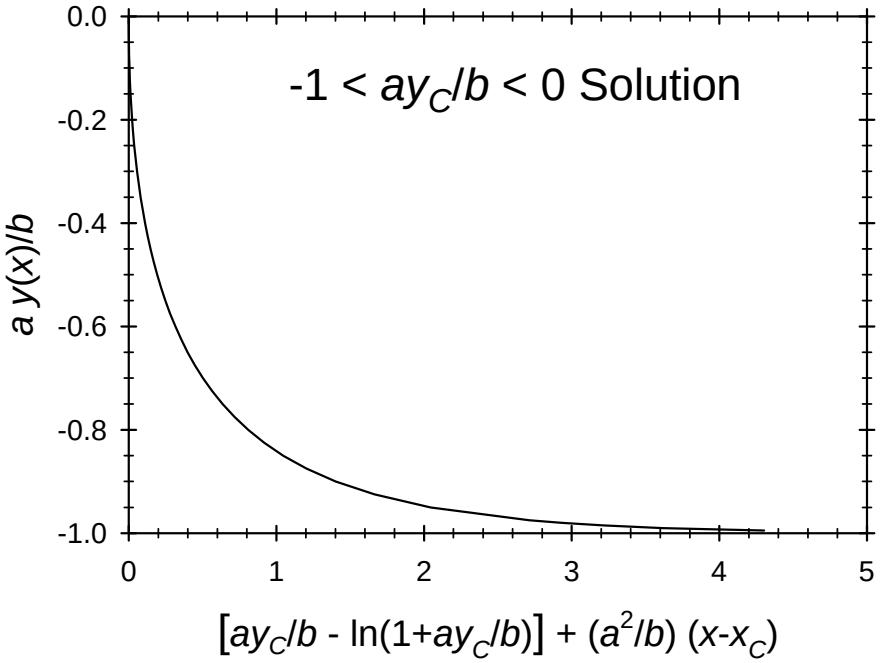

Figure 2. The $-1 < ay/b < 0$ Solution. This applies when $-1 < ay_C/b < 0$, in which case the square bracket in the horizontal axis label is positive. The $y(x)$ implied by the vertical coordinate is a uniform solution on any x-interval such that the horizontal coordinate is positive at each x within that interval. Only a finite region of the plot is shown but the solution domain includes all positive values of the horizontal coordinate.

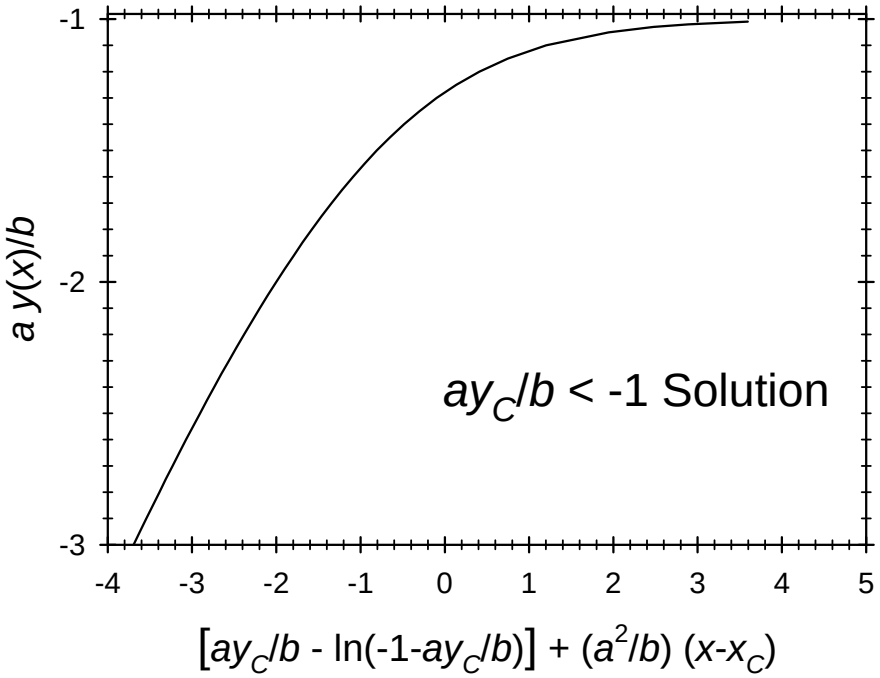

Figure 3. The *ay/b* < -1 Solution. This applies when $ay_C/b < -1$, in which case the $y(x)$ implied by the vertical coordinate is a uniform solution on any x-interval. Only a finite region of the plot is shown but the solution domain includes all values of the horizontal coordinate.

6. Application to a Semiconductor Problem

We consider the physical problem that motivated this work, which is the transport of charge carriers in a semiconductor material for a range of conditions broad enough (from low-injection to high-injection) so that simple textbook solutions do not apply. Although the governing equations that define the selected applications considered here are simplified using various physical approximations (drift-diffusion with constant mobilities and constant diffusion coefficients), even with these simplifications, the exact mathematical solutions are difficult. Other papers (e.g., [1]) have already "solved" such problems, if a number-crunching routine can be called a "solution;" however, this kind of a "solution" provides very little wisdom. The emphasis in this section is not on number crunching. Instead, the objective here is to investigate simple analytical approximations, together with their limitations, that are useful for intuitive insight.

Charge-carrier transport in a semiconductor device (e.g., a diode or a transistor) results in charge carriers being distributed a way that produces different types of regions within the device interior. One type is a space-charge region, often called a depletion region (DR) in textbooks. Another type is a quasi-neutral region (QNR). Numerical methods (computer simulations) solve equations that are sufficiently general (and complicated) so that the same set of equations applies throughout the entire device interior. This allows a complete device to be treated as a single unit, requiring boundary conditions only at the device contacts. In contrast, analytical methods replace exact charge transport equations with simpler approximations. However, the approximations that apply to a DR are different from the approximations that apply to a QNR. Because different equations are used in different device regions, the analysis partitions the device into discrete regions, solves the equations in each region, and joins the solutions together by matching boundary conditions at the interfaces between the regions. In this context, "solves the equations in each region" means that equations are derived for each region that contain enough information to solve for the electron and hole currents as functions of the carrier-density and potential boundary values at the region boundaries. When this approach is used, each region can be analyzed independently of the other regions. In particular, a selected QNR can initially be treated as an isolated unit because a later step in the analysis of a complete device assembles solutions for different device regions into a set of simultaneous equations describing the complete device. Because the QNR was the source of difficulty in an earlier analysis [1], here we focus on a QNR with the objective of obtaining analytical approximations. It should be noted that "quasi-neutral" means neutral enough so that the charge imbalance (when measured as a density of elementary charges) is much smaller than the majority-carrier density, but not necessarily neutral enough for the charge imbalance to be too small to significantly affect the electric field in the QNR. Therefore, while quasi-neutrality leads to some simplifying approximations, it is still necessary to treat carrier density and electrostatic potential as simultaneous unknowns in the transport equations; consequently, the analysis of a QNR is far from trivial.

6.1 Governing Equations for the QNR

Here we consider the simplest version of the problem, which is one-dimensional (note that a one-dimensional analysis can be extended to three dimensions using the method in Appendix B), steady-state, uniform doping, negligible carrier recombination in the interior, no photo-generation of charge carriers (until later in Section 6.5), and having constant electron and hole

mobilities. The quasi-neutral region is contained between two points x_1 and x_2 on the x-axis. In this simplest version of the problem, the continuity equations reduce to the statements that the electron current density J_e and the hole current density J_h are each constant in x. Furthermore, the currents are related to carrier density and potential via the drift-diffusion equations. These equations can be found in any textbook on semiconductors and are

$$J_e = q D_e \frac{dn}{dx} - q \mu_e n \frac{dU}{dx}, \qquad J_h = -q D_h \frac{dp}{dx} - q \mu_h p \frac{dU}{dx},$$

where q is the elementary charge, D_e and D_h are the electron and hole diffusion coefficients, μ_e and μ_h are the electron and hole mobilities, U is the electrostatic potential, and n and p are the electron and hole densities. The sign convention for the currents is such that J_h is positive when holes move in the direction of increasing x, and J_e is positive when electrons move in the direction of decreasing x. An additional equation is Poisson's equation relating the second derivative of U to the carrier densities. A region is quasi-neutral when the solution to the complete set of equations can be approximated by the solution to the set of equations obtained by replacing Poisson's equation with

$$p = P + p_0, \qquad n = P + n_0,$$

where P is the excess carrier density (taken to be the same for electrons and holes), p_0 is the equilibrium hole density, and n_0 is the equilibrium electron density. For an n-type material, we can neglect p_0 and set n_0 equal to the doping density. For a p-type material, we can neglect n_0 and set p_0 equal to the doping density. Although one of the equilibrium densities can be neglected in either application, both will be retained in the equations so that the same equations can be used for either doping type. One more equation is the Einstein relation $D = V_T \mu$, where V_T is the thermal voltage (sometimes written as KT/q and is about 0.026 volts at room temperature) and is the same for electrons as for holes. Combining the above equations gives

$$\frac{J_e}{q D_e} = \frac{dP}{dx} - \frac{P + n_0}{V_T} \frac{dU}{dx}, \qquad -\frac{J_h}{q D_h} = \frac{dP}{dx} + \frac{P + p_0}{V_T} \frac{dU}{dx}, \tag{29}$$

which is a pair of simultaneous equations used to solve for both P and U.

It was pointed out earlier that a quasi-neutral region can be treated as an isolated structure for the purpose of solving the equations, when "solving the equations" means solving for the currents in terms of carrier-density and potential boundary values. Therefore, the goal is to use (29) to solve for the two currents in terms of P boundary values and the potential difference ΔU across the quasi-neutral region. However, the same information is obtained if we solve for ΔU and one P boundary value in terms of the currents and the remaining P boundary value. When using this approach, we can imagine (for conceptual clarity) that the two currents are given, one P boundary value is given, and the objective is to solve for ΔU and the remaining P boundary value. In this approach, the equations in (29) are first-order equations subject to a given point-value condition. The method of solution has two steps. The first step solves for ΔU in terms of the two P boundary values, so all unknowns have been solved once the unknown P boundary value has been solved. The second step solves for the unknown P boundary value. To carry out

the first step, we add the two equations in (29) to get

$$\frac{p_0 - n_0}{V_T}\frac{dU}{dx} = \frac{J_e}{q D_e} - \frac{J_h}{q D_h} - 2\frac{dP}{dx}.$$

Both sides are integrable combinations and integrating between x_1 and x_2 gives

$$\frac{p_0 - n_0}{V_T}\left[U(x_2) - U(x_1)\right] = \left(\frac{J_e}{q D_e} - \frac{J_h}{q D_h}\right)(x_2 - x_1) - 2\left(P_2 - P_1\right), \tag{30}$$

where

$$P_1 \equiv P(x_1), \quad P_2 \equiv P(x_2). \tag{31}$$

We treat the case in which the doping density is not zero, so $p_0 \neq n_0$ and (30) solves for ΔU in terms of the other quantities. To carry out the second step mentioned above, we multiply the first equation in (29) by $P+n_0$, multiply the second equation by $P+p_0$, and then add the resulting equations. This gives

$$\left(P + \frac{n_0 + p_0}{2}\right)\frac{dP}{dx} = \left(\frac{J_e}{2q D_e} - \frac{J_h}{2q D_h}\right)\left(P + \frac{n_0 + p_0}{2}\right) + \frac{p_0 - n_0}{2}\left(\frac{J_e}{2q D_e} + \frac{J_h}{2q D_h}\right).$$

If we now define

$$y(x) \equiv P(x) + \frac{n_0 + p_0}{2}, \quad a \equiv \frac{J_e}{2q D_e} - \frac{J_h}{2q D_h}, \quad b \equiv \frac{p_0 - n_0}{2}\left(\frac{J_e}{2q D_e} + \frac{J_h}{2q D_h}\right) \tag{32}$$

the above equation becomes

$$y(x)\frac{dy(x)}{dx} = a\ y(x) + b. \tag{33a}$$

The boundary at which P is regarded as given can be either x_1 or x_2 and is denoted x_C, and the given boundary value is denoted P_C, so the given point-value condition for y is

$$y(x_C) = y_C \equiv P_C + \frac{n_0 + p_0}{2}. \tag{33b}$$

As pointed out in Section 5, it is sufficiently general to consider the case in which $a \neq 0$ and $b \neq 0$; i.e.,

$$\frac{J_e}{D_e} - \frac{J_h}{D_h} \neq 0, \quad \frac{J_e}{D_e} + \frac{J_h}{D_h} \neq 0 \quad \text{(the case considered)} \tag{34}$$

because other cases can be derived from this case by taking suitable limits.

To shorten the notation, let N be the doping density so

$$n_0 = \begin{cases} 0 & \text{for p-type} \\ N & \text{for n-type} \end{cases}, \qquad p_0 = \begin{cases} N & \text{for p-type} \\ 0 & \text{for n-type} \end{cases}. \tag{35a}$$

We also define

$$K_m \equiv \begin{cases} \dfrac{J_e}{2q\,D_e} & \text{for p-type} \\ -\dfrac{J_h}{2q\,D_h} & \text{for n-type} \end{cases}, \qquad K_M \equiv \begin{cases} \dfrac{J_h}{2q\,D_h} & \text{for p-type} \\ -\dfrac{J_e}{2q\,D_e} & \text{for n-type} \end{cases}. \tag{35b}$$

Note that the subscript m to the K denotes minority-carrier current, while the subscript M denotes majority-carrier current. Also note that the sign of K is either the same or opposite to the sign of the corresponding current density, depending on whether the material is p-type or n-type. This sign convention was selected so that both doping types will be described by the same equations. Substituting (35) into (32), (33b), and (34) gives

$$y(x) = P(x) + \frac{N}{2}, \quad y_C = P_C + \frac{N}{2}, \quad a = K_m - K_M, \quad b = \frac{N}{2}\left(K_m + K_M\right) \tag{36}$$

$$K_m - K_M \neq 0, \quad K_m + K_M \neq 0 \quad \text{(the case considered)}. \tag{37}$$

Several combinations of the terms in (36) that will be useful later are given by

$$\frac{a\,y(x)}{b} = \frac{K_m - K_M}{K_m + K_M}\left(1 + \frac{2P(x)}{N}\right), \qquad \frac{a\,y_C}{b} = \frac{K_m - K_M}{K_m + K_M}\left(1 + \frac{2P_C}{N}\right) \tag{38a}$$

$$\frac{a^2}{b}(x - x_C) = 2\frac{K_m - K_M}{K_m + K_M}\frac{K_m - K_M}{N}(x - x_C). \tag{38b}$$

Finally, we can express (30) in terms of the Ks as

$$\frac{N}{2V_T}\Delta U = \left(K_m - K_M\right)(x_2 - x_1) - \left(P_2 - P_1\right), \tag{39a}$$

where

$$\Delta U \equiv \begin{cases} U(x_2) - U(x_1) & \text{for p-type} \\ U(x_1) - U(x_2) & \text{for n-type} \end{cases}. \tag{39b}$$

6.2 A Forward-Biased p-n Junction Diode

The example considered here is a forward-biased p-n junction diode, as shown in Figure 4. It is convenient for notation to express the equations in terms of a modification (discussed later) of an "emitter efficiency" γ, which is defined in textbooks (e.g., in [2]) by

$$\gamma_{textbook} \equiv \frac{J_m}{J_m + J_M}, \tag{40}$$

where J_m is the minority-carrier current density and J_M is the majority-carrier current density. An emitter efficiency not only helps with notation that will come later, it also has some properties that are useful in this analysis. A derivation of these properties requires an analysis of the QNR's on both sides of the p-n junction, together with equations describing the DR (see, for example, [2]); for this discussion, however, it is enough to know what the properties are. One conclusion is that the emitter efficiency is positive and less than 1 for the forward-biased problem with no photo-generation of carriers: i.e., the electron and hole currents have the same sign. Another property, which applies to the textbook problem of low-injection conditions (the excess carrier density is much smaller than the doping density), is that the emitter efficiency is a constant. It is unique to the device construction but independent of operating conditions (to the extent that we can ignore geometry changes produced by a change in DR width with changing operating conditions) as long as operating conditions remain at low injection. Also, devices used in practical applications are constructed so that the emitter efficiency is usually very close to 1. This is accomplished by making the p-n junction "one-sided," meaning that the doping on the left side of the junction in Figure 4 is much greater than in the region that is labeled the QNR in the figure. When operating conditions change from low-injection to high-injection (the excess carrier density is much greater than the doping density), nonlinear effects result in a change in the emitter efficiency, but the emitter efficiency is still positive and less than 1 and is often (not always) close to 1 for practical devices. The emitter efficiency is a characteristic of device construction and will be regarded as a known input in this analysis. If nonlinear effects result in the emitter efficiency being a function of the current, the functional dependence is regarded as a given.

To help with the notation, we define a modified emitter efficiency γ in terms of the Ks instead of the Js and with a difference appearing in the numerator. The emitter efficiency γ is defined by

$$\gamma \equiv \frac{K_m - K_M}{K_m + K_M}. \tag{41}$$

It was previously noted that J_m and J_M have the same sign. Hence, for either doping type, K_m and K_M have the same sign, implying that $\gamma \leq 1$. Also, we confine our attention to those cases in which $\gamma_{textbook}$ is large enough to make $|J_m|/|J_M|$ large enough so that

$$|K_m| > |K_M| \quad \text{(the case considered)}. \tag{42}$$

This gives

$$0 < \gamma \leq 1. \tag{43}$$

As with $\gamma_{textbook}$, γ is a characteristic of device construction, usually close to 1 in practical applications, and will be regarded as a known input in this analysis.

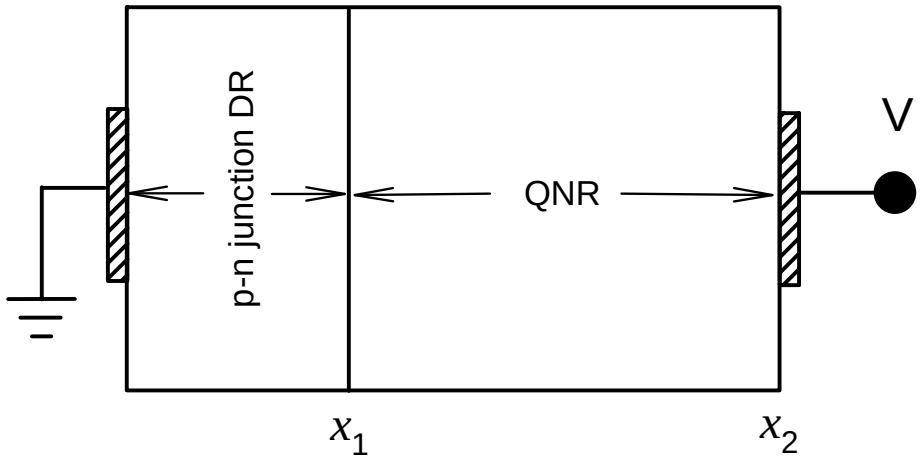

Figure 4. A Simple Diode. The diode has ideal ohmic contacts (shaded regions). The device interior on the left side of x_1 contains a p-n junction together with its space-charge region (DR). The uniformly-doped QNR is between the DRB at x_1 and the ohmic contact at x_2. The applied voltage V has a forward-biasing polarity.

To keep track of the signs of the K's, note that if the QNR is p-type, then the left side of the junction is injecting electrons to the right, so the currents are negative. The K's have the same signs as the currents for the p-type case, so they are negative. Similarly, if the QNR is n-type then the currents are positive, but the K's have the opposite signs and are negative. For both cases,

$$K_m < 0, \qquad K_M < 0. \tag{44}$$

The point x_C is taken to be at x_2 in Figure 4, where the excess carrier density has the known value $P_C = 0$ due to the ideal ohmic contact. The boundary value P_1 becomes the unknown to be solved. With these substitutions together with (41), we can write (38), when evaluated at $x = x_1$, as

$$\frac{a\,y_1}{b} = \gamma\left(1 + \frac{2P_1}{N}\right), \qquad \frac{a\,y_C}{b} = \frac{a\,y_2}{b} = \gamma \tag{45a}$$

$$\frac{a^2}{b}(x_1 - x_2) = -4\frac{\gamma^2}{\gamma+1}\frac{K_m}{N}(x_2 - x_1) = 4\frac{\gamma^2}{\gamma+1}\frac{|K_m|}{N}(x_2 - x_1). \tag{45b}$$

It is evident from (45a) together with the fact that γ is positive that the solution for y is obtained from Figure 1 because $ay_C/b > 0$. The equation for the curve in this figure is

$$\frac{a}{b}y(x) = E_1\left(\frac{a}{b}y_C - \ln\left(\frac{a}{b}y_C + 1\right) + \frac{a^2}{b}(x - x_C)\right). \tag{46}$$

Evaluating (46) at $x = x_1$ and $x_C = x_2$, and then using (45) to change notation gives

$$\gamma\left(1 + \frac{2P_1}{N}\right) = E_1\left(\gamma - \ln(1+\gamma) - 4\frac{\gamma^2}{\gamma+1}\frac{K_m}{N}(x_2 - x_1)\right). \tag{47}$$

This expresses P_1 as a function of γ and K_m, which is equivalent to expressing P_1 as a function of the two currents. An alternate format, which expresses K_m as a function of P_1 and γ, contains the same information but avoids the need for evaluating E-functions. To obtain this alternate format, note that the function T_1 defined by

$$T_1(\xi) \equiv \xi - \ln(\xi + 1) \quad \textit{when} \qquad \xi \geq 0 \tag{48}$$

is the inverse of E_1. Therefore, operating on both sides of (47) with T_1 gives

$$\gamma + \frac{2\gamma P_1}{N} - \ln\left(1 + \gamma + \frac{2\gamma P_1}{N}\right) = \gamma - \ln(1+\gamma) - 4\frac{\gamma^2}{\gamma+1}\frac{K_m}{N}(x_2 - x_1).$$

Rearranging terms gives

$$4\frac{\gamma^2}{\gamma+1}\frac{K_m}{N}(x_2 - x_1) = -\frac{2\gamma P_1}{N} + \ln\left(1 + \frac{2\gamma}{\gamma+1}\frac{P_1}{N}\right). \tag{49}$$

We now look for approximations. First consider approximations applicable to low-injection level conditions (LILC), meaning that $P_1 << N$. Inequalities that will derive the suitable approximation together with a conservative error estimate are obtained by starting with an elementary inequality for the logarithm function, which is

$$\frac{2\xi}{\xi + 2} \leq \ln(1+\xi) \leq \xi \quad \textit{when} \quad \xi \geq 0. \tag{50}$$

Using this with (49) and rearranging terms gives

$$-\frac{1}{2}\frac{P_1}{x_2 - x_1}\left(1 + \frac{P_1}{\gamma P_1 + \gamma N + N}\right) \leq K_m \leq -\frac{1}{2}\frac{P_1}{x_2 - x_1}. \tag{51}$$

Under LILC, the two bounds in (51) come together, and we obtain the low-injection level approximation (LILA) given by

$$K_m \approx -\frac{1}{2}\frac{P_1}{x_2 - x_1} \qquad \text{(LILA)}. \tag{52}$$

This approximation is not new, but the derivation found in elementary semiconductor textbooks is less rigorous. This derivation takes for granted that the drift term can be omitted from the minority carrier equation in (29). The textbook approximation is

$$\frac{J_e}{q\,D_e} \approx \frac{dP}{dx} \quad \text{for p - type,} \qquad -\frac{J_h}{q\,D_h} \approx \frac{dP}{dx} \quad \text{for n - type.}$$

When expressed in terms of the *K*s, the approximation is written as (52) for both doping types. A result obtained from the more rigorous analysis given here that is new is an estimate of the error in the LILA. This is obtained from the inequality (51). A conservative estimate of the relative (or fractional) error in the LILA, when used to estimate K_m, is the term that is added to the 1 in the parenthesis on the left side of (51). That is,

$$|\,\text{relative error in the LILA}\,| < \frac{P_1}{\gamma\,P_1 + \gamma\,N + N}. \tag{53}$$

If $\gamma \approx 1$, then P_1 can be as large as the doping density N and the relative error is still less than 1/3. The relative error is larger for smaller values of γ, but is always less than P_1/N.

