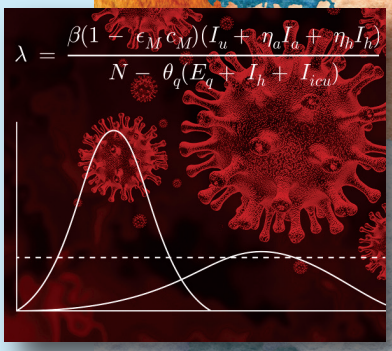
AMERICAN MATHEMATICAL SOCIETY

CURRENT EVENTS BULLETIN

Friday, January 8, 2021, 1:00 PM to 4:45 PM

Virtual Joint Mathematics Meeting

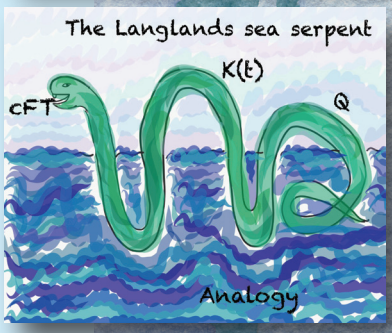


1:00 PM

Abba Gumel, *Arizona State University*

Mathematics of the Dynamics and Control of the COVID-19 Pandemic

Mathematics has a lot to say about pandemics; what could be more relevant to our moment?

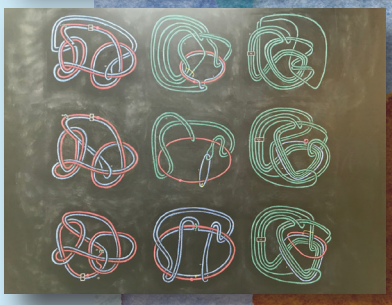


2:00 PM

Ana Caraiani, *Imperial College London*

An excursion through the land of shtukas

Geometry and the Langlands program: catch up on the deepest topic in number theory.

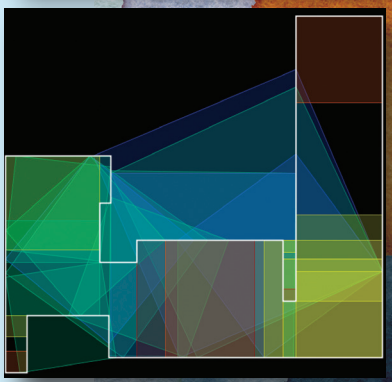


3:00 PM

Jennifer Hom, *Georgia Institute of Technology*

Getting a handle on the Conway knot

The taming of an elusive knot.



4:00 PM

Richard Evan Schwartz, *Brown University*

Rectangles, Curves, and Klein Bottles

A problem about squares and simple closed curves leads to surprising symplectic geometry.

AMS American Mathematical Society

Advancing research. Creating connections.

Organized by David Eisenbud, University of California, Berkeley

Introduction to the Current Events Bulletin

Will the Riemann Hypothesis be proved this week? What is the Geometric Langlands Conjecture about? How could you best exploit a stream of data flowing by too fast to capture? I think we mathematicians are provoked to ask such questions by our sense that underneath the vastness of mathematics is a fundamental unity allowing us to look into many different corners -- though we couldn't possibly work in all of them. I love the idea of having an expert explain such things to me in a brief, accessible way. And I, like most of us, love common-room gossip.

The Current Events Bulletin Session at the Joint Mathematics Meetings, begun in 2003, is an event where the speakers do not report on their own work, but survey some of the most interesting current developments in mathematics, pure and applied. The wonderful tradition of the Bourbaki Seminar is an inspiration, but we aim for more accessible treatments and a wider range of subjects. I've been the organizer of these sessions since they started, but a varying, broadly constituted advisory committee helps select the topics and speakers. Excellence in exposition is a prime consideration.

A written exposition greatly increases the number of people who can enjoy the product of the sessions, so speakers are asked to do the hard work of producing such articles. These are made into a booklet distributed at the meeting. Speakers are then invited to submit papers based on them to the *Bulletin of the AMS*, and this has led to many fine publications.

I hope you'll enjoy the papers produced from these sessions, but there's nothing like being at the talks -- don't miss them!

David Eisenbud, Organizer
Mathematical Sciences Research Institute
de@msri.org

For PDF files of talks given in prior years, see
http://www.ams.org/ams/current-events-bulletin.html.
The list of speakers/titles from prior years may be found at the end of this booklet.

Mathematics of a multigroup model for assessing the combined impact of masks and a potential COVID-19 vaccine

Abba B. Gumel[†,††] [*], Enahoro A. Iboi[◇], Calistus N. Ngonghala[‡,‡‡] and Gideon A. Ngwa[†††]

[†] *School of Mathematical and Statistical Sciences, Arizona State University, Tempe, Arizona, 85287, USA.*

[◇] *Department of Mathematics, Spelman College, Atlanta, Georgia, 30314, USA.*

[‡] *Department of Mathematics, University of Florida, Gainesville, FL 32611, USA.*

[‡‡] *Emerging Pathogens Institute, University of Florida, Gainesville, FL 32610, USA.*

[††] *Department of Mathematics and Applied Mathematics, University of Pretoria, Pretoria 0002, South Africa.*

[†††] *Department of Mathematics, University of Buea, P.O. Box 63, Buea, Cameroon.*

Abstract

A novel coronavirus emerged in December of 2019 (COVID-19), causing a pandemic that continues to inflict unprecedented public health and economic burden in all nooks and corners of the world. Although the control of COVID-19 has largely focused on the use of basic public health measures (primarily based on using non-pharmaceutical interventions, such as quarantine, isolation, social-distancing, face mask usage and community lockdowns), a number of exceptionally-promising vaccines are about to be approved for use in humans by the U.S. Food and Drugs Administration. We present a new mathematical model for assessing the population-level impact of the candidate vaccines, particularly for the case where the vaccination program is complemented with a social-distancing control measure at a certain compliance level. The model stratifies the total population into two subgroups, based on whether or not they habitually wear face mask in public. The resulting multigroup model, which takes the form of a compartmental, deterministic system of nonlinear differential equations, is parametrized using COVID-19 cumulative mortality data. Conditions for the asymptotic stability of the associated disease-free equilibrium, as well as expression for the vaccine-derived herd immunity threshold, are derived. This study shows that the prospect of COVID-19 elimination using any of the three candidate vaccines is quite promising, and that such elimination is more feasible if the vaccination program is combined with social-distancing control measures (implemented at moderate to high level of compliance).

Keywords: *COVID-19; vaccine; social-distancing; herd immunity; face mask; stability; reproduction number.*

1 Introduction

The novel coronavirus (COVID-19) pandemic, which started as a pneumonia of an unknown etiology late in December 2019 in the city of Wuhan, is the most devastating public health challenge mankind has faced since the 1918/1919 pandemic of influenza. The COVID-19 pandemic, which rapidly spread to essentially every nook and corner of the planet, continues to inflict devastating public health and economic challenges globally. As of December 5, 2020, the pandemic accounts for $67,021,834$ confirmed cases and $1,537,165$ cumulative mortality globally. Similarly, the United States, which recorded its first COVID-19 case on January 20, 2020, recorded $14,991,531$ confirmed cases and $287,857$ deaths (as of December 5, 2020) [1].

COVID-19, a member of the Coronavirus family of RNA viruses that cause diseases in mammals and birds, is primarily transmitted from human-to-human through direct contact with contaminated objects or surfaces and through inhalation of respiratory droplets from both symptomatic and asymptomatically-infectious humans (*albeit* there is limited evidence that COVID-19 can be transmitted *via* exhalation through normal breathing and aerosol [2]. The incubation period of the disease is estimated to lie between 2 to 14 days (with a mean of 5.1 days), and majority of individuals infected with the disease show mild or no clinical symptoms [3]. The symptoms typically include

[*]Corresponding author: Email: *agumel@asu.edu*

coughing, fever and shortness of breadth (for mild cases) and pneumonia for severe cases [3]. The people most at risk of dying from, or suffering severe illness with, COVID-19 are those with co-morbidities (such as individuals with diabetes, obesity, kidney disease, cardiovascular disease, chronic respiratory disease etc.). Younger people, frontline healthcare workers and employees who maintain close contacts (within 6 feet) with customers and other co-workers (such as meat factory workers, retail store workers etc.).

Although there are three exceptionally-promising candidate vaccines (by Pfizer, Inc., Moderna, Inc. and AstraZeneca, Inc.) and antivirals undergoing various stages of development (Pfizer has filed for FDA Emergency Use Authorization on November 20, 2020) [4], there is currently no safe and effective vaccine or antiviral that has been approved for widespread use in humans, *albeit* the approval of the aforementioned candidate vaccines is imminently expected by the end of 2020. Further, owing to its limited supply, the approved anti-COVID drug *remdesivir* is limited for use to treat individuals in hospital who display severe symptoms of COVID-19. Hence, due to the absence of safe and effective vaccines and antiviral for widespread use in humans, efforts to control and mitigate the burden of COVID-19 remain focused on non-pharmaceutical interventions (NPIs), such as quarantine, self-isolation, social (physical) distancing, the use of face masks in public, hand washing (with approved sanitizers), community lockdowns, testing and contact tracing. Of these NPIs, the use of face masks in public is considered to be the main mechanism for effectively curtailing COVID-19 [3, 5–8].

The Pfizer and Moderna vaccines, each of estimated protective efficacy of about 95% [4, 9, 10], are genetic vaccines that are developed based on stimulating a mechanism that encourages the body to produce antibodies that fights the SARS-CoV-2. Specifically, the vaccines use a synthetic messenger RNA (*m*RNA) that carries instructions for making virus spike protein to gain entry into cells when injected into muscle tissue in the upper arm. This triggers the immune system to recognize the spike protein and develop antibodies against it (so that when a human is infected with SARS-CoV-2, his/her body is able to successfully fight it) [4, 11]. Two doses are required for both the Pfizer and Moderna vaccine candidates (one to prime the immune system, and the second to boost it). For the Pfizer vaccine, the second dose will be administered 19-42 days after the first dose. Further, the Pfizer vaccine needs to be stored at a temperature of $-70^{\circ}C$. The second dose of the Moderna vaccine is administered three to four weeks after the first dose. Further, the Moderna vaccine can be stored at refrigerated temperature of (2-8$^{\circ}C$), with long-term storage conditions of $-20^{\circ}C$ for at least six months [12]. The AstraZeneca vaccine, on the other hand, has estimated protective efficacy of 70% [4, 9, 10]. It uses a replication-deficient chimpanzee viral vector that causes infections in chimpanzees and contains the genetic material of the SARS-CoV-2 virus spike protein [10]. When injected into the human, the spike protein triggers the immune system to attack the SARS-CoV-2 virus that infects the body [10]. AstraZeneca vaccine also requires two doses (one month apart) to achieve immunity, and, unlike the Pfizer and Moderna vaccines, does not have to be stored in super-cold temperatures (it can be stored at normal refrigerated temperature of (2-8$^{\circ}C$) for at least six months) [10]. Hence, owing to the imminence for the approval of the aforementioned three candidate COVID-19 vaccines by the FDA, coupled with the primary role of face masks usage, it is instructive to design new mathematical models that will allow for the realistic assessment of the combined impact of the expected COVID-19 vaccines and face masks usage in a community.

Numerous mathematical models, of various types, have been developed and used to provide insight into the transmission dynamics and control of COVID-19. The modeling types used include statistical [13], compartmental/deterministic (e.g., [3, 5, 7, 8]), stochastic (e.g., [14, 15]), network (e.g., [16]) and agent-based (e.g., [17]). The purpose of the current study is to use mathematical modeling approaches, coupled with mathematical and statistical data analytics, to assess the combined impact of the expected COVID-19 vaccines and face masks usage. A notable feature of the model to be developed is its multigroup nature. Specifically, the total population will be subdivided into two groups, namely those who habitually wear face mask in public and those who do not. Data for COVID-19 pandemic in the U.S. will be used to parametrize the model. The central goal of the study is to determine the minimum vaccine coverage level needed to effectively curtail (or eliminate) community transmission of COVID-19 in the U.S., and to quantify the reduction in the required vaccine coverage if the vaccination program is supplemented with face masks usage (under various face masks efficacy and compliance parameter space). The paper is organized as follows. The novel multigroup model is formulated in Section 2. The parameters of the model are also estimated, based on fitting the model with U.S. COVID-19 data. The model is rigorously analysed, with respect to the asymptotic stability of the disease-free equilibrium of the model, in Section 3. A condition for achieving

community-wide vaccine-derived herd immunity is also derived. Numerical simulations of the model are reported in Section 4. Discussions and concluding remarks are presented in Section 5.

2 Formulation of Mathematical Model

In order to account for heterogeneity in face masks usage in the community, the total population of individuals in the community at time t, denoted by $N(t)$, is split into the total sub-populations of individuals who do not habitually wear face mask in public (labeled "*non-mask users*"), denoted by $N_1(t)$, and the total sub-populations of those who habitually wear face mask in public (labeled "*mask users*"), represented by $N_2(t)$. That is, $N(t) = N_1(t) + N_2(t)$. Furthermore, the sub-population $N_1(t)$ is sub-divided into the mutually-exclusive compartments of unvaccinated susceptible ($S_{1u}(t)$), vaccinated susceptible ($S_{1v}(t)$), exposed ($E_1(t)$), pre-symptomatically-infectious ($P_1(t)$), symptomatically-infectious ($I_1(t)$), asymptomatically-infectious ($A_1(t)$), hospitalized ($H_1(t)$) and recovered ($R_1(t)$) individuals, so that

$$N_1(t) = S_{1u}(t) + S_{1v}(t) + E_1(t) + P_1(t) + I_1(t) + A_1(t) + H_1(t) + R_1(t).$$

Similarly, the total sub-population of the mask users, $N_2(t)$, is stratified into the compartments for unvaccinated susceptible ($S_{2u}(t)$), vaccinated susceptible ($S_{2v}(t)$), exposed ($E_2(t)$), pre-symptomatically-infectious ($P_2(t)$), symptomatically-infectious ($I_2(t)$), asymptomatically-infectious ($A_2(t)$), hospitalized ($H_2(t)$) and recovered ($R_2(t)$) individuals. Hence,

$$N_2(t) = S_{2u}(t) + S_{2v}(t) + E_2(t) + P_2(t) + I_2(t) + A_2(t) + H_2(t) + R_2(t).$$

The equations for the rate of change of the sub-populations of non-mask users is given by the following deterministic system of nonlinear differential equations (where a dot represents differentiation with respect to time t):

$$\begin{aligned}
\dot{S}_{1u} &= \Pi + \omega_v S_{1v} + \alpha_{21} S_{2u} - \lambda_1 S_{1u} - (\alpha_{12} + \xi_v + \mu) S_{1u}, \\
\dot{S}_{1v} &= \xi_v S_{1u} + \alpha_{21} S_{2v} - (1 - \varepsilon_v)\lambda_1 S_{1v} - (\alpha_{12} + \omega_v + \mu) S_{1v}, \\
\dot{E}_1 &= \lambda_1 S_{1u} + (1 - \varepsilon_v)\lambda_1 S_{1v} + \alpha_{21} E_2 - (\alpha_{12} + \sigma_1 + \mu) E_1, \\
\dot{P}_1 &= \sigma_1 E_1 + \alpha_{21} P_2 - (\alpha_{12} + \sigma_P + \mu) P_1, \\
\dot{I}_1 &= r\sigma_P P_1 + \alpha_{21} I_2 - (\alpha_{12} + \phi_{1I} + \gamma_{1I} + \mu + \delta_{1I}) I_1, \\
\dot{A}_1 &= (1 - r)\sigma_P P_1 + \alpha_{21} A_2 - (\alpha_{12} + \gamma_{1A} + \mu) A_1, \\
\dot{H}_1 &= \phi_{1I} I_1 + \alpha_{21} H_2 - (\alpha_{12} + \gamma_{1H} + \mu + \delta_{1H}) H_1, \\
\dot{R}_1 &= \gamma_{1I} I_1 + \gamma_{1A} A_1 + \gamma_{1H} H_1 + \alpha_{21} R_2 - (\alpha_{12} + \mu) R_1,
\end{aligned} \tag{2.1}$$

where, λ_1 is the *force of infection*, defined by:

$$\lambda_1 = (1 - c_s)\left[\frac{(\beta_{P_1} P_1 + \beta_{I_1} I_1 + \beta_{A_1} A_1 + \beta_{H_1} H_1)}{N_1} + (1 - \varepsilon_o)\frac{(\beta_{P_2} P_2 + \beta_{I_2} I_2 + \beta_{A_2} A_2 + \beta_{H_2} H_2)}{N_2}\right],$$

with $\beta_i \{i = P_1, I_1, A_1, H_1, P_2, I_2, A_2 \text{ and } H_2\}$ the effective contact rate for individuals in the respective $P_1, I_1, A_1, H_1, P_2, I_2, A_2$ and H_2 classes. The parameters $0 < \varepsilon_o < 1$ and $0 < \varepsilon_i < 1$ represent the outward and inward protective efficacy, respectively, of face masks to prevent the transmission of infection to a susceptible individual (ε_o) as well as prevent the acquisition of infection (ε_i) from an infectious individual, while $0 \leq c_s < 1$ is a parameter that accounts social-distancing compliance.

In (2.1), the parameter Π is the recruitment (birth or immigration) rate of individuals into the population, α_{21} is the rate of change of behavior for non-habitual face masks users to become habitual users (i.e., α_{12} is the transition rate from group 2 to group 1). Furthermore, α_{12} is the rate at which habitual face masks users choose to be non-habitual wearers. The parameter ξ_v represents the vaccination rate, and the vaccine is assumed to induce protective

efficacy $0 < \varepsilon_v < 1$ in all vaccinated individuals and wane at a rate ω_v. Natural deaths occurs in all epidemiological classes at a rate μ. Individuals in the E_1 class progress to the pre-symptomatic stage at a rate σ_1, and those in the pre-symptomatic class (P_1) transition out of this class at a rate σ_P (a proportion q of which become symptomatic, and move to the I class at a rate $q\sigma_P$, and the remaining proportion, $1-q$, move to the asymptomatically-infectious class at a rate $(1-q)\sigma_P$). Symptomatic infectious individuals are hospitalized at a rate ϕ_{1I}. They recover at a rate γ_{1I} and die due to the disease at a rate δ_{1I}. Hospitalized individuals die of the disease at the rate δ_{1H}.

