

# Using *knitr* and $\text{\LaTeX}$ for literate lab notes

Boris Veytsman

TUG'2022

# A cautionary story

How a student was frustrated trying to repeat research...

Am I sure / can understand my own work in ten years?

How do our colleagues in biomedical and experimental research solve this problem, when billions depend on it?

# Laboratory notebooks

*A Lab Notebook Is...*

- ▶ *Complete record of procedures, reagents, data, and thoughts to pass on to other researchers*
- ▶ *Explanation of why experiments were initiated, how they were performed, and the results*
- ▶ *Legal document to prove patents and defend your data against accusations of fraud*

*Philip Ryan (2012). Keeping a Lab Notebook. National Institutes of Health, Office of Intramural Training and Education.*  
*URL: [https://www.training.nih.gov/assets/Lab\\_Notebook\\_508\\_\(new\).pdf](https://www.training.nih.gov/assets/Lab_Notebook_508_(new).pdf)*

# A classic example: Linus Pauling's notebooks

72 years of work (from 1922 to 1994): <http://scarc.library.oregonstate.edu/coll/pauling/rnb/index.html>

X-ray diffraction made Oct 1922 to June 1923  
 $K_2Ni_2(5D)_3$   $a = 9.907 \text{ \AA}$   $T_{\text{K}} = 281, 13.2^\circ$   
 $NaCd_2$   $a = 3.34 \text{ \AA}$   
 $Ca(HgBr_4)$   $a = 11.14$   $Z = 32$   
 $MoS_2$   
 $(NH_4)_2FeF_6$   
 $Na_2Cd_2 + H_2O$   
 $Ca(HgBr_4)_2 I_3$   
 $(NH_4)_2FeF_6$   
 $Zn$   
 $Mg_2Si$ .  $Mg_2Si_2$ ,  $RbHg_2$

Books 1 - Chemical analysis  
from Oct 1 to Dec 21 1922:  
10.8. LiH made  
20 Oct.  $Mg_2Si_2$  made  $NaCl$   
23 Oct. Cd by electrolysis  
21 Oct.  $Ca(HgBr_4)_2$   
 $Na_2Cd_2 + H_2O$   
19 Oct.  $IC_2Mg_2(5D)_2$   
23 "  $83^\circ$   
" " C  
24 Oct.  $CaB_6 \cdot 6H_2O$  made,  
 $K_2Ni_2(5D)_3$  made  $CaCl_2$   
25 "  $Fe$   
24 "  $W$   
23 Oct.  $Ta(5D)_3$  made, white.  
15.0.  $K_2HgBr_4$  made  
" "  $Ca(HgBr_4)_2$  made  
10 Dec.  $MoS_2$  made + analyzed  
14.0. Experiments made in ~~the~~ 6 weeks

Winter Valley  $\rightarrow$   $\text{Li}_2O$ ?  
but what I decided to  
check what I could do  
with respect to which kind  
energy, explaining properties  
of metals, according to me -  
I followed well with others  
always proportional to  $A$ ,  
well as parallel aligned.

$$E = \frac{n^2 + n^2 + n^2}{2n^2} = abc$$
$$E = \frac{(a^2 + b^2 + c^2)}{abc}$$

Consider  $\frac{E}{abc}$ :  $n_1 = n_2$   
 $\frac{2n^2}{abc} = \frac{1}{abc} + \frac{1}{abc} + \frac{1}{abc}$

Take, let  $abc = 2$ ,  $2n^2/abc = 2^{3/2}$ ,  $n^2 = 2^{3/2} = 1.5874$   
then  $E = \frac{9}{2^{3/2}} = 5.6916$

Now let  $abc = 1$ ,  $n^2E = 2 + 2 + \frac{1}{2} = 5.2580$   
Now let  $abc = 3.34$ ,  $n^2E = 4.75$   
 $n^2/abc = 1.1696$ ,  $E = 2.9240$