Now consider the more general case that includes high-injection level conditions (HILC), defined by $P_1 >> N$. The bounds in (51) apply in general, but the bounds are far apart from each other under HILC and it hasn't yet been shown which of the two bounds is the more accurate estimate. We will now show that the left bound is a fairly (at worst) accurate estimate for the general case. To do this, we add and subtract a term from the right side of (49) to obtain

$$4\frac{\gamma^2}{\gamma+1}\frac{K_m}{N}(x_2 - x_1) = -\frac{2\gamma\,P_1}{N} + \frac{2\gamma}{\gamma+1}\frac{P_1}{N} - \frac{2\gamma}{\gamma+1}\frac{P_1}{N} + \ln\left(1 + \frac{2\gamma}{\gamma+1}\frac{P_1}{N}\right) = -\frac{2\gamma^2}{\gamma+1}\frac{P_1}{N} - T_1\left(\frac{2\gamma}{\gamma+1}\frac{P_1}{N}\right)$$

or

$$K_m = -\frac{1}{2}\frac{P_1}{x_2 - x_1}\left[1 + \frac{\gamma+1}{2\gamma^2}\frac{N}{P_1}\,T_1\left(\frac{2\gamma}{\gamma+1}\frac{P_1}{N}\right)\right]. \tag{54}$$

A numerical evaluation of the accuracy of the approximation

$$T_1(\xi) \approx \frac{\xi^2}{\xi+2} \quad (\text{less than } 12\% \text{ error for any } \xi > 0) \tag{55}$$

finds that the worst relative (or fractional) error is less than (but close to) 12%. The error is close to the worst relative error: i.e., close to 12%, when ξ is between 3 and 4. The relative accuracy improves as ξ decreases below 3 or increases above 4. In fact, the relative error goes to zero in either of the two limits: as $\xi \to 0^+$ or as $\xi \to +\infty$. Note that the square bracket in (54) is a sum of two positive terms; hence, when the T_1 term is replaced by an approximation, the relative error in the square bracket is less than the relative error in the approximation for the T_1 term. Therefore,

substituting (55) into (54) produces an estimate of K_m that has less than 12% error[2], and is given by

$$K_m \approx -\frac{1}{2}\frac{P_1}{x_2 - x_1}\left(1 + \frac{P_1}{\gamma P_1 + \gamma N + N}\right) \qquad \text{(less than 12\% error).} \tag{56}$$

The relative error in (56) goes to zero as the argument to T_1 in (54) either decreases towards zero or increases without bound. An approximation useful under HILC is the high-injection level approximation (HILA), which is defined to be the large-P_1 limit of (56) and is given by

$$K_m \approx -\frac{1}{2}\frac{P_1}{x_2 - x_1}\left(1 + \frac{1}{\gamma}\right) \qquad \text{(HILA).} \tag{57}$$

If $\gamma \approx 1$, the HILA becomes the same as the LILA except for a factor of 2.

The HILA (57) is not new, and there is a less rigorous derivation, as follows. Adding the two equations in (29) gives

$$\frac{J_e}{q D_e} - \frac{J_h}{q D_h} = 2\frac{dP}{dx} + \frac{p_0 - n_0}{V_T}\frac{dU}{dx}.$$

Unlike the individual currents in (29), this combination of currents does not have a large coefficient P (which is large under HILC) multiplying the dU/dx term, suggesting the approximation of omitting this term to get

$$\frac{J_e}{q D_e} - \frac{J_h}{q D_h} \approx 2\frac{dP}{dx}.$$

Integrating and then dividing by x_2-x_1 gives

$$\frac{J_e}{q D_e} - \frac{J_h}{q D_h} \approx -2\frac{P_1}{x_2 - x_1}.$$

Using (35b) to express this in terms of the Ks, and then using (41) to express the result in terms of γ and K_m produces the HILA (57) for either doping type.

A new result obtained from the more rigorous analysis given here is the simple analytical expression in (56) that not only becomes exact in either the LILC limit or HILC limit, but also provides an interpolation between the two extremes having a relative error that is guaranteed to be less than 12% for the case considered. Recall that the case considered includes not only physically-imposed conditions (that K_m and K_M have the same signs and K_m is negative) reflecting the physical problem of a forward-biased diode, but also includes another condition (that $|K_m| > |K_M|$) that was imposed to make the mathematical analysis valid. Although the

[2] Errors refer to mathematical errors. The governing equations in Section 6.1 are themselves only approximations for real devices. Accuracy claims refer to a hypothetical device that is defined by these governing equations.

derivation of (56) excluded the case in which $K_m = K_M$, the conclusion (56) is still defined and correct, and the approximation becomes exact for this case. This can be seen by noting that when $K_m = K_M$ we also have $a = 0$ and $\gamma = 0$. Comparing (56), when evaluated at $\gamma = 0$, to the $a = 0$ result (10), and using (27) and (36) to change notation in the latter result, shows that (56) is exact for this case.

Note that (56) is useful if γ can be estimated. In practice, however, this usually means that the emitter is known to have a high efficiency, so that $\gamma \approx 1$. More generally, an estimate of γ requires an analysis of another QNR on the left side of the junction in Figure 4 (the other QNR is not shown in the figure). When performing such an analysis, it is actually easier to solve for K_M than to solve for γ because the majority-carrier current in the QNR shown in the figure is a minority-carrier current in the left QNR not shown and LILC can often be assumed for calculating the minority-carrier current in the left QNR (which is typically heavily doped). For those cases in which K_M is easier to estimate than γ, (56) will be more useful if expressed in terms of K_M instead of γ. This can be done by substituting (41) into (56) and rearranging terms without changing the relative error to get

$$(P_1 + 2N)\, K_m - P_1\, K_M \approx -\frac{P_1}{x_2 - x_1}(P_1 + N) \qquad \text{(less than 12\% error)}.$$

Note that, because the right side is negative, adding the negative quantity $P_1 K_M$ to both sides can only decrease the relative error; i.e., the error will still be less than 12%. Doing so and then dividing by P_1+2N gives

$$K_m \approx -\frac{P_1}{W}\left(\frac{P_1 + N - K_M W}{P_1 + 2N}\right) \qquad \text{(alternative to (56), also less than 12\% error)}, \qquad (58)$$

where $W \equiv x_2\text{-}x_1$ is the width of the QNR.

Finally, an estimate of the voltage across the QNR might be of some interest. An estimate can be obtained by substituting (41) and (56) into (39) to get

$$\Delta U \approx \frac{2P_1}{\gamma P_1 + \gamma N + N} V_T\,. \qquad (59)$$

The voltage applied to the device terminals is divided between an applied voltage[3] across the DR, plus the voltage ΔU across the QNR. The latter voltage can be a large multiple of V_T (note that V_T is about 0.026 volts at room temperature) under HILC if γ defined by (41) is close to zero. For any other conditions (LILC and/or $\gamma \approx 1$), this voltage will be only about $2V_T$ or less. Large or small, the voltage ΔU has an important effect on majority-carrier transport, which can be seen as follows. By relating the directions of carrier flows to the signs of the Js and then relating the signs of the Ks to the signs of the Js, and noting that K_m and K_M are both negative,

[3] "Applied voltage" is distinguished from "total voltage" across the DR because the latter includes the built-in (a.k.a., equilibrium) potentials while the former does not. There is no distinction between applied and total for the QNR because the QNR has no built-in voltage.

we conclude that minority carriers move away from the DR while majority carriers move towards it. The direction of minority-carrier flow is the direction expected from diffusion that is driven by an excess-carrier boundary value P_1 at the DR boundary (DRB). However, this boundary value also drives majority-carrier diffusion, yet majority carriers move towards the DR, which is the direction opposite to the direction expected from diffusion. Note that the sign convention in (39b), together with the fact, from (59), that ΔU is positive, can be used to show that majority-carrier drift is towards the DR. The fact that the net majority-carrier flow is toward the DR implies that the voltage ΔU is enough to make majority-carrier drift stronger than majority-carrier diffusion as needed to produce this flow direction. Also, minority carriers drift in the opposite direction as majority carriers, so minority carriers drift away from the DR. This is the same direction as produced by diffusion, so drift and diffusion of minority carriers complement each other, while drift and diffusion of majority carriers oppose each other (with drift winning).

6.3 A Reverse-Biased p-n Junction Diode with a Carrier Source

The example considered here is a reverse-biased p-n junction diode containing a source of carriers as shown in Figure 5. One example of this kind of arrangement is produced when the source is a forward-biased junction (with another terminal connection to the source, not shown in Figure 5, to maintain the forward biasing of the source): i.e., a transistor. Another example of this kind of physical arrangement is the photodiode discussed later in Section 6.5. The present discussion considers a more generic case in which the physical nature of the source is unspecified because the only information that is needed is the boundary conditions it imposes. The source is assumed to be consistent with quasi-neutrality being maintained between the DRB (at x_1 in Figure 5) and the source boundary (at x_2 in Figure 5). The source is also assumed to conduct enough of the applied terminal voltage as needed to maintain a reverse-biasing condition across the DR. A supply of carriers produced by a source will modulate the excess carrier density P at the source boundary, and the boundary value, denoted P_S, is used to characterize the source. This analysis regards the boundary value P_S as a given input. It will be seen that, given the above assumptions, all relevant information regarding the source has now been specified for the purpose of calculating the current through the device.

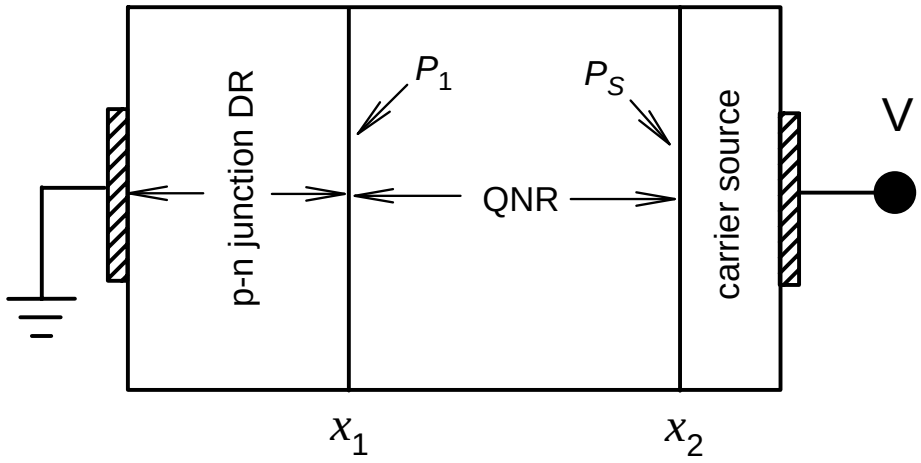

Figure 5. Reverse-Biased Diode with Carrier Source. The device interior on the left side of x_1 contains a p-n junction together with its space-charge region (DR). The uniformly-doped QNR is between the DRB at x_1 and a carrier source at x_2. The source is characterized by a boundary value P_S. The applied voltage V

has a reverse-biasing polarity.

A property of a reverse-biased DR is that it will prevent majority carriers in the QNR from entering it. Similarly, the reverse-biased DR will not supply minority carriers to the QNR. Also, the case considered has no carrier generation anywhere on the left side of the point x_2 in Figure 5 (thermal generation is also excluded: i.e., there is no dark current), so the DR does not supply majority carriers to the QNR. The only current is the result of minority carriers emitted by the source and traveling through the QNR to enter the DR. To avoid the need for taking mathematical limits, we assume that this current is not zero: i.e., the source is "turned on." If the QNR is p-type, minority carriers are electrons moving to the left, so J_e is positive. The Ks have the same signs as the Js for the p-type case, so K_m is positive. If the QNR is n-type, minority carriers are holes moving to the left, so J_h is negative. The Ks have the opposite signs as the Js for the n-type case, so K_m is positive. For both doping types we have

$$K_M = 0, \qquad K_m > 0. \tag{60}$$

We will select x_C to be x_1, where $P = P_1$ (an unknown, but approximations will follow); therefore evaluating (38) at $x = x_2$, while using $K_M = 0$, gives

$$\frac{a\,y_2}{b} = 1 + \frac{2P_S}{N}, \qquad \frac{a\,y_C}{b} = \frac{a\,y_1}{b} = 1 + \frac{2P_1}{N} \tag{61a}$$

$$\frac{a^2}{b}(x_2 - x_1) = 2\,\frac{K_m}{N}\,W\,, \tag{61b}$$

where $W \equiv x_2\text{-}x_1$ is the width of the QNR. It is evident from (61a) that this problem is characterized by $ay_C/b > 0$, so the solution for y is again obtained from Figure 1. Again, the equation for the curve in this figure is (46). Evaluating (46) at $x = x_2$ and $x_C = x_1$, and then using (61) to change notation gives

$$1 + \frac{2P_S}{N} = E_1\left(1 + \frac{2P_1}{N} - \ln\left(2 + \frac{2P_1}{N}\right) + 2W\,\frac{K_m}{N}\right). \tag{62}$$

This expresses P_S as a function of P_1 and K_m. An alternate format, which expresses K_m as a function of P_S and P_1, contains the same information but avoids the need for evaluating E-functions. To obtain this alternate format, operate on both sides with the inverse function T_1 and rearrange terms to get

$$\frac{2W\,K_m}{N} = \frac{2P_S}{N} - \frac{2P_1}{N} - \ln\left(\frac{P_S + N}{P_1 + N}\right). \tag{63}$$

Note that (63) is an exact result (for a hypothetical device defined by the governing equations in Section 6.1) that expresses K_m in terms of P_1 and P_S via elementary functions and is simple enough that an approximation might seem unnecessary. However, a particular approximation has the advantage of more clearly showing how sensitive K_m is to errors produced by replacing P_1

with zero. This approximation is obtained using steps similar to those used in Section 6.2. Adding and subtracting a term on the right side of (63) gives

$$\frac{2W\,K_m}{N} = \frac{2P_1+N}{N}\,\frac{P_S-P_1}{P_1+N} + \frac{P_S-P_1}{P_1+N} - \ln\left(1+\frac{P_S-P_1}{P_1+N}\right). \tag{64}$$

A result that can be taken for granted for now, because we will come back to this later, is that $P_S > P_1$. Therefore, the quantity $(P_S\text{-}P_1)/(P_1\text{+}N)$ is in the domain of T_1; hence, (64) can be written as

$$\frac{2W\,K_m}{N} = \frac{2P_1+N}{N}\,\frac{P_S-P_1}{P_1+N} + T_1\left(\frac{P_S-P_1}{P_1+N}\right). \tag{65}$$

The right side of (65) is the sum of two positive terms. Also, the first term is larger than the T_1 term, which can be seen by comparing terms in (64). Therefore, if the T_1 term is replaced by an approximation, the relative error in the approximation for the right side of (65) will be less than half of the relative error in the approximation for the T_1 term. Therefore, substituting (55) into (65) produces an estimate of K_m that has less than 6% error (again, error refers to a hypothetical device that is defined by the governing equations in Section 6.1), and is given by

$$K_m \approx \frac{P_S-P_1}{W}\,\frac{P_S+P_1+N}{P_S+P_1+2N} \qquad \text{(less than 6\% error)}. \tag{66}$$

To finish the analysis, it is necessary to include some information that is not derivable from a QNR analysis because it requires an analysis of the DR. The conclusions are stated without proof. Under LILC, a reverse-biased DR resembles a sink for minority carriers in the sense that $P_1 << N$, and it is customary to use $P_1 \approx 0$ for such conditions. However, under HILC (where HILC can be defined in terms of currents but can also be recognized by $P_S >> N$) the DR can be populated by a high density of carriers[4]. For this case, it can happen that $P_1 > N$. However, regardless of the injection level, it is still true that $P_1 << P_S$. Therefore, (66) can be reduced to

$$K_m \approx \frac{P_S}{W}\,\frac{P_S+N}{P_S+2N} \tag{67}$$

without significantly compromising the accuracy, whether conditions be low injection, high injection, or in-between. The conclusion is that a real reverse-biased DR can be approximated by the ideal (and hypothetical) case in which $P_1 = 0$.

Elementary textbooks treat the case in which $P_S << N$; for this case, (67) further reduces to

$$K_m \approx \frac{P_S}{2W} \quad \textit{when } P_S << N. \tag{68}$$

[4] The name "depletion region" is used for historical reasons even though the region is not always depleted of charge carriers. It is still a space-charge region that is distinguishable from a QNR.

Note that (68) is the LILA discussed in Section 6.2 except that the driving boundary value P_1 in (52) is replaced by the source boundary value P_S here, and there is a sign change because the driving boundary value is now on the right side of the QNR (Figure 5) instead of the left side.

Finally, an estimate of the voltage across the QNR might be of some interest. Using (39) with x_2-$x_1 = W$, $P_2 = P_S$, $K_M = 0$, and treating the ideal case in which $P_1 = 0$, gives

$$\Delta U = \left[K_m W - P_S\right]\frac{2V_T}{N}. \tag{69}$$

Substituting (63) into (69), again treating the ideal case in which $P_1 = 0$, gives

$$\frac{\Delta U}{V_T} = -\ln\left(1 + \frac{P_S}{N}\right). \tag{70}$$

Note that even if P_S is two orders of magnitude larger than N (an extreme case of high-injection-level conditions), the absolute value of this voltage is still only a few V_T. Although small, it is still important to majority-carrier flow. The majority-carrier current is zero because drift and diffusion of majority-carriers oppose each other with perfect cancellation. Note that majority-carrier drift and majority-carrier diffusion opposing each other implies that drift and diffusion of minority carriers complement each other. Another way to see that drift enhances (instead of opposing) minority-carrier flow is by noting that the LILA (68) is what the minority-carrier current would be if the flow were purely by diffusion, while the more general approximation (67) includes both drift and diffusion. The more general approximation (67) has the same sign but a larger absolute value than the pure diffusion current, indicating that drift and diffusion are in the same direction for minority carriers.

6.4 Carrier Source Producing Majority- and Minority-Carrier Flow to Contact

The example considered here is a variant of the forward-biased diode in Section 6.2. Recall that majority carriers flow towards the DR in the diode in Figure 4. The variant problem considered here replaces the forward biased DR with a different carrier source that, under the biasing conditions considered, results in both minority carriers and majority carriers moving from the source towards the contact at the other end of the device. The device is shown in Figure 6. An example of this kind of physical arrangement is the photodiode discussed later in Section 6.5. The present discussion considers a more generic case in which the physical construction of the source is unspecified. One parameter used to characterize the source is the boundary value of P, denoted P_S, at the source boundary. This analysis regards P_S as a given input. Like the forward-biased diode, but unlike the reverse-biased diode with a source, specifying the carrier-density boundary value is not enough information for calculating the currents. For the forward-biased diode, an estimate of either γ or K_M was also needed. For the problem considered here, the majority-carrier current is the second piece of information that will be regarded as given.

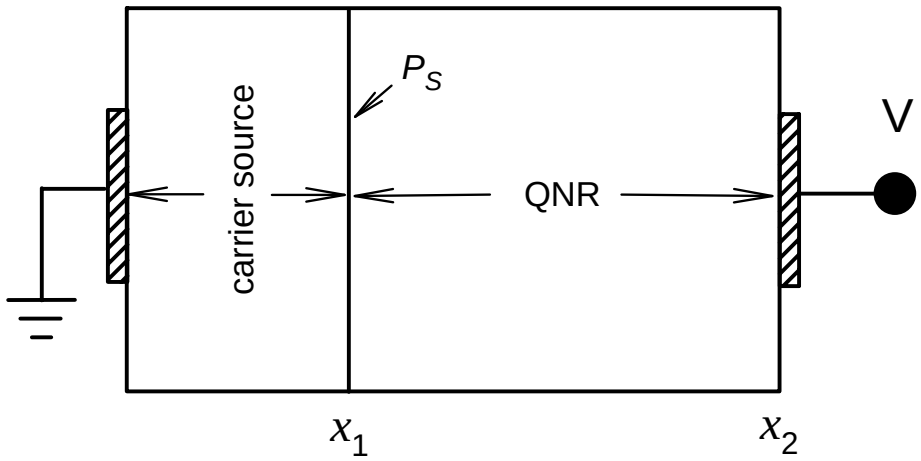

Figure 6. Carrier Source Producing Majority- and Minority-Carrier Flow to Contact. The case considered is such that the carrier source together with the biasing voltage have the property that minority carriers and majority carriers both flow towards the contact on the right.