Similarly, the equations for the rate of change of the sub-populations of mask users is given by:

$$\begin{aligned}
\dot{S}_{2u} &= \omega_v S_{2v} + \alpha_{12} S_{1u} - \lambda_2 S_{2u} - (\alpha_{21} + \xi_v + \mu) S_{2u}, \\
\dot{S}_{2v} &= \xi_v S_{2u} + \alpha_{12} S_{1v} - (1-\varepsilon_v)\lambda_2 S_{2v} - (\alpha_{21} + \omega_v + \mu) S_{2v}, \\
\dot{E}_2 &= \lambda_2 S_{2u} + (1-\varepsilon_v)\lambda_2 S_{2v} + \alpha_{12} E_1 - (\alpha_{21} + \sigma_2 + \mu) E_2, \\
\dot{P}_2 &= \sigma_2 E_2 + \alpha_{12} P_1 - (\alpha_{21} + \sigma_P + \mu) P_2, \\
\dot{I}_2 &= q\sigma_P P_2 + \alpha_{12} I_1 - (\alpha_{21} + \phi_{2I} + \gamma_{2I} + \mu + \delta_{2I}) I_2, \\
\dot{A}_2 &= (1-q)\sigma_P P_2 + \alpha_{12} A_1 - (\alpha_{21} + \gamma_{2A} + \mu) A_2, \\
\dot{H}_2 &= \phi_{2I} I_2 + \alpha_{12} H_1 - (\alpha_{21} + \gamma_{2H} + \mu + \delta_{2H}) H_2, \\
\dot{R}_2 &= \gamma_{2I} I_2 + \gamma_{2A} A_2 + \gamma_{2H} H_2 + \alpha_{12} R_1 - (\alpha_{21} + \mu) R_2,
\end{aligned} \tag{2.2}$$

where,

$$\lambda_2 = (1-c_s)(1-\varepsilon_i)\left[\frac{(\beta_{P_1} P_1 + \beta_{I_1} I_1 + \beta_{A_1} A_1 + \beta_{H_1} H_1)}{N_1} + (1-\varepsilon_o)\frac{(\beta_{P_2} P_2 + \beta_{I_2} I_2 + \beta_{A_2} A_2 + \beta_{H_2} H_2)}{N_2}\right].$$

Thus, Equations (2.1) and (2.2) represent the multi-group model for assessing the population impact of face masks usage and vaccination on the transmission dynamics and control of COVID-19 in a community. The flow diagram of the model $\{(2.1), (2.2)\}$ is depicted in Figure 1 (the state variables and parameters of the model are described in Tables 5 and 6, respectively).

Some of the main assumptions made in the formulation of the multi-group model $\{(2.1), (2.2)\}$ include the following:

1. Homogeneous mixing (i.e., we assumed a well-mixed population, where every member of the community is equally likely to mix with every other member of the community).

2. Exponentially-distributed waiting time in each epidemiological compartment.

3. The anti-COVID vaccine is imperfect. That is, the vaccine offers partial protective immunity (with efficacy $0 < \varepsilon_v < 1$), which wanes over time (at a rate ω_v). Further, it is assumed that the vaccine does not offer any therapeutic benefit (such as slowing progression to active disease or increasing recovery rate in breakthrough infections).

4. Although there is no definitive data on COVID-19 immunity, we assume that natural recovery from infection confers permanent immunity against reinfection.

5. Endemicity assumption: although epidemic models (with no demographics) are typically used for studying the dynamics of new epidemics, such as COVID-19, we assume that, for the purpose of vaccination program, COVID-19 has attained endemic status. This is to account for the fact that the vaccine will be administered to every member of the community (including newborns) for an extended period of time (perhaps years). The implication of this assumption is that human demography (as represented by the recruitment parameter, Π, and the natural death parameter, μ) must be incorporated into the model.

The multi-group model $\{(2.1), (2.2)\}$ is an extension of the two-group mask-use model in [5] by, *inter alia*:

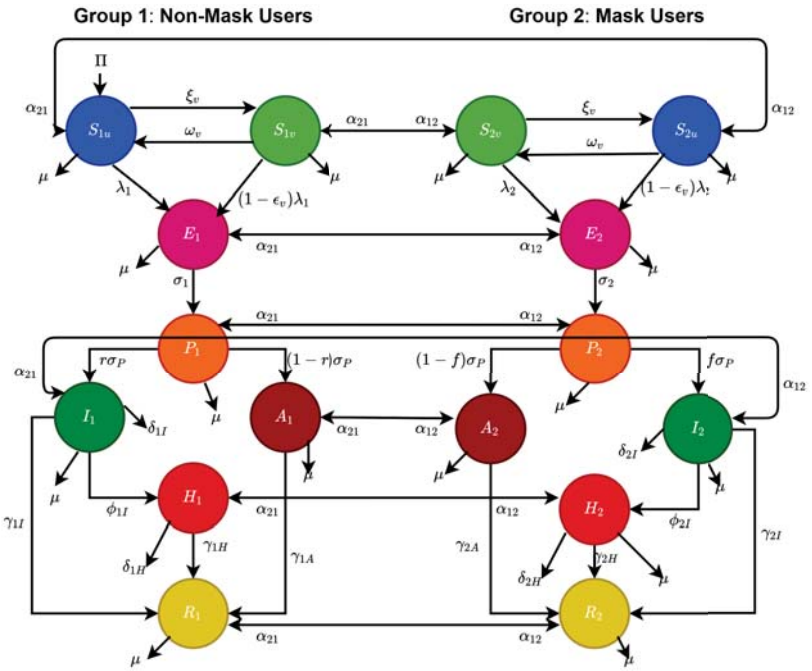


Figure 1: Flow diagram of the model {(2.1), (2.2)}.

(i) allowing for back-and-forth transitions between the two groups (mask-users and non-mask-users), to account for human behavioral changes *vis a vis* decision to either be (or not to be) a face mask user in public;

(ii) incorporating an imperfect vaccine, which offers protective efficacy ($0 < \varepsilon_v < 1$) against acquisition of COVID-19 infection, which may wane over time (at a rate ω_v);

(iii) allowing for disease transmission by pre-symptomatic and asymptomatically-infectious individuals.

2.1 Data Fitting and Parameter Estimation

In this section, cumulative mortality data for the US (from January 22, 2020 to November 16, 2020) will be used to fit the model (2.1)-(2.2) in the absence of vaccination and estimate some of its key parameters. In particular, the parameters to be estimated from the data are the community transmission rate for individuals who do not wear face masks in public (β_1), the transmission rate for individuals who habitually wear face masks in public (β_2), the inward efficacy of masks in preventing disease acquisition by susceptible individuals who habitually wear face masks (ε_i), the outward efficacy of masks to prevent the spread of disease by infected individuals who habitually wear face masks (ε_o), the proportion of individuals in the community who comply to social-distancing measures while in public (c_s), the rate at which people who do not wear masks adopt a mask-wearing habit (α_{12}), the rate at which those who habitually wear face masks stop wearing masks in public (α_{21}), and the mortality rates of symptomatic infectious and hospitalized individuals (δ_i and δ_h, respectively). It should be mentioned that modification parameters η_P, η_I, η_A, and η_H relating to disease transmission by pre-symptomatic infectious, symptomatic infectious, asymptomatic infectious and hospitalized individuals, respectively, are introduced in the forces of infection λ_1 and λ_2, so that $\beta_j = \eta_j \beta_k$ ($j \in \{P_k, I_k, A_k, H_k\}, k \in \{1, 2\}$). The model fitting was carried out using MATLAB R2020b and the process involved minimizing the sum of the square differences between each observed cumulative mortality data point and the corresponding mortality point obtained from the model (2.1)-(2.2) in the absence of vaccination [3, 18, 19]. The choice of mortality over case data is motivated by the fact that

mortality data for COVID-19 is more reliable than case data (see [7] for details). The estimated values of the fitted parameters are tabulated in Table 1(a). The fitting of the model to the observed cumulative and daily mortality data is depicted in Figure 2 (a). Furthermore, Figure 2 (b) compares the simulations of the model using the fitted (estimated) and fixed parameters (given in Tables 1 (a) and (b)-(c)) with the observed daily COVID-19 mortality for the US, showing a good fit.

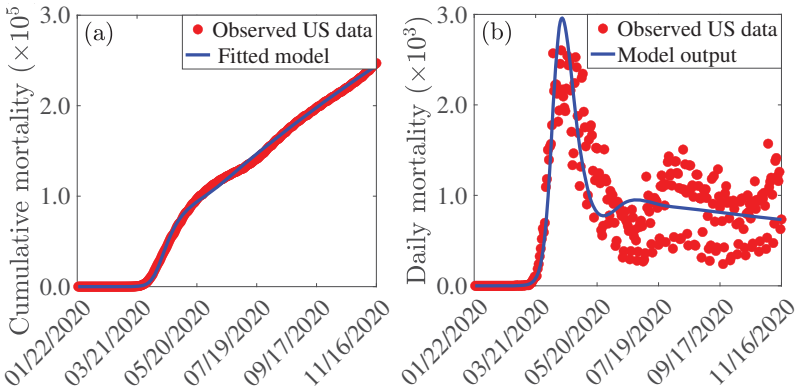


Figure 2: (a) Observed cumulative mortality data for the US (red dots) and predicted cumulative mortality for the US from the model (2.1)-(2.2) (with no vaccination) for the period from January 22 to November 16, 2020. (b) Simulations of the model (2.1)-(2.2) using the fixed parameters in Table 1(b)-(c) and the estimated parameters from the cumulative COVID-19 mortality data for the US presented in Table 1(a). We started the simulations of the pandemic near the disease-free equilibrium for the US. In particular, we used the following initial conditions (with mask usage compliance initially set at 1% of the total current US population): $(S_1^0, E_1^0, P_1^0, I_1^0, A_1^0, H_1^0, R_1^0, S_2^0, E_2^0, P_2^0, I_2^0, A_2^0, H_2^0, R_2^0) = (0.99 \times 336218660 - 1, 0, 0, 1, 0, 0, 0, 0.01 \times 336218660, 0, 0, 0, 0, 0, 0)$.

Table 1: Baseline parameter values for the model (2.1)-(2.2). (a) Estimated (fitted) parameter values for the model in the absence of vaccination, using COVID-19 mortality data for the US for the period from January 22, 2020 to November 16, 2020. (b)-(c) Baseline values of the remaining fixed parameters of the model (2.1)-(2.2) extracted from the literature or estimated using information from the literature.

(a) Fitted parameters

Parameter	Value
β_1	0.6566/day
β_2	0.5249/day
c_s	0.3051
ε_o	0.6304
ε_i	0.9965
α_{12}	0.0459/day
α_{21}	0.0010/day
δ_i	0.0008/day
δ_h	0.0025/day

(b) Fixed parameters

Parameter	Value	Source
σ_1	$1/2.5$/day	[20, 21]
σ_2	$1/2.5$/day	[20, 21]
σ_p	$1/2.5$/day	[20, 21]
$r\ (q)$	$0.2(0.2)$	[22, 23]
ϕ_{1I}	$1/6$/day	[24]
ϕ_{2I}	$1/6$/day	[24]
γ_I	$1/10$/day	[17, 25]
γ_A	$1/5$/day	[24]
γ_H	$1/8$/day	[17]

(c) Fixed parameters

Parameter	Value	Source
Π	1.2×10^4/day	Estimated
μ	$1/(79 \times 365)$/day	Estimated
η_P	1.25	Assumed
η_I	1.0	Assumed
η_A	1.50	Assumed
η_H	0.25	Assumed
ω_v	0/day	Assumed
ξ_v	2.97×10^{-4}/day	Assumed
ε_v	0.70	[9, 10]

3 Mathematical Analysis

Since the model $\{(2.1), (2.2)\}$ monitors the temporal dynamics of human populations, all state variables and parameters of the model are non-negative. Consider the following biologically-feasible region for the model:

$$\Omega = \left\{ (S_{1u}, S_{1v}, S_{2u}, S_{2v}, E_1, E_2, P_1, P_2, I_1, I_2, A_1, A_2, H_1, H_2, R_1, R_2) \in \mathbf{R}_+^{16} : N(t) \leq \frac{\Pi}{\mu} \right\}. \tag{3.1}$$

Theorem 3.1. *The region Ω is positively-invariant with respect to the model $\{(2.1), (2.2)\}$.*

Proof. Adding all the equations of the model $\{(2.1), (2.2)\}$ gives

$$\dot{N} = \Pi - \mu N - \delta_{1I} I_1 - \delta_{1H} H_1 - \delta_{2I} I_2 - \delta_{2H} H_2. \tag{3.2}$$

Recall that all parameters of the model $\{(2.1), (2.2)\}$ are non-negative. Thus, it follows, from (3.2), that

$$\dot{N} \leq \Pi - \mu N. \tag{3.3}$$

Hence, if $N > \frac{\Pi}{\mu}$, then $\dot{N} < 0$. Furthermore, by applying a standard comparison theorem [26] on (3.3), we have:

$$N(t) \leq N(0)e^{-\mu t} + \frac{\Pi}{\mu}(1 - e^{-\mu t}).$$

In particular, $N(t) \leq \frac{\Pi}{\mu}$ if $N(0) \leq \frac{\Pi}{\mu}$. Thus, every solution of the model $\{(2.1), (2.2)\}$ with initial conditions in Ω remains in Ω for all time $t > 0$. In other words, the region Ω is positively-invariant and attracts all initial solutions of the model $\{(2.1), (2.2)\}$. Hence, it is sufficient to consider the dynamics of the flow generated by $\{(2.1), (2.2)\}$ in Ω (where the model is epidemiologically- and mathematically well-posed) [27]. □

3.1 Asymptotic Stability of Disease-free Equilibrium

The model $\{(2.1), (2.2)\}$ has a unique disease-free equilibrium (DFE), obtained by setting all the infected compartments of the model to zero, given by (where $S_{1u}^* > 0$, $S_{1v}^* > 0$, $S_{2u}^* > 0$ and $S_{2v}^* > 0$; their expressions are too lengthy, hence not presented here)

$$\mathbb{E}_0 : \ (S_{1u}^*, S_{1v}^*, S_{2u}^*, S_{2v}^*, E_1^*, E_2^*, P_1^*, P_2^*, I_1^*, I_2^*, A_1^*, A_2^*, H_1^*, H_2^*, R_1^*, R_2^*) =$$

$$\left(\frac{\Pi + \omega_v S_{1v}^* + \alpha_{21} S_{2u}^*}{\alpha_{12} + \xi_v + \mu}, \ \frac{\xi_v S_{1u}^* + \alpha_{21} S_{2v}^*}{\alpha_{12} + \omega_v + \mu}, \ \frac{\omega_v S_{2v}^* + \alpha_{12} S_{1u}^*}{\alpha_{21} + \xi_v + \mu}, \ \frac{\xi_v S_{2u}^* + \alpha_{12} S_{1v}^*}{\alpha_{21} + \omega_v + \mu}, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0 \right).$$

The local asymptotic stability property of the DFE ($\mathbb{E}_0$) can be explored using the *next generation operator* method [28, 29]. In particular, using the notation in [28], it follows that the associated non-negative matrix (F) of new infection terms, and the M-matrix (V), of the linear transition terms in the infected compartments, are given, respectively, by (where the entries f_i and g_i, $i = 1, \cdots, 8$, of the non-negative matrix F, are given in Appendix I):

$$F = \begin{bmatrix} 0 & f_1 & f_2 & f_3 & f_4 & 0 & f_5 & f_6 & f_7 & f_8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & g_1 & g_2 & g_3 & g_4 & 0 & g_5 & g_6 & g_7 & g_8 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and,

$$V = \begin{bmatrix} K_1 & 0 & 0 & 0 & 0 & -\alpha_{21} & 0 & 0 & 0 & 0 \\ -\sigma_1 & K_2 & 0 & 0 & 0 & 0 & -\alpha_{21} & 0 & 0 & 0 \\ 0 & -r\sigma_p & K_3 & 0 & 0 & 0 & 0 & -\alpha_{21} & 0 & 0 \\ 0 & -(1-r)\sigma_p & 0 & K_4 & 0 & 0 & 0 & 0 & -\alpha_{21} & 0 \\ 0 & 0 & -\phi_{1I} & 0 & K_5 & 0 & 0 & 0 & 0 & -\alpha_{21} \\ -\alpha_{12} & 0 & 0 & 0 & 0 & K_6 & 0 & 0 & 0 & 0 \\ 0 & -\alpha_{12} & 0 & 0 & 0 & 0 & K_7 & 0 & 0 & 0 \\ 0 & 0 & -\alpha_{12} & 0 & 0 & 0 & -q\sigma_p & K_8 & 0 & 0 \\ 0 & 0 & 0 & -\alpha_{12} & 0 & 0 & -(1-q)\sigma_p & 0 & K_9 & 0 \\ 0 & 0 & 0 & 0 & -\alpha_{12} & 0 & 0 & -\phi_{2I} & 0 & K_{10} \end{bmatrix},$$

where $K_1 = \alpha_{12} + \sigma_1 + \mu, K_2 = \alpha_{12} + \sigma_P + \mu, K_3 = \alpha_{12} + \phi_{1I} + \gamma_{1I} + \mu + \delta_{1I}, K_4 = \alpha_{12} + \gamma_{1A} + \mu, K_5 = \alpha_{12} + \gamma_{1H} + \mu + \delta_{1H}, K_6 = \alpha_{21} + \sigma_2 + \mu, K_7 = \alpha_{21} + \sigma_P + \mu, K_8 = \alpha_{21} + \phi_{2I} + \gamma_{2I} + \mu + \delta_{2I}, K_9 = \alpha_{21} + \gamma_{2A} + \mu$ and $K_{10} = \alpha_{21} + \gamma_{2H} + \mu + \delta_{2H}$.

For mathematical tractability, the computations will be carried out for the special case of the model $\{(2.1), (2.2)\}$ in the absence of the back-and-forth transitions between the no-mask and mask-user groups (i.e., the special case of the model with $\alpha_{12} = \alpha_{21} = 0$). Hence, from now on, we set $\alpha_{12} = \alpha_{21} = 0$. It follows that the *control reproduction number* of the model $\{(2.1), (2.2)\}$ (with $\alpha_{12} = \alpha_{21} = 0$), denoted by $\mathcal{R}_c$, is given by (where ρ is the spectral radius):

$$\mathcal{R}_c = \rho(FV^{-1}) = \max\{\mathcal{R}_{c_1},\ \mathcal{R}_{c_2}\}, \tag{3.4}$$

where,

$$\mathcal{R}_{c_1} = (a_{11} + a_{22}) + \sqrt{(a_{22} - a_{11})^2 + 4\,a_{21}a_{12}},\ \text{and}\ \mathcal{R}_{c_2} = (a_{11} + a_{22}) - \sqrt{(a_{22} - a_{11})^2 + 4\,a_{21}a_{12}}, \tag{3.5}$$

with $a_{11},\ a_{12},\ , a_{21}$ and a_{22} defined in Appendix II. The result below follows from Theorem 2 of [28].

Theorem 3.2. *The DFE* ($\mathbb{E}_0$) *of the special case of the model* $\{(2.1), (2.2)\}$, *with* $\alpha_{12} = \alpha_{21} = 0$, *is locally-asymptotically stable if* $\mathcal{R}_c < 1$, *and unstable If* $\mathcal{R}_c > 1$.

The threshold quantity $\mathcal{R}_c$ is the *control reproduction number* of the model $\{(2.1), (2.2)\}$. It measures the average number of new COVID-19 cases generated by a typical infectious individual introduced into a population where a certain fraction of the population is protected (*via* the use of interventions, such as face mask, social-distancing and/or vaccination). The epidemiological implication of Theorem 3.2 is that a small influx of COVID-19 cases will not generate an outbreak in the community if the control reproduction number ($\mathcal{R}_c$) is brought to, and maintained at a, value less than unity.