$\frac{a^2}{c^2} = 1.0215$ ,  $c_2 = 1.581$

$\frac{2n^2}{abc} = 5.1380$

Winter 1922  $\rightarrow$   $\text{Li}_2O$ ?  
Calculation p. 105 - very Figure 3.  
 $- \Delta G(g \text{ state} \rightarrow \text{solid state}) = 25 + 7.06 \log \frac{100}{T_{\text{K}}}$   
For  $\frac{T_{\text{K}}}{100} = 1$  we get  $\frac{-\Delta G}{7.06} = \frac{1}{12.3215} \log \frac{100}{T_{\text{K}}} = -1.1325$   
Hence  $- \Delta G(\text{state} \rightarrow \text{solid}) = 20.00 - 14.2 = 5.70$   
But this value is quite arbitrary - it depends on  
the standard state for the gas. Similarly,  
Hückel's argument (p. 3) is faulty.

# Lab notes as literate science

**Knuth's insight:** Your code is for computer. Your prose is for humans. ⇒ Literate programming<sup>1</sup>.

**Research situation:** A paper (preprint, presentation) is just an *advertisement* of the research, but not the research.  
Research is a reproducible *environment* which includes computation and publication<sup>2</sup>. ⇒ Lab notes as literate science

---

<sup>1</sup> Donald E. Knuth (1992). *Literate Programming*. CSLI Lecture Notes 27. California: Stanford.

<sup>2</sup> Jill P. Mesirov (2010). "Accessible Reproducible Research". In: *Science* 327.5964, pp. 415–416. ISSN: 0036-8075. DOI: 10.1126/science.1179653. URL: <http://science.sciencemag.org/content/327/5964/415>.

# How do we keep lab notes?

The classic way: bunches of physical notebooks

- ▶ Very versatile: you can put there anything! *But*
- ▶ You cannot search efficiently (where is my grep?)
- ▶ Too many dead trees.
- ▶ Not too easy to keep after a couple of decades.

The modern way: electronic records

- ▶ Can be indexed, searched, compact! *But*
- ▶ Can we make them as versatile as physical ones?
- ▶ Can we make writing them as fast as scribbling?

# What is in my lab notes? (1)

## ► Thoughts and ideas:

*It seems that cell diffusion inside a tissue is quite different if a different matrix around the tissue was used. This fact is quite inexplicable from the conventional picture of diffusion borrowed from the molecular physics. Indeed, how would a molecule inside a vessel “know” what is the vessel made of? One expects the measured diffusion not depend on the walls around the molecules.*