We consider a source and biasing condition having the property that minority and majority carriers both flow to the right in Figure 6. To avoid the need for taking mathematical limits, we assume that the currents are not zero: i.e., the source is "turned on." By relating the directions of carrier flows to the signs of the *J*s, and then relating the signs of the *K*s to the signs of the *J*s, the stated property can be written as

$$K_m < 0, \qquad K_M > 0, \qquad K_m + K_M \neq 0 \qquad \text{(the case considered)}, \tag{71}$$

where the last condition was imposed to avoid mathematical difficulties.

We will select x_C to be x_2, where $P = 0$ due to the ideal ohmic contact. Hence, evaluating (38) at $x = x_1$ gives

$$\frac{a\,y_1}{b} = \frac{K_m - K_M}{K_m + K_M}\left(1 + \frac{2P_S}{N}\right), \qquad \frac{a\,y_C}{b} = \frac{a\,y_2}{b} = -\frac{K_M - K_m}{K_M + K_m} \tag{72a}$$

$$\frac{a^2}{b}(x_1 - x_2) = 2\frac{K_m - K_M}{K_m + K_M}\frac{K_M - K_m}{N}\,W, \tag{72b}$$

where $W \equiv x_2\text{-}x_1$ is the width of the QNR. Note that the inequalities in (71) allow two possibilities. The first possibility is $K_M + K_m < 0$, in which case (71) and (72) imply $ay_C/b > 0$. The second possibility is $K_M + K_m > 0$, in which case (71) and (72) imply $ay_C/b < -1$. The two cases are considered separately.

6.4.1 Case 1: $K_M + K_m < 0$

For the first case, $ay_C/b > 0$ (in fact, we have the stronger condition $ay_C/b > 1$), so the solution for y is again obtained from Figure 1. Again, the equation for the curve in this figure is (46). Note that the analysis of the forward-biased diode that started with (46) and ended up with the approximation (58) also applies here, except that the error estimate that accompanies (58)

was derived by noting that $K_M < 0$, which is not the case considered here. To obtain an error estimate derived from alternate considerations, it is necessary to start over. Evaluating (46) at $x = x_1$ and $x_C = x_2$, and then using (72) to change notation gives

$$\frac{K_m - K_M}{K_m + K_M}\left(1 + \frac{2P_S}{N}\right) = E_1\left(-\frac{K_M - K_m}{K_M + K_m} - \ln\left(\frac{2K_m}{K_M + K_m}\right) - \frac{2(K_M - K_m)^2}{K_M + K_m}\frac{W}{N}\right). \tag{73}$$

This expresses P_S as a function of K_M and K_m. An alternate format, which expresses K_m as a function of P_S and K_M, might be more useful for some practical applications, but approximations will be needed to express this alternate format in terms of elementary functions. To obtain this alternate format, operate on both sides of the above equation with the inverse function T_1 defined by (48) and rearrange terms to get

$$-2\frac{(K_m - K_M)^2}{K_m + K_M}\frac{W}{N} = 2\frac{K_m - K_M}{K_m + K_M}\frac{P_S}{N} - \ln\left(1 + \frac{K_m - K_M}{K_m}\frac{P_S}{N}\right). \tag{74}$$

Adding and subtracting a term on the right gives

$$-2\frac{(K_m - K_M)^2}{K_m + K_M}\frac{W}{N} = \frac{(K_m - K_M)^2}{K_m + K_M}\frac{P_S}{N\,K_m} + T_1\left(\frac{K_M - K_m}{-K_m}\frac{P_S}{N}\right), \tag{75}$$

where the argument to T_1 is positive and, therefore, in the domain of T_1. Now multiply both sides of (75) by the expression

$$\frac{K_m + K_M}{(K_m - K_M)^2}(K_m\,P_S - K_M\,P_S + 2K_m\,N)\frac{N}{P_S}$$

and subtract $2K_MW$ from both sides of the resulting equation to get

$$\begin{aligned} -2(P_S + 2N)K_m\frac{W}{P_S} = &\left[P_S - \frac{K_M}{K_m}P_S + 2N - 2K_M\,W\right] \\ &+ \left\{\frac{\dfrac{K_M - K_m}{-K_m}\dfrac{P_S}{N} + 2}{\left[\dfrac{K_M - K_m}{-K_m}\dfrac{P_S}{N}\right]^2}\,T_1\left(\frac{K_M - K_m}{-K_m}\frac{P_S}{N}\right)\right\}\frac{K_m + K_M}{K_m}P_S\,. \end{aligned} \tag{76}$$

The approximation (55) will be used for the curly bracket in (76); however, to obtain bounds for the relative error that this will produce on the right side of (76), it is necessary to determine whether the two expressions on the right have the same sign or opposite signs. The curly bracket is positive, and the coefficient on its right is also positive because K_m is negative and (for Case 1) the sum K_m+K_M is negative. The sign of the square bracket on the right is less obvious (it

contains both positive and negative terms because $K_m < 0$ and $K_M > 0$) but can be determined by noting that the error in (55) is one-sided in the sense that T_1 satisfies the inequality

$$T_1(\xi) \le \frac{\xi^2}{\xi + 2} \quad (\text{for any } \xi \ge 0), \tag{77}$$

which can be verified by combining (48) with (50). Writing (77) as

$$\frac{\xi + 2}{\xi^2} T_1(\xi) \le 1 \quad (\text{for any } \xi > 0)$$

and using this with (76) and rearranging terms produces the left bound in

$$-\frac{P_S}{W}\left(\frac{P_S + N - K_M W}{P_S + 2N}\right) \le K_m < -K_M \ . \tag{78}$$

The bound on the right is immediately implied by the definition of Case 1 conditions.

Case 1 is a sub-case in which $K_M > 0$. Using this with (78) gives

$$0 < K_M < \frac{P_S}{W}\left(\frac{P_S + N - K_M W}{P_S + 2N}\right).$$

Rearranging terms gives

$$0 < K_M W < \frac{P_S}{2} \quad (\text{necessary for Case 1 conditions}). \tag{79}$$

We have derived the inequalities that will be needed to quantify the accuracy of an approximation that will be derived later for (76). However, another inequality will be useful for other applications later, so we temporarily interrupt the logic flow to derive this other inequality. Combining the right inequality in (50) with (74) gives

$$2\frac{(K_m - K_M)^2}{K_m + K_M}\frac{W}{N} + 2\frac{K_m - K_M}{K_m + K_M}\frac{P_S}{N} \le \frac{K_m - K_M}{K_m}\frac{P_S}{N}$$

or

$$2\frac{(K_m - K_M)^2}{K_m + K_M}\frac{W}{N} \le \frac{K_m - K_M}{K_m}\frac{P_S}{N} - 2\frac{K_m - K_M}{K_m + K_M}\frac{P_S}{N} = -\frac{(K_m - K_M)^2}{K_m (K_m + K_M)}\frac{P_S}{N}.$$

Therefore,

$$\frac{1}{K_m + K_M} \leq -\frac{1}{K_m\left(K_m + K_M\right)}\frac{P_S}{2W}.$$

Note that K_m is negative. For Case 1 conditions, the quantity K_m+K_M is also negative, so $K_m(K_m+K_M)$ is positive. Multiplying the above inequality by this positive quantity preserves the direction of the inequality, and the result is

$$K_m \leq -\frac{P_S}{2W} \quad \text{(implied by Case 1 conditions).} \tag{80}$$

Returning to (76), we now quantify the accuracy of an approximation for (76). In view of (79), the square bracket in (76) is now seen to be positive. Therefore, the right side of (76) is a sum of two positive terms; hence, if the curly bracket is replaced by an approximation, the relative error in the right side is less than the relative error in the approximation for the curly bracket. Therefore, substituting (55) into (76) (i.e., replacing the curly bracket with 1) produces an estimate of K_m that has less than 12% error (again, error refers to a hypothetical device defined by the governing equations in Section 6.1) and is given by

$$K_m \approx -\frac{P_S}{W}\left(\frac{P_S + N - K_M W}{P_S + 2N}\right) \qquad \text{(less than 12\% error).} \tag{81}$$

Elementary textbooks treat the case in which $P_S << N$. Note that the relative error in (81) goes to zero in the limit as $P_S \to 0$; therefore, an exact limit can be obtained by taking the limit of this approximation. To take this limit, we utilize the inequality (79) to conclude that the K_M term in (81) can be omitted along with the P_S when they are added to N. The result is

$$K_m \approx -\frac{P_S}{2W} \quad \textit{when } P_S << N\,, \tag{82}$$

with the relative error becoming zero in the small-P_S limit. Note that (82) is the LILA discussed in Section 6.2, and was derived from the assumption that minority-carrier flow is purely diffusion, except that the driving boundary value P_1 in (52) is replaced by the source boundary value P_S here. The condition $P_S << N$ is one sufficient condition, but not the only sufficient condition for the LILA to be an accurate approximation. A second sufficient condition can be obtained by recognizing that the definition of the case considered restricts the allowed values of K_M according to (79), but it can come arbitrarily close to either bound. In particular, we can take the limit as K_M approaches $P_S/2W$ from below while holding P_S at a fixed positive value. In this limit, we see that the two bracketing bounds for K_m in (78) come together. The conclusion is

$$K_m \to -\frac{P_S}{2W} \quad \textit{as} \quad K_M \to \frac{P_S}{2W} \ \textit{from below with fixed } P_S > 0. \tag{83}$$

In other words, regardless of whether P_S is large or small compared to N, the LILA (i.e., the approximation $K_m \approx -P_S/2W$) is accurate if K_M is sufficiently close to $P_S/2W$. This assertion could have been anticipated from physical arguments. The condition that K_M is nearly equal to $P_S/2W$ is a statement that the majority-carrier current is nearly a pure diffusion current, implying a weak electric field, which implies that the minority-carrier current is also nearly a pure diffusion current, which is the LILA.

Finally, an estimate of the voltage across the QNR might be of some interest. An estimate can be obtained by substituting (81) into (39) to get

$$\Delta U \approx \frac{2P_S - 4K_M W}{P_S + 2N} V_T \,. \tag{84}$$

In view of (79), we estimate ΔU to be less than $2V_T$. Although small, it still has an important influence on minority-carrier flow when P_S is large enough that the right side of (82) is significantly different than the right side of (81). This can be seen by noting that the LILA (82) is what the minority-carrier current would be if the flow were purely by diffusion, while the more general approximation (81) includes both drift and diffusion. When different, drift is important. Also note that (80) implies that the drift-diffusion current has the same sign but a larger absolute value as the pure diffusion approximation (82). This is another case in which drift and diffusion of minority carriers are in the same direction.

6.4.2 Case 2: $K_M + K_m > 0$

The forward-biased diode with a high emitter efficiency, the reverse-biased diode with a source, and a carrier source that emits majority carriers under Case 1 conditions all have five characteristics in common. The first is $|K_M| < |K_m|$. The second characteristic is that the solution is represented by Figure 1: i.e., the solution is an E_1 function. The third characteristic is that the voltage across the QNR is only a few V_T or less. The fourth characteristic is that the equation for the minority-carrier current reduces to the LILA when the driving boundary value for the excess carrier density is much less than N. The fifth characteristic is that minority-carrier drift and minority-carrier diffusion are in the same direction. All of these characteristics change for the carrier source under Case 2 conditions, defined by $K_M + K_m > 0$. The present problem is more difficult than those previously considered, but it is important because it will be encountered in the analysis of a photo-diode in Section 6.5. As previously stated, this is a case in which $ay_C/b < -1$, so the solution for y is obtained from Figure 3. The equation for the curve in this figure is

$$\frac{a}{b} y(x) = E_3\left(\frac{a}{b} y_C - \ln\left(-\frac{a}{b} y_C - 1 \right) + \frac{a^2}{b}(x - x_C) \right). \tag{85}$$

Evaluating (85) at $x = x_1$ and $x_C = x_2$, and then using (72) to change notation, gives

$$\frac{K_m - K_M}{K_m + K_M}\left(1 + \frac{2P_S}{N}\right) = E_3\left(-\frac{K_M - K_m}{K_M + K_m} - \ln\left(\frac{-2K_m}{K_M + K_m} \right) - \frac{2(K_M - K_m)^2}{K_M + K_m} \frac{W}{N} \right). \tag{86}$$

A useful inequality is obtained by operating on both sides with the inverse function T_3 defined by

$$T_3(\xi) \equiv \xi - \ln(-\xi - 1), \qquad when \qquad \xi < -1$$

and then rearranging terms to get

$$-2\frac{(K_M - K_m)^2}{K_m + K_M}\frac{W}{N} = 2\frac{K_m - K_M}{K_m + K_M}\frac{P_S}{N} - \ln\left(1 + \frac{K_m - K_M}{K_m}\frac{P_S}{N}\right). \tag{87}$$

The inequality is obtained by combining (50) with (87) while recognizing that $K_m+K_M > 0$ (for Case 2) when rearranging terms. The result is

$$-\frac{P_S}{2W} \le K_m \le -\frac{P_S}{W}\left(\frac{P_S + N - K_M W}{P_S + 2N}\right) \quad \text{(necessary for Case 2).} \tag{88}$$

Combining this with the inequality $K_m+K_M > 0$ gives a second inequality

$$K_M W > \frac{P_S}{2} \quad \text{(necessary for Case 2 conditions).} \tag{89}$$

Note that (89) ensures that the upper bound in (88) is greater than the lower bound.

Note that (86) expresses P_S as a function of K_M and K_m. An alternate format, which expresses K_m as a function of P_S and K_M, might be more convenient for some practical applications. As an attempt to invert (86) to solve for K_m as a function of P_S and K_M, we could start with (86) and use the same rearrangement of terms that was used to obtain (76). This produces (76) again, which is repeated below for easy reference:

$$-2(P_S + 2N)K_m\frac{W}{P_S} = \left[P_S - \frac{K_M}{K_m}P_S + 2N - 2K_M W\right] + \left\{\frac{\frac{K_M - K_m}{-K_m}\frac{P_S}{N} + 2}{\left[\frac{K_M - K_m}{-K_m}\frac{P_S}{N}\right]^2} T_1\left(\frac{K_M - K_m}{-K_m}\frac{P_S}{N}\right)\right\}\frac{K_m + K_M}{K_m}P_S\,. \tag{90}$$

Unlike Case 1, the curly bracket multiplied by its (negative) coefficient in (90) is negative for Case 2. The square bracket on the right is seen to be positive because the left side is positive. Because the two terms have opposite signs, the relative error in the right side, produced by replacing the curly bracket with an approximation, might not be less than the relative error in the approximation for the curly bracket. In fact, it is easy to show that if we substitute the usual approximation (55) into (90), we will obtain a very poor approximation for K_m. This can be seen by noting that these steps will produce the same approximation for K_m that was obtained for Case

1 and is given by (81). However, for Case 2 it is possible for K_M to be large enough to make this approximation for K_m positive, which is a very poor approximation for a negative quantity.

This author was not able to find an approximation for K_m, as a function of P_S and K_M, that was judged (by this author) to be an acceptable compromise between accuracy and simplicity under Case 2 conditions. However, Case 2 was considered here because of a practical application in Section 6.5 (a photodiode), and for this application it is not necessary to invert (86) to solve for K_m. Additionally, physical insight (and verification of the statements made in the introduction to Case 2) can be obtained from asymptotic forms together with an illustrative plot, which are discussed below.

Regarding K_m as a function of P_S and K_M, we can investigate limiting values (when they exist) of K_m as the point (P_S, K_M) approaches a selected point while moving along a selected path in a plane. However, limit points and paths followed must be consistent with Case 2 conditions in order to obtain results valid for these conditions. In particular, the path followed must be consistent with (89), although we can take the limit as the left side of (89) approaches the right side from above. Also, Case 2 is a special case in which (71) applies. Whether a limiting value of K_m is approached from above or from below can be recognized (for the limits derived here) by combining the left inequality in (71) with the left inequality in (88) to get

$$-\frac{P_S}{2W} \le K_m < 0 \quad \text{(necessary for Case 2).} \tag{91}$$

One simple limit is easily obtained from (88). If we hold P_S fixed at an arbitrary positive value and take the limit as K_M approaches $P_S/2W$, the two bracketing bounds for K_m in (88) come together. The conclusion is

$$K_m \to -\frac{P_S}{2W} \quad as \quad K_M \to \frac{P_S}{2W} \ \textit{from above with fixed} \ P_S > 0. \tag{92a}$$

Another simple limit is evident from (91). If P_S approaches zero from above with the point (P_S, K_M) moving along any path consistent with (89), we see from (91) that $K_m \to 0$. A special case of such a path is with a fixed $K_M > 0$ with P_S small enough to satisfy (89) at all points on the path. The conclusion is

$$K_m \to 0 \quad as \quad P_S \to 0 \ \textit{from above with fixed} \ K_M > 0. \tag{92b}$$

A third limit is less obvious and the derivation is deferred to Appendix C. The result is

$$K_m \to 0 \quad as \quad K_M \to +\infty \ \textit{with fixed} \ P_S > 0. \tag{92c}$$

The simple limits in (92) are used in Appendix C to derive more refined results that are asymptotic limits.[5] The results derived in Appendix C are

$$K_m + \frac{P_S}{2W} \xrightarrow{A} \frac{\frac{2P_S}{N} - \ln\left(1+\frac{2P_S}{N}\right)}{\frac{2P_S}{N} + \ln\left(1+\frac{2P_S}{N}\right)}\left(K_M - \frac{P_S}{2W}\right)$$

$$as \quad K_M \to \frac{P_S}{2W} \ from\ above\ with\ fixed\ P_S > 0 \tag{93a}$$

$$K_m \xrightarrow{A} -\frac{P_S}{N}\exp\left(\frac{2P_S}{N}\right)\frac{K_M}{\exp\left(\frac{2K_M W}{N}\right)-1}$$

$$as\ either \quad K_M \to +\infty\ with\ fixed\ P_S > 0$$

$$or\ as \quad P_S \to 0\ from\ above\ with\ fixed\ K_M > 0. \tag{93b}$$

Asymptotic limits provide information about local behaviors of K_m as a function of P_S and K_M. A visual perception of the global behavior can be obtained by looking at some plots. One way to construct a plot of K_m as a function of K_M for any given fixed value of P_S is to use parametric equations. For this purpose, we define a parameter α by

$$\alpha \equiv \frac{K_M - K_m}{K_M + K_m}. \tag{94}$$

Note that the inequalities $K_m < 0$ and $K_M + K_m > 0$ imply that $\alpha > 1$. It is not difficult to show that the pair of equations consisting of (87) and (94) is equivalent to the pair of equations

$$K_m = -\frac{P_S}{2W}\frac{\alpha-1}{\alpha}\left[1+\frac{N}{2\alpha P_S}\ln\left(1+\frac{2\alpha}{\alpha-1}\frac{P_S}{N}\right)\right] \tag{95a}$$

$$K_M = \frac{P_S}{2W}\frac{\alpha+1}{\alpha}\left[1+\frac{N}{2\alpha P_S}\ln\left(1+\frac{2\alpha}{\alpha-1}\frac{P_S}{N}\right)\right]. \tag{95b}$$

By selecting a $P_S > 0$ and then assigning a value (exceeding 1) to α, the two currents can be calculated from (95) to produce one point in a plot of K_m versus K_M corresponding to the selected P_S. Repeating this for different values of α produces a set of points used to construct a curve corresponding to the selected P_S. The bounds (89) and (91), and the limit (92a), suggest that it might be convenient to express the K's in units of $P_S/2W$ when plotting K_m as a function of K_M.

[5] An asymptotic limit, denoted $\xrightarrow{A}$, refers to relative (or fractional) differences between two functions. The statement $f(x) \xrightarrow{A} g(x)$ as $x \to a$ means that the relative difference between $f(x)$ and $g(x)$ goes to zero as $x \to a$. A more formal definition is (A11) in Appendix A.

A plot constructed from (95) and using these units is shown in Figure 7 for each of two example values of P_S/N (10 and 0.1). The solid curves are the exact results produced by (95). The dashed curves are the asymptotic forms given by (93).

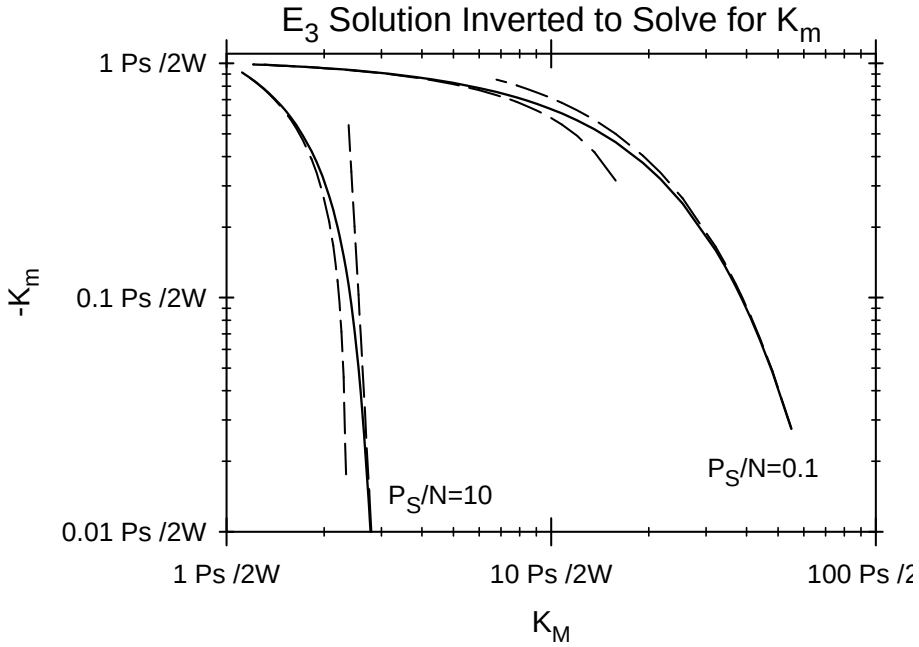

Figure 7. K_m versus K_M when $K_m < 0$, $K_M > 0$, and $K_M+K_m > 0$. The physical arrangement in Figure 6 has an E_3 solution when $K_M+K_m > 0$. Inverting this solution to solve for K_m produces the plots (solid curves) shown for each of two example values of P_S/N. The dashed curves that join the solid curves on the left side are plots of the right side of (93a) minus $P_S/2W$. The dashed curves that join the solid curves on the right side are plots of the right side of (93b).