In the absence of public health interventions (i.e., in the absence of vaccination, face mask usage and social-distancing), the control reproduction number ($\mathcal{R}_c$) reduces to the *basic reproduction number* (denoted by $\mathcal{R}_0$), given by:

$$\mathcal{R}_0 = \mathcal{R}_c|_{c_s=\varepsilon_0=\varepsilon_i=\varepsilon_v=S^*_{1v}=S^*_{2v}=0} = \max\{\mathcal{R}_1,\ \mathcal{R}_2\}, \tag{3.6}$$

where,

$$\mathcal{R}_1 = (b_{11} + b_{22}) + \sqrt{(b_{22} - b_{11})^2 + 4\,b_{21}b_{12}},\ \text{and}\ \mathcal{R}_2 = (b_{11} + b_{22}) - \sqrt{(b_{22} - b_{11})^2 + 4\,b_{21}b_{12}}, \tag{3.7}$$

with $b_{11},\ b_{12},\ b_{21}$ and b_{22} defined in Appendix II.

3.2 Vaccine-induced Herd Immunity Threshold

Herd immunity is a measure of the minimum percentage of the number of individuals in a community that is susceptible to a disease that need to be protected (i.e., become immune) so that the disease can be eliminated from the population. There are two main ways to achieve herd immunity, namely through acquisition of natural immunity (following natural recovery from infection with the disease) or by vaccination. Vaccination is the safest and fastest way to achieve herd immunity [30, 31]. For vaccine-preventable diseases, such as COVID-19, not every susceptible member of the community can be vaccinated, for numerous reasons (such as individuals with certain underlying medical conditions, infants, pregnant women, or those who opt out of being vaccinated for various reasons etc.) [8]. So, the question, in the context of vaccine-preventable diseases, is what is the minimum proportion of individuals that can be vaccinated we need to vaccinate in order to achieve herd immunity (so that those individuals that cannot be vaccinated will become protected owing to the community-wide herd-immunity). In this section, a condition for achieving vaccine-derived herd immunity in the U.S. will be derived.

It is convenient to define (where N_1^* and N_2^* represent the total size of the sub-population of Group 1 and Group 2 at disease-free equilibrium, respectively):

$$q_1 = (1-c_s)\left[\frac{S_{1u}^* + (1-\varepsilon_v)S_{1v}^*}{N_1^*}\right] \text{ and } q_2 = (1-c_s)\left[\frac{S_{2u}^* + (1-\varepsilon_v)S_{2v}^*}{N_2^*}\right]. \tag{3.8}$$

Using Equation (3.8), the expressions for a_{11}, a_{12}, a_{21} and a_{22} in Appendix II can be re-written as:

$$a_{11} = q_1 b_{11},\ \ a_{12} = q_1(1-\varepsilon_0)b_{12},\ \ a_{21} = q_2(1-\varepsilon_i)b_{21},\ \ a_{22} = q_2(1-\varepsilon_i)(1-\varepsilon_0)b_{22}. \tag{3.9}$$

Furthermore, using (3.9) in (3.4) gives:

$$\mathcal{R}_c = [q_1 b_{11} + q_2(1-\varepsilon_i)(1-\varepsilon_0)b_{22}] + \sqrt{[q_2(1-\varepsilon_i)(1-\varepsilon_0)b_{22} - q_1 b_{11}]^2 + 4q_1 q_2 b_{12} b_{21}(1-\varepsilon_i)(1-\varepsilon_0)}. \tag{3.10}$$

Let $f_{1v} = S_{1v}^*/N_1^*$ and $f_{2v} = S_{2v}^*/N_2^*$ be the proportions of susceptible individuals in Groups 1 and 2, respectively, that have been vaccinated at the disease-free equilibrium ($\mathbb{E}_0$). Hence, (3.8) can be re-written (in terms of f_{1v} and f_{2v}) as:

$$q_1 = (1-c_s)(1-f_{1v}\varepsilon_v) \text{ and } q_2 = (1-c_s)(1-f_{2v}\varepsilon_v). \tag{3.11}$$

In order to compute the expression for the herd immunity threshold associated with the model $\{(2.1), (2.2)\}$, it is convenient to let $f_v = \max\{f_{1v},\ f_{2v}\}$. Using this definition in Equation (3.10) gives:

$$\mathcal{R}_c = (1-c_s)(1-f_v\varepsilon_v)\left\{[b_{11} + (1-\varepsilon_i)(1-\varepsilon_0)b_{22}] + \sqrt{[(1-\varepsilon_i)(1-\varepsilon_0)b_{22} - b_{11}]^2 + 4b_{12}b_{21}(1-\varepsilon_i)(1-\varepsilon_0)}\right\}. \tag{3.12}$$

Setting $\mathcal{R}_c$, in Equation (3.12), to unity and solving for f_v gives the herd immunity threshold (denoted by f_v^c):

$$\begin{aligned} f_v &= \frac{1}{\epsilon_v}\left\{1 - \frac{1}{(1-c_s)[b_{11} + (1-\varepsilon_i)(1-\varepsilon_0)b_{22}] + \sqrt{[(1-\varepsilon_i)(1-\varepsilon_0)b_{22} - b_{11}]^2 + 4b_{12}b_{21}(1-\varepsilon_i)(1-\varepsilon_0)}}\right\} \\ &= f_v^c. \end{aligned} \tag{3.13}$$

It follows from (3.13) that $\mathcal{R}_c < (>)1$ if $f_v > (<) f_v^c$. Further, $\mathcal{R}_c = 1$ whenever $f_v = f_v^c$. This result is summarized below:

Theorem 3.3. *Consider the special case of the model $\{(2.1), (2.2)\}$ with $\alpha_{12} = \alpha_{21} = 0$. Vaccine-induced herd immunity (i.e., COVID-19 elimination) can be achieved in the U.S., using an imperfect anti-COVID vaccine, if $f_v > f_v^c$ (i.e., if $\mathcal{R}_c < 1$). If $f_v < f_v^c$ (i.e., if $\mathcal{R}_c > 1$), then the vaccination program will fail to eliminate the COVID-19 pandemic in the U.S.*

The epidemiological implication of Theorem 3.3 is that the use of an imperfect anti-COVID vaccine can lead to the elimination of the COVID-19 pandemic in the U.S. if the sufficient number of individuals residing in the U.S. is vaccinated, such that $f_v > f_v^c$. The Vaccination program will fail to eliminate the pandemic if the vaccine coverage level is below the aforementioned herd immunity threshold (i.e., if $f_v < f_v^c$). Although vaccination, no matter the coverage level, is always useful (i.e., vaccination will always reduce the associated reproduction number, $\mathcal{R}_c$, thereby reducing disease burden, even if the program is unable to bring the reproduction number to a value less than unity), elimination can only be achieved if the herd immunity threshold is reached (i.e., disease elimination is only feasible if the associated reproduction number of the model is reduced to, and maintained at, a value less than unity). The pandemic will persist in the U.S. if $\mathcal{R}_c > 1$. Figure 3 depicts the cumulative mortality of COVID-19 in the U.S. for various steady-state vaccination coverage levels (denoted by f_v). This figure shows a decrease in cumulative mortality with increasing vaccination coverage. In particular, a marked decrease in cumulative mortality is recorded when herd immunity (i.e., $f_v > f_v^c$) is reached in the population.

Furthermore, Figure 4 depicts a contour plot of the control reproduction number ($\mathcal{R}_c$) of the model, as a function of vaccination efficacy (ε_v) and coverage (f_v). This figure shows that the reproduction number decreases with increasing values of vaccination efficacy and coverage. For instance, this figure shows that, with the baseline level of social-distancing and face-mask usage in the U.S., although the AstraZeneca vaccine (with estimated efficacy of 75%) can significantly reduces the reproduction number (from $\mathcal{R}_c \approx 4.5$ to about $\mathcal{R}_c \approx 1.5$ (hence, greatly reduce disease burden), it is unable to lead to the elimination of the disease even if every member of the U.S. population is vaccinated. However, such elimination is feasible using the AstraZeneca vaccine if the coverage level of social-distancing is increased from the baseline (Table 2). For instance, if 60% of the U.S. population observe social-distancing in public, the AstraZeneca vaccine can lead to COVID-19 elimination in the U.S. if about 89% of the populace is vaccinated. The vaccination coverage needed to achieve elimination (using AstraZeneca vaccine) decreases to a mere 35% if 80% of Americans will socially-distant in public (Table 2). If, on the other hand, either the Moderna or Pfizer vaccine (with estimated efficacy of about 95%) is used, Figure 4 shows that, based on the current baseline level of social-distancing coverage, vaccinating about 83% of the population will lead to the elimination of the pandemic in the U.S. The vaccine coverage level needed to eliminate the pandemic (using either of the Pfizer or Moderna vaccine) dramatically decreases to 26% if 80% social-distancing coverage can be reached (Table 2).

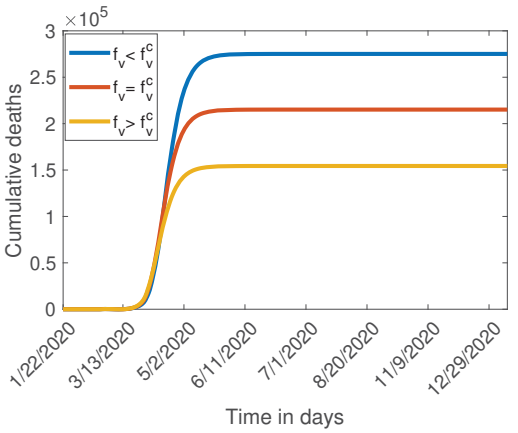


Figure 3: Simulations of the special case of the model $\{(2.1), (2.2)\}$, with $\alpha_{12} = \alpha_{21} = 0$, showing the cumulative COVID-19 mortality in the U.S., as a function of time. (a) $f_v < f_v^c\,(r = 0.5)$ (b) $f_v = f_v^c\,(r = 0.7)$ (c) $f_v > f_v^c\,(r = 0.9)$. Other parameter values used in the simulations are as given in Table 1, with $\alpha_{12} = \alpha_{21} = 0$.

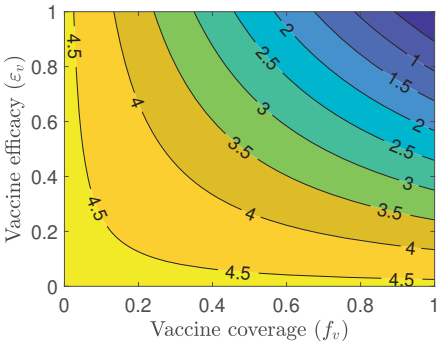


Figure 4: Contour plot of the control reproduction number ($\mathcal{R}_c$) of the model $\{(2.1), (2.2)\}$, with $\alpha_{12} = \alpha_{21} = 0$, as a function of vaccine coverage (f_v) and vaccine efficacy (ε_v), for the US. Parameter values used are as given in Table 1, with $\alpha_{12} = \alpha_{21} = 0$.

Table 2: Herd immunity threshold (f_v) for the U.S. for various levels of social-distancing coverage (c_s). Parameter values used are as given in Table 1, with $\alpha_{12} = \alpha_{21} = 0$.

	Herd threshold	**Herd threshold**	**Herd threshold**	**Herd threshold**
Vaccine type (efficacy)	$c_s = 30\%$	$c_s = 40\%$	$c_s = 60\%$	$c_s = 80\%$
AstraZeneca ($\varepsilon_v = 70\%$)	$f_v = 112\%$	$f_v = 107\%$	$f_v = 89.1\%$	$f_v = 35.3\%$
Pfizer & Moderna ($\varepsilon_v = 95\%$)	$f_v = 82.5\%$	$f_v = 78.9\%$	$f_v = 65.7\%$	$f_v = 26\%$

4 Numerical Simulations: Assessment of Control Strategies

The model $\{(2.1), (2.2)\}$ will now be simulated to assess the population-level impact of the various intervention strategies described in this study. In particular, the singular and combined impact of social-distancing, face mask usage and the three candidate vaccines (by AstraZeneca, Moderna and Pfizer) on curtailing (or eliminating) the burden of the COVID-19 pandemic in the U.S. will be assessed. Unless otherwise stated, the simulations will be carried out using the estimated and baseline values of the parameters of the model tabulated in Table 1. Further, the baseline initial condition for the face mask use group (assumed to be 1%) will be used.

4.1 Assessing the impact of mask-use

The model (2.1)-(2.2) is simulated to assess the community-wide impact of using face-masks alone on the pandemic in the United States. Specifically, we simulate the model using the baseline values of the parameters in Table 1 and various initial values of the number of individuals who habitually wear face masks in public, right from the very beginning of the pandemic (denoted by $N_2(0)$). It should be noted that the parameters associated with other interventions (e.g., vaccination-related and social-distancing-related parameters) are kept at their baseline values given in Table 1. The simulation results obtained, depicted in Figure 5, generally show that the early adoption of face masks measures play a vital role in curtailing the COVID-related mortality in the U.S., particularly for the case when mask-wearers do not opt to give up mask wearing (i.e., when $\alpha_{21} \neq 0$). For the case where the parameters associated with the back-and-forth transitions between the masking and non-masking groups (i.e., α_{12} and α_{21}) are maintained at their baseline values (given in Table 1), this figure shows that the size of the initial number of individuals who wear face masks, right from the beginning of the pandemic, has only marginal impact on the cumulative COVID-related mortality in the U.S., as measured in relation to the cumulative mortality recorded when the initial population of mask wearers is at the 1% baseline level (Figure 5 (a)). On the other hand, for the case when mask-wearers remain mask-wearers since the very beginning of the pandemic (i.e., $\alpha_{21} = 0$), while

non-maskers in Group 1 can change their behavior and become mask-wearers (i.e., $\alpha_{12} \neq 0$), the initial number of individuals who adopt masking from the beginning of the pandemic has a more pronounced effect on the cumulative mortality (Figure 5 (b)), in relation to the baseline. In particular, if 25% of the U.S. population adopt mask-wearing right from the beginning of the pandemic (and remain mask-wearers), up to 6% of COVID-related mortality can be averted, in relation to the 1% baseline mask-wearing at the beginning of the pandemic (green curve, Figure 5 (b)). Further, the reduction in cumulative mortality rises to 11% (in relation to the baseline) could be achieved if half of Americans opted to wear face masks since the very beginning of the pandemic (blue curve, Figure 5 (b)). For the case when no back-and-forth transitions between the two (mask-wearing and non-mask-wearing) groups is allowed (i.e., when $\alpha_{12} = \alpha_{21} = 0$), our simulations show a far more dramatic effect of face mask usage in reducing COVID-19 mortality (Figure 5 (c)). In particular, we showed that up to 92% cumulative mortality can be averted, in comparison to the baseline, if 25% of the U.S. population adopted mask-wearing mandate right from the beginning of the pandemic (green curve, Figure 5 (c)). Furthermore, 95% of the cumulative mortality could have been prevented if 50% of the U.S. population were wearing face masks since the beginning of the pandemic (blue curve, Figure 5 (c)).

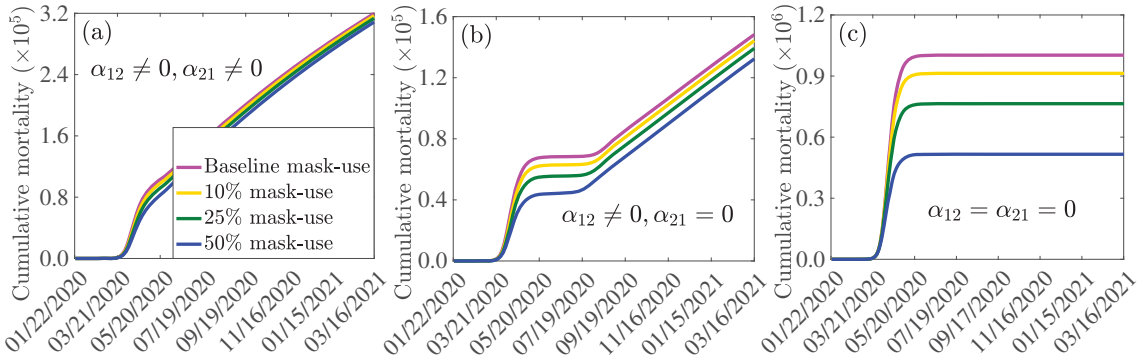


Figure 5: Assessment of the singular impact of face mask usage on COVID-19 pandemic in the U.S. Simulations of the model (2.1)-(2.2), showing cumulative COVID-induced mortality, as a function of time, for (a) face mask transition parameters (α_{12} and α_{21}) maintained at their baseline values, (b) mask-wearers strictly adhere to wearing masks ($\alpha_{21} = 0$) and non-mask-wearers transit to mask wearing at their baseline rate ($\alpha_{12} \neq 0$), (c) non-mask wearers and mask-wearers do not change their behavior (i.e., $\alpha_{12} = \alpha_{21} = 0$). Mask use change is implemented in terms of changes in the initial population of individuals who wear face-masks (i.e., in terms of changes in the initial total population size in Group 2, $N_2(0)$). Parameter values used in the simulations are as given by the baseline values in Table 1, with different values of α_{12} and α_{21}.

4.2 Assessing the impact of social-distancing

In this section, we carry out numerical simulations to assess whether social-distancing alone (implemented right from the very beginning of the pandemic) might be sufficient to contain the COVID-19 pandemic in the U.S. To achieve this objective, the model $\{(2.1), (2.2)\}$ is simulated using the parameters in Table 1 with various levels of the social-distancing compliance parameter (c_s) and all other control-related parameters (e.g., initial face mask coverage and efficacy, vaccination rate and efficacy etc.) are maintained at their baseline values.

The simulation results obtained, depicted in Figure 6, show that the cumulative mortality (Figure 6 (a)) and daily mortality (Figure 6 (b)) decrease with increasing social-distancing compliance. In the absence of social-distancing (i.e., $c_s = 0$), the simulations show that the U.S. could record up to $422,013$ cumulative deaths by September 12, 2021 (Figure 6 (a), red curve). For this (social-distancing-free) scenario, the U.S. will record a peak daily mortality of about $6,585$ deaths on March 21, 2020 (Figure 6 (b), red curve). It is further shown that, if 30% of the U.S. population will be observing social-distancing in public, up to 24% reduction can be recorded in the cumulative mortality, in relation to the cumulative mortality recording for the social-distancing-

free scenario (Figure 6 (a), magenta curve). Similarly, up to 51% reduction can be achieved in daily mortality (Figure 6 (b), magenta curve), and the pandemic would have peaked a month later (in April 2020; the daily mortality at this peak would have been $3,247$). Further dramatic reduction in COVID-19 mortality is recorded as social-distancing compliance is further increased. For instance, if 60% of the U.S. population adhere to the social-distancing measures, about 62% of the cumulative deaths recorded (for the case with $c_s = 0$) would have been averted (Figure 6 (a), green curve). For this scenario, 87% of the daily deaths would have been prevented and the pandemic would have peaked in June 2020 (the daily mortality at this peak would have been 864). Finally, if 75% of the U.S. population complied with the social-distancing measures, right from the beginning of the pandemic, the COVID-19 pandemic would have failed to generate a major outbreak in the U.S. (Figure 6, blue curves). In particular, the cumulative mortality for the entire U.S. by September 21, 2021 will be about $20,000$. Thus, in summary, the simulations in Figure 6 show that COVID-19 could have been effectively suppressed in the U.S. using social-distancing at moderate to high compliance levels.