## ► Equations

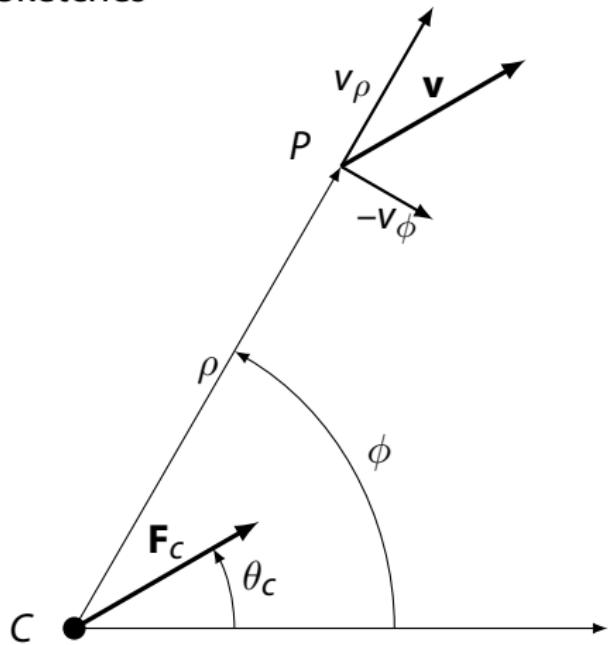
$$\mathfrak{a} = -\frac{l(l+1)c}{r},$$

$$\mathfrak{b} = -\frac{dc}{dr} - \frac{c}{r},$$

$$\mathfrak{c} = -\frac{a}{r} + \frac{db}{dr} + \frac{b}{r}.$$

# What is in my lab notes? (2)

## ► Sketches



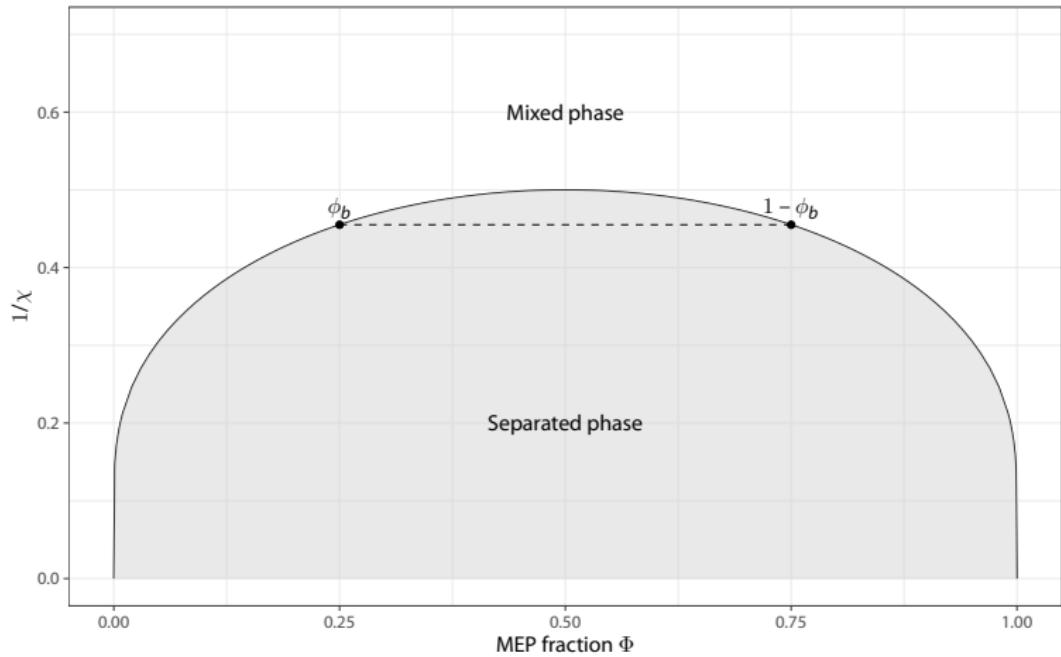
# What is in my lab notes? (3)

## ► Program snippets...

```
chiPhi <- tibble(phi=c(seq(0,0.01, by=0.001),
                      seq(0.01,0.99,by=0.01),
                      seq(0.99,1, by=0.001))) %>%
  mutate(chi= 1/(1-2*phi)*log((1-phi)/phi)) %>%
  filter(!is.nan(chi))
chiPhib <- chiPhi %>% filter(phi==0.25 | phi==0.75) %>%
  mutate(label=c('$\\phi_b$', '$1-\\phi_b$'))
ggplot(chiPhi) + geom_line(aes(phi, 1/chi)) +
  geom_polygon(data=chiPhi %>% add_row(phi=c(0,1), chi=c(Inf,Inf)),
               aes(phi,1/chi), fill='lightgray', alpha=0.5) +
  ylim(0,.7) + xlab("MEP fraction $\\Phi$") + ylab("$1/\\chi$")
  annotate("text", x=0.5, y=0.6, label="Mixed phase") +
  annotate("text", x=0.5, y=0.2, label="Separated phase") +
  geom_point(data=chiPhib, aes(phi, 1/chi)) +
  geom_line(data=chiPhib, aes(phi, 1/chi), linetype='dashed') +
  geom_text(data=chiPhib, aes(x=phi, y=1/chi, label=label),
            nudge_y=0.025)
```

## What is in my lab notes? (4)

- ▶ ...and their results



# Example: Computer algebra notebooks (1)

Many (all) commercial systems have them. Here is a free wxMaxima (<https://wxmaxima-developers.github.io/wxmaxima/>)

The screenshot shows the wxMaxima application window titled "untitled". The interface includes a toolbar with various icons and a menu bar. The main area displays a notebook titled "wxMaxima example". The notebook contains the following content:

```
We start with the standard identity

(%i2) integrate(sin(x)/x, x, 0, inf) = integrate(sin(x)/x, x, 0, inf);
(%o2) ∫_0^∞ sin(x)/x dx = π/2

We can plot the function $\sin(x)/x$:

(%i4) wxplot2d(sin(x)/x, [x, 0.001, 20]);
(%o4)
```

A 2D plot is shown in the notebook, displaying the function  $\frac{\sin(x)}{x}$  for  $x \geq 0.001$ . The x-axis ranges from 0.001 to 20, and the y-axis ranges from 0 to 1. The curve starts at approximately (0, 1) and decays towards the x-axis as x increases.