The vertical coordinate in Figure 7 is $-K_m$ (a positive quantity) in units of $P_S/2W$. Another motivation for using these units is that the approximation $-K_m \approx P_S/2W$ is the LILA first discussed in Section 6.2 and more recently discussed in Section 6.4 under Case 1 conditions. This approximation is found in elementary textbooks and is derived from a pure diffusion analysis of minority carriers. One observation from Figure 7 is that the condition $P_S << N$ is neither necessary nor sufficient for the LILA to be an accurate approximation under Case 2 conditions. Regardless of whether P_S is small or large compared to N, the approximation is accurate if K_M is sufficiently close to $P_S/2W$, and inaccurate otherwise. (How close is "sufficiently close;" e.g., how close is close enough for the LILA to have less than 20% error depends on P_S/N.) This could have been anticipated from the same physical arguments that were used for Case 1 conditions. If K_M is nearly equal to $P_S/2W$, then the majority-carrier current is nearly a pure diffusion current, implying a weak electric field, which implies that the minority-carrier current is also nearly a pure diffusion current, which is the LILA. Like Case 1, Case 2 also has the property that the approximation $K_m \approx -P_S/2W$ is accurate if K_M is sufficiently close to $P_S/2W$ (this assertion is also implied by the limit (92a)). Unlike Case 1, the condition $P_S << N$ is not an alternate sufficient condition for the LILA to be accurate under Case 2 conditions. Even when $P_S << N$, the LILA will still be a poor approximation if K_M is sufficiently large. Another difference between Case 1 and Case 2 can be seen by noting that the actual minority-carrier current (which includes both drift and diffusion) satisfies (91) under Case 2 conditions, which

has the same sign but a smaller absolute value compared to what the minority-carrier current would be if the flow were purely by diffusion (in which case we would have $-K_m = P_S/2W$). The conclusion is that minority-carrier drift opposes minority-carrier diffusion, but diffusion produces the larger current. Case 2 is the first example in this paper in which minority-carrier drift opposes minority-carrier diffusion. However, the cancellation can be nearly perfect in the sense that the absolute value of the drift-diffusion current ($-K_m$) can be a very small multiple of $P_S/2W$. This occurs at the larger values of K_M , as seen in Figure 7.

Finally, an estimate of the voltage across the QNR might be of some interest. An estimate can be obtained by writing (39a) as

$$\Delta U = \left[\left(K_m - K_M \right) W + P_S \right] \frac{2V_T}{N} \ . \tag{96}$$

An inequality derived in Appendix C is

$$K_M - K_m > \frac{P_S}{W} .$$

Using this with (96), we conclude that ΔU is negative. Using (39b), while paying attention to the direction of drift for each type of carrier, we conclude that this polarity is consistent with an earlier conclusion that minority-carrier drift opposes minority-carrier diffusion (hence, drift and diffusion are in the same direction for majority carriers). Unlike cases previously considered, ΔU under Case 2 conditions can have an arbitrarily large absolute value if K_M is sufficiently large. This can be seen by using the simple limit (92c) with (96) to obtain

$$\Delta U \to -\left[K_M W - P_S \right] \frac{2V_T}{N} \quad as \quad K_M \to +\infty \ with \ fixed \ P_S > 0 . \tag{97}$$

6.4.3 Combining Case 1 with Case 2

The two cases are each sub-cases of

$$K_m < 0, \qquad K_M > 0, \qquad K_m + K_M \neq 0, \qquad P_S > 0 , \tag{98}$$

with Case 1 defined by $K_m+K_M < 0$, and the solution utilizes the E_1 function via (73). Case 2 is defined by $K_m+K_M > 0$, and the solution utilizes the E_3 function via (86). A number of properties have already been derived; however, with all of the mathematical detail, it might be difficult to "see the forest for the trees." It is, therefore, instructive to construct some plots that give a visual illustration of properties that have already been derived. Plots are constructed by selecting a positive K_M (it is convenient to express K_M in the units of $N/2W$), which will be held fixed for each plotted curve. We then select a set of negative values for K_m (again, the units $N/2W$ are convenient). For each selected value (excluding $K_m = -K_M$), calculate P_S/N from (73) if $K_m < -K_M$, or from (86) if $K_m > -K_M$. Plotting P_S/N on the horizontal and K_m on the vertical axis, we obtain a plot of K_m versus P_S/N for fixed K_M. Example plots are shown in Figure 8. The curve labels are the selected values of K_M in units of $N/2W$. The dashed line is given by the equation $K_m = -P_S/2W$.

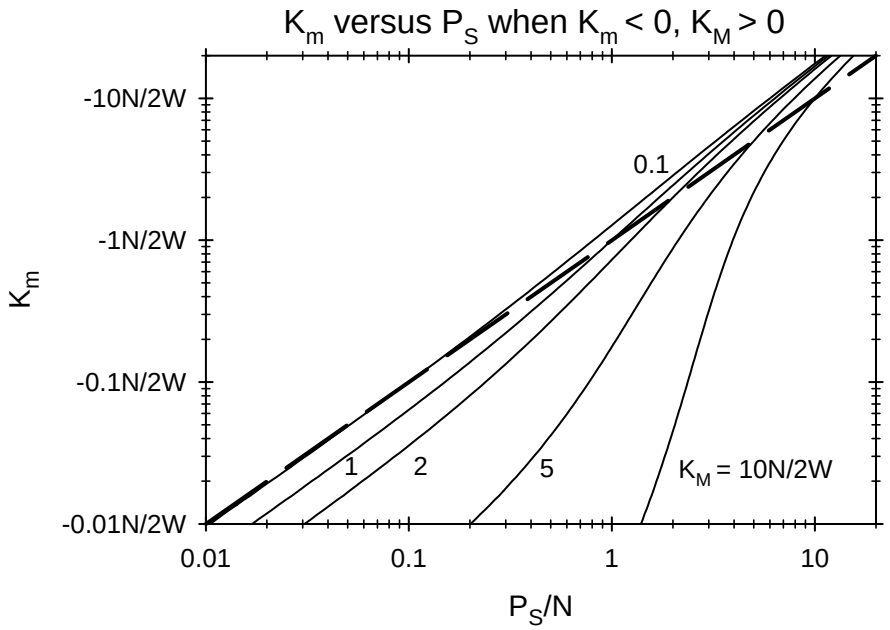

Figure 8. K_m versus P_S for a Carrier Source Producing Majority- and Minority-Carrier Flow to Contact. K_m in units of $N/2W$ is plotted against P_S/N for each of several values of K_M. Curve labels are K_M in units of $N/2W$. Case 1 (with solution given by the E_1-function) and Case 2 (with solution given by the E_3-function) are both represented in the same plot. Any point on a solid curve that is to the left of the intersection between that curve and the dashed line is a Case 2 point (minority-carrier drift opposes minority-carrier diffusion and a large QNR voltage is possible). Points on the right are Case 1 points (minority-carrier drift compliments minority-carrier diffusion and the QNR voltage is only a few V_T or less).

The first observation from Figure 8 is that each solid curve crosses the dash line at the point where $P_S/N = 2WK_M/N$. The point where the curves cross satisfies $K_m = -P_S/2W$; however, because this point also satisfies $P_S/N = 2WK_M/N$, we have the simultaneous conditions $K_m = -P_S/2W = -K_M$, implying $K_m+K_M = 0$ at this point.[6] Another observation from Figure 8 is that each solid curve is below the dashed line at points on the left side of the intersection. A (K_m, P_S/N) point that is below the dashed line is a point that is below the point ($-P_S/2W$, P_S/N). Noting that there is a negative sign in the vertical axis in Figure 8, the implication is $K_m > -P_S/2W$ at points on the left side of the intersection. Also, a point on the left side of the intersection is a point where $P_S/N < 2WK_M/N$. We, therefore, have the simultaneous conditions

$$K_m > -\frac{P_S}{2W}, \quad K_m > \frac{P_S}{2W}, \quad \textit{implying} \quad K_m + K_M > 0 \quad \text{on left of intersection.} \tag{99a}$$

Similarly,

$$K_m < -\frac{P_S}{2W}, \quad K_m < \frac{P_S}{2W}, \quad \textit{implying} \quad K_m + K_M < 0 \quad \text{on right of intersection.} \tag{99b}$$

[6] The condition $K_m+K_M = 0$ could have been allowed in the theory by using limits to define functions. It was excluded for analytical convenience (so we won't have to do that), and also to avoid numerical errors when constructing the plots in Figure 8 and in other numerical work to follow (excluding problem points is an alternative to redefining functions to expand their domains). The exclusion is for convenience, not out of necessity, and we imagine that the condition is allowed when discussing implications from Figure 8.

It is seen from (99) that any point (K_m,P_S/N) on a solid curve that is on the left side of the intersection with the dashed line is a Case 2 point. Recall that Case 2 is a condition in which it is possible to obtain a large voltage (many V_T) across the QNR. In contrast, any point on the right side is a Case 1 point, and the voltage across the QNR cannot be more than a few V_T. Therefore, by comparing any point on any solid curve to the dashed line, we can see at a glance whether a large QNR voltage is or is not possible. Recall, however, that all of these conclusions are based on the background hypotheses that all cases considered satisfy (98).

Note that (99) has only two possibilities when equalities are excluded, or three possibilities if equalities are included. The three possibilities are

$$\begin{aligned} &\textit{Either} \\ &\quad K_m = -\frac{P_S}{2W} \quad \textit{and} \quad K_m = \frac{P_S}{2W}, \\ &\textit{or} \\ &\quad K_m < -\frac{P_S}{2W} \quad \textit{and} \quad K_m < \frac{P_S}{2W}, \\ &\textit{or} \\ &\quad K_m > -\frac{P_S}{2W} \quad \textit{and} \quad K_m > \frac{P_S}{2W}. \end{aligned} \tag{100}$$

Again, this is subject to the background hypotheses that $K_m < 0$ and $K_M > 0$. This result was already proven more formally,[7] but the plots in Figure 8 provide a more visual illustration. Note that (100) could have been anticipated from physical arguments. For example, suppose $K_m < -P_S/2W$. Then the minority-carrier current has a larger absolute value than it would have in a pure diffusion flow driven by a boundary value P_S, which implies that drift compliments diffusion for minority carriers, which implies that drift opposes diffusion for majority carriers, which implies that the majority-carrier current is less than it would be in a pure diffusion process driven by a boundary value P_S: i.e., $K_M < P_S/2W$. Therefore, the condition $K_m < -P_S/2W$ is accompanied by the condition $K_M < P_S/2W$. Similarly, the condition $K_m > -P_S/2W$ is accompanied by the condition $K_M > P_S/2W$.

6.5 Photo-Diode with Localized Source

6.5.1 Overview

We consider a reverse-biased p-n junction diode that is exposed to photons (or any other ionizing radiation, but photons are the most common application) that are able to ionize resident atoms to produce electron-hole pairs. The purpose of the device is to detect photon irradiation and measure its intensity. Photon irradiation intensity is inferred by measuring the photo-generation rate (the rate of production of electron-hole pairs), and the photo-generation rate is measured as follows. The photo-generation liberates mobile electrons and holes in equal numbers. Some of the electrons liberated on the p-side of the metallurgical junction between the

[7] Excluding the equalities and given the background hypotheses, Cases 1 and 2 are mutually exclusive and all inclusive. It was already shown that Case 1 implies the second line in (100) and Case 2 implies the third. Equalities (the first line in (100)) can be included by taking limits.

n-material and p-material move (via drift-diffusion) across the junction to the n-side. Similarly, some of the holes liberated on the n-side move across the junction to the p-side. This produces a current through the device, with the current being an indirect measure of the photo-generation rate. If the photo-generation is confined to only one side of the junction (because the irradiation does not penetrate to both sides), the current consists of minority carriers in the irradiated side crossing the junction to the other side (where the same carriers are now called majority carriers). An illustration is shown in Figure 9. In this example, carrier generation is not only confined to one side of the junction, it is confined to within a narrow region on one side of the junction (a hypothetical case) for reasons that will be explained later. The reverse-biased DR blocks the majority-carrier current, as seen in the QNR that is on the right side of the junction, but on the left side of the generation site, so the total current is the minority-carrier current moving through this QNR towards the junction. Not all liberated minority carriers participate in this motion. Some move (via diffusion) from the generation site to the contact on the right. This is not desirable if the goal is to have a sensitive detector of photon irradiation, because these carriers are not contributing to the current at the junction.[8] Ideally, the current would be equal to the rate of electron-hole pair generation (times the elementary charge), but in reality the current is less than that for the reasons just given.

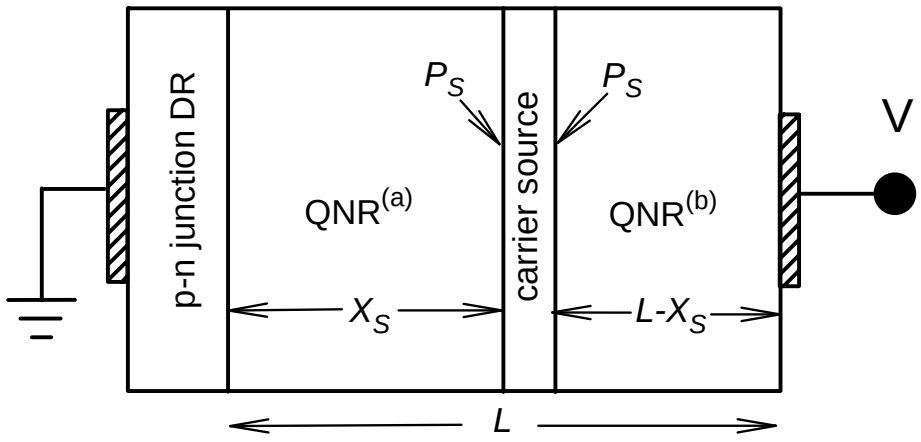

Figure 9. Photo-Diode with a Localized Carrier Generation Site. The case considered is that in which the applied voltage reverse biases the DR and there is a localized generation site. The QNR(a) is the portion of the QNR on the DR side of the generation site and is the same as the QNR in Figure 5 where it was found that the voltage across this region (a few V_T or less) produces a drift that compliments diffusion for minority carriers, and enhances minority-carrier flow towards the DRB on the left. The QNR(b) is the portion of the QNR on the contact side of the generation site and is the same as the QNR in Figure 6 where it was found that either of two possibilities can be encountered. One possibility is that the voltage across this region has a polarity such that drift opposes diffusion of minority carriers and impedes minority-carrier flow towards the contact on the right. In the extreme case, a large voltage (many V_T) stops virtually all minority carriers from moving to the right, implying that virtually all generated minority carriers move towards the DRB.

[8] A different point of view looks at the contact on the right instead of the junction on the left of Figure 9. The current at this location is seen as a flow of carriers moving to the right. This includes all majority carriers, but also some minority carriers. This minority-carrier flow cancels a portion of the majority carrier flow and reduces the current. From either point of view, we see that it is undesirable to have minority carriers moving to the right if the goal is to obtain the largest possible current.

One measure of how close the current is to the ideal limit is a "charge-collection efficiency," defined here to be the current divided by the rate of electron-hole pair generation (times the elementary charge). The charge-collection efficiency is 1 for the ideal case and less than 1 for a real case. Note that the charge-collection efficiency will depend on the spatial distribution of photo-generation. The simplest case to analyze is that in which the region of carrier generation (the source region in Figure 9) is very narrow; the charge-collection efficiency, therefore, is a function of the source location, instead of a functional of the spatial distribution of carrier generation. Such a localized source region is a hypothetical case (but not entirely hypothetical because it actually is possible to control the location of carrier generation: e.g., with a two-photon absorption method as in [4]), but still provides insight in some theoretical investigations. In particular, the charge-collection efficiency associated with a point source, expressed as a function of the spatial coordinates (three coordinates in a three-dimensional device) of the source was a topic of interest in some previous charge-collection investigations (e.g., in [5] and [6]). However, previous investigations did not attempt to calculate the charge-collection efficiency from an exact mathematical analysis of drift-diffusion equations. This will be done here for the simple arrangement in Figure 9, which is a one-dimensional version of the point-source problem. The QNR has a width L, and the source location is at a distance X_S from the DRB. The portion of the QNR that is on the left of the source in Figure 9, denoted $\text{QNR}^{(a)}$, has width X_S, and the portion of the QNR that is on the right, denoted $\text{QNR}^{(b)}$, has width L-X_S. The width of the source region is small enough (effectively zero) for the excess carrier density to be essentially constant within the source region; therefore, a common boundary value P_S applies to each boundary of the source region.

The analysis given here assumes that the applied voltage V in Figure 9 is able to maintain a reverse biasing condition across the DR. However, only a portion of the applied voltage appears across the DR, with another portion across the left QNR and the remainder across the right QNR (as will be seen later, the latter voltage can be many V_T under certain conditions). Therefore, in order to obtain a reverse biasing condition across the DR, as needed for the analysis given here to be valid, it is not enough that the applied voltage have a reverse-biasing polarity. It must also have enough strength to supply the QNR voltages and still have enough left over to reverse bias the DR. As long as it is given that this condition is satisfied, we are given all of the information that is needed regarding the applied voltage.[9]

Before starting a quantitative analysis, we can use results derived for previous examples to reach some qualitative conclusions about the photo-diode in Figure 9. Note that the $\text{QNR}^{(a)}$ in Figure 9 is the same as the QNR in Figure 5, where it was already found in Section 6.3 that the voltage across this region produces a drift that compliments diffusion for minority carriers and enhances minority-carrier flow towards the DRB on the left. This voltage is only a few V_T or less (recall that V_T is about 0.026 volts at room temperature), but can still have an important effect on the current. Also note that the $\text{QNR}^{(b)}$ in Figure 9 is the same as the QNR in Figure 6 because minority and majority carriers both flow to the right in either case. The analysis, in Section 6.4, of the QNR in Figure 6 applies to the $\text{QNR}^{(b)}$ in Figure 9 and was divided into two sub-cases. A quantitative analysis given later will show that either sub-case can apply to the photo-diode in Figure 9, depending on doping type and on the numerical values of various parameters, so each

[9] One complication is being ignored. The voltage across the DR affects its width which in turn affects the width of the left QNR in Figure 9. Analysis of this effect is avoided here by taking the QNR width as given.

sub-case is a possibility when doping type and numerical values are unspecified. One possibility (Case 1 in Section 6.4.1) is that the voltage across the QNR$^{(b)}$ (a few V_T or less) produces a drift that compliments diffusion for minority carriers and enhances minority-carrier flow towards the contact on the right. The other possibility (Case 2 in Section 6.4.2) is that the voltage across this region (which might be a few V_T or less, but could be much larger) has a polarity such that drift opposes diffusion of minority carriers and impedes minority-carrier flow towards the contact on the right. In the extreme case, a large voltage (many V_T) — i.e., a strong electric field — stops virtually all minority carriers from moving to the right, implying that virtually all generated minority carriers move towards the DRB. Therefore, we can anticipate that there might be situations in which virtually all minority carriers are collected at the DRB (the charge-collection efficiency is unity) even when geometrical considerations are such that this would not be the case if minority-carrier flow were purely by diffusion. However, a quantitative analysis is needed to identify these situations.

6.5.2 Governing Equations

To start a quantitative analysis, we first list the relevant equations. Recall that the QNR$^{(a)}$ in Figure 9 is the same as the QNR in Figure 5, which was analyzed in Section 6.3. The result (63) applies to the photo-diode in Figure 9 by merely replacing W with X_S and applying a superscript "(a)" to K_m. However, one approximation will be used. It was argued in Section 6.3 that as long as $P_1 << P_S$ (which we know to be true from experience with numerical solutions of diodes) then P_1 is a small perturbation, even when it is larger than the doping density. A real reverse-biased DR can then be approximated by an ideal (and hypothetical) case in which $P_1 = 0$. In this analysis we will consider such an ideal DR, but the remainder of the mathematical analysis will be exact. For the ideal reverse-biased DR, (63) written in the notation appropriate for Figure 9 becomes

$$\frac{2X_S K_m^{(a)}}{N} = \frac{2P_S}{N} - \ln\left(1 + \frac{P_S}{N}\right). \tag{101}$$

Similarly, the QNR$^{(b)}$ in Figure 9 is the same as the QNR in Figure 6, which was analyzed in Section 6.4. There are two cases exhibiting different properties, but both types of properties are implicitly implied by one common equation. This was listed as (74) under Case 1 conditions and as (87) under Case 2 conditions. This equation applies to the QNR$^{(b)}$ in Figure 9 by merely replacing W with L-X_S and applying a superscript "(b)" each K, which gives

$$2\left(K_M^{(b)} - K_m^{(b)}\right)\frac{L - X_S}{N} = \frac{2P_S}{N} + \frac{K_M^{(b)} + K_m^{(b)}}{K_M^{(b)} - K_m^{(b)}}\ln\left(1 + \frac{K_m^{(b)} - K_M^{(b)}}{K_m^{(b)}}\frac{P_S}{N}\right). \tag{102}$$

Some properties of (102) are not obvious from a casual inspection, but were derived in Section 6.4. These are (100) except for a change in notation by inserting "b" superscripts and replacing W with L-X_S. However, in order for these properties to be valid, it is essential that $K_m^{(b)} < 0$ and $K_M^{(b)} > 0$.

To complete the set of equations, we need to relate the currents on the two sides of the source. Carrier generation within a QNR was not previously considered; however, to relate these

currents, we need the more general continuity equations that include carrier generation and result in the currents being functions of x. These equations are

$$\frac{d\,J_e(x)}{dx} = -q\,g(x), \quad \frac{d\,J_h(x)}{dx} = q\,g(x), \tag{103}$$

where $g(x)$ is the carrier-generation rate density (not to be confused with the g in Section 4). In our application, $g(x)$ is positive within the source region and is zero within the QNRs on either side of the source. Integrating gives

$$J_e^{(b)} - J_e^{(a)} = -qG, \quad J_h^{(b)} - J_h^{(a)} = qG, \tag{104}$$

where G (not to be confused with the G in Section 4) is the integral of g over the source region and is the total (i.e., over the entire source region) rate of electron-hole pair generation per unit area. The quantity qG has the same units as a current density (charge per area per time). Using (35b) to convert to the K notation, and using the fact that majority-carrier flow is blocked by the DR (i.e., $K_M^{(a)} = 0$), we can write (104) as

$$K_m^{(a)} - K_m^{(b)} = \frac{G}{2D_m}, \quad K_M^{(b)} = \frac{G}{2D_M}, \tag{105}$$

where D_M is the diffusion coefficient for majority carriers and D_m is the diffusion coefficient for minority carriers.