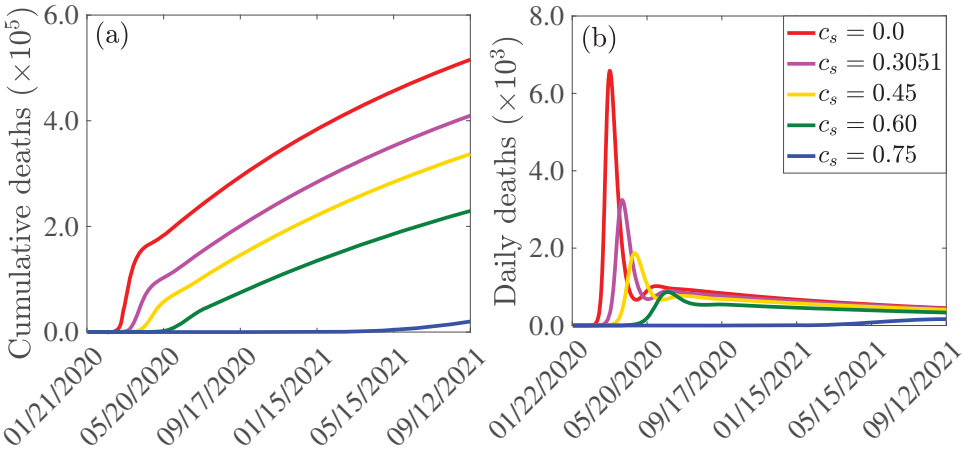


Figure 6: Assessment of the singular impact of social-distancing on COVID-19 pandemic in the U.S. Simulations of the model (2.1)-(2.2) showing (a) cumulative mortality, as a function of time; (b) daily mortality, as a function of time, for various compliance levels of the social-distancing parameter (c_s). Initial conditions used are: $(S_1^0, E_1^0, P_1^0, I_1^0, A_1^0, H_1^0, R_1^0, S_2^0, E_2^0, P_2^0, I_2^0, A_2^0, H_2^0, R_2^0) = (0.99 \times 336218660 - 1, 0, 0, 1, 0, 0, 0, 0.01 \times 336218660, 0, 0, 0, 0, 0, 0)$. Parameter values used in the simulations are as given by the baseline values in Table 1.

4.3 Assessment of combined impact of vaccination and social-distancing

The model (2.1)-(2.2) will now be simulated to assess the community-wide impact of the combined vaccination and social-distancing interventions. Although the vaccines are expected to be available by the end of the year 2020 or early in 2021, we assume that there will be some time lag before the vaccines are made widely available to the general public. This is because the vaccines will initially be targeted to the people most at risk (notably the frontline healthcare workers, nursing home residents and staff, essential workers, people with underlying conditions etc.) before being made available to the general. For simulation purposes, we assume that the vaccines will be available to the general public by March 15, 2021.

We consider the three vaccines currently on the verge of being approved by the FDA for use in humans, namely the AstraZeneca vaccine (with estimated efficacy of 70%) and the Moderna and Pfizer vaccines (each with estimated efficacy of about 95%). Simulations are carried out using the baseline parameter values in Table 1, with various values of the vaccination coverage parameter (ξ_v). For these simulations, parameters and initial conditions related to the other intervention (face mask usage) are maintained at their baseline values. Since the Moderna and Pfizer vaccines have essentially the same estimated efficacy ($\approx 95\%$), we group them together in the simulations.

The simulation results obtained for the AstraZeneca vaccine, depicted in Figures 7 (a)-(c)), show that, in the absence of vaccination (and with social-distancing at baseline compliance level), approximately $1,388$ will be recorded on August 31, 2021 (red curves of Figures 7 (a)-(c)). Furthermore, this figure shows a marked reduction in daily mortality with increasing vaccination coverage (ξ_v). This reduction further increases if vaccination is combined with social-distancing (particularly with high enough compliance). For instance, with social-distancing compliance maintained at its baseline value ($c_s = 0.3015$), vaccinating at a rate of 0.00074 *per* day (which roughly translates to vaccinating $250,000$ people every day) resulted in a reduction of the projected daily mortality on August 31, 2021 by 14% (in comparison to the case when no vaccination is used; magenta curve in Figure 7 (a)). In fact, up to 78% of the projected daily mortality for August 31, 2021 could be averted if, for this vaccination rate, 60% social-distancing compliance is attained (magenta curve in Figure 7 (c)). If the vaccination rate is further increased to, for instance, $\xi_v = 0.0015$ *per* day (corresponding to vaccinating about $500,000$ people every day), our simulations show a reduction of 26% in the projected daily mortality on August 31, 2020 if social-distancing is maintained at its baseline level (gold curve, Figure 7 (a)). This reduction increases to 85% if the vaccination program is supplemented with social-distancing with 60% compliance (gold curve, Figures 7 (c)). If 1 million people are vaccinated *per* day (i.e., $\xi_v = 0.003$) *per* day, our simulations show that the use of AstraZeneca vaccine could lead to up to 46% reduction in the projected daily mortality on August 31, 2021 if the vaccination program is combined with social-distancing at baseline compliance level. Further reductions in the projected daily mortality are recorded when either the Moderna or Pfizer vaccine (with moderate to high vaccination coverage) is used (Figures 7 (d)-(f)), particularly if combined with social-distancing with high compliance (blue curves in Figures 7 (d)-(f)). These results are summarized in Table 3.

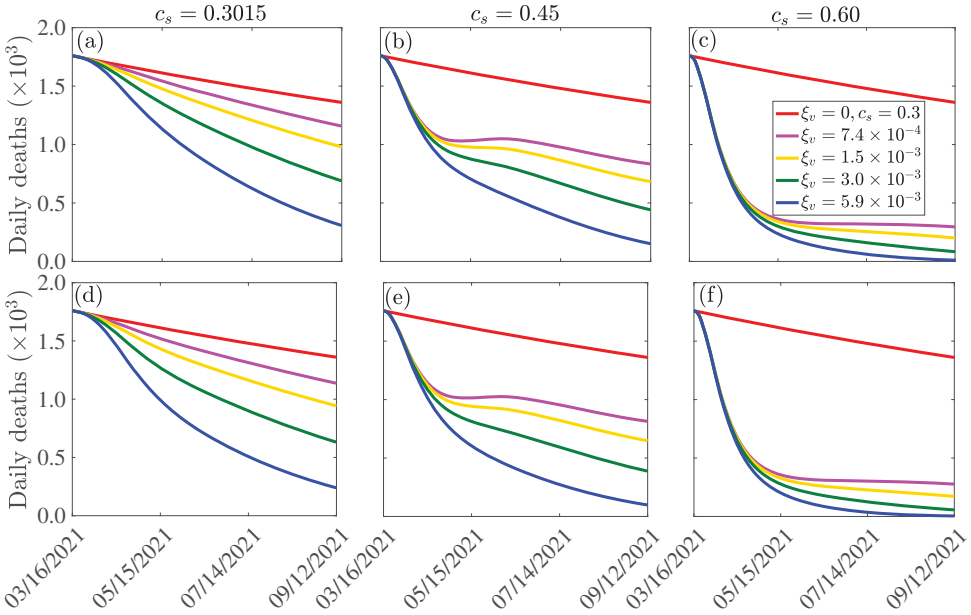


Figure 7: Assessment of the combined impact of vaccination and social-distancing on daily mortality. Simulations of the model (2.1)-(2.2), depicting daily mortality as a function of time, for various vaccine types and social-distancing compliance (c_s). (a)-(c): AstraZeneca vaccine. (d)-(f): Pfizer or Moderna vaccine. The vaccination rates $\xi_v = 7.4 \times 10^{-4}, 1.5 \times 10^{-3}, 3.0 \times 10^{-3}, 5.9 \times 10^{-3}$ correspond, respectively, to vaccinating approximately $2.5 \times 10^5, 5.0 \times 10^5, 1.0 \times 10^6, 2.0 \times 10^6$ people *per* day. Other parameter values of the model are as presented in Table 1.

Table 3: Percentage reduction in projected daily mortality on August 31, 2021, in relation to the daily mortality in the absence of vaccination ($1,383$ COVID-19 deaths on August 31, 2021), for different types of COVID-19 vaccines: AstraZeneca vaccine (efficacy $\varepsilon_v = 0.7$); Pfizer and/or Moderna vaccine (efficacy $\varepsilon_v = 0.95$), and various compliance levels of social-distancing (c_s) and number of individuals vaccinated *per* day.

Number of people vaccinated *per* day	**Reduction with** $c_s = 0.3051$		**Reduction with** $c_s = 0.45$		**Reduction with** $c_s = 0.60$	
	$\varepsilon_v = 70\%$	$\varepsilon_v = 95\%$	$\varepsilon_v = 70\%$	$\varepsilon_v = 95\%$	$\varepsilon_v = 70\%$	$\varepsilon_v = 95\%$
250,000	14%	15%	38%	39%	78%	79%
500,000	26%	29%	48%	51%	85%	87%
1,000,000	46%	51%	65%	69%	93%	95%
2,000,000	74%	80%	87%	91%	98.8%	99.5%

4.4 Impact of vaccination and social-distancing on time to pandemic elimination

The model (2.1)-(2.2) will now be simulated to assess the community-wide impact of the combined vaccination and social-distancing interventions on the expected time the implementation of these interventions will take to result in the elimination of the pandemic in the U.S. (i.e., time needed for the number of new COVID-19 cases to be essentially zero). As in Section 4.3, we consider the three candidate vaccines (the AstraZeneca, Moderna and the Pfizer vaccines). The model is simulated to generate a time series of new daily COVID-19 cases in the U.S., for various vaccination coverage and social-distancing compliance levels. The results obtained, for the AstraZeneca vaccine, depicted in Figures 8 (a)-(c), show a marked decrease in the time-to-elimination with increasing vaccination coverage and social-distancing compliance. In particular, vaccinating $250,000$ people *per* day, with the AstraZeneca vaccine, will result in COVID-19 elimination in the U.S. by late October of 2025, if the social-distancing compliance is kept at its current baseline level of 30.51% (red curve of Figure 8 (a)). For this scenario, the elimination will be reached in early October 2025 using either the Moderna or the Pfizer vaccine. If the vaccination rate is further increases, such as vaccinating 1 million people every day, COVID-19 elimination is achieved much sooner. For instance, for this scenario (i.e., $\xi_v = 0.003$ *per* day), the pandemic can be eliminated, using the AstraZeneca vaccine, by mid July of 2022 if the vaccination program is combined with social-distancing at 60% compliance (green curve of Figure 8 (c)). Here, too, using the Moderna or the Pfizer vaccine can lead to the elimination of the pandemic a little sooner (by mid June 2022) if social-distancing is maintained at 60% (green curve, Figure 8 (f)). A summary of time-to-elimination for the aforementioned, and other, scenarios is given in Table 4. In conclusions, these simulations show that any of the three candidate vaccines considered in this study will lead to the elimination of the U.S. The time-to-elimination depends on the vaccination rate and the compliance level of social-distancing. The pandemic can be eliminated by as early as June of 2022 if moderate to high vaccination rate (e.g., 1 million people are vaccinated *per* day) and social-distancing compliance (e.g., $c_s = 0.6$) are attained and maintained.

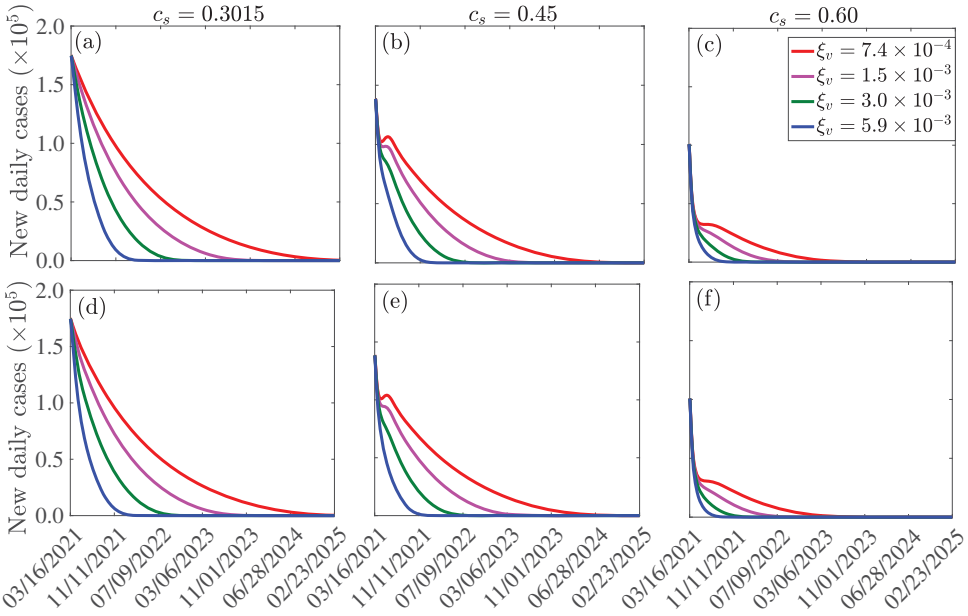


Figure 8: Effect of vaccination and social-distancing on time-to-elimination. Simulations of the model (2.1)-(2.2), depicting the impact of three candidate vaccines against COVID-19 (the AstraZeneca vaccine, and the Pfizer or Moderna vaccine) and social-distancing, on time-to-elimination of the pandemic in the U.S. (a)-(c): AstraZeneca vaccine. (d)-(f): Moderna or Pfizer vaccine. The social-distancing compliance is $c_s = 0.3051$ for (a) and (d), $c_s = 0.45$ for (b) and (e), and $c_s = 0.60$ for (e) and (f). The vaccination rates $\xi_v = 7.4 \times 10^{-4}, 1.5 \times 10^{-3}, 3.0 \times 10^{-3}, 5.9 \times 10^{-3}$ correspond, respectively, to vaccinating approximately $2.5 \times 10^5, 5.0 \times 10^5, 1.0 \times 10^6, 2.0 \times 10^6$ people *per* day. The values of the other parameters of the model used in the simulation are as given in Table 1.

Table 4: Time to eliminate the COVID-19 pandemic in the U.S., for various values of the vaccination rate (ξ_v) using the three candidate vaccines (AstraZeneca vaccine with efficacy $\varepsilon_v = 70\%$; the Moderna or Pfizer vaccine with efficacy $\varepsilon_v = 95\%$) and various levels of social-distancing compliance (c_s). Parameter values used are as given in Table 1.

Number of people vaccinated *per* day	**Reduction with** $c_s = 0.3051$		**Reduction with** $c_s = 0.45$		**Reduction with** $c_s = 0.60$	
	$\varepsilon_v = 70\%$	$\varepsilon_v = 95\%$	$\varepsilon_v = 70\%$	$\varepsilon_v = 95\%$	$\varepsilon_v = 70\%$	$\varepsilon_v = 95\%$
250,000	10/26/2025	10/09/2025	07/08/2025	07/15/2025	07/03/2024	06/06/2024
500,000	05/19/2024	05/14/2024	01/10/2024	12/26/2023	04/25/2023	04/02/2023
1,000,000	03/26/2023	03/06/2023	12/31/2022	12/11/2022	07/14/2022	06/18/2022
2,000,000	06/24/2022	06/09/2022	04/24/2022	04/05/2022	01/14/2022	12/21/2021

5 Discussion and Conclusions

Since its emergence late in December of 2019, the novel Coronavirus pandemic continues to inflict devastating public health and economic burden across the world. As of December 5, 2020, the pandemic accounted for over 67 million confirmed cases and over 1.5 million deaths globally. Although control efforts against the pandemic have focused on the use of non-pharmaceutical interventions, such as social-distancing, face mask usage, quarantine, self-isolation, contact-tracing, community lockdowns, etc., a number of highly promising (safe and highly-efficacious)

anti-COVID vaccines are currently on the verge of being approved by the Food and Drug Administration (FDA) for use in humans. In particular, two vaccine manufacturers, Moderna Inc. and Pfizer Inc., filed for Emergency Use Authorization with the FDA in November 2020 (each of the two vaccines has estimated protective efficacy of about 95%). Furthermore, AstraZeneca vaccine, developed by the pharmaceutical giant, AstraZeneca, and University of Oxford, UK, is undergoing Phase III of clinical trials with very promising results (estimated efficacy of 70%). Mathematics (modeling, analysis and data analytics) has historically been used to provide robust insight into the transmission dynamics and control of infectious diseases, dating back to the pioneering works of the likes of Daniel Bernoulli in the 1760s (on smallpox immunization), Sir Ronald Ross and George Macdonald between the 1920s and 1950s (on malaria modeling) and the compartmental modeling framework developed by Kermack and McKendrick in the 1920s [32–34]. The purpose of our study is to use mathematical modeling approaches, coupled with rigorous analysis, to assess the potential population-level impact of the wide scale deployment of any (or combination of) the aforementioned candidate vaccines in curtailing the burden of the COVID-19 pandemic in the U.S. We also seek to assess the impact of other non-pharmaceutical interventions, such as face mask and social-distancing, implemented singly or in combination with any of the three vaccines, on the dynamics and control of the pandemic in the U.S.

We developed a novel mathematical model, which stratifies the total population into two subgroups of individuals who habitually wear face masks in public and those who do not. The resulting two group COVID-19 vaccination model, which takes the form of a deterministic system of nonlinear differential equations, was initially fitted using observed cumulative COVID-induced mortality data for the U.S. The model allows for the assessment of social-distancing measures on combating the spread of the pandemic. The model was then rigorously analysed to gain insight into its dynamical features. In particular, we showed that the disease-free equilibrium of the model is locally-asymptotically stable whenever a certain epidemiological threshold, known as the *control reproduction number* (denoted by $\mathcal{R}_c$), is less than unity. The implication of this result is that (for the case when $\mathcal{R}_c < 1$), a small influx of COVID-infected individuals will not generate an outbreak in the community.

The expression for the reproduction number ($\mathcal{R}_c$) was used to compute the nationwide vaccine-induced *herd immunity* threshold. The herd immunity threshold represents the minimum proportion of the U.S. population that needs to be vaccinated to ensure elimination of the pandemic. Simulations of our model shows, for the current baseline level of social-distancing in the U.S. (at 30%), herd immunity cannot be achieved in the U.S. using the AstraZeneca vaccine. However, achieving such herd immunity threshold is feasible using either the Moderna or the Pfizer vaccine if at least 83% of the U.S. residents are vaccinated. Our simulations further showed that the level of herd immunity needed to eliminate the pandemic decreases, for each of the three vaccines, with increasing social-distancing compliance. In particular, if 80% of American residents adhere to social-distancing, vaccinating only 35% and 26% with the AstraZeneca or Moderna/Pfizer vaccine, respectively, will generate the desired herd immunity. In other words, this study shows that the prospect of achieving vaccine-derived herd immunity, using any of the three candidate vaccines, is very promising, particularly if the vaccination program is complemented with social-distancing measures with moderate to high compliance levels.