At the bottom of the notebook, it says "Maxima is ready for input." and "ping image, 1200x1200, 144 ppi".

# Example: Computer algebra notebooks (2)

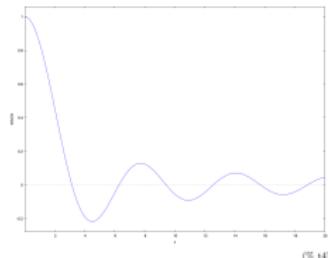
## WXMAXIMA EXAMPLE

We start with the standard identity

$$(\%) \text{12) } \int_0^\infty \frac{\sin(x)}{x} dx = \frac{\pi}{2}; \quad (\%) \text{a2}$$

We can plot the function  $\frac{\sin(x)}{x}$ :

(%) 14) wxplot2d(sin(x)/x, [x, 0.001, 20]);



(%) a4)

# Example: Jupyter notebooks (1)

The great Jupyter project (<https://jupyter.org/>)

A screenshot of a Jupyter Notebook interface running in a web browser. The title bar shows 'jupyter - Jupyter Notebook'. The address bar shows 'localhost:8888/notebooks/jupyter.ipynb'. The menu bar includes File, Edit, View, Insert, Cell, Kernel, Help, Trusted, and Python 3 (ipykernel). The toolbar below the menu has icons for New, Open, Save, Run, Cell, Kernel, Help, and Markdown. The main area contains a code cell with the title '# Example of jupyter notebook' and the text 'Let us make a graph of  $\frac{\sin(x)}{x}$ '. Below it is an In [3] cell containing Python code to generate a plot:

```
In [3]: import numpy as np
import matplotlib.pyplot as plt
x = np.arange(0.01, 20, 0.01)
y = np.sin(x)/x
plt.plot(x, y)
```

The Out[3] cell shows the resulting plot, which is a blue line graph of the function  $y = \frac{\sin(x)}{x}$  for  $x \in [0.01, 20]$ . The x-axis ranges from 0.0 to 20.0 with major ticks every 2.5 units. The y-axis ranges from -0.2 to 1.0 with major ticks every 0.2 units. The plot shows the characteristic oscillations of the sinc function, starting at approximately (0, 1) and decaying towards zero as x increases.

# Example: Jupyter notebooks (2)

```
jupyter  
July 17, 2022
```

## 1 Example of jupyter notebook

Let us make a graph of  $\frac{\sin(x)}{x}$

```
[3]:  
import numpy as np  
import matplotlib.pyplot as plt  
x = np.arange(0.01, 20., 0.01)  
y = np.sin(x)/x  
plt.plot(x, y)
```

[3]: [<matplotlib.lines.Line2D at 0x10e239f10>]

[ 1: ]

1

## My (humble) opinions about the examples

wxMaxima: Good for documenting equation manipulations. Not much convenient for everything else.

Jupyter: Good interface, especially when you play with code.  
Can incorporate many languages other than Python.  
*But:*

- ▶ Only a subset of  $\text{\LaTeX}$  implemented. No label-ref, bibliography, etc.
- ▶ No support for sketches other than plots.

Common feature:  $\text{\LaTeX}$  backend. Why not use  $\text{\LaTeX}$  from the beginning?

# My setup

Ideas:

1. I need the features of  $\text{\LaTeX}$ : bibliographies, numbering, etc.
2. A bunch of `tex` files is easily searched by `grep` and `find`.

A problem: I sometimes play with code and do a lot of plots.

Solution: Use `knitr`.

# An aside: $\text{\LaTeX}$ and Markdown

---

```
## An aside: LaTeX and Markdown
```

Many people use Markdown for

- \* notes,
  - \* reports,
  - \* documents,
  - \* some math:  $\int_0^\infty \sin x/x, dx = \pi/2$ .
- 

Markdown: easy to learn, but limited possibilities.