Finally, the charge-collection efficiency, denoted Ω, is defined to be the absolute value of the total current density divided by qG. The majority-carrier flow is zero in QNR$^{(a)}$, so the total current is the minority-carrier current in this region. This gives

$$\Omega \equiv \frac{\left|J_m^{(a)}\right|}{qG} = \frac{2D_m K_m^{(a)}}{G}. \tag{106}$$

Other quantities that might be of interest are P_S and the voltage across QNR$^{(b)}$ (because the voltage across QNR$^{(a)}$ is only a few V_T or less, this voltage is less interesting). Note that when solving the simultaneous equations (101) through (106) to calculate Ω, the solution to the system of equations already includes a solution for P_S. However, if the solution to the system of equations is presented in graphical format (e.g., a plot of Ω versus X_S and/or a plot of P_S versus X_S), there will be less clutter if Ω is the only quantity plotted. If we want to know what P_S is when Ω is the only quantity plotted, we can calculate it by first estimating Ω from a plot (a set of plots will be given later) and then calculating P_S from Ω. This is done by noting that that the QNR$^{(a)}$ in Figure 9 is the same as the QNR in Figure 5, which was analyzed in Section 6.3. The result (62) applies to the photo-diode in Figure 9 by merely replacing W with X_S and applying a superscript "(a)" to K_m. For the ideal reverse-biased DR ($P_1 = 0$), (62) written in the notation appropriate for Figure 9 becomes

$$\frac{P_S}{N} = \frac{1}{2} E_1\left(1 - \ln(2) + 2X_S \frac{K_m{}^{(a)}}{N}\right) - \frac{1}{2}.$$

Using (106) to eliminate $K_m{}^{(a)}$ gives

$$\frac{P_S}{N} = \frac{1}{2} E_1\left(1 - \ln(2) + \Omega \frac{X_S}{L} \frac{G\,L}{D_m\,N}\right) - \frac{1}{2}. \tag{107}$$

Similarly, the voltage across QNR$^{(b)}$, denoted $\Delta U^{(b)}$, can be calculated from Ω and P_S. This is done by writing (39a) in the notation appropriate for QNR$^{(b)}$ to get

$$\Delta U^{(b)} = \left[\left(K_m{}^{(b)} - K_M{}^{(b)}\right)(L - X_S) + P_S\right] \frac{2V_T}{N}. \tag{108}$$

Using (105) and (106) to eliminate the Ks gives

$$\frac{\Delta U^{(b)}}{2V_T} = \left(\Omega - 1 - \frac{D_m}{D_M}\right) \frac{G\,L}{2D_m\,N} \left(1 - \frac{X_S}{L}\right) + \frac{P_S}{N}. \tag{109a}$$

Any situation in which there is a large (many V_T) voltage across the QNR is a situation in which there is a large voltage across QNR$^{(b)}$ (there is only a few V_T across QNR$^{(a)}$), in which case $\Delta U^{(b)}$ is approximately the same as the voltage across the entire QNR. However, the latter voltage has the desirable property of satisfying a simpler equation (as seen below); this simplicity will add clarity to some future discussions. To derive this total voltage, first write (69) in the notation appropriate for QNR$^{(a)}$ to get

$$\Delta U^{(a)} = \left[K_m{}^{(a)}\, X_S - P_S\right] \frac{2V_T}{N}$$

and add this to (108) to get

$$\Delta U^{(a)} + \Delta U^{(b)} = \left[K_m{}^{(a)} X_S + \left(K_m{}^{(b)} - K_M{}^{(b)}\right)(L - X_S)\right] \frac{2V_T}{N}.$$

Now use (105) to substitute for the (b) currents to get

$$\Delta U^{(a)} + \Delta U^{(b)} = \left[2K_m{}^{(a)} L - \left(\frac{1}{D_m} + \frac{1}{D_M}\right)(L - X_S) G\right] \frac{V_T}{N}.$$

Finally, use (106) to substitute for $K_m{}^{(a)}$ to get

$$-\frac{\Delta U^{(a)} + \Delta U^{(b)}}{V_T} = \left[\left(1 + \frac{D_m}{D_M}\right)\left(1 - \frac{X_S}{L}\right) - \Omega\right] \frac{G\,L}{D_m\,N}. \tag{109b}$$

All of the equations needed to calculate Ω have been listed in (101) through (106). After Ω has been calculated, the auxiliary equations (107) and (109) can be used to calculate other quantities that might be of interest.

6.5.3 Small-G Limit and Large-G Limit

A limiting case that deserves some attention is the small-G limit. This is the case treated in elementary textbooks, which solve for the current by assuming that minority-carrier flow is a pure diffusion process; i.e., the current is calculated by assuming that the carrier density satisfies a linear diffusion equation. However, this calculated current requires some clarification. The relationship between P_S and G given by (107) is derived from drift-diffusion, not pure diffusion, so the relationship between P_S and G is not consistent with a linear diffusion equation. Therefore, the current calculated from the linear diffusion equation that is driven by the boundary value P_S (the case considered in all earlier discussions of the LILA) is not the same as the current calculated from a linear diffusion equation that is driven by a carrier generation rate G.[10] In particular, the conclusion under Case 2 in Section 6.4.2, that a small P_S is not a sufficient condition for the calculated diffusion current to be an accurate approximation for the actual current, was referring to a calculated current that is driven by the boundary value P_S. It might still be true (in fact it is true, as seen later) that a sufficiently small G results in the calculated diffusion current being an accurate approximation for the actual current when the calculated current is driven by the carrier generation rate G. It is not difficult to show that the linear diffusion equation driven by the carrier generation rate predicts the charge-collection efficiency to be given by

$$\Omega = 1 - \frac{X_S}{L} \qquad \text{(pure diffusion)}. \tag{110}$$

The charge-collection efficiency will converge to the right side of (110) in the small-G limit, which agrees with elementary textbooks. It is not necessary to give a formal proof of this assertion because graphical solutions presented later will show that Ω converges to the right side of (110) in the small-G limit.

Another limit of interest is the large-G limit. Taking it for granted that P_S increases without bound when G increases without bound with all other parameters fixed, the large-G limit is also a large-P_S limit. In this limit, the linear term in (101) dominates the logarithmic term; hence, the logarithmic term can be omitted. Now consider the logarithmic term in (102). Recognizing that $K_m^{(b)}$ is negative and $K_M^{(b)}$ is positive, the coefficient to the logarithmic term has an absolute value that is less than 1. Therefore, in the large-P_S limit, the logarithmic singularity in (102) is dominated by linear singularities (e.g., in the $2P_S/N$ term), subject to a qualification. This qualification is that the limiting case is not one in which the $K_m^{(b)}$, appearing in the denominator inside the logarithm, goes to zero. We, therefore, confine our attention to those values of X_S such that Ω is bounded below 1 (i.e., there is an $\varepsilon > 0$ such that $\Omega \le 1-\varepsilon$) as $P_S/N \to \infty$ (the existence of

[10] In addition to these two "calculated" diffusion currents, there is a third "true" diffusion current, but it is associated with a point in the device instead of being a terminal quantity. In the un-normalized notation (J-current notation) this current at a point x is $\pm qD\mathrm{d}P(x)/\mathrm{d}x$ where P is the actual excess carrier density, i.e., calculated from the nonlinear drift-diffusion equation. This current is the same whether driven by P_S or by G because the two driving terms are consistent from the point of view of the nonlinear drift-diffusion equation.

such an X_S can be taken for granted without proof because conclusions will be verified by plots presented later). With such an X_S selected, the limiting case allows us to neglect the logarithmic term in (102). Combining (101), without the logarithmic term, and (102), also without the logarithmic term, with (105) and (106) gives

$$\Omega = \left(\frac{D_m}{D_M} + 1 \right) \frac{L - X_S}{L}.$$

Recall that there is a qualification. The limit applies only to those values of X_S large enough for Ω to be bounded below 1 in the limit as $P_S \to \infty$. The equation above is erroneous at any X_S at which the right side exceeds 1; i.e., at any $X_S < X_S'$ where X_S' is given by

$$\frac{X_S'}{L} \equiv \frac{1}{1 + \frac{D_M}{D_m}}. \tag{111a}$$

For $X_S < X_S'$, we set Ω equal to 1. Therefore, the final result for this limiting case is

$$\Omega = \begin{cases} \left(\frac{D_m}{D_M} + 1 \right) \frac{L - X_S}{L} & if \quad X_S > X_S' \\ 1 & if \quad X_S \le X_S' \end{cases} \qquad \text{(large - } G \text{ limit).} \tag{111b}$$

6.5.4 Numerical Algorithm

Limiting cases were considered above. We now consider the more general case. As previously stated, all of the equations needed to calculate Ω have been listed in (101) through (106). After Ω has been calculated, the auxiliary equations (107) and (109) can be used to calculate other quantities that might be of interest. Due to the complexity of (102), it will be necessary to resort to a numerical solution for Ω. In view of this fact, it is reasonable to ask whether the analysis given here has any advantage over the more traditional practice of using a computer simulation (performed by a computer code that numerically solves boundary value problems) to investigate the device in Figure 9. The advantage of the analysis here is explained as follows. A simulation does not treat the generic case. Each numerical example is a special case. Therefore, when using simulations to construct a plot of Ω versus X_S, the simulator must be given a set of six input parameters, consisting of doping type (n-type or p-type), N, L, D_e, D_h, and G. A plot of Ω versus X_S produced by one set of input parameters tells us nothing (without an analysis such as given here) about how this plot compares to a plot produced by a different set of input parameters. To make comparisons, it is necessary to perform repeat simulations for different choices of these six input parameters. Similarities between different plots are discovered (if the investigator is lucky enough to notice them) instead of derived, so conclusions are reliable only for the specific numerical examples that were simulated, and the examples will never be an exhaustive set when there are six input parameters. In contrast, the analysis given here will show that the original six input parameters can be grouped into a reduced set of two input parameters (but the independent variable X_S will be replaced by X_S/L). Different choices of the original six parameters that produce the same reduced set of two parameters are described by

the same plot of Ω versus X_S/L, so similarities between plots are derived rather than discovered. Also, with only two parameters in the reduced set, it is not difficult to construct a family of plots that can be used (via interpolations) to represent all possible numerical examples of practical interest.

A reduced set of parameters is easy to recognize if we first define normalized (and dimensionless) measures of the currents (each denoted with an *I*) and a normalized (and dimensionless) measure of the carrier generation rate (denoted *H*, but not to be confused with the *H* in Sections 1 through 5) by

$$I_m^{(a)} \equiv \frac{2K_m^{(a)}}{N} X_S, \quad I_m^{(b)} \equiv \frac{2K_m^{(b)}}{N}(L - X_S), \quad I_M^{(b)} \equiv \frac{2K_M^{(b)}}{N}(L - X_S) \tag{112a}$$

$$H \equiv \frac{G\,L}{D_m\,N}. \tag{112b}$$

Using (112), we can write (101), (102), (105), and (106) as

$$I_m^{(a)} = \frac{2P_S}{N} - \ln\left(1 + \frac{P_S}{N}\right) \tag{113}$$

$$I_M^{(b)} - I_m^{(b)} - \frac{2P_S}{N} - \frac{I_M^{(b)} + I_m^{(b)}}{I_M^{(b)} - I_m^{(b)}} \ln\left(1 + \frac{I_m^{(b)} - I_M^{(b)}}{I_m^{(b)}} \frac{P_S}{N}\right) = 0 \tag{114}$$

$$I_m^{(a)} \frac{L}{X_S} - I_m^{(b)} \frac{L}{L - X_S} = H, \qquad I_M^{(b)} \frac{L}{L - X_S} = \frac{D_m}{D_M} H \tag{115}$$

$$\Omega = \frac{I_m^{(a)}}{H} \frac{L}{X_S}. \tag{116}$$

Now suppose two dimensionless input parameters *H* and D_m/D_M are given, and the dimensionless independent variable X_S/L is given. The system of equations (113) through (116) then contains five dimensionless unknowns to be solved, consisting of P_S/N, $I_m^{(a)}$, $I_m^{(b)}$, $I_M^{(b)}$, and Ω. The five equations can solve for all of these, so only the two previously mentioned input parameters need be given, if suitable constraints are imposed, as discussed below.

To avoid mathematical difficulties, constraints that must be observed by the user when entering inputs are

$$H > 0, \qquad \frac{D_m}{D_M} > 0, \qquad 0 < \frac{X_S}{L} < 1. \tag{117}$$

It is necessary for *H* to be strictly positive because a charge-collection efficiency can't be defined when *H* is zero except by taking a limit. Note that either doping type is represented, so minority

carriers can be either electrons or holes; therefore, D_m/D_M can be less than 1 or greater than 1. Other constraints are encountered within the algorithm that numerically solves the system of equations. These are $P_S > 0$, $I_m^{(a)} > 0$, $I_m^{(b)} < 0$, and $I_M^{(b)} > 0$. The second constraint in this list is implied by the first together with (113) (using an inequality for logarithms). The fourth constraint is implied by (115). Therefore, the only constraints that have to be forced are

$$P_S > 0 \tag{118}$$

$$I_m^{(b)} < 0. \tag{119}$$

Because of (115), we conclude that (119) is equivalent to $I_m^{(a)} < HX_S/L$ (i.e., $\Omega < 1$). Also, (119) is essential for eliminating extraneous solutions. Without this constraint, there will be a solution to (113) through (115) that is extraneous because it violates (119) even though it satisfies (118). This solution consists of $I_M^{(b)} = I_m^{(b)}$, which satisfies (114),[11] with $I_M^{(b)}$ and $I_m^{(a)}$ selected to satisfy (115), and then P_S selected to satisfy (113) (such a positive P_S exists). The possibility of an algorithm finding an extraneous solution can be eliminated by imposing a suitable upper bound on P_S that forces the condition (119). To find this bound, note that (119) is equivalent to $I_m^{(a)} < HX_S/L$. Using this with (113) gives

$$\frac{2P_S}{N} - \ln\left(1 + \frac{P_S}{N}\right) < \frac{X_S}{L} H.$$

Subtracting some terms from both sides and using the definition (48) of the T_1 function, we can write this as

$$T_1\left(1 + \frac{2P_S}{N}\right) < \frac{X_S}{L} H + 1 - \ln(2).$$

Operating on both sides with the inverse function E_1 preserves the direction of the inequality, because E_1 is strictly increasing; the result is

$$1 + \frac{2P_S}{N} < E_1\left(\frac{X_S}{L} H + 1 - \ln(2)\right)$$

or

$$\frac{P_S}{N} < \frac{1}{2} E_1\left(\frac{X_S}{L} H + 1 - \ln(2)\right) - \frac{1}{2}. \tag{120}$$

The strict inequality in (120) must be observed when selecting a trial P_S in order to avoid a mathematical error. Reversing the steps used to derive (120) we find that if P_S/N is set equal to the right side of (120) we will obtain $I_m^{(a)} = HX_S/L$, leading to $\Omega = 1$ and $I_m^{(b)} = 0$. The latter condition produces an undefined logarithm in (114). Root-finding algorithms using the bisection

[11] The right side of (113) can be evaluated at $I_M^{(b)}$-$I_m^{(b)} = 0$ by defining it by its limit to obtain a continuous function. Doing so will show that (113) is satisfied when $I_M^{(b)} = I_m^{(b)}$ regardless of the value of P_S.

method often ask the user for a closed interval to select trial values of a variable from (i.e., endpoint values for the interval should be selected to avoid mathematical errors), as opposed to strict inequalities. The lower bound zero for P_S/N is okay in the sense of avoiding math errors, but the upper bound given by the right side of (120) is not. To remedy this, suppose (for example) that (120) is replaced with

$$\frac{P_S}{N} \le \frac{1}{2} E_1\left(\frac{X_S}{L} \frac{H}{1.0001} + 1 - \ln(2) \right) - \frac{1}{2}.$$

Reversing the steps used to derive (120) we find that if P_S/N satisfies the above inequality we will obtain $\Omega \le 0.9999$. Conversely, for any user-defined H, D_m/D_M, and X_S/L for which the correct Ω is less than 0.9999, this correct Ω can be found by using the above inequality to constrain the trial values of P_S/N. If this constraint causes a bisection method to "crash," because the correct solution P_S violates this inequality, the crash is interpreted to mean that $0.9999 < \Omega \le 1$; i.e., Ω is virtually indistinguishable from 1. With this interpretation, the trial values of P_S/N can be selected from the closed interval given by

$$0 \le \frac{P_S}{N} \le \frac{1}{2} E_1\left(\frac{X_S}{L} \frac{H}{1.0001} + 1 - \ln(2) \right) - \frac{1}{2}. \tag{121}$$

The numerical algorithm used here is as follows. Assign values consistent with (117) to the parameters and independent variable shown in (117). Select a trial value for P_S/N consistent with (121). Then calculate $I_m^{(a)}$ from (113) and calculate $I_m^{(b)}$ and $I_M^{(b)}$ from (115). Finally, the trial value for P_S/N is tested by determining if (114) is satisfied. If not, select another trial value for P_S/N and test again. This can be done by commercial software using a bisection method for updating trial values. If the algorithm crashes because the left side of (114) does not change sign when P_S/N is in the interval given by (121), set Ω equal to 1. Otherwise, calculate Ω from (116).

6.5.5 Discussion of the Results

The results of the above calculations are shown in Figures 10 through 16, which show plots of Ω versus X_S/L. Different figures refer to different D_m/D_M ratios. Different curves within a figure refer to different values of G. In each figure, the smallest selected G is small enough for the small-G limit (110) to be an accurate approximation; hence, it is also an accurate approximation for all smaller G. Similarly, the largest selected G is large enough for the large-G limit (111) to be an accurate approximation; hence, it is also an accurate approximation for all larger G. In general, the curves are bracketed between the small-G and large-G limits.

The values of qG shown in Figures 10 through 16 are in the units of qD_mN/L. To get an order-of-magnitude estimate of this unit, let us consider a specific example. Suppose the QNR width L is 10 μm (an arbitrary choice, but numbers are easily scaled for other values of L), and suppose the material is p-type (so D_m is the electron diffusion coefficient) with a doping density N equal to $10^{16}/\text{cm}^3$. The electron mobility in silicon with this doping density is about 1100 cm^2/V-s, which gives $D_m = V_T\mu_e = 28.6$ cm^2/s. Using these numbers gives $qD_mN/L \approx 0.5$ micro-amps/μm^2. For this example, the curve labeled $qG = 100\ qD_mN/L$ refers to a generation rate of 50 micro-amps/μm^2, or 5,000 amps/cm^2. Note that Figures 10 through 16 show that, in all cases, the

small-G limit (110) is an excellent approximation when $qG = 0.1\ qD_mN/L$ and a fairly good approximation when $qG = 1\ qD_mN/L$. Also, in all cases, the large-G limit (111) is an excellent approximation when $qG = 1000\ qD_mN/L$ and a fairly good approximation when $qG = 100\ qD_mN/L$.

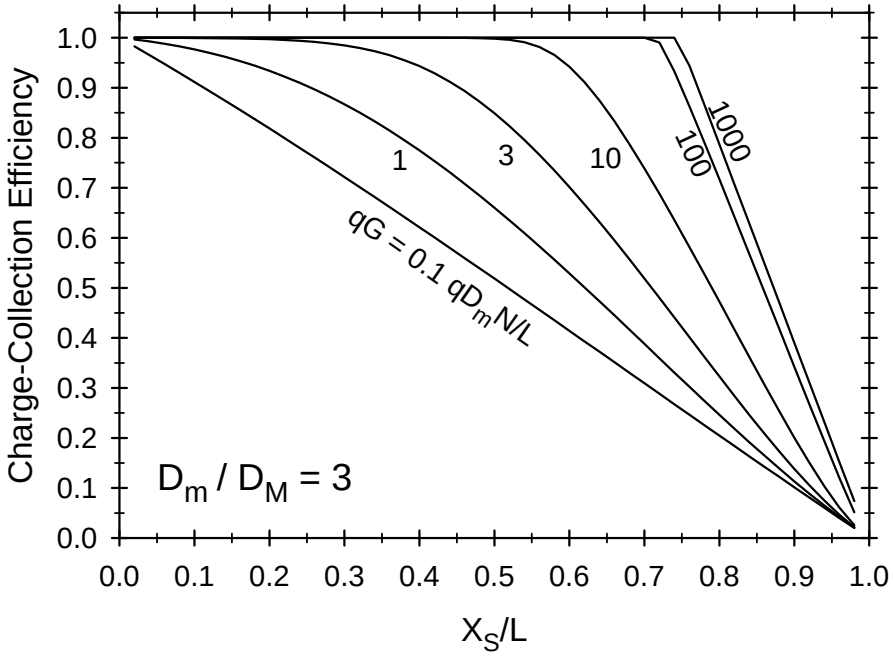

Figure 10. Charge-Collection Efficiency versus X_S/L when $D_m/D_M = 3$. The curve labels are the values of qG in units of qD_mN/L. The smallest-G curve is accurately approximated by (110), which can be used for all smaller G. The largest-G curve is accurately approximated by (111), which can be used for all larger G.