This study also shows that the use of any of the three vaccines (i.e., the AstraZeneca, Pfizer, or Moderna vaccine) will dramatically reduce the burden of the COVID-19 pandemic in the U.S. (as measured in terms of cumulative or daily COVID-induced mortality). The level of reduction achieved increases with increasing daily vaccination coverage. Furthermore, the effectiveness of the vaccination program (using any of the three candidate vaccines), to reduce COVID-19 burden, is significantly enhanced if the vaccination program is complemented with other interventions, such as social-distancing (at moderate to high compliance levels). Our study further highlights the fact that early implementation of masks adoption (i.e., face mask adoption from the very beginning of the pandemic) plays a crucial role in effectively combating the burden of the COVID-19 pandemic (as measured in terms of reduction in cumulative COVID-related mortality) in the U.S. It was further shown that the level of such reduction is very sensitive to the rate at which mask-wearers opt to abandon mask-wearing (i.e., reverting to the group of non-mask wearers). In other words, our study emphasize the fact that early implementation or adoption of mask mandate, together with (strict) compliance to this mandate, plays a major role in effectively curtailing, or halting, the COVID-19 pandemic in the U.S.

We further showed that the time-to-elimination of COVID-19 in the U.S., using a vaccine (and a non-pharmaceutical

intervention), depends on the vaccination rate (i.e., number of people vaccinated everyday) and the level of compliance of social-distancing measures in the country. Specifically, our study shows that the COVID-19 pandemic can be eliminated in the U.S. by early June of 2022 if moderate to high vaccination rate (e.g., 1 million people vaccinated *per* day) and social-distancing compliance (e.g., 60% social-distancing compliance) are achieved and maintained. It should, however, be mentioned that the time-to-elimination is sensitive to the level of community transmission of COVID-19 in the population (it is also sensitive to the effectiveness and coverage (compliance) levels of the other (non-pharmaceutical) interventions, particularly face mask usage and social-distancing compliance, implemented in the community). Specifically, our study was carried out during the months of November and December of 2020, when the United States was experiencing a devastating third wave of the COVID-19 pandemic (recording an average of $200,000$ confirmed cases *per* day, together with record numbers of hospitalizations and COVID-induced mortality). This explains the somewhat *longer* estimated time-to-elimination of the pandemic, using any of the three vaccines, for the case where social-distancing compliance is low. The estimate for the time-to-elimination (using any of the three vaccines) will be shorter if the community transmission is significantly reduced (as will be vividly evident from the reduced values of the transmission- and mortality-related parameters of the re-calibrated version of our model).

Acknowledgments

One of the authors (ABG) acknowledge the support, in part, of the Simons Foundation (Award #585022) and the National Science Foundation (Award #1917512). CNN acknowledges the support of the Simons Foundation (Award #627346).

Table 5: Description of the state variables of the model {(2.1), (2.2)}.

State variable	Description
S_{1u}	Population of non-vaccinated susceptible individuals who do not habitually wear face masks
S_{2u}	Population of non-vaccinated susceptible individuals who habitually face masks
S_{1v}	Population of vaccinated susceptible individuals who do not habitually wear face masks
S_{2v}	Population of vaccinated susceptible individuals who habitually wear face masks
E_1	Population of exposed (newly-infected) individuals who do not habitually wear face masks
E_2	Population of exposed (newly-infected) individuals who habitually wear face masks
P_1	Population of pre-symptomatic infectious individuals who do not habitually wear face masks
P_2	Population of pre-symptomatic infectious individuals who habitually wear face masks
I_1	Population of symptomatically-infectious individuals who do not habitually wear face masks
I_2	Population of symptomatically-infectious individuals who habitually wear face masks
A_1	Population of asymptomatically-infectious individuals who do not habitually wear face masks
A_2	Population of asymptomatically-infectious individuals who habitually wear face masks
H_1	Population of hospitalized individuals who do not habitually wear face masks
H_2	Population of hospitalized individuals who habitually wear face masks
R_1	Population of recovered individuals who do not habitually wear face masks
R_2	Population of recovered individuals who habitually wear face masks

Table 6: Description of the parameters of the model {(2.1), (2.2)}.

Parameters	**Description**
Π	Recruitment (birth or immigration) rate into the population
μ	Natural mortality rate
$\beta_{P1}(\beta_{P2})$	Effective contact rate for pre-symptomatic individuals who do not wear (wear) face masks
$\beta_{I1}(\beta_{I2})$	Effective contact rate for infectious symptomatic individuals who do not wear (wear) face masks
$\beta_{A1}(\beta_{A2})$	Effective contact rate for symptomatically-infectious individuals who do not wear (wear) face masks
$\beta_{H1}(\beta_{H2})$	Effective contact rate for hospitalized individuals who do not wear (wear) face masks
$0 < \epsilon_0 < 1$	Outward protective efficacy of face masks
$0 < \epsilon_i < 1$	Inward protective efficacy of face masks
ω_v	Vaccine waning rate
α_{12}	Rate at which non-habitual face masks wearers choose to become habitual wearers
α_{21}	Rate at which habitual face masks wearers choose to become non-habitual wearers
ξ_v	*Per capita* vaccination rate
$0 < \varepsilon_v < 1$	Protective efficacy of the vaccine
$\sigma_1(\sigma_2)$	Rate at which exposed individuals who do not wear (wear) face masks progress to the corresponding pre-symptomatic infectious stage
σ_P	Rate at which pre-symptomatic infectious individuals progress to symptomatically-infectious or asymptomatically-infectious stage
$r(q)$	Proportion of pre-symptomatic infectious individuals who do not wear (wear) face masks that become symptomatically-infectious
$\phi_{1I}(\phi_{2I})$	Hospitalization rate for symptomatically-infectious individuals who do not wear (wear) face masks
$\gamma_{1A}(\gamma_{2A})$	Recovery rate for asymptomatically-infectious individuals who do not wear (wear) face masks
$\gamma_{1I}(\gamma_{2I})$	Recovery rate for symptomatically-infectious individuals who do not wear (wear) face masks
$\gamma_{1H}(\gamma_{2H})$	Recovery rate for hospitalized individuals who do not wear (wear) face masks
$\delta_{1I}(\delta_{2I})$	Disease-induced mortality rate for symptomatically-infectious individuals who do not wear (wear) face masks
$\delta_{1H}(\delta_{2H})$	Disease-induced mortality rate for hospitalized individuals who do not wear (wear) face masks

Appendix I: Entries of the Non-negative Matrix F

$$f_1 = (1-c_s)\beta_{P_1}\left[\frac{S_{1u}^* + (1-\epsilon_v)S_{1v}^*}{N_1^*}\right], f_2 = (1-c_s)\beta_{I_1}\left[\frac{S_{1u}^* + (1-\epsilon_v)S_{1v}^*}{N_1^*}\right], f_3 = (1-c_s)\beta_{A_1}\left[\frac{S_{1u}^* + (1-\epsilon_v)S_{1v}^*}{N_1^*}\right],$$

$$f_4 = (1-c_s)\beta_{H_1}\left[\frac{S_{1u}^* + (1-\epsilon_v)S_{1v}^*}{N_1^*}\right], f_5 = (1-c_s)(1-\epsilon_0)\beta_{P_2}\left[\frac{S_{1u}^* + (1-\epsilon_v)S_{1v}^*}{N_1^*}\right],$$

$$f_6 = (1-c_s)(1-\epsilon_0)\beta_{I_2}\left[\frac{S_{1u}^* + (1-\epsilon_v)S_{1v}^*}{N_1^*}\right], f_7 = (1-c_s)(1-\epsilon_0)\beta_{A_2}\left[\frac{S_{1u}^* + (1-\epsilon_v)S_{1v}^*}{N_1^*}\right],$$

$$f_8 = (1-c_s)(1-\epsilon_0)\beta_{H_2}\left[\frac{S_{1u}^* + (1-\epsilon_v)S_{1v}^*}{N_2^*}\right], g_1 = (1-c_s)(1-\epsilon_i)\beta_{P_1}\left[\frac{S_{2u}^* + (1-\epsilon_v)S_{2v}^*}{N_2^*}\right],$$

$$g_2 = (1-c_s)(1-\epsilon_i)\beta_{I_1}\left[\frac{S_{2u}^* + (1-\epsilon_v)S_{2v}^*}{N_2^*}\right], g_3 = (1-c_s)(1-\epsilon_i)\beta_{A_1}\left[\frac{S_{2u}^* + (1-\epsilon_v)S_{2v}^*}{N_2^*}\right],$$

$$g_4 = (1-c_s)(1-\epsilon_i)\beta_{H_1}\left[\frac{S_{2u}^* + (1-\epsilon_v)S_{2v}^*}{N_2^*}\right], g_5 = (1-c_s)(1-\epsilon_i)(1-\epsilon_0)\beta_{P_2}\left[\frac{S_{2u}^* + (1-\epsilon_v)S_{2v}^*}{N_2^*}\right],$$

$$g_6 = (1-c_s)(1-\epsilon_i)(1-\epsilon_0)\beta_{I_2}\left[\frac{S_{2u}^* + (1-\epsilon_v)S_{2v}^*}{N_2^*}\right], g_7 = (1-\epsilon_i)(1-\epsilon_0)\beta_{A_2}\left[\frac{S_{2u}^* + (1-\epsilon_v)S_{2v}^*}{N_2^*}\right],$$

$$g_8 = (1-\epsilon_i)(1-\epsilon_0)\beta_{H_2}\left[\frac{S_{2u}^* + (1-\epsilon_v)S_{2v}^*}{N_2^*}\right].$$

Appendix II

$$a_{11} = \frac{K_3K_5K_6K_7K_8K_9K_{10}\sigma_1[f_3\sigma_p(1-r) + f_1K_4] + rK_4K_6K_7K_8K_9K_{10}\sigma_1\sigma_p(f_2K_5 + f_4\phi_1)}{2\prod_{i=1}^{10} K_i},$$

$$a_{12} = \frac{\sigma_1K_1K_2K_3K_4K_5\left[rg_2K_4K_5\sigma_p + (1-r)g_3K_3K_5\sigma_p + rg_4K_4\phi_1\sigma_p + g_1K_3K_4K_5\right]}{4\prod_{i=1}^{10} K_i},$$

$$a_{21} = \frac{\sigma_2K_6K_7K_8K_9K_{10}\left[qf_6K_9K_{10}\sigma_p + (1-q)f_7K_8K_{10}\sigma_p + qf_8K_9\phi_2\sigma_p + f_5K_8K_9K_{10}\right]}{4\prod_{i=1}^{10} K_i},$$

$$a_{22} = \frac{K_1K_2K_3K_4K_5K_8K_{10}\sigma_2[g_7\sigma_p(1-q) + g_5K_9] + qK_1K_2K_3K_4K_5K_9\sigma_2\sigma_p(g_6K_{10} + g_8\phi_2)}{2\prod_{i=1}^{10} K_i}.$$

$$b_{11} = \frac{K_3K_5K_6K_7K_8K_9K_{10}\sigma_1[\beta_{A_1}\sigma_p(1-r)+\beta_{P_1}K_4]+rK_4K_6K_7K_8K_9K_{10}\sigma_1\sigma_p(\beta_{I_1}K_5+\beta_{H_1}\phi_1)}{2\prod_{i=1}^{10}K_i},$$

$$b_{12} = \frac{\sigma_1K_1K_2K_3K_4K_5\left[r\beta_{I_1}K_4K_5\sigma_p+(1-r)\beta_{A_1}K_3K_5\sigma_p+r\beta_{H_1}K_4\phi_1\sigma_p+\beta_{P_1}K_3K_4K_5\right]}{4\prod_{i=1}^{10}K_i},$$

$$b_{21} = \frac{\sigma_2K_6K_7K_8K_9K_{10}\left[q\beta_{I_2}K_9K_{10}\sigma_p+(1-q)\beta_{A_2}K_8K_{10}\sigma_p+q\beta_{H_2}K_9\phi_2\sigma_p+\beta_{P_2}K_8K_9K_{10}\right]}{4\prod_{i=1}^{10}K_i},$$

$$b_{22} = \frac{K_1K_2K_3K_4K_5K_8K_{10}\sigma_2[\beta_{A_2}\sigma_p(1-q)+\beta_{P_2}K_9]+qK_1K_2K_3K_4K_5K_9\sigma_2\sigma_p(\beta_{I_2}K_{10}+\beta_{H_2}\phi_2)}{2\prod_{i=1}^{10}K_i}.$$

References

[1] "Center for Systems Science and Engineering at Johns Hopkins University. COVID-19," (2020).
Online Version

[2] Centers for Disease Control and Prevention, "Coronavirus disease 2019 (COVID-19)," National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases (Accessed on March 4, 2020).
Online Version

[3] C. N. Ngonghala, E. Iboi, S. Eikenberry, M. Scotch, C. R. MacIntyre, M. H. Bonds, and A. B. Gumel, "Mathematical assessment of the impact of non-pharmaceutical interventions on curtailing the 2019 novel coronavirus," Mathematical Biosciences. **325**, 108364 (2020).

[4] Pfizer, "Pfizer and BioNTech to Submit Emergency Use Authorization Request Today to the U.S. FDA for COVID-19 Vaccine," (2020).
Online Version

[5] S. E. Eikenberry, M. Muncuso, E. Iboi, T. Phan, E. Kostelich, Y. Kuang, and A. B. Gumel, "To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic," Infectious Disease Modeling **5**, 293–308 (2020).

[6] C. N. Ngonghala, E. Iboi, S. Eikenberry, M. Scotch, C. R. MacIntyre, M. H. Bonds, and A. B. Gumel, "Mathematical assessment of the impact of non-pharmaceutical interventions on curtailing the 2019 novel coronavirus," Mathematical Biosciences 108364 (2020).

[7] C. N. Ngonghala, E. Iboi, and A. B. Gumel, "Could masks curtail the post-lockdown resurgence of covid-19 in the US?" Mathematical Biosciences **329**, 108452 (2020).

[8] E. A. Iboi, C. N. Ngonghala, and A. B. Gumel, "Will an imperfect vaccine curtail the COVID-19 pandemic in the US?" Infectious Disease Modelling **5**, 510–524 (2020).

[9] National Institute of Health, "Promising Interim Results from Clinical Trial of NIH-Moderna COVID-19 Vaccine," (2020).
Online Version

[10] AstraZeneca, "AZD1222 Vaccine Met Primary Efficacy Endpoint in Preventing COVID-19," (2020).
Online Version

[11] Graham Lawton, "Everything you Need to Know About the Pfizer/BioNTech COVID-19 Vaccine," (2020).
Online Version

[12] Moderna, "Moderna Announces Longer Shelf Life for its COVID-19 Vaccine Candidate at Refrigerated Temperatures," (2020).
Online Version

[13] A. Srivastava and G. Chowell, "Understanding spatial heterogeneity of COVID-19 pandemic using shape analysis of growth rate curves," medRxiv (2020).

[14] J. Hellewell, S. Abbott, A. Gimma, N. I. Bosse, C. I. Jarvis, T. W. Russell, J. D. Munday, A. J. Kucharski, W. J. Edmunds, F. Sun, et al., "Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts," The Lancet Global Health **8**, E488–E496 (2020).

[15] A. J. Kucharski, T. W. Russell, C. Diamond, Y. Liu, J. Edmunds, S. Funk, R. M. Eggo, F. Sun, M. Jit, J. D. Munday, et al., "Early dynamics of transmission and control of COVID-19: a mathematical modelling study," The Lancet Infectious Diseases **20**, 553–558 (2020).

[16] L. Xue, S. Jing, J. C. Miller, W. Sun, H. Li, J. G. Estrada-Franco, J. M. Hyman, and H. Zhu, "A data-driven network model for the emerging covid-19 epidemics in Wuhan, Toronto and Italy," Mathematical Biosciences **326**, 108391 (2020).

[17] N. M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, et al., "Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand," London: Imperial College COVID-19 Response Team, March **16** (2020).

[18] H. T. Banks, M. Davidian, J. R. Samuels, and K. L. Sutton, *An Inverse Problem Statistical Methodology Summary*, 249–302 (Springer Netherlands, Dordrecht, 2009).
Online Version

[19] G. Chowell, "Fitting dynamic models to epidemic outbreaks with quantified uncertainty: a primer for parameter uncertainty, identifiability, and forecasts," Infectious Disease Modelling **2**, 379–398 (2017).

[20] C. Zhou, "Evaluating new evidence in the early dynamics of the novel coronavirus COVID-19 outbreak in Wuhan, China with real time domestic traffic and potential asymptomatic transmissions," medRxiv (2020).

[21] N. M. Linton, T. Kobayashi, Y. Yang, K. Hayashi, A. R. Akhmetzhanov, S.-m. Jung, B. Yuan, R. Kinoshita, and H. Nishiura, "Incubation period and other epidemiological characteristics of 2019 novel coronavirus infections with right truncation: a statistical analysis of publicly available case data," Journal of Clinical Medicine **9**, 538 (2020).

[22] World Health Organization, "Coronavirus disease 2019 (COVID-19): situation report, 46," WHO (2020).

[23] Z. Wu and J. M. McGoogan, "Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China: summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention," JAMA (2020).

[24] S. Kissler, C. Tedijanto, E. Goldstein, Y. Grad, and M. Lipsitch, "Projecting the transmission dynamics of SARS-CoV-2 through the postpandemic period," Science (2020).
Online Version

[25] L. Zou, F. Ruan, M. Huang, L. Liang, H. Huang, Z. Hong, J. Yu, M. Kang, Y. Song, J. Xia, et al., "SARS-CoV-2 viral load in upper respiratory specimens of infected patients," New England Journal of Medicine **382**, 1177–1179 (2020).

[26] V. Lakshmikantham and A. Vatsala, "Theory of differential and integral inequalities with initial time difference and applications," in "Analytic and Geometric Inequalities and Applications," 191–203 (Springer, 1999).

[27] H. W. Hethcote, "The mathematics of infectious diseases," SIAM Review **42**, 599–653 (2000).

[28] P. van den Driessche and J. Watmough, "Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission," Mathematical Biosciences **180**, 29–48 (2002).

[29] O. Diekmann, J. A. P. Heesterbeek, and J. A. Metz, "On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations," Journal of Mathematical Biology **28**, 365–382 (1990).

[30] R. M. Anderson and R. M. May, "Vaccination and herd immunity to infectious diseases," Nature **318**, 323–329 (1985).

[31] R. M. Anderson, "The concept of herd immunity and the design of community-based immunization programmes," Vaccine **10**, 928–935 (1992).

[32] D. Bernoulli, "Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir," Histoire de l'Acad., Roy. Sci. 1–45 (1760).

[33] R. Ross, *The prevention of malaria* (John Murray, 1911).

[34] W. O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character **115**, 700–721 (1927).

