The “Story” of the Diode Lab

This lab is constructed as a kind of “journey of discovery” that is a bit like doing a “real” physics experiment.

We start out with a quantitative mathematical model of something. In this case, it’s a semiconductor PN junction.

A semiconductor junction has a built-in electric field from the doping gradient, which makes an energy barrier that electrons or holes have to jump to make a current flow. The probability of finding one on the other side of the energy barrier goes like \( \exp(\Delta E/kt) = \exp(qV/kt) \).

The expected current-voltage relationship is then

\[
I(V,T) = I_0(T) \times \left[ \exp\left(\frac{qV}{kT}\right) - 1 \right]
\]

We started out with a simple check at room temperature and a medium resistor value. A semi-log plot of the data gave the expected straight line for positive voltage.

So the theoretical exponential relation is basically true. But is it quantitatively true? We do have a definite prediction from the theory:

\[
V_0 = kT/q \approx 0.025V
\]

What did you find for \( V_0 \)?
What if the Data and Theory Disagree?

What you should NOT do (at least not in MY class!) is to simply call the difference “experimental error!”

First, do anything you can to check whether you did the measurement right. If the disagreement is big, usually this is the problem. Consistency checks, repeating the measurement, having someone else look at what you did.

Also, check whether you really understand the theory, and that it really applies to your situation. This can also cause big disagreements.

Next, estimate/calculate the error on your measurement. Maybe it really is consistent with the theory within the measurement error. (your error was << a factor of 2!)

Estimate the error on the theoretical expectation. Usually you need some measured input to the theoretical calculation (like the room temperature here), and sometimes that has a significant error (again, can’t explain a factor of 2 this way!)

Finally, you “check the literature” for similar measurements (i.e., ask your neighbors what they got for the slope, and for the theoretical expectation).

If other experimenters independently replicate your results, then you should conclude that the theory is wrong!
Improving the Diode Law

There is another process that is possible in semiconductors that we have not included in the theory yet.

It is possible for an electron to travel partway up the energy barrier from one side, and a hole to travel partway up the energy barrier from the other side.

Then the electron can fill up the hole and both just vanish! This is called recombination.

One unit of charge has moved from one side of the energy barrier to the other, so some current has flowed.

But it’s possible with each charge only going halfway up the potential barrier. The result is

\[ I = I_0 \times \left[ \exp\left(\frac{qV/2}{kT}\right) - 1 \right] = I_0 \times \left[ \exp\left(\frac{qV}{2kT}\right) - 1 \right] \]

So for this process we predict \( V_0 = 2kT/q \approx 0.050V \)
Improving the Diode Law (2)

In silicon diodes, the recombination process turns out to dominate, and we get \( V_0 = 2kT/q \approx 0.050V \).

But in other materials, like Germanium, the direct process dominates and we get \( V_0 = kT/q \approx 0.025V \).

In reality, both processes occur, so we need to add them up, with different \( I_0 \) constants in front.

When adding exponentials, usually one term dominates for part of the range, and the other dominates elsewhere.

We no longer have a straight line on the semi-log plot, and local slope is between the two extreme slopes.
Improving the Diode Law (3)

If we go up higher in voltage and current, we should see a transition between $\exp(qV/2kT)$ and $\exp(qV/kT)$.

Of course we can’t expect the current to increase exponentially forever. At some point the diode will start acting more like a resistor, with current growing only linearly with voltage.

There isn’t a clear prediction from the theory whether the resistor-like downturn will occur before or after the transition from $\exp(qV/2kT)$ to $\exp(qV/kT)$.

The 100K resistor value that we started with was chosen specifically to avoid showing you the region where these effects start to show up, to avoid confusing you with too many things at once.

The current is basically limited by the resistor once the diode starts up the exponential. But if we used a much smaller resistor, we could get to much larger currents.

That is part of the point of the second lab session.
Diodes at Negative Voltage

When the voltage is negative, the current should be independent of voltage with value $-I_0$, where this is the same $I_0$ as we get from the exponential part of the data.

Your data with the 100K resistor at negative voltage probably didn’t show the expected negative current.

But as we discussed last time, the value if $I_0$ is pretty small. When this flows through the 100K resistor, you expect a tiny voltage. So far we can’t really see if this is true, due to the offset and resolution of the digitizer voltage measurement.

Measuring the offset of the digitizer and correcting for helps a bit, but the intrinsic resolution of the voltage measurement doesn’t change.

The solution to this is to use a larger resistor to make the same current into a bigger voltage, so we can measure it accurately despite the digitizer resolution. The drawback: the maximum current and voltage that we can reach goes down. But if we combine data from different resistors, we can have the best of both worlds.