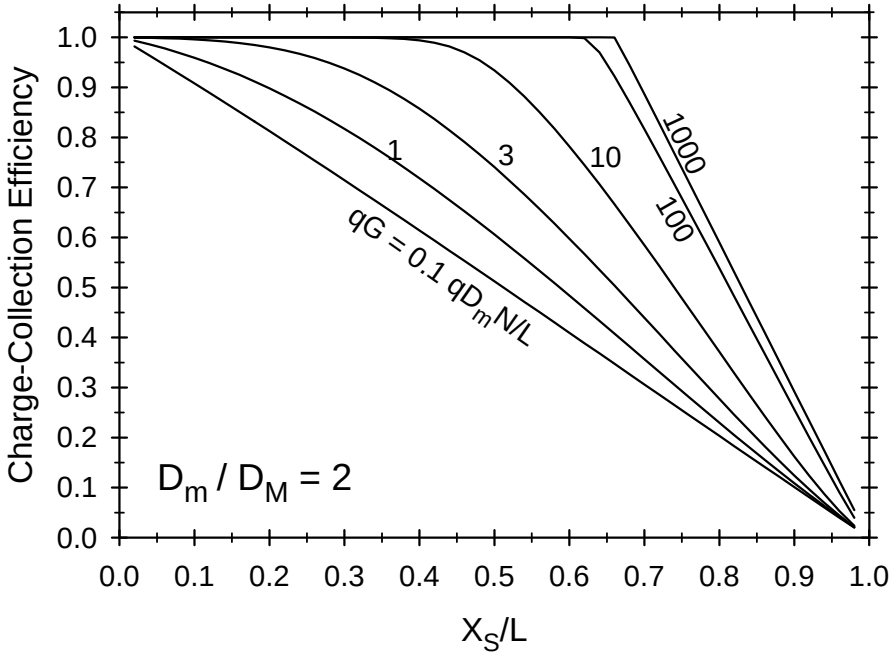

Figure 11. Charge-Collection Efficiency versus X_S/L when $D_m/D_M = 2$. The curve labels are the values of qG in units of qD_mN/L. The smallest-G curve is accurately approximated by (110), which can be used for all smaller G. The largest-G curve is accurately approximated by (111), which can be used for all larger G.

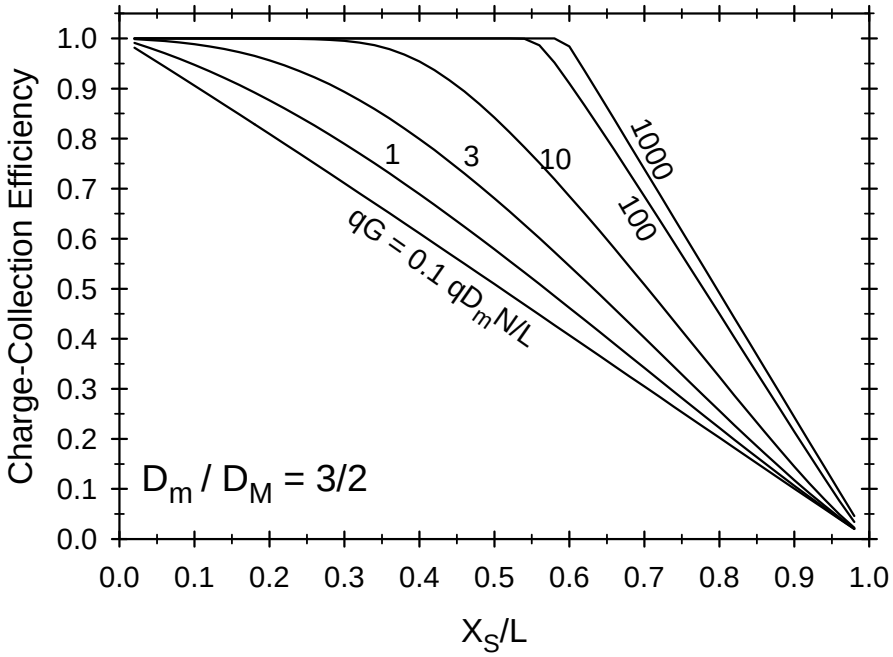

Figure 12. Charge-Collection Efficiency versus X_S/L when $D_m/D_M = 3/2$. The curve labels are the values of qG in units of qD_mN/L. The smallest-G curve is accurately approximated by (110), which can be used for all smaller G. The largest-G curve is accurately approximated by (111), which can be used for all larger G.

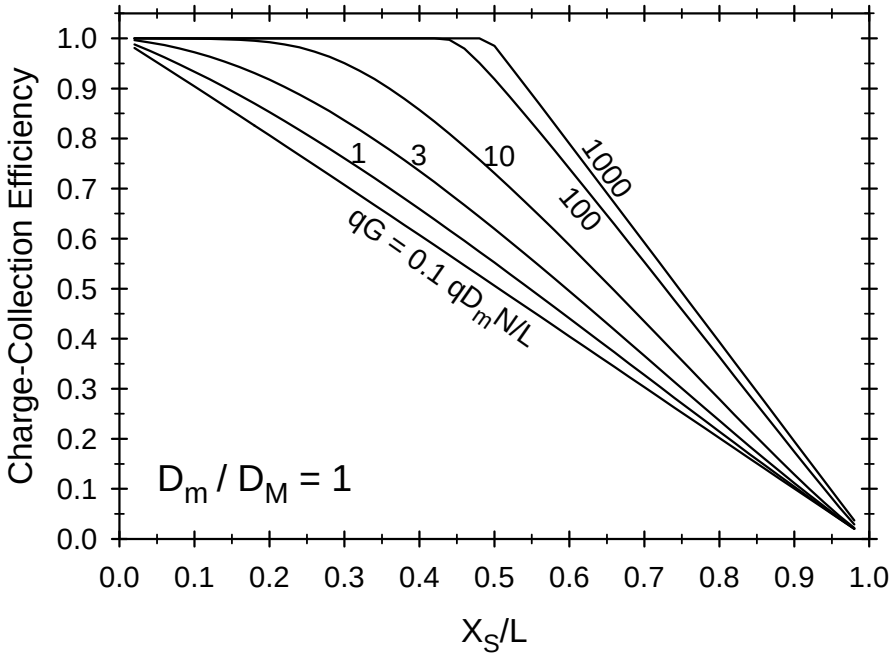

Figure 13. Charge-Collection Efficiency versus X_S/L when $D_m/D_M = 1$. The curve labels are the values of qG in units of qD_mN/L. The smallest-G curve is accurately approximated by (110), which can be used for all smaller G. The largest-G curve is accurately approximated by (111), which can be used for all larger G.

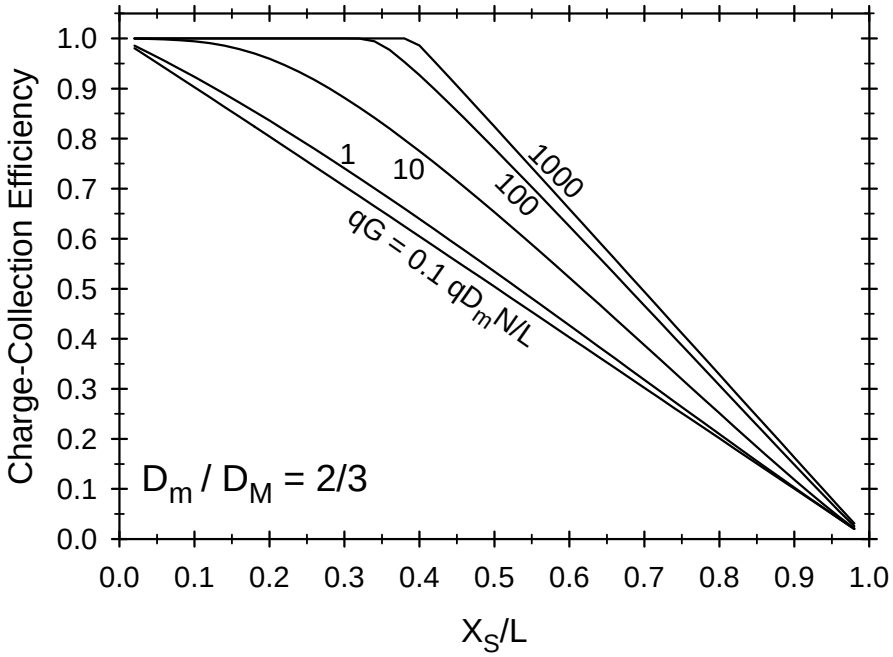

Figure 14. Charge-Collection Efficiency versus X_S/L when D_m/D_M = 2/3. The curve labels are the values of qG in units of qD_mN/L. The smallest-G curve is accurately approximated by (110), which can be used for all smaller G. The largest-G curve is accurately approximated by (111), which can be used for all larger G.

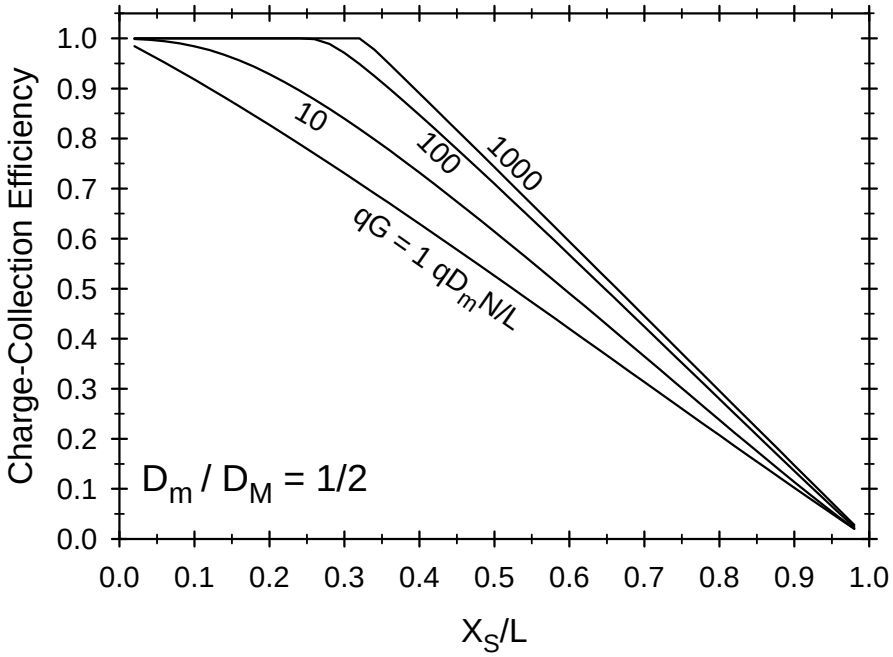

Figure 15. Charge-Collection Efficiency versus X_S/L when D_m/D_M = ½. The curve labels are the values of qG in units of qD_mN/L. The smallest-G curve is accurately approximated by (110), which can be used for all smaller G. The largest-G curve is accurately approximated by (111), which can be used for all larger G.

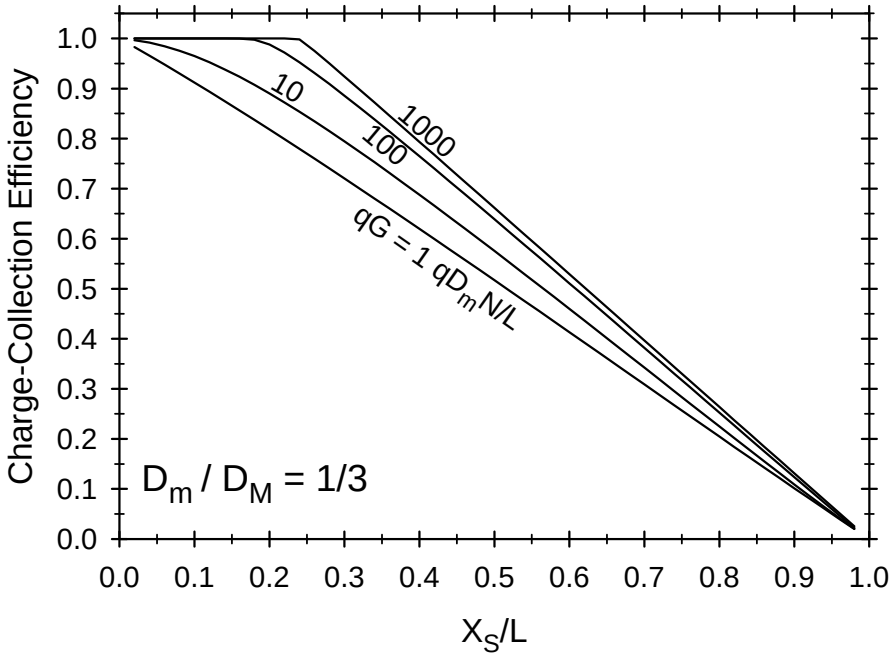

Figure 16. Charge-Collection Efficiency versus X_S/L when D_m/D_M = 1/3. The curve labels are the values of qG in units of qD_mN/L. The smallest-G curve is accurately approximated by (110), which can be used for all smaller G. The largest-G curve is accurately approximated by (111), which can be used for all larger G.

One observation from Figures 10 through 16 is that Ω increases when G is increased with a fixed D_m/D_M and a fixed X_S/L. This is seen by comparing different curves within the same figure. By comparing curves having the same label (and at the same X_S/L) but in different figures, we conclude that Ω increases when D_m/D_M increases with a fixed GL/D_mN and a fixed X_S/L. Note that G is not held fixed in this comparison. Although not obvious from a casual inspection of the plots, numerical evaluations will show that Ω increases when D_m/D_M increases with a fixed G and a fixed X_S/L. Because a p-type material has a larger D_m/D_M than an n-type material with the same doping, Ω is larger for the p-type than the n-type.

The large-G limit is particularly interesting. To discuss this in more detail, define Ω* to be this limiting case. That is,

$$\Omega^*(X_S) \equiv \begin{cases} \left(\frac{D_m}{D_M}+1\right)\frac{L-X_S}{L} & \text{if} \quad X_S > X_S' \\ 1 & \text{if} \quad X_S \le X_S' \end{cases}, \tag{122a}$$

where

$$X_S' \equiv \frac{L}{1+\frac{D_M}{D_m}}. \tag{122b}$$

In general, we have $\Omega(X_S) < \Omega^*(X_S)$; however at a sufficiently large G, we have $\Omega(X_S) \approx \Omega^*(X_S)$. It is interesting to consider the voltage across the QNR when this approximation applies. Note that if we replace Ω with Ω^* in (109b), the square bracket will become zero when $X_S > X_S'$. The actual Ω is slightly less than Ω^*, so the actual value for the square bracket is not exactly zero, and even a small deviation from zero can be significant when multiplied by a large coefficient GL/D_mN, but the voltage given by (109b) is much closer to zero than it would be if it were evaluated at some X_S that is less than X_S'. For example, if we take the limit as $X_S \to 0$ while using $\Omega \to 1$ we obtain

$$-\frac{\Delta U^{(a)} + \Delta U^{(b)}}{V_T} \to \frac{G\,L}{D_M\;N}\,.$$

If $qG = 100\; qD_mN/L$ (for example) and $D_m/D_M = 2$ (for example), the voltage across the QNR is 200 V_T, or 5.2 volts.[12] As previously pointed out, nearly all of this voltage is across QNR$^{(b)}$. Note that if G is the same for both cases being compared, the p-type QNR will have a larger voltage than the n-type because D_M is smaller for the p-type. Also, the critical depth X_S' (the depth at which $\Omega \approx 1$ when G is large) given by (122b) is larger for the p-type because of the larger D_m/D_M ratio.

Let us now vary X_S while still considering the case in which G is sufficiently large so that $\Omega(X_S) \approx \Omega^*(X_S)$. If we vary X_S within an interval in which $X_S > X_S'$, we find that Ω increases linearly with decreasing X_S (the sloped portion of the upper-most curve in any of the Figures 10 through 16), while the QNR voltage remains so small that it is calculated to be zero when replacing Ω with Ω^* in (109b). If we further decrease X_S so we are now in the interval in which $X_S < X_S'$, Ω is virtually constant and virtually equal to 1 for all such X_S, while the absolute value of the QNR voltage, calculated from (109b), increases with decreasing X_S. A physical explanation was already given in an earlier investigation [3] and is as follows. In the large-G limit, the carrier density between the source and the DRB will always be large enough to make this region highly conductive and the electric field will be weak in this highly conductive region. Different cases differ in the region between the source and contact. If the source is sufficiently close to the contact, the carrier density between the source and contact is large enough to make this region highly conductive so the electric field will be weak throughout the QNR. Note that even a weak electric field can produce a large enough minority-carrier drift current, in a highly conductive region, to nearly compensate for minority-carrier diffusion and drive most minority carriers towards the DR. This occurs when X_S is close to X_S' and Ω is nearly equal to 1. Therefore, any X_S greater than X_S' is characterized by a weak electric field (due to a large conductivity) throughout the QNR, but Ω can be anything between 0 or 1 depending on whether

[12] Recall that the analysis assumes that the power supply voltage is sufficient to reverse bias the DR. In this example the power supply voltage must exceed 5.2 volts in order to provide the QNR voltage and still have enough left over to reverse bias the DR. Otherwise the DR will become forward biased and produce a forward current that competes with the reverse current associated with the carrier generation source. This is the physical explanation of a competing current. A mathematical explanation refers to boundary conditions. When forward biased, the DRB is no longer sink-like for excess carriers. Instead, the boundary value of P at the DRB can become large enough to influence the current. An extreme case is an open circuit condition in which there is always a forward current that exactly compensates for the reverse current. The text assumes that the power supply is maintaining a reverse-biasing condition.

X_S is closer to L or closer to X_S'. However, if the source is moved further from the contact, so it is now within a distance X_S' from the DRB, a different situation occurs. The carrier density between source and contact is no longer large throughout this region. Instead, this region divides into two sub-regions. The sub-region adjacent to the contact is characterized by the excess carrier density being nearly zero, so the conductivity is now controlled by the doping density. This sub-region was called the high-resistance region (HRR) in [3] because the conductivity is much less than in the adjacent sub-region where the carrier density is large. This adjacent sub-region was called the ambipolar region (AR) in [3]. Nearly all of the QNR voltage is across the HRR, so another characteristic of the HRR is a strong electric field. The intense electric field in the HRR adjacent to the contact prevents minority carriers from reaching the contact, so they are all driven to the DRB. This explains why Ω remains at a constant value of 1 when X_S is varied between 0 and X_S'. Also, moving the source further from the contact (closer to the DRB) causes the HRR to become wider. This increases the voltage across the QNR, so the absolute value of the QNR voltage increases with decreasing X_S when X_S is less than X_S'.

The previous paragraph gives a literal description of the relevant physics in the large-G limit, but a much simpler (and hypothetical) device mimics the actual case in the sense of producing the same end result (122). The single-event-effects community uses the term “sensitive volume” (SV) to describe a device region having the property that all charge liberated within is collection (i.e., the charge-collection efficiency is 1 within such a region). The same charge collection-efficiency function given by (122) would also be produced by a hypothetical device containing a SV adjacent to the DRB and having a width X_S'.[13] The linear behavior of Ω versus source location X_S, when X_S is outside the SV, is the same behavior that would be produced if charge collection from a source outside the SV was from pure diffusion from the source to the SV boundary. Therefore, the simpler physical picture that mimics the actual case in the large-G limit is one in which the device contains a SV that collects charge liberated within with a 100% efficiency, while charge liberated outside the SV is collected with an efficiency that is consistent with pure diffusion from the source to the SV boundary.

An unfortunate characteristic of such visualization models is the risk of interpreting them too literally and reaching erroneous conclusions. That charge collection can be described (in the large-G limit) in terms of a SV was derived for the case in which the carrier generation is from a localized source. If the SV model is interpreted literally, we might expect it to apply to an arbitrary spatial distribution of carrier generation. For a more specific example, consider two localized sources, one within the SV and one outside. If the SV model is interpreted as a literal description of charge-collection physics, charge liberated by the source within the volume might be expected to be collected with a 100% efficiency, while charge liberated outside might be expected to be collected with an efficiency that is consistent with diffusion. A physically correct interpretation is needed to predict what will really happen. In this more correct interpretation, we will not interpret the SV model literally, but the statement “a point is within the SV” can still be given a meaning by defining it to mean that the point is closer to the DRB than the point X_S'

[13] The SV width should not be confused with the width of the AR discussed earlier. The former depends only on device construction while the latter depends also on the location and strength of the source. If the source is outside the SV, there is no HRR so all of the QNR$^{(b)}$ is the AR. If the source is inside the SV, the QNR$^{(b)}$ divides into an HRR and AR but the widths of these regions vary continuously as the source location is varied. The depth of the source relative to the DRB, the SV width, and the AR width are three distinct quantities.

given by (122b). First consider the case in which only the generation site within the SV is present. The physically correct interpretation notes that carrier generation from this site leads to a strong electric field near the contact (i.e., the formation of an HRR). This prevents the liberated minority carriers from reaching the contact; hence, they all move to the DRB, producing a 100% collection efficiency. Let us now add the other point source, which is outside the SV. To avoid the necessity of analyzing synergetic effects, assume that the other source is much weaker so that it has only a small affect on the electric field, so the electric field with both sources present is essentially the same as the field just described. This prevents minority carriers liberated by either source from reaching the contact; hence, they all move to the DRB, producing a 100% collection efficiency. In other words, the presence of the generation site within the SV results in a 100% collection efficiency for the other generation site, even though this other site is outside the SV. This is contrary to what might be expected from a literal interpretation of the SV model.