AN EXCURSION THROUGH THE LAND OF SHTUKAS

ANA CARAIANI

Abstract. Vincent Lafforgue made a deep breakthrough in the global Langlands program over function fields: he gave a general construction of the "automorphic to Galois" direction of the Langlands correspondence. This connects spectral data attached to Hecke operators on the automorphic side with arithmetic data coming from representations of the absolute Galois group of the function field. Lafforgue dreamed up additional symmetries, known as excursion operators, on the automorphic side, and used them as a guide towards the correct Galois representation. The goal of this survey is to explain this result and some key ingredients in its proof. We also mention several exciting, even more recent developments in the field.

1. Introduction

The Langlands program is a "grand unified theory" of mathematics: an intricate network of conjectures that touch on number theory, representation theory, harmonic analysis, and even parts of theoretical physics. At its heart lies the principle of reciprocity, or the *Langlands correspondence*, which is like a magical bridge that connects different mathematical worlds.

The principle of reciprocity goes back to the eighteenth century to the celebrated *law of quadratic reciprocity*, discovered by Euler and Legendre and proved by Gauss. We can use this law to answer basic questions about whole numbers.

Question 1.1. Let $\ell \geq 5$ be a prime number. Can 3 be the last digit of a perfect square in base ℓ?

If the answer is "Yes" we say that 3 is a *quadratic residue* modulo ℓ. If p and ℓ are distinct odd prime numbers, we can define the Legendre symbol

$$\left(\frac{p}{\ell}\right) = \begin{cases} 1 \text{ if } p \text{ is a quadratic residue modulo } \ell; \\ -1 \text{ if } p \text{ is not a quadratic residue modulo } \ell. \end{cases}$$

The law of quadratic reciprocity says that

$$\left(\frac{p}{\ell}\right) \cdot \left(\frac{\ell}{p}\right) = (-1)^{\frac{p-1}{2} \cdot \frac{\ell-1}{2}}.$$

In particular, the number of solutions to the quadratic equation

$$x^2 \equiv p \pmod{\ell}$$

when $p \neq \ell$ is either 0 or 2 and only depends on the residue of ℓ modulo $4p$. As ℓ varies over prime numbers, this number is periodic with period $4p$.

Set $p = 3$. Using quadratic reciprocity, we now see that the answer to Question 1.1 depends only on ℓ modulo 12, i.e., on where ℓ lands on the 12-hour clock! For example, 13, 37, 61 and 1093 are all congruent to 1 modulo 12, so the answer is "Yes" for all these primes. On the other hand, 3 can never be the last digit of a

perfect square modulo 5. Since 17, 29, 41 and 1637 are all congruent to 5 modulo 12, the answer is "No" for all these primes.

In the twentieth century, we have come to view quadratic reciprocity as a one-dimensional reciprocity law, an instance of Langlands reciprocity for the group GL_1 over the rational numbers $\mathbb{Q}$. To generalize this, we first reformulate Question 1.1: how do we characterize the set of primes ℓ such that the polynomial $x^2 - 3$ splits as a product of distinct linear factors modulo ℓ?

This question can also be formulated for polynomials of higher degree, leading to the notion of a *reciprocity law* as in [Wym72]. One can even formulate reciprocity laws for polynomials in more than one variable. These higher degree and higher dimensional reciprocity laws, discovered in the twentieth century, go beyond the original number-theoretic setting, with the tantalizing potential of connecting *different* areas of mathematics.

Example 1.2. The infinite product

$$F(q) := q \prod_{n=1}^{\infty} (1-q^n)^2 (1-q^{11n})^2$$

and the Diophantine equation

$$E : y^2 + y = x^3 - x^2$$

seem to know about each other in a mysterious way.

Indeed, if we consider N_ℓ to be the number of solutions of the congruence

$$y^2 + y \equiv x^3 - x^2 \pmod{\ell},$$

then we obtain the following list of values for primes $\ell \neq 11$:

ℓ	2	3	5	7	13	17	19	23	29
N_ℓ	4	4	4	9	9	19	19	24	29

If we look at the coefficients a_ℓ of q^ℓ in the expansion of $F(q)$ we obtain:

ℓ	2	3	5	7	13	17	19	23	29
a_ℓ	-2	-1	1	-2	4	-2	0	-1	0

Notice that we always have

$$a_\ell + N_\ell = \ell.$$

We see that $F(q)$ is essentially a generating series for the solutions modulo ℓ of the Diophantine equation E!

The generating series $F(q)$ is the Fourier expansion of a *modular form*, a highly symmetric holomorphic function on the upper-half plane. This naturally lives in the world of harmonic analysis. The Diophantine equation E describes an *elliptic curve* defined over $\mathbb{Q}$. This naturally lives in the world of arithmetic algebraic geometry.

The reciprocity law that relates $F(q)$ and E is an instance of the modularity of elliptic curves over $\mathbb{Q}$, which was famously the cornerstone of the proof of Fermat's Last Theorem [Wil95, TW95]. This is a two-dimensional reciprocity law, an instance of Langlands reciprocity for the group GL_2 over $\mathbb{Q}$.

There is a deep and fruitful analogy between the arithmetic of the integers, with the special role played by prime numbers, and the geometry of curves defined over finite fields, where prime numbers are replaced by the points of the curve. The former setting is that of *number fields*, such as the field of rational numbers $\mathbb{Q}$, while the latter is the setting of *function fields*, such as the field of rational functions $\mathbb{F}_p(t)$ with $\mathbb{F}_p := \mathbb{Z}/p\mathbb{Z}$.

While the concept of Langlands reciprocity can be traced back to the work of Euler, Legendre and Gauss in the number field setting, a parallel set of conjectures and results about Langlands reciprocity developed in the function field setting from the second half of the nineteenth century on. This culminated in the breakthrough construction of the correspondence (in both directions) for the group GL_n: by Drinfeld in the case $n = 2$ [Dri80, Dri87b, Dri88, Dri87a] and by Laurent Lafforgue for arbitrary n [Laf02].

In the past decade, Vincent Lafforgue [Laf18a] gave a highly original and completely general construction of the *automorphic to Galois direction* of the Langlands correspondence in the function field setting. His construction applies to arbitrary connected reductive groups: it gives a more elegant proof in the case of GL_n, and treats symplectic, unitary and exceptional groups, among others, with the same method. The goal of this survey is to explain the statement of this beautiful theorem of V. Lafforgue and discuss some of the key ideas that go into its proof.

Finally, we emphasize that there are many exciting, even more recent developments in the field that were inspired by or build on [Laf18a]. This includes important results in the function field setting, such as the work of Böckle–Khare–Harris–Thorne [BHKT19], which establishes a general potential automorphy result (in the Galois to automorphic direction of the Langlands correspondence), the work of Genestier–Lafforgue on the local Langlands correspondence [GL17], and the work of Lafforgue–Zhu [LZ18] that goes in the direction of the Arthur–Kottwitz conjectures.

In a development that was hard to foresee even a decade ago, this also includes striking results in the number field setting, such as the work of Xiao–Zhu [XZ17] on the Tate conjecture for the cohomology of Shimura varieties, and the spectacular upcoming work of Fargues–Scholze [FS20] on the geometrization of the local Langlands correspondence.

1.3. **Organization.** In § 2, we discuss the analogy between number fields and function fields. In § 3, we discuss the two sides of the global Langlands correspondence in a more systematic way and work through concrete examples of reciprocity. We conclude this section by discussing the statement of the main theorem of [Laf18a].

The technical heart of this survey consists of § 4 and § 5. In § 4, we give some sense of the geometry of moduli spaces of shtukas. These are highly symmetric geometric objects, which provide a link between the two sides of the Langlands correspondence. In § 5 we explain how to extract a Galois representation from a system of eigenvalues attached to certain excursion operators that act on the automorphic side. The excursion operators are defined using moduli spaces of

shtukas, but we formalize the information we need from § 4, so that it can be taken largely as a black box.

1.4. **Further references.** In addition to the original research papers, we recommend the following surveys on the Langlands correspondence. For more technical surveys that describe the main result of [Laf18a], see [Hei18], [Str17] (which has a particular detailed overview of the case of GL_1), and [Laf18b] (which also discusses some further developments in the function field setting). For a historical overview of Langlands reciprocity in the number field setting see [Eme20] and [Wei16]. For a cutting edge survey on the Langlands program that combines ideas from the number field and function field settings, see [Sch18].

1.5. **Acknowledgements.** I am grateful to Toby Gee, Steven Sivek, and Matteo Tamiozzo for useful conversations and for their comments on an earlier version of this text.

2. Number fields and function fields

To discuss the global Langlands correspondence, we will need two main players: a global field F and a reductive group G. In this section, we focus on the former. The global field F can be a *number field*, i.e., the field of rational numbers $\mathbb{Q}$ or a finite extension obtained by adjoining the roots of a polynomial with rational coefficients, such as the real quadratic field $\mathbb{Q}(\sqrt{3})$ or the imaginary quadratic field $\mathbb{Q}(i)$. Alternatively, the global field can be a *function field*, i.e., the field of rational functions on a smooth, projective and geometrically connected curve X defined over the finite field $\mathbb{F}_q$ of order $q = p^f$.

Example 2.1. Take X to be the projective line $\mathbb{P}^1_{\mathbb{F}_q}$: this can be identified with the space of lines in 2-dimensional affine space $\mathbb{A}^2_{\mathbb{F}_q}$ passing through the origin 0. As an algebraic curve, $\mathbb{P}^1_{\mathbb{F}_q}$ can be constructed by gluing two copies of the affine line $\mathbb{A}^1_{\mathbb{F}_q}$, with rings of functions $\mathbb{F}_q[t]$ and $\mathbb{F}_q[t^{-1}]$, along the common open subset with ring of functions $\mathbb{F}_q[t, t^{-1}]$. In this case, the function field of X is $F = \mathbb{F}_q(t)$, the field of rational functions in one variable t over $\mathbb{F}_q$.

Instead of studying global fields directly, we can first consider their completions, which have a simpler structure and will play an auxiliary role. More precisely, a *place* v of a global field F is a non-trivial multiplicative norm $|\ |: F \to \mathbb{R}_{\geq 0}$, given up to equivalence. The equivalence relation identifies $|\ |$ and $|\ |^s$ for $s \in \mathbb{R}_{>0}$. The completion of F with respect to a place v is called a *local field*.

For example, if $F = \mathbb{Q}$, the places are:

- The infinite or archimedean place, where the norm is the usual absolute value and the completion is the field of real numbers $\mathbb{R}$.
- The finite or non-archimedean places, which correspond to prime numbers p. The norm is such that two rational numbers are close if they are congruent modulo a high power of p. For example, we can take the norm that sends $r \in \mathbb{Q}^\times$ to $p^{-n_p(r)}$, where $n_p(r) \in \mathbb{Z}$ is the exponent of p in the prime factorization of r.

  The corresponding completion is the field of p-adic numbers $\mathbb{Q}_p$. The elements in $\mathbb{Q}_p$ of norm ≤ 1 form the ring of p-adic integers $\mathbb{Z}_p$. This ring

can also be constructed as the inverse limit

$$\mathbb{Z}_p = \varprojlim_n \mathbb{Z}/p^n\mathbb{Z}$$

and then we can recover $\mathbb{Q}_p$ by inverting p: $\mathbb{Q}_p = \mathbb{Z}_p[1/p]$.

If F is the function field of a curve X over $\mathbb{F}_q$, the places of F are in bijection with the closed points of X, which are all defined over $\mathbb{F}_q$ or some finite extension thereof. More precisely, the residue field at a closed point v of X is a finite extension $k(v)$ of $\mathbb{F}_q$, and we set $\deg(v)$ to be the degree of this extension. We also define

$$n_v : F^\times \to \mathbb{Z}$$

to be the order of vanishing of a rational function at v. We then obtain a multiplicative norm on F by

$$f \in F^\times \mapsto q^{-\deg(v) n_v(f)}.$$

The completion F_v with respect to this norm has the following geometric interpretation: it is the field of functions on a punctured formal neighbourhood of v in X. The elements $\mathcal{O}_v \subset F_v$ of norm ≤ 1 can be identified with the ring of functions in a formal neighbourhood of v in X.

For example, let $X = \mathbb{P}^1_{\mathbb{F}_5} = \mathbb{A}^1_{\mathbb{F}_5} \cup \{\infty\}$. The places of the function field F of X are:

- The place ∞, where the completion is the field of Laurent series $\mathbb{F}_5(\!(t^{-1})\!)$ and the ring of integers is the power series ring $\mathbb{F}_5[\![t^{-1}]\!]$. The place ∞ is already defined over $\mathbb{F}_5$, so the degree is equal to 1.
- The place 0, where the completion is the field of Laurent series $\mathbb{F}_5(\!(t)\!)$ and the ring of integers is the power series ring $\mathbb{F}_5[\![t]\!]$. The place 0 is also defined over $\mathbb{F}_5$, so the degree is equal to 1. We see that the roles of 0 and ∞ are completely symmetric, so unlike in the number field setting, there is nothing special about ∞.
- More generally, any monic irreducible polynomial in $\mathbb{F}_5[t]$ defines a place of F. For example, the polynomial $t^2 - 3$ is irreducible in $\mathbb{F}_5$ (since we have seen that 3 is not a square modulo 5), so this defines a place of degree 2 with residue field equal to $\mathbb{F}_{25}$.

This is the precise sense in which there is an analogy between prime numbers and the closed points on a curve defined over a finite field. Note that, in the case of a curve over a finite field, all the residue fields have the same prime characteristic p, and all the norms are non-archimedean. Moreover, we can form the product $X \times_{\mathbb{F}_q} X$ of the curve with itself, to obtain a surface over $\mathbb{F}_q$, whereas it is not clear how to do this with $\mathbb{Z}$, or what to take the product over. All these properties add extra flexibility to the function field setting.

Remark 2.2. To illlustrate the power of the analogy between number fields and function fields, recall the Riemann zeta function, defined by the convergent series

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s}$$

for a real number $s > 1$. By the unique factorisation of integers into primes, this admits an alternative Euler product formula

$$\zeta(s) = \prod_p \left(1 - \frac{1}{p^s}\right)^{-1},$$

where p runs over prime numbers. By analogy, if X is an algebraic variety over the finite field $\mathbb{F}_p$ (so X is cut out by polynomial equations modulo p), then we can define a zeta function

$$\zeta(X, s) := \prod_v \left(1 - \frac{1}{\#k(v)^s}\right)^{-1},$$

where v runs over the closed points of the algebraic variety X, and $k(v)$ is the residue field at the point v, a finite extension of $\mathbb{F}_p$. This product can be shown to converge when $s > \dim_{\mathbb{F}_p} X$.

The analogy between number fields and function fields led Weil to conjecture that zeta functions of smooth projective varieties behave in many ways like the Riemann zeta function [Wei49]. For example, he conjectured that they admit a functional equation and satisfy an analogue of the Riemann hypothesis. The Weil conjectures are now theorems due to Dwork, Grothendieck and his collaborators, and Deligne. In turn, they inspired much of the development of modern algebraic geometry and had a lasting impact on mathematics.

The analogue of the Riemann hypothesis in the function field setting was established by Deligne in [Del74], whereas the original Riemann hypothesis in the number field setting remains one of the most important open problems in mathematics. Deligne's proof exploited the additional flexibility of the more geometric setting of function fields. We will see later on that V. Lafforgue's construction of the global Langlands correspondence also relies on this additional flexibility.

3. The global Langlands correspondence

In its modern, higher-dimensional incarnation, reciprocity matches spectral data, such as systems of eigenvalues obtained from the topology of highly symmetric manifolds, to arithmetic data, such as the number of solutions to polynomial equations modulo primes. The spectral data, such as the set of coefficients of the generating series $F(q)$ in Example 1.2, lives on the *automorphic side*. The arithmetic data, such as the data coming from the Diophantine equation E, lives on the *Galois side*. The goal of this section is to discuss the two sides in a more systematic way.

3.1. **The Galois side.** Let F be a global field, which can be a number field or a function field. An important question in number theory and arithmetic geometry is to understand the structure of the *absolute Galois group* of F, where we view all finite separable extensions of F inside the same algebraic closure $\overline{F}$ and consider its group of automorphisms, which we denote by $\Gamma_F = \mathrm{Gal}(\overline{F}/F)$. This can be identified with the inverse limit

$$\Gamma_F = \varprojlim_{F'} \mathrm{Gal}(F'/F),$$

where F' runs over all finite Galois extensions of F, making Γ_F into a *profinite* topological group, that looks like a Cantor set. When F is a global field, this is an extremely mysterious and highly non-abelian group.

For a finite field $\mathbb{F}_q$ of order $q = p^f$, we can similarly define

$$\Gamma_{\mathbb{F}_q} := \mathrm{Gal}(\overline{\mathbb{F}}_q/\mathbb{F}_q) = \varprojlim_{\mathbb{F}} \mathrm{Gal}(\mathbb{F}/\mathbb{F}_q),$$

where $\mathbb{F}$ runs over all finite Galois extensions of $\mathbb{F}_q$. The structure theory of finite fields shows that this Galois group is isomorphic to the profinite completion $\widehat{\mathbb{Z}}$ of $\mathbb{Z}$, and is topologically generated by one element, the *Frobenius automorphism*. Indeed, the multiplicative map $x \mapsto x^q$ has the magical property that

$$(x+y)^q \equiv x^q + y^q \pmod{p},$$

so in characteristic p it is additive as well. This means that the map defines a field automorphism of $\overline{\mathbb{F}}_q$, and its fixed subfield is precisely $\mathbb{F}_q$. From now on, we denote the Frobenius automorphism by Frob_q.

Unlike the case of finite fields, it is hard to study the Galois groups of global fields directly, so instead we appeal to an idea that has been extremely fruitful in mathematics, and we study their representations. More precisely, we study the continuous, finite-dimensional representations of Γ_F; these are called *Galois representations*.

Example 3.2. The following is an expanded version of the example discussed in [Eme13]. In the spirit of Question 1.1, we can ask to characterize the rational prime numbers ℓ such that the cubic polynomial

$$f(x) = x^3 - x - 1$$

splits into distinct linear factors modulo ℓ. We first rephrase this as a question about Galois representations.

Let $K/\mathbb{Q}$ be the *splitting field* of the polynomial $f(x)$, i.e., the smallest field extension of $\mathbb{Q}$ over which $f(x)$ splits into linear factors. It can be shown that K is a cubic extension of the imaginary quadratic field $\mathbb{Q}(\sqrt{-23})$[1]. The Galois group $\mathrm{Gal}(K/\mathbb{Q})$ is isomorphic to S_3, the symmetric group on 3 elements, and it acts on the 3 roots of the polynomial via the permutation representation.

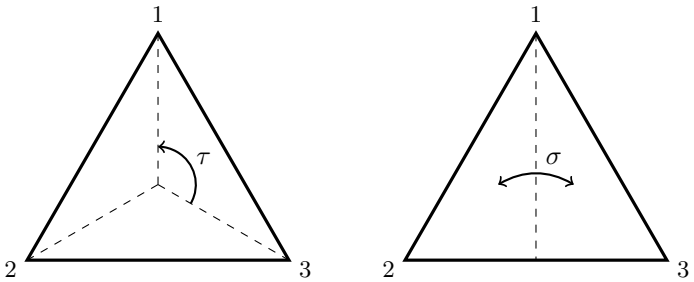


FIGURE 1. The symmetries of the equilateral triangle

To visualize this, consider the symmetries of the equilateral triangle. There is the rotation τ by 120 degrees around its center, and the reflection σ around the vertical axis passing through its center. The group S_3 is generated by τ and σ,

[1]The prime 23 is special here because the discriminant of the cubic polynomial $f(x)$ is equal to -23. Modulo 23, the polynomial $f(x)$ decomposes as $(x-10)^2(x-3)$ and has a repeated root.

with the relations $\tau^3 = 1$, $\sigma^2 = 1$ and $\sigma\tau = \tau^2\sigma$. The action of S_3 on the vertices of the triangle is precisely the permutation representation.