For real diodes, at sufficiently large negative voltage the current starts to increase dramatically due to “reverse avalanche” or “reverse breakdown” or “Zener effect.” We don’t apply enough reverse voltage to see this happen.
Distortion at Low Resistor Values

When you used low resistor values, you probably noticed that the triangular wave was distorted. This should have made you wonder if there was a problem.

It’s not a problem, it’s due to the 50 ohm resistor built into the output of the signal generator:

![Diagram of signal generator and diode connection](image)

When V1-V2 > 0.5 volts or so, the diode starts to act like a wire. So the total signal generator voltage drop is divided between circuit resistor and the 50 ohm internal resistor.

Only the relation between the “internal” signal generator voltage and the “external” signal generator voltage does change. And we don’t actually use this relationship.

The V1 and V2 voltages that you measure are still correct. And you don’t need to add 50 to the circuit resistor!
Distortion at High Resistor Values

There is “stray capacitance” in any real circuit. In this case, it’s partly in the coaxial cables, and partly in the inputs to the oscilloscope or digitizer.

To help understand what is going on, imagine replacing the diode by a switch, that is controlled by the sign of the current that tries to flow through it.

![Circuit Diagram](image)

When the signal generator voltage is rising, the diode conducts like a closed switch, and the capacitor charges up rapidly.

When the signal generator voltage is falling, the diode acts like an open switch, and stops conducting. So the only way the capacitor can discharge is though the resistor.
Distortion at High Resistor Values (2)

The time-constant is $R$ times $C$. The capacitance is fixed (and not very large), but if $R$ is big, the time to discharge can be comparable to the period of the triangle wave, and we get a distorted signal.

The reason the lab procedure said to set the frequency to 1 Hz is to avoid this problem.

But even at 1 Hz, the 10 MegOhm resistor data shows some effect. Your data will show two different current-voltage curves, one from the rising part of the triangle wave, and one from the falling part.

The best thing to do is to take data at 0.1 Hz for the 10 Meg resistor (and slow down the data acquisition rate by a factor of 10 as well).

If you didn’t do that, you can still use the data from the rising part. Just delete the lines from the 10 Meg data file where the voltage is falling instead of rising.
Gnuplot and Multiple Data Files

Gnuplot can easily show data from multiple files on a single plot. It’s no harder than plotting a function on top of data. Just type things separated by commas

```
plot 'file1.dat' using 1:2, 'file2.dat' using 1:2
```

You can also plot one file, then replot with the next files

```
plot 'file1.dat' using 1:2
replot 'file2.dat' using 1:2
replot 'file3.dat' using 1:2
```

Of course you can do more complicated “using” stuff than just “1:2”

Beware that if you use symbolic names like R for the resistor value, that you should use different symbolic names for each file! If you redefine a symbolic name then replot, Gnuplot will use the current value for the symbol!

Just like a function curve is plotted with a different color than the data, each data file will be plotted with a different color (and symbol, so you can tell them apart on a black and white printed graph).

When you plot lots of files, the “key” area of the plot can get too crowded or overlap with the data. You can move it around by `set key below` or `set key left` or several other possibilities. Do `help set key` to see the options.
Gnuplot and Multiple Data Files (2)

For your own lab notebook, the Gnuplot convention of showing the file name and “using” string in the key-area is a useful one. But for making pretty plots for a formal report, you probably want “reader-friendly” labels. You can do this by

```plaintext
plot 'file1.dat' using 1:2 title "10 Meg"
replot 'file2.dat' using 1:2 title "100K"
```

You don’t have to combine your data files into a single file to make a combined plot.

You do have to combine the data into a single file to make a combined fit.

There is a way to plot such a combined file with different colors, symbols, and labels for the different parts. Just insert two or more blank lines between the parts. Then use the “index” keyword to specify which part to plot how:

```plaintext
plot 'combined.dat' index 0 using 1:2 title "10 Meg"
replot 'combined.dat' index 1 using 1:2 title "100K"
```

If you leave out the “index” part, the blank lines just get ignored.
Plotting Multiple Diode Files

When you plot your data from multiple resistor values on a single semi-log graph, the different data files should map out overlapping parts of a single line.

If you use a wrong resistor value, a piece of the line will be vertically shifted.

For the largest resistor values, the 10 MegOhm internal resistance of the digitizer is significant, and you need to combine it with the “real” resistor to make things line up.

The 10 Meg digitizer resistance in parallel with the 10 Meg resistor in the circuit looks like a 5 Meg resistor. If the oscilloscope was also connected, it’s resistance is only 1 Meg, which is an even bigger correction!