An alternate model also predicts the upper curves in any of the Figures 10 through 16, but is a more literal description of charge-collection physics than the SV model. This is ambipolar diffusion with a cutoff. The derivation of this model can be shown to be mathematically equivalent to the derivation already used to obtain (111) (e.g., neglecting logarithmic terms in the former derivation is equivalent to an ambipolar approximation); however, here we use different terminology, so that physical interpretations become more obvious. The intention here is not to be mathematically rigorous by quantifying the accuracy of various approximations. This is not necessary because the conclusion — that $\Omega \to \Omega^*$ in the large-G limit — has already been established when constructing the plots in Figures 10 through 16. The intention is to merely identify what the physical approximations are that will produce this conclusion. We start here with an arbitrary generation function g and later specialize to a localized source. The drift-diffusion equations (29) still apply, except that the currents are now functions of x. Adding equations and using (35b) to express the result in the K notation gives

$$K_m(x) - K_M(x) = \frac{dP(x)}{dx} + \frac{p_0 - n_0}{2V_T}\frac{dU(x)}{dx} \qquad for \quad 0 < x < L\,, \tag{123}$$

where the QNR is taken to be between $x = 0$ and $x = L$. Also, a linear combination of the equations in (103), when expressed in the K notation is

$$\frac{d[K_m(x) - K_M(x)]}{dx} = -\frac{g(x)}{D^*} \qquad for \quad 0 < x < L\,, \tag{124}$$

where the ambipolar diffusion coefficient D^* is defined by

$$D^* \equiv \frac{2D_m D_M}{D_m + D_M}\,. \tag{125}$$

The majority-carrier current is blocked at the DRB, so the total current, denoted J_T, is equal to $J_m(0)$. In the K notation, the absolute value of the total current is given by $|J_T| = 2qD_mK_m(0)$, so (123) gives

$$\left|J_T\right| = 2qD_m\left[\frac{dP(\xi)}{d\xi} + \frac{p_0 - n_0}{2V_T}\frac{dU(\xi)}{d\xi}\right]_{\xi=0}. \tag{126}$$

Unlike the individual currents in (29), the combination of currents given by (123) does not have a large coefficient P (which is large under HILC) multiplying the dU/dx term on the right, suggesting the approximation of omitting this term. Omitting this term in (123) while using (124) gives

$$\frac{d^2P(x)}{dx^2} \approx -\frac{g(x)}{D^*} \qquad for \quad 0 < x < L. \tag{127a}$$

Also, (126) reduces to

$$\left|J_T\right| \approx 2qD_m\left[\frac{dP(\xi)}{d\xi}\right]_{\xi=0}. \tag{127b}$$

We will call (127) the "ambipolar diffusion approximation." An equivalent statement of this approximation is $P \approx P^*$ and $|J_T| \approx J_T^*$, where P^* is defined by the ambipolar diffusion equation

$$\frac{d^2P^*(x)}{dx^2} = -\frac{g(x)}{D^*} \qquad for \quad 0 < x < L, \quad P^*(0) = P^*(L) = 0, \tag{128}$$

and J_T^* is defined by

$$J_T^* \equiv 2qD_m\left[\frac{dP^*(\xi)}{d\xi}\right]_{\xi=0}. \tag{129}$$

We now specialize to the case of a localized generation site at a location X_S and strength G. Solving (128) for this case gives

$$P^*(x) = x\frac{G}{D^*}\frac{L - X_S}{L} \qquad for \quad 0 \le x \le X_S,$$

$$P^*(x) = (L - x)\frac{G}{D^*}\frac{X_S}{L} \qquad for \quad X_S \le x \le L,$$

and (129) becomes

$$J_T^* = 2qD_m\frac{G}{D^*}\frac{L - X_S}{L} = qG\left(\frac{D_m}{D_M} + 1\right)\frac{L - X_S}{L}. \tag{130}$$

This is the ambipolar approximation for the current without a cutoff. However, depending on the source location X_S and on the ratio D_m/D_M, it is possible for this current to exceed qG. This

unphysical result implies that the omission of the electric field (the *dU*/*dx* term) from (123) is invalid under these conditions. This, in turn, is interpreted as an indication that an HRR has formed adjacent to the contact, with a strong electric field forcing all minority carriers to move to the DRB. In other words, if the ambipolar diffusion approximation for the current exceeds the carrier generation rate, the actual current is expected to be equal to the carrier generation rate. When using the ambipolar diffusion approximation to define a charge-collection efficiency function, denoted Ω*, we include a cutoff (i.e., we prevent it from exceeding 1) by defining it by

$$\Omega^*(X_S) \equiv \begin{cases} \dfrac{J_T^*}{qG} & \text{if} \quad \dfrac{J_T^*}{qG} \le 1 \\ 1 & \text{if} \quad \dfrac{J_T^*}{qG} > 1. \end{cases} \tag{131}$$

Using (130), we find that the definition (131) is the same as (122), which was already shown to be the correct result in the large-*G* limit.

7. Conclusions

This paper treated two subjects. The first, treated in Sections 1 through 5, was a mathematical investigation of the boundary-value problem (1) for the purpose of obtaining some mathematical tools. These tools consist of the T-functions and E-functions defined in Appendix A. It was shown that the uniform solution (defined to be a solution that has no zeros or sign changes) to (1) is given by (10) if $a = 0$ and (9) is satisfied, or given by (12) if $b = 0$ and (11) is satisfied, or given by (26) if (25) is satisfied. These tools will be useful for any scientific or engineering application that encounters (1).

The second main subject of this paper, treated in Section 6, used the above tools for a specific application. This is the physical problem of drift-diffusion of charge carriers in a quasi-neutral portion (a QNR) of a semiconductor device. The assumed governing equations that define the physical problem were simplified by using a number of approximations (constant mobilities, constant diffusion coefficients related to mobilities by the Einstein relation, no carrier recombination in the interior of the QNR, and ideal boundary conditions); however, conclusions were rigorously derived for the hypothetical device that is defined by these simplified governing equations. Some of these conclusions are:

a) *Analytical Approximations:* Analytical approximations for the minority-carrier current were given for each of several arrangements in which the driving term was a boundary value. One of these applies to the forward-biased diode when operating conditions are such that K_m and K_M are both negative and $|K_m| \geq |K_M|$. The result is (58). Another applies to a reverse-biased diode with a carrier source. The result is (66). A third applies to a carrier source producing majority- and minority-carrier flow to a contact under Case 1 conditions ($K_m + K_M < 0$). The result is (81). Each approximation not only becomes exact in either of two limits (the low-injection level limit and the high-injection level limit), but also provides an interpolation between the two extremes having a relative error that is guaranteed to be less than 12% under the applicable conditions.

b) *Location of an Intense Electric Field:* Consider charge collection in a device consisting of a reverse-biased depletion region (DR) having a boundary (DRB, which is a sink for excess carriers) above a contact (another sink). The QNR is between these boundaries. Carriers are liberated (e.g., by photo-generation) at a generation site within the QNR. Because the drift-diffusion equations are nonlinear, there is a synergism between drift and diffusion. The analysis has shown that this produces a week electric field (always) at locations above the generation site (source) and a strong electric field (sometimes, depending on doping type, source strength and source location) at locations below the source. This was also seen in a computer simulation by Hsieh *et al.* [7] of a transient problem in which carriers were liberated by an ion traveling part way through the device, entering at the top, and stopping in the interior (the column of liberated carriers is called an “ion track”). The simulation produced a plot of equipotential surfaces within the device interior. Paying attention to the spacing between equipotential surfaces in the figure in [7], it is seen that there is a weak electric field along the track and a strong electric field below the track, as predicted by the theory given here. This recognition is contrary to conclusions in the older literature. Some investigators failed to recognize that the electric field is weakest (not strongest) at locations where the equipotential surfaces

are furthest apart, so some of the older literature (e.g., [8]) misinterpreted the Hsieh plot and concluded that an upper part of the device contains an intense electric field, making it, in effect, an extension of the DR. Although a strong-field extension of the DR is fictional, it was perceived by some investigators and given the name "funnel."[14] The more rigorous analysis given here shows that a strong-field extension of the DR is fictional.

c) *P-Type Produces a Greater Charge-Collection Efficiency than N-Type*: Consider the same arrangement as in Item (b). The theory shows that a p-type QNR produces a greater amount of collected charge compared to the n-type. A casual explanation — that minority carriers are more mobile in p-type compared to n-type — is naive because it fails to recognize that, in a pure diffusion process, charge collection is determined by the location of the source relative to the sink boundaries and is independent of the diffusion coefficient. Also, explanations in the older literature (e.g., in [8]) have no grasp of the relevant physics and were based on a misinterpretation of computer simulation results (see Item (b) above) in addition to other misconceptions (e.g., that an ion track is nonconductive and that an electric current requires a charge separation). In contrast, the prediction given here is rigorously derived from analysis of the nonlinear drift-diffusion equations.

d) *P-Type Exhibits a Stronger Electric Field than N-Type*: Considering the same arrangement as in Items (b) and (c), the analysis has shown that the electric field at locations below the source tends to be stronger in the p-type QNR compared to the n-type. This is consistent with Item (c) above. The stronger electric field in the p-type produces a stronger opposition to the downward flow of minority carriers and forces more of them to move to the DRB above.

e) *Plots of Charge-Collection Efficiency*: A photodiode exposed to an ionizing environment, which creates a source of free electron-hole pairs within the QNR, was considered, and a charge-collection efficiency Ω was defined to be the collected charge (produced by a minority-carrier flow from the QNR into the DR) divided by the amount of charge liberated by the source. The analysis considered a localized source, which makes Ω dependent on source location in addition to source strength and parameters describing device construction. The original governing equations contain seven parameters, consisting of doping type (n-type versus p-type), doping density N, source location X_S (measured from the DRB), source strength G, length (or depth) L of the QNR, electron diffusion coefficient, and hole diffusion coefficient. The analysis has shown that these seven parameters can be lumped into just three parameters, consisting of X_S/L, GL/D_mN, and D_m/D_M (where D_m is the minority carrier diffusion coefficient and D_M is the majority-carrier diffusion coefficient [note that the diffusion coefficient ratio is also a mobility ratio according to the Einstein relation]) that must be specified in order to uniquely determine Ω. This means that all possible numerical examples of practical interest can be

[14] Occasionally a paper in the literature will show an equipotential surface, plotted by a computer simulation, that the author has identified as a funnel because it appears to be isolated from other equipotential surfaces shown in the plot, without realizing that the choice is arbitrary because the set of plotted surfaces is an artifact of the contour values that were selected for plotting. Pictures of funnels that are most frequently seen in the literature are in cartoons and are an artist conception.

represented (via interpolations) by a finite number of plots of Ω as a function of X_S/L. Such plots are included in this report (Figures 10 through 16).[15] There is a caveat. The analysis requires the DR to be reverse-biased. There are situations in which there can be a large voltage (several volts) across the QNR. It is, therefore, not enough for the external power supply to have the correct polarity; it must also have enough strength to supply the QNR voltage, with this voltage calculated from (109b), and still have enough voltage left to reverse bias the DR. Otherwise the DR will become forward biased and produce a forward current that competes with the reverse current associated with the carrier generation source. An extreme case is an open circuit condition in which there is always a forward current that exactly compensates for the reverse current. As long as the power supply voltage maintains a reverse-biasing condition, the power supply voltage becomes irrelevant, except for its effect on DR width, which, in turn, affects the QNR length. This issue was avoided in the analysis by measuring depth from the DRB and taking the QNR length L as given. Note that the plots represent a one-dimensional device (1D), but a 3D problem can be converted into a 1D problem using concepts in Appendix B (details are in [3]). This allows 1D results to be applied to 3D problems.

f) *Sensitive Volumes*: The single-event-effects community uses the term "sensitive volume" (SV) to describe a device region having the property that all charge liberated within is collection (i.e., the charge-collection efficiency is 1 within such a region). Funnels perceived in the literature have been interpreted not only as strong-field extensions of the DR, but also as SVs. While the former interpretation is fictional, the latter interpretation has a basis in reality in the sense that a SV is created within the QNR when the carrier generation rate G is sufficiently large. This is shown by the upper curves in any of the Figures 10 through 16. The charge-collection efficiency Ω is virtually 1 down to a critical depth X_S' given by (122b), with this depth defining the SV. Note that D_m/D_M is greater than 1 for a p-type material and less than 1 for n-type; hence, (122b) shows that the p-type material has the larger SV. The linear behavior of Ω versus source location X_S, when X_S is below this critical depth, is the same behavior that would be produced if charge collection from a source below this depth was from pure diffusion from the source to the lower SV boundary. In summary, charge liberated within the SV is collected with a virtually 100% efficiency, while charge liberated below the SV is collected with an efficiency that is consistent with pure diffusion from the source to the lower SV boundary. The same caveat pointed out in Item (e) regarding biasing conditions also applies to this discussion. Note that some earlier and less rigorous attempts to estimate a SV depth (e.g., in [9]) estimate the depth to be a multiple of the DR depth and independent of the QNR depth. In contrast, the more rigorous analysis given here finds that the SV depth is a fraction of the QNR depth[16] and independent of the DR depth. However, it is important to be aware that the SV model is a symbolic model and is

[15] The range of values $1/3 \le D_m/D_M \le 3$ is expected to include all cases of practical interest for silicon devices. If another material is considered, in which the diffusion-coefficient ratio goes outside this range, readers can use the same algorithms explained in detail in Section 6.5 to construct additional plots.

[16] Recall that the analysis given here is one-dimensional. A three-dimensional version would replace the normalized rectangular coordinate X_S/L, which is normalized to make the QNR depth equal to 1, with a different (curvilinear) coordinate [3] that measures distance from the DRB and is normalized to make the QNR depth equal to 1. The statement that the sensitive volume depth is a definite fraction of the QNR depth would use this curvilinear coordinate to define the depths.

limited to generation from a localized source. This limitation is discussed in more detail in Section 6.5.5.

g) *Ambipolar Diffusion with a Cutoff*: An alternate model also predicts the upper curves in any of the Figures 10 through 16, but is a more literal description of charge-collection physics than the SV model. This is ambipolar diffusion with a cutoff. It is derived by starting with the linear combination of the drift-diffusion equations that is least sensitive (in the large-G regime) to errors produced by neglecting the electric field in the QNR. Omitting the electric field in this equation produces the ambipolar diffusion approximation (127), which does not yet have a cutoff. There are conditions under which the ambipolar diffusion approximation for the current exceeds the carrier generation rate. This implies that the omission of the QNR electric field is invalid. This, in turn, is interpreted as an indication that an intense electric field has formed adjacent to the contact, forcing all minority carriers to move to the DRB. In other words, if the ambipolar diffusion approximation for the current exceeds the carrier generation rate, the actual current is expected to be equal to the carrier generation rate. The final approximation for the current, called “ambipolar diffusion with a cutoff,” is equal to the ambipolar diffusion approximation (without the cutoff) when this is less than the carrier generation rate and equal to the carrier generation rate when the ambipolar diffusion approximation (without the cutoff) exceeds the carrier generation rate. This model predicts the same charge-collection efficiency function, given by (122), that was predicted by the SV model and shown as the upper curves in any of the Figures 10 through 16. The same caveat pointed out in Item (e) regarding biasing conditions also applies to this discussion.

Appendix A: Definitions and Properties of Three Special Functions

Three functions defined below will be seen (later) to satisfy a differential equation relevant to this work. These functions are defined to be inverses of three other functions that are easier to define, so we start by defining these other functions. Starting with the function T defined by

$$T(\xi) \equiv \xi - \ln|\xi + 1|, \qquad \xi \neq -1 \tag{A1}$$

and plotted in Figure A1, we partition the domain to obtain the three functions T_1, T_2, and T_3 with domains and ranges given by

$$T_1 : [0, \infty) \leftrightarrow [0, \infty), \quad T_2 : (-1, 0] \leftrightarrow [0, \infty), \quad T_3 : (-\infty, -1) \leftrightarrow (-\infty, \infty) \tag{A2}$$

and with the mapping rules given by

$$T_1(\xi) = \xi - \ln(\xi + 1), \qquad T_1(\xi) \geq 0, \qquad \xi \geq 0 \tag{A3a}$$

$$T_2(\xi) = \xi - \ln(\xi + 1), \qquad T_2(\xi) \geq 0, \qquad -1 < \xi \leq 0 \tag{A3b}$$

$$T_3(\xi) = \xi - \ln(-\xi - 1), \qquad -\infty < T_3(\xi) < \infty, \qquad \xi < -1. \tag{A3c}$$

Note that each function in (A3) is defined when ξ is in the indicated interval. Also, differentiating shows that T_1 is strictly increasing, T_2 is strictly decreasing, and T_3 is strictly increasing on their respective domains. Therefore, each function has an inverse. The inverses are denoted E_1, E_2, and E_3. They satisfy

$$E_1 : [0, \infty) \leftrightarrow [0, \infty), \quad E_1 = T_1^{-1}$$

$$E_2 : [0, \infty) \leftrightarrow (-1, 0], \quad E_2 = T_2^{-1}$$

$$E_3 : (-\infty, \infty) \leftrightarrow (-\infty, -1), \quad E_3 = T_3^{-1}$$

with mapping rules given by

$$x = E_1(x) - \ln(E_1(x) + 1), \qquad x \geq 0, \qquad E_1(x) \geq 0. \tag{A4a}$$

$$x = E_2(x) - \ln(E_2(x) + 1), \qquad x \geq 0, \qquad -1 < E_2(x) \leq 0 \tag{A4b}$$

$$x = E_3(x) - \ln(-E_3(x) - 1), \qquad -\infty < x < \infty, \qquad E_3(x) < -1. \tag{A4c}$$

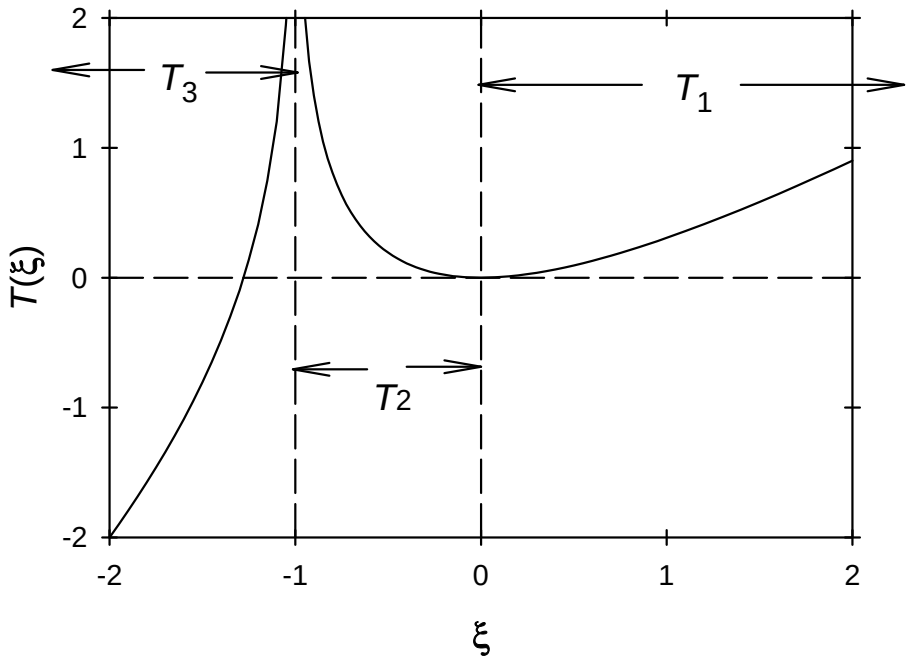

Figure A1. The Function *T*. The function $T(\xi) \equiv \xi - \ln|\xi+1|$ is defined for all $\xi \neq -1$, but only a finite region of the plot is shown. Outside the plotted region; $T(\xi) \to +\infty$ as $\xi \to -1$, $T(\xi) \to +\infty$ as $\xi \to +\infty$, and $T(\xi) \to -\infty$ as $\xi \to -\infty$. By partitioning the domain as shown in the plot, we construct three invertible functions; T_1 with domain $[0, +\infty)$, T_2 with domain $(-1, 0]$, and T_3 with domain $(-\infty, -1)$.

These mapping rules state that T_i is a left inverse of E_i (for $i = 1, 2, 3$). Other mapping rules, which state that T_i is a right inverse of E_i (for $i = 1, 2, 3$), are

$$\xi = E_1\left(\xi - \ln\left(\xi + 1\right)\right) \quad for \quad \xi \geq 0 \tag{A4d}$$

$$\xi = E_2\left(\xi - \ln\left(\xi + 1\right)\right) \quad for \quad -1 < \xi \leq 0 \tag{A4e}$$

$$\xi = E_3\left(\xi - \ln\left(-\xi - 1\right)\right) \quad for \quad \xi < -1. \tag{A4f}$$

Differential equations satisfied by these functions can be derived by implicitly differentiating the above equations to obtain

$$E_1(x)\frac{dE_1(x)}{dx} = E_1(x) + 1 \quad for \quad x > 0 \tag{A5a}$$

$$E_2(x)\frac{dE_2(x)}{dx} = E_2(x) + 1 \quad for \quad x > 0 \tag{A5b}$$

$$E_3(x)\frac{dE_3(x)}{dx} = E_3(x) + 1 \quad for \quad -\infty < x < +\infty. \tag{A5c}$$

In other words, each E-function is a solution to a common differential equation. Note that translations of these functions are also solutions. For example, if c is an arbitrary constant then $E_1(x+c)$ is a solution for any $x > -c$.

A plot of each E-function can be constructed by working backwards. For example, to construct a plot of $E_1(x)$ versus x, we select a non-negative value for $E_1(x)$ and then use (A4a) to solve for x. This procedure produces the plots shown in Figures A2, A3, and A4. The same procedure used to construct plots can also find special values or limiting values that are given by

$$E_1(0) = 0 \tag{A6a}$$

$$E_1(x) \text{ increases without bound as } x \text{ increases without bound} \tag{A6b}$$

$$E_2(0) = 0 \tag{A6c}$$

$$\lim_{x \to +\infty} E_2(x) = -1 \tag{A6d}$$

$$E_3(x) \text{ decreases without bound as } x \text{ decreases without bound} \tag{A6e}$$

$$E_3(-2) = -2 \tag{A6f}$$

$$\lim_{x \to +\infty} E_3(x) = -1. \tag{A6g}$$

Substituting (A6f) into (A5c) gives

$$\left.\frac{dE_3(x)}{dx}\right|_{x=-2} = \frac{1}{2}. \tag{A6h}$$

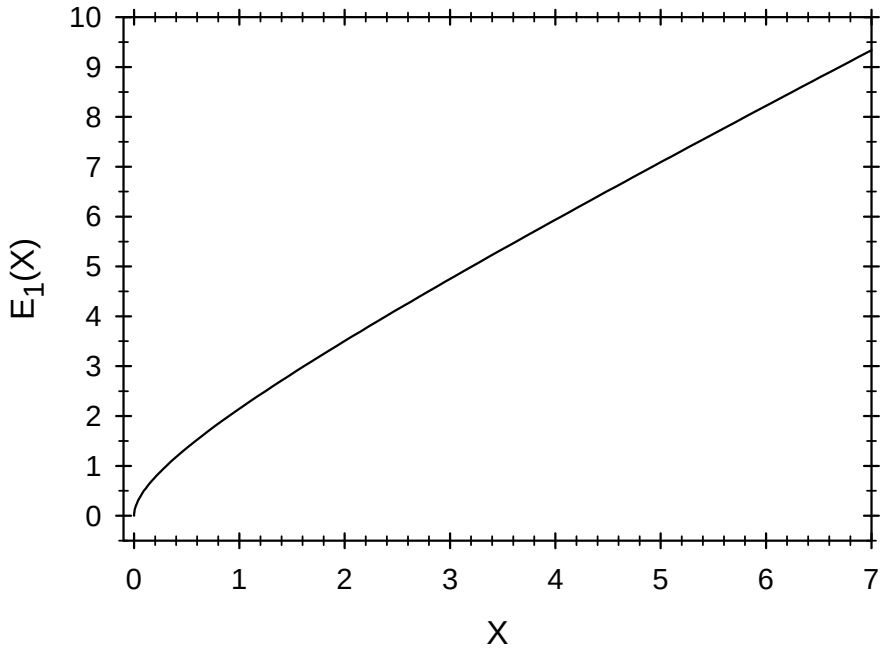

Figure A2. The Function E_1. This is the inverse of T_1 shown in Figure A1.