For the polynomial $f(x)$ to split into distinct linear factors modulo ℓ, we need the equation $f(x) \equiv 0 \pmod{\ell}$ to have 3 distinct roots valued in $\mathbb{F}_\ell$. If $\ell \neq 23$, the 3 roots will be distinct, and a priori valued in some finite field extension of $\mathbb{F}_\ell$. To test whether the roots are actually valued in $\mathbb{F}_\ell$, it is equivalent to check whether they are fixed under the action of the Frob_ℓ-automorphism of $\overline{\mathbb{F}}_\ell$ given by $x \mapsto x^\ell$.

An essential, but slightly subtle point is that the Frobenius automorphism Frob_ℓ, a priori thought of as an element of the absolute Galois group of $\mathbb{F}_\ell$, determines a well-defined conjugacy class in $\mathrm{Gal}(K/\mathbb{Q})$. To see why this might be true, we must consider the relationship between the Galois groups of finite fields, local fields and global fields. This can be encoded in the diagram

$$\begin{array}{ccccc} \mathrm{Gal}(\overline{\mathbb{Q}}_\ell/\mathbb{Q}_\ell) & \hookrightarrow & \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) & \twoheadrightarrow & \mathrm{Gal}(K/\mathbb{Q}). \\ \downarrow & & & & \\ \mathrm{Gal}(\overline{\mathbb{F}}_\ell/\mathbb{F}_\ell) & & & & \end{array} \tag{3.2.1}$$

The first inclusion is determined by a choice of embedding $\overline{\mathbb{Q}} \hookrightarrow \overline{\mathbb{Q}}_\ell$, so this inclusion is well-defined only up to conjugacy. We choose a lift of Frob_ℓ from $\mathrm{Gal}(\overline{\mathbb{F}}_\ell/\mathbb{F}_\ell)$ to $\mathrm{Gal}(\overline{\mathbb{Q}}_\ell/\mathbb{Q}_\ell)$. By (3.2.1), this gives rise to a conjugacy class in $\mathrm{Gal}(K/\mathbb{Q})$. Finally, when $\ell \neq 23$, one can show that this conjugacy class is independent of the choice of lift.

To check that Frob_ℓ fixes the 3 roots, it is equivalent to check that the corresponding conjugacy class in $\mathrm{Gal}(K/\mathbb{Q})$ has trace equal to 3 under the permutation representation of S_3. The group S_3 admits a unique 2-dimensional irreducible representation up to conjugation, given by

$$\tau \mapsto \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}, \qquad \sigma \mapsto \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The 3-dimensional permutation representation decomposes as the direct sum of the trivial 1-dimensional representation and the 2-dimensional irreducible one. Write

$$\rho_K : \mathrm{Gal}(K/\mathbb{Q}) \simeq S_3 \hookrightarrow \mathrm{GL}_2(\mathbb{C})$$

for the corresponding 2-dimensional Galois representation. Note that this representation has finite image in $\mathrm{GL}_2(\mathbb{C})$. Therefore, it is continuous for the discrete topology on $\mathbb{C}$, and so for any topology on $\mathbb{C}$.

Summarizing the discussion so far, we see that $f(x)$ splits into distinct linear factors modulo ℓ if and only if $\rho_K(\mathrm{Frob}_\ell)$ has trace equal to 2. In order to characterize the primes ℓ for which $f(x)$ modulo ℓ splits as a product of distinct linear factors, it is enough to find a generating function for the values $\{\mathrm{tr}\rho_K(\mathrm{Frob}_\ell)\}$, and select those primes for which the value is equal to 2.

Example 3.3. The cubic equation $y^2 + y = x^3 - x^2$ (or rather its projectivization) represents an *elliptic curve* E defined over $\mathbb{Q}$. This is a smooth, projective curve of genus 1 with a specified point, the point at ∞. The elliptic curve can be endowed with a group structure defined over $\mathbb{Q}$. For any prime number p, the p^n-torsion points

$$E(\mathbb{C})[p^n] \simeq (\mathbb{Z}/p^n\mathbb{Z})^2$$

have algebraic numbers as their coordinates. This induces an action of the absolute Galois group $\Gamma_{\mathbb{Q}}$ on $(\mathbb{Z}/p^n\mathbb{Z})^2$, and the actions are compatible as n varies. By taking the inverse limit and inverting p, we obtain a continuous 2-dimensional Galois representation

$$\rho_E : \Gamma_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Q}_p).$$

Note that, unlike Example 3.2, the coefficients of ρ_E are the field of p-adic numbers $\mathbb{Q}_p$. This representation does not have finite image, and the topology of $\Gamma_{\mathbb{Q}}$ interacts with the topology of $\mathbb{Q}_p$.

Recall that N_ℓ is the number of solutions to the congruence

$$y^2 + y \equiv x^3 - x^2 \pmod{\ell}.$$

Using a (version of the) Lefschetz fixed point formula, one can show that, for a prime $\ell \neq 11$, we have

$$1 + N_\ell = \#E(\mathbb{F}_\ell) = 1 + \ell - \mathrm{tr}\rho_E(\mathrm{Frob}_\ell).$$

The summand 1 corresponds to the point at ∞. In order to compute N_ℓ, it is enough to find a generating function for the values $\{\mathrm{tr}\rho_E(\mathrm{Frob}_\ell)\}$.

More generally, if $X/\mathbb{Q}$ is an algebraic variety and ℓ is a prime number, its ℓ-adic étale cohomology groups $H^i_{\text{ét}}(X_{\overline{\mathbb{Q}}}, \mathbb{Q}_\ell)$ are continuous, finite-dimensional representations of $\Gamma_{\mathbb{Q}}$ on $\mathbb{Q}_\ell$-vector spaces[2]. Étale cohomology is a cohomology theory for algebraic varieties defined by Grothendieck, which behaves like singular cohomology for manifolds and is a rich source of Galois representations.

The fundamental notion that this theory relies on is that of an *étale morphism*, which is an algebraic analogue of the notion of local isomorphism in topology. A morphism between smooth varieties is étale if and only if its differential at every point is an isomorphism of tangent spaces. A finite separable extension of fields gives rise to an étale morphism. Therefore, the étale theory subsumes Galois theory.

Example 3.4. Let X be a smooth, projective and geometrically connected curve over a finite field $\mathbb{F}_q$, and let F be the function field of X. A finite étale and Galois cover of the curve $\widetilde{X} \to X$ determines a continuous representation of Γ_F on a finite set. Indeed, the function field $\widetilde{F}$ of $\widetilde{X}$ is a finite Galois extension of F and determines a finite quotient $\mathrm{Gal}(\widetilde{F}/F)$ of the profinite group Γ_F.

More generally, the perspective on Galois theory developed by Grothendieck shows that there is an equivalence of categories between:

- The category of finite sets equipped with a continuous action of the profinite group Γ_F.
- The category of finite separable F-algebras.

If $U \subseteq X$ is an open and dense subset, the absolute Galois group Γ_F has a profinite quotient $\pi_1(U, \bar{\eta})$, the étale fundamental group of U (this also depends on a choice of base point $\bar{\eta}$, i.e., of an algebraic closure $\overline{F}$ of F). The action of Γ_F on some finite set factors through $\pi_1(U, \bar{\eta})$ if and only if the corresponding finite F-algebra $\widetilde{F}$ extends to a finite étale cover $\widetilde{U} \to U$.

[2]This does indeed generalize the construction via p^n-torsion points in the special case of elliptic curves: if $X = E$, the representation of $\Gamma_{\mathbb{Q}}$ on $H^1_{\text{ét}}(E_{\overline{\mathbb{Q}}}, \mathbb{Q}_p)$ is the dual of the representation ρ_E, which can be identified with the étale homology of E.

Example 3.5. If $X = \mathbb{P}^1_{\mathbb{F}_q}$ with $q = p^f$ and $p \geq 3$, we can take U to be the open dense subset with ring of functions $\mathbb{F}_q[t, t^{-1}]$ and let $\widetilde{U} \to U$ be the degree 2 cover with ring of functions $\mathbb{F}_q[s, t, t^{-1}]/(s^2 - t)$. The cover is étale because the derivative $\frac{d(s^2-t)}{ds} = 2s$ is invertible over the whole of $\widetilde{U}$. We obtain a character of Γ_F of degree 2.

Let v denote a place of F, which is just a closed point of the curve X. Let $U \subseteq X$ be an open and dense subset that contains v. We have a diagram analogous to (3.2.1)

$$\begin{array}{ccccc} \mathrm{Gal}(\overline{F}_v/F_v) & \hookrightarrow & \mathrm{Gal}(\overline{F}/F) & \twoheadrightarrow & \pi_1(U, \bar{\eta}). \\ \downarrow & & & & \\ \mathrm{Gal}(\overline{k(v)}/k(v)) & & & & \end{array} \tag{3.5.1}$$

As in the number field case, the first inclusion is well-defined only up to conjugacy. The residue field $k(v)$ is a finite extension of $\mathbb{F}_q$, so the absolute Galois group $\mathrm{Gal}(\overline{k(v)}/k(v)) \simeq \widehat{\mathbb{Z}}$ is topologically generated by the Frobenius automorphism $x \mapsto x^{q^{\deg(v)}}$. We denote by Frob_v an element of $\mathrm{Gal}(\overline{F}_v/F_v)$ that lifts this Frobenius automorphism. Since $v \in U$, one can show that the image of Frob_v in the étale fundamental group $\pi_1(U, \bar{\eta})$ is a well-defined conjugacy class, independent of the choice of lift.

Now consider a Galois representation

$$\rho : \Gamma_F \to \mathrm{GL}_n(\overline{\mathbb{Q}}_\ell),$$

that factors through $\pi_1(U, \bar{\eta})$ for some open and dense $U \subseteq X$. As v runs over all places of X contained in U, we obtain a well-defined infinite set $\{\mathrm{tr}\rho(\mathrm{Frob}_v)\}$ of numbers in $\overline{\mathbb{Q}}_\ell$. This set represents the kind of *arithmetic data* we see on the Galois side of the Langlands correspondence.

Question 3.6. Can we find a different way to generate the numbers $\{\mathrm{tr}\rho(\mathrm{Frob}_v)\}$?

3.6.1. *Langlands parameters.* Galois representations are, roughly, the objects on the Galois side of the Langlands correspondence. To be as general as possible, we want to allow the Galois representations to be valued in general connected reductive groups rather than just GL_n. This brings us to the second main player needed to formulate the global Langlands correspondence.

We let G be a connected reductive algebraic group defined over $\mathbb{Q}$ (in the number field setting) or over $\mathbb{F}_q$ (in the function field setting). For example, we can take G to be:

(1) the general linear group GL_n, consisting of $n \times n$ invertible matrices;
(2) the special linear group SL_n, consisting of $n \times n$ matrices of determinant 1;
(3) the projective general linear group PGL_n, which is the quotient of GL_n by the subgroup of non-zero scalars, embedded diagonally;
(4) the symplectic group Sp_{2n}, consisting of $2n \times 2n$ matrices that preserve the standard symplectic form on a $2n$-dimensional vector space;
(5) the special orthogonal group SO_n, consisting of $n \times n$ matrices of determinant 1 and who are equal to their transpose inverse.

In this survey, we will assume our group is *split* over $\mathbb{Q}$ (or over $\mathbb{F}_q$), which means that it contains a split maximal torus defined over $\mathbb{Q}$ (or over $\mathbb{F}_q$), equal to a product of some number of copies of the multiplicative group $\mathbb{G}_m := \mathrm{GL}_1$. All the examples above are split groups: for instance, GL_n contains the maximal torus of diagonal matrices. This restriction still exhibits some of the richness of the general theory, as we will see in the final paragraphs of this section, while making the exposition substantially simpler.

The group G determines, through an explicit combinatorial recipe that involves its root datum, a split reductive group $\widehat{G}$ over $\mathbb{Q}_\ell$ called *the Langlands dual group* of G. The group $\widehat{G}$ has a canonical description in terms of algebraic geometry[3] and, in some sense, controls the representation theory of G. Taking the Langlands dual of $\widehat{G}$ recovers the original group G (as a group over $\mathbb{Q}_\ell$).

G	GL_n	SL_n	PGL_n	Sp_{2n}	SO_{2n+1}	SO_{2n}
$\widehat{G}$	GL_n	PGL_n	SL_n	SO_{2n+1}	Sp_{2n}	SO_{2n}

FIGURE 2. Examples of connected reductive groups and their Langlands dual groups.

Assume now that we are in the function field setting and that ℓ is a prime number different from the characteristic p of $\mathbb{F}_q$.

Definition 3.6.2. *A* global Langlands parameter *for G is a conjugacy class of homomorphisms*

$$\rho : \Gamma_F \to \widehat{G}(\overline{\mathbb{Q}}_\ell)$$

that factor through the étale fundamental group $\pi_1(U, \bar{\eta})$ for some open dense subset $U \subseteq X$, that take values in $\widehat{G}$ over a finite extension of $\mathbb{Q}_\ell$, and that are continuous and semisimple.

Given a finite-dimensional algebraic representation W of $\widehat{G}$ and a global Langlands parameter ρ, we can consider the Galois representation $W \circ \rho$ and keep track of the arithmetic data $\{\mathrm{tr}(W \circ \rho)(\mathrm{Frob}_v)\}$, as v runs over all places of X contained in U. By varying W, we can encode the semi-simple conjugacy class of each $\rho(\mathrm{Frob}_v)$ in $\widehat{G}(\overline{\mathbb{Q}}_\ell)$, as v runs over all places of X contained in U.

Example 3.7. If $G = \mathrm{GL}_2$ then $\widehat{G} = \mathrm{GL}_2$. Let $\rho : \Gamma_F \to \widehat{G}(\overline{\mathbb{Q}}_\ell)$ be a global Langlands parameter. We can take W to be the standard representation of GL_2 on a two-dimensional $\overline{\mathbb{Q}}_\ell$-vector space, in which case

$$\mathrm{tr}(W \circ \rho)(\mathrm{Frob}_v) = \mathrm{tr}\rho(\mathrm{Frob}_v).$$

On the other hand, we can also consider the one-dimensional representation W given by the determinant, in which case

$$\mathrm{tr}(W \circ \rho)(\mathrm{Frob}_v) = \det \rho(\mathrm{Frob}_v).$$

We see that, by varying W, we can recover the coefficients of the characteristic polynomial of each $\rho(\mathrm{Frob}_v)$, which in turn recover the semi-simple conjugacy class of each $\rho(\mathrm{Frob}_v)$ in $\mathrm{GL}_2(\overline{\mathbb{Q}}_\ell)$.

[3]This canonical description uses the so-called geometric Satake equivalence, which appears again in § 4 and § 5.

3.8. **The automorphic side.** The objects on the automorphic side of the global Langlands correspondence can seem more mysterious than the objects on the Galois side, though in getting to know them better, one discovers that they have many remarkable properties. To help build some intuition for the notion of an *automorphic form*, we start by discussing the number field setting.

The generating functions that correspond to Examples 3.2 and 3.3 are modular forms, which are examples of automorphic forms for the groups SL_2 or GL_2 over $\mathbb{Q}$. A *modular form* is a holomorphic function on the upper-half complex plane

$$\mathbb{H} := \{z \in \mathbb{C} \mid \mathrm{Im}\ z > 0\}$$

that satisfies many symmetries and a growth condition. The symmetries are given in terms of the action of the group $\mathrm{SL}_2(\mathbb{Z}) \subset \mathrm{SL}_2(\mathbb{R})$ on $\mathbb{H}$ by Möbius transformations:

$$z \mapsto \frac{az+b}{cz+d} \quad \text{for } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R}).$$

Below is a picture of a fundamental domain for the action of $\mathrm{SL}_2(\mathbb{Z})$ on $\mathbb{H}$. Notice that this fundamental domain is non-compact in the direction $\mathrm{Im}\ z \to \infty$; to compactify it, one needs to add a *cusp*, which corresponds to the point $i\infty$.

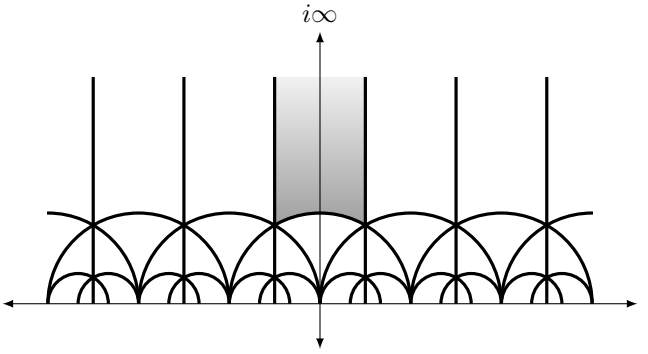


FIGURE 3. A fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$ acting on $\mathbb{H}$

Because the action of $\mathrm{SL}_2(\mathbb{R})$ on $\mathbb{H}$ is transitive, we can identify

$$\mathbb{H} = \mathrm{SL}_2(\mathbb{R})/\mathrm{SO}_2(\mathbb{R}),$$

where $\mathrm{SO}_2(\mathbb{R})$ is the stabilizer of the point $i \in \mathbb{H}$, and is a maximal compact subgroup of $\mathrm{SL}_2(\mathbb{R})$. A modular form f of weight $k \geq 1$ and level Γ, where $\Gamma \subseteq \mathrm{SL}_2(\mathbb{Z})$ is a subgroup cut out by congruence conditions, satisfies the automorphy condition

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^k f(z).$$

Because of this automorphy condition, modular forms can be identified with certain differential forms on the quotient

$$\Gamma\backslash\mathbb{H} = \Gamma\backslash\mathrm{SL}_2(\mathbb{R})/\mathrm{SO}_2(\mathbb{R}). \tag{3.8.1}$$

Cusp forms are those modular forms that vanish at the cusps; they are those automorphic forms that are genuinely new for the group SL_2, rather than those that essentially arise from a smaller group, the maximal torus GL_1 of diagonal matrices inside SL_2.

Example 3.9. For a prime p, define the congruence subgroup

$$\Gamma_0(p) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \mid c \equiv 0 \pmod{p} \right\}$$

Note that $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma_0(p)$. This means that a modular form f of some weight k and level $\Gamma_0(p)$ satisfies $f(z+1) = f(z)$, which implies that we can describe it using its Fourier expansion.