Even for the 1 Meg resistor, 10 Meg in parallel matters.
Plotting Multiple Diode Files (2)

When you plot the data from all the resistor values, the page can get pretty messy.

Make separate plots for positive and negative voltage.

Use a logarithmic vertical scale for positive voltage.
Use a linear vertical scale for negative voltage.

There will be lots of points in horizontal stripes, far away from the diagonal line. These points are probably still consistent with the line, because they are from small resistor values and have large vertical error bars.

It’s useful to make two versions of this plot, one with errorbars to show that it’s all consistent, and another one without the errorbars so it’s cleaner.

You could also make copies of the files and use a text editor to remove the points that make the plot ugly.

When you make a “pulls” plot these points will not mess things up, because you divide the large deviation by the comparably large error.

You should look at the highest positive voltage points, and see if the trend is to be above the straight line (indicating that \( n = 1 \) is taking over) or below the line (indicating that something else like internal resistance is becoming important).
Plotting Multiple Diode Files (3)

Use a linear vertical scale for negative voltage. Plot the data with errorbars.

Autoscaling will make the points from high resistors invisibly close to zero, even though they have the smallest errors and are therefore the most meaningful.

Do manual vertical scaling: `set yrange [-1e-8:1e+8]`

The data from the highest resistor values should be significantly below zero, many times the size of the errorbar, and many times the typical offset (when converted into current).

The negative current value should be consistent with the $I_0$ value that you got from fitting the exponential in the positive voltage data.
Combining Diode Data Files

Gnuplot can’t fit multiple data files. So to fit the whole data set, we need to combine the files together into one.

We need to tell Gnuplot which resistor value to use for which data points. But Gnuplot only thinks about one line of the data file at a time, so we have to get the resistor value into each line of the file (AddCols, Gnumeric, Excel).

There are versions of AddCols for the physics.ubc.ca Sun machine, and for Windows machines, on the course web page. Each version only runs on its own type of machine. The Hebb 42 machines have Gnumeric and/or Excel.

AddCols just appends your information as text, so you have to make sure that the text will be meaningful to Gnuplot. It must be a number, not a math expression.

If you go from Windows to Linux and back, the different line-endings may cause an extra line containing only your added columns to be written. Just delete it in a text editor.

On the Sun or Linux, you can combine the files using “cat”. On Windows, you will have to combine the files using a text editor, or Gnumeric or Excel.

In either case, you might want to add a few blank lines at the start or end of the files and use “index” with Gnuplot as mentioned above.
Fitting Combined Data Files

When you fit the combined data, it’s OK to leave in all the points which look far from the line, as long as you tell Gnuplot what the errors on each point are.

This will vary from file to file because the resistor values are different, and you may also have used different digitizer inputs which have different quantization errors for different files.

You have to make sure that Gnuplot has enough information on each line of the file to calculate the right error.

That error should include the contribution of the “horizontal” error times the function slope. The lab procedure gives you some hints on how to do that in a way that will work on a combined file.

You can use “index” with fitting too, if you want to fit just a part of the combined data file.
An increase $\Delta k$ that corresponds to the flux increase $\Delta \Phi$ is equivalent to a displacement of an extended state by $\Delta k A$ in the $y$ direction. By the Stokes theorem and the definition of the vector potential we have $\Delta \Phi = L_0 A$. Thus $\Delta k \Phi$ causes a motion of the entire electron gas in the $y$ direction.

By $\Delta \Phi = N_e V y$ and $\Delta \Phi = h/2e$, we have

$$I = e(\Delta \Phi / \Delta \Phi_e) = e N_e V y / h = (N_e h / 2e) V y,$$

so that the Hall resistance is

$$\rho_H = V y / I = h / 2 e^2,$$

as in (2).

**Fractional Quantized Hall Effect (FQHE).** A quantized Hall effect has been reported for similar systems at fractional values of the index $n$, by working at lower temperatures and higher magnetic fields. In the extreme quantum limit the lowest Landau level is only partially occupied, and the integral QHE created above should not occur. It has been observed, however, that the Hall resistance $\rho_H$ is quantized in units of $36 e^2 / h$ when the occupation of the lowest Landau level is $1/3$ and $2/3$, and $\rho_H$ vanishes for these occupations. Similar breaks have been reported for occupations of $5/2, 7/2, 4/5$, and $2/7$.

**p-n JUNCTIONS**

A p-n junction is made from a single crystal modified in two separate regions. Acceptor or impurity atoms are incorporated into one part to produce the p region in which the majority carriers are holes. Donor impurity atoms in the other part produce the n region in which the majority carriers are electrons. The interface region may be less than $10^{-8}$ cm thick. Away from the junction region on the p side there are (−) ionized acceptor impurity atoms and an equal concentration of free holes. On the n side there are (+) ionized donor atoms and an equal concentration of free electrons. Thus the majority carriers are holes on the p side and electrons on the n side, Fig. 13.