$\text{\LaTeX}$ : more difficult to learn, but huge possibilities:  
references, bibliographies, sketches, plots...

Preaching to the choir:  $\text{\LaTeX}$  is a good investment!



Yihui Xie (2015). *Dynamic Documents with R and knitr*. Second edition. Boca Raton; London; New York: Chapman and Hall/CRC. ISBN: 978-1498716963

A great tool for literate programming and literate science  
(Boris Veytsman (2014). "Book review: Dynamic Documents with R and knitr, by Yihui Xie". In: *TUGboat* 35.1, pp. 115–119. URL: <http://tug.org/TUGboat/tb35-1/tb109reviews-xie.pdf>).

## knitr example (1)

We start from the standard identity

```
\begin{equation}
```

$$\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2}$$

```
\end{equation}
```

We also add a simple plot

```
<<device='tikz', fig.width=8, fig.height=3>>=
data <- tibble(x=seq(0.01, 20, by=0.01)) %>%
    mutate(y=sin(x)/x)
ggplot(data) + geom_line(aes(x,y))
@
```

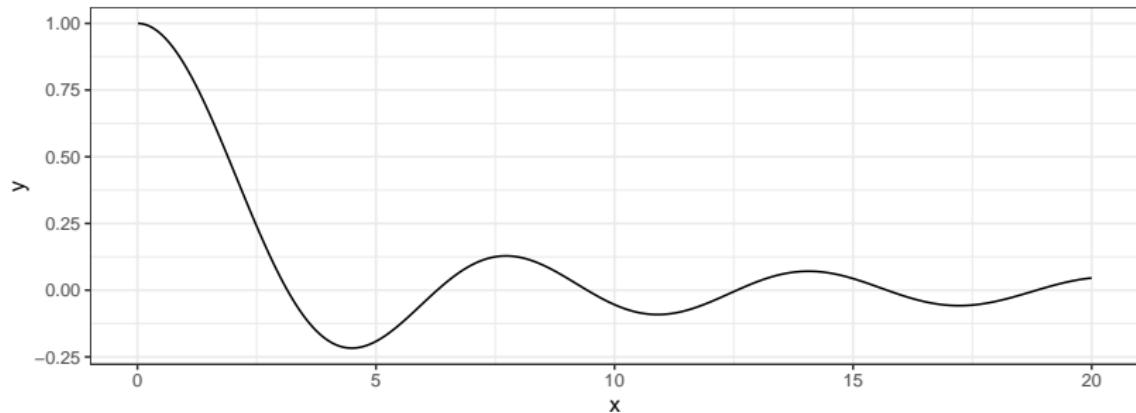
## knitr example (2)

We start from the standard identity

$$\int_0^\infty \frac{\sin x}{x} dx = \frac{\pi}{2} \quad (1)$$

We also add a simple plot

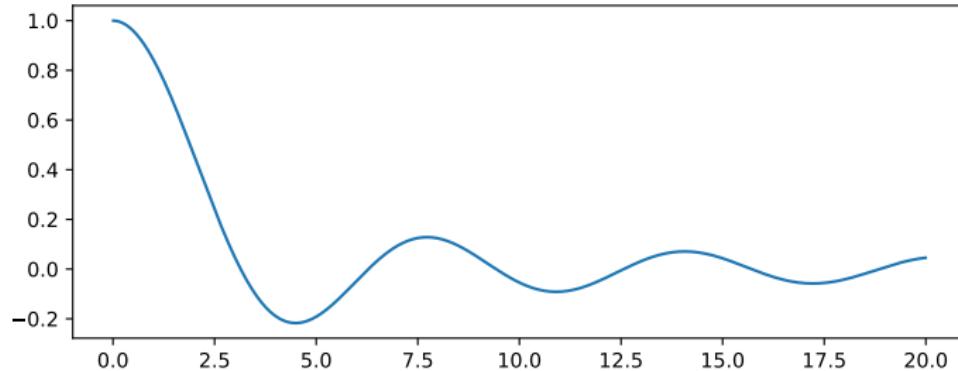
```
data <- tibble(x=seq(0.01, 20, by=0.01)) %>%
  mutate(y=sin(x)/x)
ggplot(data) + geom_line(aes(x,y))
```