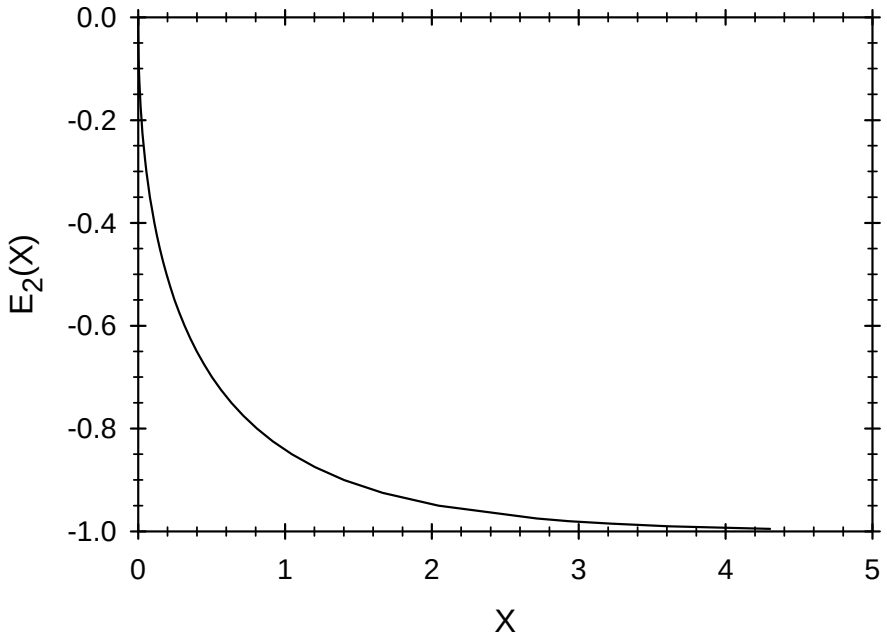

Figure A3. The Function E_2. This is the inverse of T_2 shown in Figure A1.

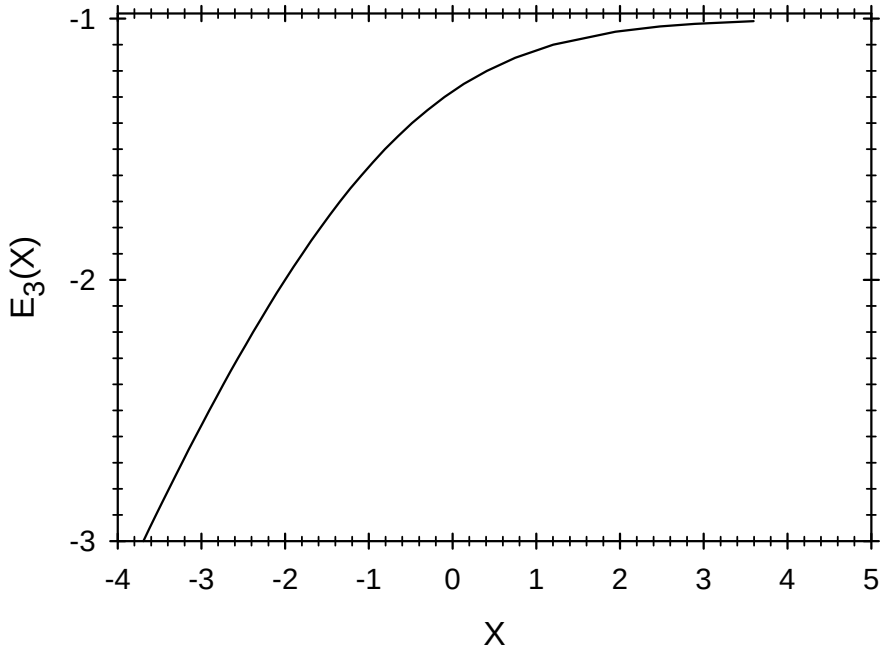

Figure A4. The Function E_3. This is the inverse of T_3 shown in Figure A1.

The simple limits in (A6) can be combined with (A4) to derive other limits that describe the E-functions with greater resolution. For example, consider the expression $E_1(x) - x - \ln(1+x)$, which is defined for any $x \geq 0$. Using (A4a) to substitute for x, we can write this expression as

$$E_1(x) - x - \ln(x+1) = \ln\left[\frac{1+E_1(x)}{1+E_1(x)-\ln\left(1+E_1(x)\right)}\right].$$

The limit as x increases without bound (if the limit exists) is given by

$$\lim_{x\to+\infty}\left\{E_1(x) - x - \ln(x+1)\right\} = \lim_{x\to+\infty}\left\{\ln\left[\frac{1+E_1(x)}{1+E_1(x)-\ln\left(1+E_1(x)\right)}\right]\right\}.$$

But (A6b) implies that the large-x limit on the right is also the large-E_1 limit, so the equation can be written as

$$\lim_{x\to+\infty}\left\{E_1(x) - x - \ln(x+1)\right\} = \lim_{\xi\to+\infty}\left\{\ln\left[\frac{1+\xi}{1+\xi-\ln\left(1+\xi\right)}\right]\right\}.$$

The logarithm function is continuous at any positive argument, so the limit of the logarithm is the logarithm evaluated at the limiting value of the argument (which is positive). This gives

$$\lim_{x\to+\infty}\left\{E_1(x) - x - \ln(x+1)\right\} = 0. \qquad \text{(A7)}$$

The symbolism "$\to$" will be used to denote limits. The precise definition of the notation is

$$f(x) \to g(x) \quad \text{as} \quad x \to a \quad \text{if and only if} \quad \lim_{x \to a}\{f(x) - g(x)\} = 0. \tag{A8}$$

An analogous definition applies to the case where x approaches $+\infty$ or $-\infty$. The right side is intentionally written as the limit of a difference instead of a difference between limits so that the notation will still make sense even if f and g do not have limits. The only requirement for the notation to make sense is that the difference $f-g$ have a limit. In the notation of (A8), we can write (A7) as

$$E_1(x) \to x + \ln(x+1) \quad \text{as} \quad x \to +\infty. \tag{A9}$$

Similar steps give

$$E_3(x) \to x + \ln(|x+1|) \quad \text{as} \quad x \to -\infty. \tag{A10}$$

When considering limits in which an E-function approaches zero, the behavior is described with better resolution when described in terms of asymptotic behavior instead of the more conventional limit. The symbol "$\xrightarrow{A}$" will be used to denote asymptotic behavior, which refers to relative or fractional differences between two functions. The statement that some function $f(x)$ asymptotically approaches some function $g(x)$ as x approaches some value a is denoted $f(x) \xrightarrow{A} g(x)$ as $x \to a$ and defined by

$$f(x) \xrightarrow{A} g(x) \quad \text{as} \quad x \to a \quad \text{if and only if} \quad \lim_{x \to a}\left[\frac{f(x) - g(x)}{g(x)}\right] = 0.$$

An equivalent definition is

$$f(x) \xrightarrow{A} g(x) \quad \text{as} \quad x \to a \quad \text{if and only if} \quad \lim_{x \to a}\left[\frac{f(x)}{g(x)}\right] = 1. \tag{A11}$$

An analogous definition applies to the case where x approaches $+\infty$ or $-\infty$. The asymptotic behavior of $E_1(x)$ as $x \to 0$ can be deduced from (A4a) after some inequalities have been derived. By differentiating the expression $\ln(1+\xi) - \xi + \xi^2/2$ with respect to ξ, we find that the derivative is positive for $\xi > -1$, so the expression is increasing. Therefore, the expression is larger at $\xi > 0$ then at $\xi = 0$, and it is smaller at $\xi > 0$ then at $\xi = 0$. That is,

$$\ln(1+\xi) \geq \xi - \frac{1}{2}\xi^2 \quad \textit{when} \quad \xi \geq 0 \tag{A12a}$$

$$\ln(1+\xi) \leq \xi - \frac{1}{2}\xi^2 \quad \textit{when} \quad 0 \geq \xi > -1. \tag{A12b}$$

Similar steps give

$$\ln(1+\xi) \le \xi - \frac{1}{2}\frac{\xi^2}{\xi+1} \quad when \quad \xi \ge 0 \tag{A12c}$$

$$\ln(1+\xi) \ge \xi - \frac{1}{2}\frac{\xi^2}{\xi+1} \quad when \quad 0 \ge \xi > -1. \tag{A12d}$$

Using $E_1(x) \ge 0$, the applicable inequalities in (A12) are

$$E_1(x) - \frac{1}{2}\frac{{E_1}^2(x)}{E_1(x)+1} \ge \ln\left(E_1(x)+1\right) \ge E_1(x) - \frac{1}{2}{E_1}^2(x).$$

Combining with (A4a) gives

$$1 \ge \frac{2x}{{E_1}^2(x)} \ge \frac{1}{E_1(x)+1}.$$

As $x \to 0$, so $E_1(x) \to 0$. The far right and far left sides of the above inequality come together and we conclude that

$$\lim_{x \to 0+}\left[\frac{2x}{{E_1}^2(x)}\right] = 1,$$

where the limit is a one-sided limit (x approaches zero from above) because E_1 is not defined on negative arguments. Neither x nor $E_1(x)$ are negative, so the above limit can also be written as

$$\lim_{x \to 0+}\left[\frac{E_1(x)}{\sqrt{2x}}\right] = 1.$$

That is,

$$E_1(x) \xrightarrow{A} \sqrt{2x} \qquad as \quad x \to 0^+. \tag{A13}$$

Similar steps give

$$E_2(x) \xrightarrow{A} -\sqrt{2x} \qquad as \quad x \to 0^+. \tag{A14}$$

When considering limits in which an E-function approaches -1, the behavior is described with the greatest resolution when described in terms of the asymptotic behavior of 1 plus the E-function. The asymptotic behavior of $1+E_2(x)$ as $x \to +\infty$ can be deduced by considering the expression $[1+E_2(x)]/\exp(-(1+x))$, which is defined for any $x \ge 0$. Note that (A4b) can be written as

$$\frac{1+E_2(x)}{\exp(-(1+x))} = \exp\big(1+E_2(x)\big).$$

As x increases without bound, $E_2(x)$ approaches -1 and the right side approaches 1. That is,

$$\lim_{x\to+\infty}\left\{\frac{1+E_2(x)}{\exp(-(1+x))}\right\} = 1,$$

which can also be written as

$$1+E_2(x) \xrightarrow{A} \exp(-(1+x)) \qquad as \quad x\to+\infty. \tag{A15}$$

Similar steps give

$$1+E_3(x) \xrightarrow{A} -\exp(-(1+x)) \qquad as \quad x\to+\infty. \tag{A16}$$

Appendix B: Extending the QNR Analysis to Three Dimensions

A three-dimensional version of the problem in Section 6 can be reduced to a one-dimensional problem by using the method in [3]. The trick is to define a generalized coordinate $\Phi(x,y,z)$ to be a solution to Laplace's equation with suitable boundary conditions. Laplace's equation is easier to solve (at least approximately) than most other equations, so Φ is considered to be a known function of the spatial coordinates. The three-dimensional problem expresses potential and carrier density as functions of Φ in the same way that the one-dimensional problem expresses them in terms of x. In the three-dimensional problem, like the one-dimensional problem, carrier density and potential become functions of a single coordinate, but the coordinate is now Φ instead of *x*. Solving for the carrier density and potential in terms of Φ leads to the same equations that must be solved in the one-dimensional problem.

Appendix C: Limits Applicable To Case 2 in Section 6.4.2

The equation considered is (86) subject to the constraints

$$K_m < 0, \quad K_m + K_M > 0, \quad P_S > 0. \tag{C1a}$$

Some implications from (C1a) together with (86) were already derived in the main text and are

$$K_M W > \frac{P_S}{2} \tag{C1b}$$

$$-\frac{P_S}{2W} \le K_m < 0\,. \tag{C1c}$$

A positive W and a positive N are regarded as given, so (86) is regarded as an equation that relates three parameters: P_S, K_m, and K_M. Of these three parameters, we can select any pair to be called the independent variables and regard (86) as an equation that implicitly solves for the third (dependent) variable. Various types of limits can be considered. For example, we can take the limit as the point (K_m, K_M) approaches a selected point by moving along a selected path in a plane and use (86) to define the limiting value (if it exists) of P_S. Alternately, we can take the limit as K_M approaches a selected value with P_S fixed and use (86) to define the limiting value (if it exists) of K_m. We will start with the first example. It is convenient to shorten the notation by defining α by

$$\alpha \equiv \frac{K_M - K_m}{K_M + K_m}\,. \tag{C2}$$

Note that (C1) and (C2) imply that

$$\alpha > 1\,. \tag{C3}$$

Using (C2), we can write (86) as

$$1 + \frac{2P_S}{N} = -\frac{1}{\alpha} E_3\left(-\alpha - \ln\left(\frac{-2K_m}{K_M + K_m} \right) - 2\alpha\,(K_M - K_m)\,\frac{W}{N} \right). \tag{C4}$$

First consider the limit as K_m approaches zero from below with a fixed $K_M > 0$. We see from (C2) that $\alpha \to 1$, so the only term in the argument of E_3 in (C4) that becomes singular is the logarithmic term; the result is that the argument increases without bound, so E_3 evaluated at this argument approaches -1. We conclude from (C4) that

$$P_S \to 0 \quad as \quad K_m \to 0 \ \textit{from below with fixed}\ K_M > 0. \tag{C5}$$

Note that the point (K_m, K_M) could also follow some other paths and still produce $P_S \to 0$. For example, we could consider a path in which K_m and K_M go to zero together. The limiting value of

α would be unique to the selected path (if the limit exists); therefore, the derivation given here would not apply, but the conclusion that $P_S \to 0$ is still correct because it is implied by (C1b).

Now consider the limit as the point (K_m, K_M) moves along a path for which $K_M \to +\infty$ while K_m is bounded below zero (i.e., there is an $\varepsilon > 0$ such that $K_m \le -\varepsilon$). This is another case in which $\alpha \to 1$. Also, the logarithmic term in the argument of E_3 in (C4) has a singularity as it did in the previous limit considered. However, there is now a competing singularity, a first-order singularity, in the last term in the argument of E_3 in (C4). As long as K_m is bounded below zero, the logarithmic term is a logarithmic singularity. The first-order singularity wins, so the entire argument of E_3 in (C4) goes to $-\infty$. The E_3 function evaluated at this argument also goes to $-\infty$, and we conclude from (C4) that

$$P_S \to +\infty \quad as \quad K_M \to +\infty \; with \; K_m \; bounded \; below \; 0. \tag{C6}$$

We can now use (C6) in a proof by contradiction to show that

$$K_m \to 0 \quad as \quad K_M \to +\infty \; with \; fixed \; P_S > 0. \tag{C7}$$

To prove (C7), let $K_M \to +\infty$ with fixed $P_S > 0$. Assume (for a contradiction) that K_m is bounded below zero. According to (C6), P_S increases without bound, contradicting the given condition that P_S is fixed. Therefore, the assumption that K_m is bounded below zero is false, which proves (C7). Other limits were already proven in the main text and are

$$K_m \to 0 \quad as \quad P_S \to 0 \; from \; above \; with \; fixed \; K_M > 0 \tag{C8}$$

$$K_m \to -\frac{P_S}{2W} \quad as \quad K_M \to \frac{P_S}{2W} \; from \; above \; with \; fixed \; P_S > 0. \tag{C9}$$

Having derived some simple limits, we can use these to derive several asymptotic limits. The first result is obtained by noting that (86) is equivalent to (87). This can be written, using (C2), as

$$\ln\left(1 + \frac{K_m - K_M}{K_m}\frac{P_S}{N}\right) = 2\left[\left(K_M - K_m\right)\frac{W}{N} - \frac{P_S}{N}\right]\alpha. \tag{C10}$$

Taking the exponential function of both sides and rearranging terms gives

$$K_m = \left[K_m + \left(K_m - K_M\right)\frac{P_S}{N}\right]\exp\left\{2\left[\left(K_m - K_M\right)\frac{W}{N} + \frac{P_S}{N}\right]\alpha\right\}. \tag{C11}$$

Now consider the limit as $K_M \to +\infty$ with fixed $P_S > 0$. From (C7), we have $K_m \to 0$ and (C2) gives $\alpha \to 1$. Using these facts with the definition (A11) (in Appendix A) of an asymptotic limit, it is easy to show that (C11) gives

$$K_m \xrightarrow{A} -K_M \frac{P_S}{N}\exp\left(\frac{2P_S}{N} - \frac{2K_M W}{N}\right) \quad as \quad K_M \to +\infty \; with \; fixed \; P_S > 0. \tag{C12}$$

The second result is obtained by performing another rearrangement of terms on (C11) to get

$$K_m = (K_M - K_m)\frac{P_S}{N}\frac{\exp\left\{2\left[(K_m - K_M)\frac{W}{N} + \frac{P_S}{N}\right]\alpha\right\}}{\exp\left\{2\left[(K_m - K_M)\frac{W}{N} + \frac{P_S}{N}\right]\alpha\right\} - 1}. \tag{C13}$$

Now consider the limit as $P_S \to 0$ with fixed $K_M > 0$. From (C8) we have $K_m \to 0$ and (C2) gives $\alpha \to 1$. Using these facts with the definition (A11) (in Appendix A) of an asymptotic limit, it is easy to show that (C13) gives

$$K_m \xrightarrow{A} K_M \frac{P_S}{N}\frac{\exp\left(-\frac{2K_M W}{N}\right)}{\exp\left(-\frac{2K_M W}{N}\right) - 1} = -K_M \frac{P_S}{N}\frac{1}{\exp\left(\frac{2K_M W}{N}\right) - 1}$$

$$as \quad P_S \to 0 \ \textit{from above with fixed} \ K_M > 0. \tag{C14}$$

Note that the expression on the right side of (93b) in the main text asymptotically approaches the right side of (C12) in the limit indicated in (C12) and asymptotically approaches the right side of (C14) in the limit indicated in (C14). Therefore, the same expression in (93b) applies to both limits. Another inequality needed in the main text is obtained by combining (C1a) with (C13) to conclude that the exponential function in (C13) is less than 1, so the curly bracket in (C13) is negative. But α is positive, so the square bracket in (C13) is negative. This gives

$$K_M - K_m > \frac{P_S}{W}. \tag{C15}$$

The third asymptotic limit is obtained by going back to (C10) and using (C2) to eliminate α and then rearranging terms to get

$$K_m + \frac{P_S}{2W} = \frac{\frac{2W}{N}(K_M - K_m) - \ln\left(1 + \frac{K_m - K_M}{K_m}\frac{P_S}{N}\right)}{\frac{2W}{N}(K_M - K_m) + \ln\left(1 + \frac{K_m - K_M}{K_m}\frac{P_S}{N}\right)}\left(K_M - \frac{P_S}{2W}\right). \tag{C16}$$

Now consider the limit as $K_M \to P_S/2W$ from above with fixed $P_S > 0$. From (C9), we have $K_m \to -P_S/2W$. Using this fact with the definition (A11) (in Appendix A) of an asymptotic limit, it is easy to show that (C16) gives

$$K_m + \frac{P_S}{2W} \xrightarrow{A} \frac{\frac{2P_S}{N} - \ln\left(1+\frac{2P_S}{N}\right)}{\frac{2P_S}{N} + \ln\left(1+\frac{2P_S}{N}\right)}\left(K_M - \frac{P_S}{2W}\right)$$

$$as \quad K_M \to \frac{P_S}{2W} \ from\ above\ with\ fixed\ P_S > 0. \tag{C17}$$

REFERENCES

[1] L.D. Edmonds, "High-level injection in n^+-p junction silicon devices," *J. Appl. Phys.*, vol. 97, p. 124506, 2005.

[2] S.M. Sze, *Physics of Semiconductor Devices*, 2nd ed., Wiley Interscience, New York, p. 141, 1981.

[3] L. Edmonds, *A Theoretical Analysis of Steady-State Photocurrents in Simple Silicon Diodes*, Jet Propulsion Laboratory Publication 95-10, March 1995.

[4] D. McMorrow, W.T. Lotshaw, J.S. Melinger, S. Buchner, and R.L. Pease, "Subbandgap laser-induced single event effects: carrier generation via two-photon absorption," *IEEE Trans. Nucl. Sci.*, vol. 49, pp. 3002-3008, Dec. 2002.

[5] L.D. Edmonds, "Proton SEU cross sections derived from heavy-ion test data," *IEEE Trans. Nucl. Sci.*, vol. 47, no. 5, pp. 1713-1728, Oct. 2000.

[6] L.D. Edmonds, "A method for correcting cosine-law errors in SEU test data," *IEEE Trans. Nucl. Sci.*, vol. 49, no. 3, pp. 1522-1538, June 2002.

[7] C.M. Hsieh, P.C. Murley, and R.R. O'Brien, "A field-funneling effect on the collection of alpha-particle-generated carriers in silicon devices," *IEEE Elect. Dev. Lett.*, vol. 2, no. 4, pp. 103-105, April 1981.

[8] F.B. McLean and T.R. Oldham, "Charge funneling in n- and p-type SI substrates," *IEEE Trans. Nucl. Sci.*, vol. 29, no. 6, pp. 2018-2023, Dec. 1982.

[9] L.D. Edmonds, "A simple estimate of funneling-assisted charge collection," *IEEE Trans. Nucl. Sci.*, vol. 38, no. 2, pp. 828-833, April 1991.