Holes concentrated on the p side would like to diffuse to fill the crystal uniformly. Electrons would like to diffuse from the n side. But diffusion will upset the local electrical neutrality of the system.

---

A small charge transfer by diffusion leaves behind on the $p$ side an excess of ($-$) ionized acceptors and on the $n$ side an excess of ($+$) ionized donors. This charge double layer creates an electric field directed from $n$ to $p$ that inhibits diffusion and thereby maintains the separation of the two carrier types. Because of this double layer the electrostatic potential in the crystal takes a jump in passing through the region of the junction.

In thermal equilibrium the chemical potential of each carrier type is everywhere constant in the crystal, even across the junction. For holes

$$k_B T \ln p(x) + \varphi(x) = \text{constant}, \quad (27a)$$

where $p$ is the hole concentration and $\varphi$ the electrostatic potential. Thus $p$ is low where $\varphi$ is high. For electrons

$$k_B T \ln n(x) - \varphi(x) = \text{constant}, \quad (27b)$$

and $n$ will be low where $\varphi$ is low.

The total chemical potential is constant across the crystal. The effect of the concentration gradient exactly cancels the electrostatic potential, and the net particle flow of each carrier type is zero. However, even in thermal equilibrium there is a small flow of electrons from $n$ to $p$ where the electrons end their lives by recombination with holes. The recombination current $J_{re}$ is balanced by a current $J_{in}$ of electrons which are generated thermally in the $p$ region and which are pushed by the built-in field to the $n$ region. Thus in zero external applied electric field

$$J_{re}(0) + J_{in}(0) = 0, \quad (28)$$
for otherwise electrons would accumulate indefinitely on one side of the barrier.

**Rectification**

A p-n junction can act as a rectifier. A large current will flow if we apply a voltage across the junction in one direction, but if the voltage is in the opposite direction only a very small current will flow. If an alternating voltage is applied across the junction the current will flow chiefly in one direction—the junction has rectified the current (Fig. 14).

For back voltage bias a negative voltage is applied to the p region and a positive voltage to the n region, thereby increasing the potential difference between the two regions. Now practically no electrons can climb the potential energy hill from the low side of the barrier to the high side. The recombination current is reduced by the Boltzmann factor:

\[
J_w(V \text{ back}) = J_w(0) \exp(-eV/k_BT).
\]

The Boltzmann factor controls the number of electrons with enough energy to get over the barrier.

The thermal generation current of electrons is not particularly affected by the back voltage because the generation electrons flow downhill (from p to n) anyway:

\[
J_w(V \text{ back}) = J_w(0).
\]

We saw in (26) that \( J_w(0) = J_w(0); \) thus the generation current dominates the recombination current for a back bias.

When a forward voltage is applied, the recombination current increases because the potential energy barrier is lowered, thereby enabling more electrons to flow from the n side to the p side:

\[
J_w(V \text{ forward}) = J_w(0) \exp(nV/k_BT).
\]

Again the generation current is unchanged:

\[
J_w(V \text{ forward}) = J_w(0).
\]

The hole current flowing across the junction behaves similarly to the electron current. The applied voltage which lowers the height of the barrier for electrons also lowers it for holes, so that large numbers of electrons flow from the n region under the same voltage conditions that produce large hole currents in the opposite direction.

The electric currents of holes and electrons are additive, so that the total forward electric current is

\[
l = I_T[\exp(eV/k_BT) - 1],
\]

where \( I_T \) is the sum of the two generation currents. This equation is well satis-
Solar Cells and Photovoltaic Detectors

Let us shine light on a p-n junction, one without an external bias voltage. Each absorbed photon creates an electron and a hole. When these carriers diffuse to the junction, the built-in electric field of the junction separates them at the energy barrier. The separation of the carriers produces a forward voltage across the barrier, forward, because the electric field of the positively charged carriers is opposite to the built-in field of the junction.

The appearance of a forward voltage across an illuminated junction is called the photovoltaic effect. An illuminated junction can deliver power to an external circuit. Large area p-n junctions of silicon are used to convert solar photons to electrical energy.

Schottky Barrier

When a semiconductor is brought into contact with a metal, there is formed in the semiconductor a barrier layer from which charge carriers are severely depleted. The barrier layer is also called a depletion layer or exhaustion layer.

In Fig. 15 an n-type semiconductor is brought into contact with a metal. The Fermi levels are coincident after the transfer of electrons to the conduction band of the metal. Positively charged donor ions are left behind in this region.