# Not only R!

Here we use `engine='python'` magic

```
import numpy as np
import matplotlib.pyplot as plt
x = np.arange(0.01, 20, 0.01)
y = np.sin(x)/x
plt.plot(x,y)
```



# Details, tips and tricks

- ▶ Start a project with a directory, README and Makefile or Rstudio proj (or arara rules).
- ▶ You may need separate directories for data, etc.
- ▶ Number notes like 001-introduction.rnw, 002-hypothesis.rnw, etc.
- ▶ Always use version control!

# Examples of my lab notes

## 1 Introduction

In Chapter 1, we consider a number of strange models. For example, we found that the  $\langle \cos(\theta) \rangle$  quantity depends on the geometry of the system. It is an important exercise to understand what is happening here. In this note, we will discuss a similar system that gives some counterintuitive results.

We consider a system consisting of two particles with mass  $m$  and  $M$  and the interaction potential  $V(r) = k/r^2 - A/r^{1/2}$ .

Let us assume that  $m \ll M$ , so that  $v_{\perp} \ll V$ .

For this system, we find

The classical mechanics equations of motion are

$$\ddot{r} = -\frac{\partial V}{\partial r}, \quad \dot{\theta} = \frac{L}{r^2}, \quad (1)$$

with

$$L = \sqrt{\frac{2}{m}V + \frac{1}{2}m\dot{r}^2 + \frac{1}{2}I\dot{\theta}^2}, \quad (2)$$

Then we take the limit  $M \rightarrow \infty$ . We do this for the particles, which are close to one another than the gravitation. It is clear from the equations that the angular momentum is conserved, so that the particle with angular momentum  $L$ , and the component of velocity

$$v_r = \sqrt{\frac{2}{m}\left(V - \frac{1}{2}I\dot{\theta}^2\right)}, \quad (3)$$
$$v_\theta^2 = L^2/m^2r^2. \quad (4)$$

We expect the motion plotted in Figure 1 with  $I = 0$ ,

$$r(t) = \frac{1}{2}\left(\frac{L^2}{m^2} + \frac{1}{2}I\dot{\theta}^2\right)\left(1 - \frac{2}{m^2}\int_0^t V(r')dr'\right)^{-1/2}, \quad (5)$$

and

$$\dot{\theta}(t) = \frac{L}{r(t)^2}. \quad (6)$$

A good question: how many points are used to choose get a good curve for  $r(t)$ ? Answer: 1000. This is a good example for the benefit of the more relating to the needed duration of  $t^2 = \text{quality}^2$ .

2

## 2 Simulations

### 2.1 Numerical simulations

Let us define simulation functions. We write them in C++ to speed up the computation.

For the standard header

```
#include <iostream>
#include <cmath>
```

using namespace std;

We generate  $n$  orbitals in a ring  $1 \times N$  with the orientation defined by  $\theta$ :

```
#define N 1000
#define pi 3.1415926535897932384626433832795028841971693993751058209749445923078164062862089986280348253421170679
```

```
#define m 1.0
#define M 1000.0
#define k 1.0
#define A 0.001
#define L 1.0
#define I 0.0
#define V(r) (k/r - A/r^(1.0/2.0))
#define dr 0.001
#define dt 0.01
```

```
double angle(double theta) {
    if(theta < 0) theta += 2*pi;
    return theta;
}

double dotproduct(double v1[], double v2[]) {
    double sum = 0;
    for(int i=0; i<3; i++) sum += v1[i]*v2[i];
    return sum;
}

double crossproduct(double v1[], double v2[]) {
    double sum = 0;
    for(int i=0; i<3; i++) sum += v1[i]*v2[(i+1)%3];
    return sum;
}

double dotproduct(double v1[], double v2[]) {
    double sum = 0;
    for(int i=0; i<3; i++) sum += v1[i]*v2[i];
    return sum;
}

double crossproduct(double v1[], double v2[]) {
    double sum = 0;
    for(int i=0; i<3; i++) sum += v1[i]*v2[(i+1)%3];
    return sum;
}
```

2



Figure 2: Motion on the surface (a) and motion not on the surface (b).

where  $\theta$  is the rotation about  $z$ -axis. The magnet is located in all space, both here and in plots (Figure 1). However, we did not want to distract the reader from the main point: the effect of the magnet on the motion of the particles.

For the sake of numerical energy, let us note that the value of the total is conserved for all cases. However, the surface may be non-convex. These are the reasons why the motion is different. We can see this in the first plot on the right. The motion is not on the surface, so it goes around the ring, and it is not bounded by the first local potential energy, so, the dynamics is given by

$$\frac{d\theta}{dt} = \frac{L^2}{m^2r^2} - \frac{2}{m^2}\int_0^t V(r')dr' \quad (8)$$

where  $L$  is the angular momentum, and  $r$  is the distance considered when we take the integral. This gives

$$\frac{d\theta}{dt} = \frac{L^2}{m^2r^2} - \frac{2}{m^2}\int_0^t V(r')dr' \quad (9)$$

and the corresponding conservation of angular momentum is given by

$$R_2 = \frac{L^2}{m^2r^2} \quad (10)$$

For the sake of the second case we must take the potential energy of interaction between the magnet and the ring. This is the reason why we do not simulate the motion of the surface. Therefore for case (b), the change of the coordinate system is given by  $r \rightarrow r - R_2$ .

3

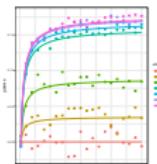


Figure 3: Velocity vs. time corresponding to the prediction from (5) and (6).

3



Figure 4: Angle and distance.

## 2.2 Kinetic discussion

In Chapter 1, we discussed a model for flow in fiber optics. In this project we will come to a reasonable explanation of different phenomena. In particular, we will consider the case when there is a shear force, and we will discuss the effect of the magnet on the motion taking into account the velocity in the plane  $T \times z$  (see Figure 3).

The obvious formula

$$\dot{r} = \sqrt{\frac{2}{m}V + \frac{1}{2}m\dot{v}_\perp^2 + \frac{1}{2}I\dot{\theta}^2}, \quad (11)$$

and the angle between this force and relative velocity is  $\pi/2$  (Figure 4). Note that equation (11) immediately requires that  $I = 0$  and  $v_\perp = 0$  in the computer simulations, since  $\dot{r} = 0$  for all  $v_\perp$  and  $\dot{\theta}$ .

The obvious formula

$$\dot{\theta} = \frac{L}{r^2}, \quad (12)$$

is self-evident.

$$\dot{v}_\perp = \frac{L}{r^3} \frac{dr}{dt}, \quad (13)$$

where  $L$  is the strength of the magnet, and all of this neglect one variable  $A$ , which is  $0$ , and  $r$  is the scaling value. We are interested mainly in  $v_\perp \ll r$  for the following reason:

4

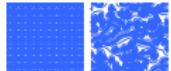


Figure 5: Completely aligned.

When we increase the value parameter  $k$ , the force of the field will increase, and we will see more and more alignment. For example, in Figure 5, we see a completely aligned field. When we decrease the value of  $k$ , the field becomes less strong, and we see less alignment.

## 4 Conclusion

It seems the ordered cells, both in their location and alignment, is very important for the propagation of light in optical fibers. Due to the previous quadratic charges, giving them a linear tangential force.

5

# Problems & Solutions

1. Limitations of PDF format: movies & interactive plots are not easy to do! There are solutions, but how reproducible are they?  
Flash debacle...
2. Speed:
  - ▶ I write prose with the speed I think—*good!*
  - ▶ I program in knitr with the same speed as in IDE—*good!*
  - ▶ I write equations in  $\text{\TeX}$  slightly slower than with a pen—*ok!*
  - ▶ I write sketches in TikZ (and in PSTricks) much slower than with a pen—*bad!*.

Solutions for the sketching speed I am considering:

- ▶ Doodle with a pen, then scan and use `\includegraphics`.
- ▶ Use a program with PDF output.
- ▶ Write TikZ faster.

# Final exhortation (standing on the shoulders of a giant)

GO FORTH now and create *beautiful, clear and reproducible laboratory notes!*