Setting $q := e^{2\pi i z}$, the power series expansion of the infinite product

$$G(q) = q \prod_{n=1}^{\infty} (1-q^n)(1-q^{23n})$$

is the Fourier expansion of a cusp form $g(z)$ of weight 1 and level $\Gamma_0(23)$. We will see that this is the "mirror image" on the automorphic side of the Galois representation ρ_K of Example 3.2 under the global Langlands correspondence for GL_2 over $\mathbb{Q}$.

The power series expansion of the infinite product

$$F(q) = q \prod_{n=1}^{\infty} (1-q^n)^2(1-q^{11n})^2$$

is the Fourier expansion of a cusp form $f(z)$ of weight 2 and level $\Gamma_0(11)$. We will see that this is the mirror image on the automorphic side of the Galois representation ρ_E of Example 3.3 under the global Langlands correspondence for GL_2 over $\mathbb{Q}$.

The data we would like to keep track of on the automorphic side is spectral data, i.e., systems of eigenvalues. These are the eigenvalues of a commutative algebra of *Hecke operators* that act on spaces of modular forms. Hecke operators encode the symmetries of the tower of quotients of $\mathbb{H}$ obtained by imposing various congruence conditions. For a fixed level Γ, we can let ℓ run over all but finitely many primes and obtain in each case a diagram

(3.9.1)
$$\begin{array}{ccc} & (\Gamma \cap \Gamma_0(\ell)) \backslash \mathbb{H} & \\ \swarrow & & \searrow \\ \Gamma \backslash \mathbb{H} & & \Gamma \backslash \mathbb{H} \end{array}$$

where the map on the left is the natural projection, and the map on the right is given by $z \mapsto \left(\begin{smallmatrix} \ell & 0 \\ 0 & 1 \end{smallmatrix}\right) z$ followed by the natural projection. This diagram defines a Hecke operator T_ℓ on the space of modular forms of level Γ (and arbitrary weight k).

Example 3.10. There is a natural bijection between $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ and the set of lattices $\Lambda \subset \mathbb{C}$, taken up to homothety. This gives the following moduli interpretation for the diagram (3.9.1) in the case $\Gamma = \mathrm{SL}_2(\mathbb{Z})$. In that case, $\Gamma \cap \Gamma_0(\ell)$ is just $\Gamma_0(\ell)$ and the quotient $\Gamma_0(\ell)\backslash\mathbb{H}$ parametrizes pairs of lattices $(\Lambda' \overset{\phi}{\hookrightarrow} \Lambda)$ such that the quotient Λ/Λ' has cardinality ℓ. The map on the left sends such a tuple to Λ' and the map on the right sends it to Λ. The formula for the Hecke operator T_ℓ is

$$T_\ell(f)(\Lambda) = \sum_{\Lambda', \phi} f(\Lambda').$$

for a modular or cusp form f.

For varying ℓ, the different Hecke operators commute; we can diagonalize them simultaneously and consider those modular forms that are simultaneous eigenvectors; these are called *eigenforms*. To an eigenform, we can associate the spectral data consisting of the Hecke eigenvalues.

Example 3.11. Both cusp forms $g(z)$ and $f(z)$ from Example 3.9 are eigenforms. Write

$$g(z) = \sum_{n=1}^{\infty} a_n(g)q^n = q - q^2 - q^3 + \dots.$$

By direct computation, one can show that the eigenvalue of the Hecke operator T_ℓ on $g(z)$ is equal to the ℓth Fourier coefficient $a_\ell(g)$. Similarly, if we write

$$f(z) = \sum_{n=1}^{\infty} a_n(f)q^n = q - 2q^2 - q + \dots,$$

the eigenvalue of the Hecke operator T_ℓ is equal to the ℓth Fourier coefficient $a_\ell(f)$. In fact, for any cuspidal eigenform that is normalized such that the coefficient of q equals 1, the Fourier coefficients recover the Hecke eigenvalues.

We now try to understand what kind of objects are automorphic forms in the function field setting. Recall the double quotient (3.8.1) which can be rewritten as

$$\mathrm{GL}_2(\mathbb{Z})\backslash \mathrm{GL}_2(\mathbb{R})/\mathbb{R}_{>0}\mathrm{SO}_2(\mathbb{R}).$$

An automorphic form for GL_2 over $\mathbb{Q}$ can be thought of as a function on the quotient $\mathrm{GL}_2(\mathbb{Z}) \setminus \mathrm{GL}_2(\mathbb{R})$. This parametrizes finite free (or projective) $\mathbb{Z}$-modules M of rank 2 together with a trivialization $M \otimes_{\mathbb{Z}} \mathbb{R} \simeq \mathbb{R}^2$. In more geometric terms, finite projective $\mathbb{Z}$-modules of rank 2 are the same as rank 2 vector bundles over Spec $\mathbb{Z}$.

Analogously, an automorphic form for GL_n defined over a curve $X/\mathbb{F}_q$ is a function on the set of isomorphism classes of rank n vector bundles on X. More generally, an automorphic form for a group G is a function on the set of isomorphism classes of G-bundles on X. This set is denoted by $\mathrm{Bun}_G(\mathbb{F}_q)$, and we will see later on that it also has an algebro-geometric structure.

Remark 3.12. Assume G is a semisimple group, such as SL_n. To make the analogy with the number field case even more striking, choose any point v on the X, assumed for simplicity to be a point of degree 1. We can trivialize a G-bundle on X in a formal neighborhood of v and also on the open subset $X \setminus \{v\}$. To recover the original G-bundle, we need to specify how to glue these two trivial G-bundles in a punctured formal neighborhood of v. The transition function gives an element of $G(F_v)$, where F_v is the completion of F at v and can be identified with a Laurent series ring $\mathbb{F}_q(\!(t)\!)$. On the other hand, we must forget the two original trivializations. The trivialization in the formal neighborhood of v gives an element of $G(\mathcal{O}_v)$, where $\mathcal{O}_v$ is the power series ring $\mathbb{F}_q[\![t]\!]$, a maximal compact subgroup of $G(F_v)$. The other trivialization gives an element of $G(\mathcal{O}(X \setminus \{v\}))$. We obtain the following uniformization result

$$\mathrm{Bun}_G(\mathbb{F}_q) = G(\mathcal{O}(X \setminus \{v\}))\backslash G(F_v)/G(\mathcal{O}_v),$$

which is formally analogous to the double quotient (3.8.1) we saw in the definition of modular forms.

For technical reasons, we impose a condition relative to the center Z of G: we choose a finite index subgroup $\Xi \subseteq \mathrm{Bun}_Z(\mathbb{F}_q)$ and consider functions on $\mathrm{Bun}_G(\mathbb{F}_q)/\Xi$. We write

$$\mathfrak{H} := C_c^{\mathrm{cusp}}\left(\mathrm{Bun}_G(\mathbb{F}_q)/\Xi, \overline{\mathbb{Q}}_\ell\right)$$

for the vector space of *cuspidal automorphic forms* for G with coefficients in $\overline{\mathbb{Q}}_\ell$. This is a finite dimensional $\overline{\mathbb{Q}}_\ell$-vector space. The cuspidality condition picks out those automorphic forms that are genuinely new for G, rather than those that can be obtained from automorphic forms on smaller groups.

In order to read off spectral data from $C_c^{\mathrm{cusp}}\left(\mathrm{Bun}_G(\mathbb{F}_q)/\Xi, \overline{\mathbb{Q}}_\ell\right)$, we must now define an action of Hecke operators on this vector space. These can be defined using the diagram

(3.12.1)
$$\begin{array}{ccc} & \mathcal{H}_{v,W} & \\ \swarrow & & \searrow \\ \mathrm{Bun}_G(\mathbb{F}_q) & & \mathrm{Bun}_G(\mathbb{F}_q), \end{array}$$

where v runs over closed points of the curve X and W runs over irreducible algebraic representations of the Langlands dual group $\widehat{G}$ over $\mathbb{Q}_\ell$. The diagram (3.12.1) has the following moduli interpretation: $\mathcal{H}_{v,W}$ parametrizes tuples

$$(\ \mathcal{E} - \stackrel{\phi}{-} \to \mathcal{E}'\ ),$$

where $\mathcal{E}$ and $\mathcal{E}'$ are G-bundles on X, and ϕ is a *modification* at v bounded by W. This means that ϕ is an isomorphism between the restrictions of $\mathcal{E}$ and $\mathcal{E}'$ to $X \setminus \{v\}$, such that the order of the poles of ϕ at v is bounded, and it turns out that this bound can be expressed naturally in terms of an irreducible algebraic representation of $\widehat{G}$, such as W. The map on the left sends the tuple $(\ \mathcal{E} - \stackrel{\phi}{-} \to \mathcal{E}'\ )$ to $\mathcal{E}'$ and the map on the right sends it to $\mathcal{E}$. The diagram (3.12.1) defines a Hecke operator $T_{v,W}$ on the space of cuspidal automorphic forms $\mathfrak{H}$.

Example 3.13. Let $G = \mathrm{GL}_2$, in which case we also have $\widehat{G} = \mathrm{GL}_2$. Assume for simplicity that v has degree 1. The type of the singularity of ϕ is determined by a double coset

$$\mathrm{GL}_2(\mathbb{F}_q[\![t]\!]) \backslash \mathrm{GL}_2(\mathbb{F}_q(\!(t)\!)) / \mathrm{GL}_2(\mathbb{F}_q[\![t]\!]).$$

Every fractional ideal in $\mathbb{F}_q(\!(t)\!)$ can be generated by a single element of the form t^d with $d \in \mathbb{Z}$. By the elementary divisor theorem, the set of such double cosets can be identified with the set of diagonal matrices

$$\begin{pmatrix} t^{d_1} & 0 \\ 0 & t^{d_2} \end{pmatrix}, d_1 \geq d_2 \in \mathbb{Z},$$

which in turn can be identified with the set of dominant cocharacters of G. If we set $(d_1, d_2) = (1, 0)$, this imposes the following condition on the modification $(\ \mathcal{E} - \stackrel{\phi}{-} \to \mathcal{E}'\ )$: it should realize $\mathcal{E}'$ as a sub-G-bundle of $\mathcal{E}$ such that $\mathcal{E}/\mathcal{E}'$ is the skyscraper sheaf supported at v with fiber a one-dimensional $\mathbb{F}_q$-vector space.

On the other hand, the cocharacter $(1, 0)$ of G can be thought of as a character of $\widehat{G} = \mathrm{GL}_2$, and the standard two-dimensional representation Std is the associated

highest weight representation. We obtain a Hecke operator $T_{v,\mathrm{Std}}$, given by the formula

$$T_{v,\mathrm{Std}}(f)(\mathcal{E}) = \sum_{(\mathcal{E}',\phi)} f(\mathcal{E}')$$

for a cuspidal function f, where $\mathcal{E}' \subset \mathcal{E}$ runs over degree 1 modifications at v. This is the function field analogue of the Hecke operator T_p from Example 3.10.

For fixed v and varying W, the Hecke operators $T_{w,W}$ form a commutative algebra, called the Hecke algebra, that is isomorphic to the Grothendieck ring of representations of $\widehat{G}$ with coefficients in $\mathbb{Q}_\ell$[4]. This is called the *arithmetic Satake isomorphism* and it plays an essential role in formulating the Langlands correspondence. As v varies among places of X, the different Hecke operators commute with each other, so that it makes sense to decompose the space of cuspidal automorphic forms $\mathfrak{H}$ into Hecke eigenspaces. The spectral data that can be read off the automorphic side consists of systems of eigenvalues for the Hecke operators $\{T_{W,v}\}$ acting on $\mathfrak{H}$.

3.14. **The correspondence.** To make the global Langlands correspondence as concrete as possible, consider first the case of GL_2 over $\mathbb{Q}$.

(1) The cusp form $g(z)$ of weight 1 and level $\Gamma_0(23)$ corresponds to the Galois representation

$$\rho_K : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \twoheadrightarrow S_3 \hookrightarrow \mathrm{GL}_2(\mathbb{C}).$$

This implies that, for any prime $\ell \neq 23$, we have an equality between spectral data and arithmetic data

$$a_\ell(g) = \mathrm{tr}\rho_K(\mathrm{Frob}_\ell). \tag{3.14.1}$$

Therefore, for any $\ell \neq 23$, we can read off the splitting behaviour of the polynomial

$$x^3 - x - 1 \pmod{\ell}$$

from the Fourier coefficient $a_\ell(g)$. By expanding the infinite product

$$q\prod_{n=1}^{\infty}(1-q^n)(1-q^{23n}),$$

using for example [LMF20, Newform orbit 23.1.b.a], we see that the first prime for which the polynomial splits into distinct linear factors is $\ell = 59$.

(2) The cusp form $f(z)$ of weight 2 and level $\Gamma_0(11)$ corresponds to the Galois representation

$$\rho_E : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{Q}_\ell)$$

obtained from the elliptic curve

$$y^2 + y = x^3 - x^2.$$

This implies that, for any prime $\ell \neq p$ we have an equality between spectral data and arithmetic data

$$a_\ell(f) = \mathrm{tr}\rho_E(\mathrm{Frob}_\ell) = 1 + \ell - \#E(\mathbb{F}_\ell). \tag{3.14.2}$$

[4]This holds if we assume that $\sqrt{p} \in \mathbb{Q}_\ell$, otherwise we enlarge our coefficient field to be the unramified extension of $\mathbb{Q}_\ell$ of degree 2, which does contain $\sqrt{p}$.

The following classical bound for elliptic curves was first established by Hasse (and fits within the framework of the Weil conjectures)

$$|1 + \ell - \#E(\mathbb{F}_\ell)| \leq 2\sqrt{\ell} \text{ whenever } \ell \neq 11.$$

Together with (3.14.2), this implies that

$$|a_\ell(f)| \leq 2\sqrt{\ell} \text{ whenever } \ell \neq 11.$$

This is a special case of the Ramanujan–Petersson conjecture for $\mathrm{GL}_2/\mathbb{Q}$, a famous conjecture with applications throughout mathematics and computer science, which bounds the size of the Fourier coefficients of cuspidal automorphic forms.

Remark 3.15. The relations (3.14.1) and (3.14.2) are called *Eichler–Shimura relations*. They are proved using a geometric argument, that relies on reinterpreting the quotients $\Gamma\backslash\mathbb{H}$ as solutions to certain moduli problems.

Going back to the function field setting, the following is a preliminary statement of the main theorem proved in [Laf18a]. This establishes the automorphic to Galois direction of the global Langlands correspondence for a general connected reductive group G.

Theorem 3.16. *There exists a canonical decomposition of the space of cuspidal automorphic forms on G with coefficients in $\overline{\mathbb{Q}}_\ell$*

$$\mathfrak{H} = \bigoplus_\rho \mathfrak{H}_\rho, \tag{3.16.1}$$

where the RHS is indexed by global Langlands parameters

$$\rho : \Gamma_F \to \widehat{G}(\overline{\mathbb{Q}}_\ell)$$

as in Definition 3.6.2, which factor through $\pi_1(X, \bar{\eta})$.

This decomposition is stable under the action of the Hecke operators at all places v of X and matches spectral data with arithmetic data as expected. Explicitly, for each place v and each finite dimensional algebraic representation W of $\widehat{G}$, the eigenvalue of the Hecke operator $T_{W,v}$ on $\mathfrak{H}_\rho$ is equal to $\mathrm{tr}(W \circ \rho)(\mathrm{Frob}_v)$.

Remark 3.17.

(1) In fact, V. Lafforgue proves a more general version of Theorem 3.16 that incorporates a non-trivial level structure at a finite set of places $N \subset X$. This means that he considers a space of cuspidal automorphic forms with deeper level at places in N, and decomposes this space in terms of Langlands parameters that factor through the profinite group $\pi_1(X \setminus N, \bar{\eta})$. The compatibility between the spectral data and the arithmetic data makes sense and holds only at places $v \in X \setminus N$. For simplicity of exposition, we restrict ourselves to the case without level structure.

(2) In the case of $G = \mathrm{GL}_n$, this theorem was proved by Drinfeld for $n = 2$ [Dri80] and L. Lafforgue for arbitrary n [Laf02]. Their proofs rely on the Arthur–Selberg trace formula, a difficult method in harmonic analysis and representation theory. V. Lafforgue gave a unified proof that works for general G and that avoids the trace formula completely.

The compatibility between the spectral data seen on the automorphic side, consisting of the eigenvalues of the Hecke operators $\{T_{W,v}\}$, and the arithmetic data $\{\mathrm{tr}(W \circ \rho)(\mathrm{Frob}_v)\}$ seen on the Galois side is not always enough to characterize the decomposition in (3.16.1). This is a subtle, but crucial point. If $G = \widehat{G} = \mathrm{GL}_n$, the compatibility between the spectral data and the arithmetic data does suffice to characterize the decomposition. There are two reasons for this: firstly, by the Cebotarev density theorem, the set $\{\mathrm{Frob}_v\}$ is dense in the profinite group $\pi_1(X, \bar{\eta})$, as v runs over all the places of X. Secondly, representations valued in $\mathrm{GL}_n(\overline{\mathbb{Q}}_\ell)$ are determined by their traces.

As we have seen in Example 3.7, the arithmetic data recovers the semi-simple conjugacy class of each $\rho(\mathrm{Frob}_v)$ in $\widehat{G}(\overline{\mathbb{Q}}_\ell)$. However, it is sometimes possible to have two global Langlands parameters

$$\rho_i : \pi_1(X, \bar{\eta}) \to \widehat{G}(\overline{\mathbb{Q}}_\ell), \text{ for } i = 1, 2,$$

such that $\rho_1(\mathrm{Frob}_v)$ and $\rho_2(\mathrm{Frob}_v)$ are conjugate in $\widehat{G}(\overline{\mathbb{Q}}_\ell)$ for every v, but the representations ρ_1 and ρ_2 are not conjugate in $\widehat{G}(\overline{\mathbb{Q}}_\ell)$. This can happen, for example, if $G = \mathrm{SL}_n$ and $\widehat{G} = \mathrm{PGL}_n$, for $n > 2$. On the automorphic side, this phenomenon corresponds to a failure of multiplicity one, cf. [Bla94, Lap99]. On the Galois side, this phenomenon has been studied in [Lar94, Lar96]: it is formulated in terms of representations that are locally (element-wise) conjugate but not globally conjugate.

This leads us to the key new idea introduced in [Laf18a] and exploited to prove Theorem 3.16. One needs to enlarge the algebra of Hecke operators that act on the automorphic side to a commutative algebra $\mathfrak{B}$ of *excursion operators* that contains all the information needed to recover a Langlands parameter. Heuristically, one should think of the algebra $\mathfrak{B}$ as the algebra of regular functions on the coarse moduli space of Galois representations[5]. A system of simultaneous eigenvalues for the excursion operators is a maximal ideal of $\mathfrak{B}$ and thus determines, by geometric invariant theory, a semi-simple Galois representation.

If we believed in a refined "geometric" version of the Langlands correspondence, such a correspondence would relate the space of cuspidal automorphic forms $\mathfrak{H}$ to a coherent sheaf on the corresponding moduli stack of Galois representations. This leads to a heuristic explanation for why $\mathfrak{H}$ could admit an action by the excursion algebra $\mathfrak{B}$ (see also [Laf18b, Remark 8.5]). V. Lafforgue constructed the action of $\mathfrak{B}$ on $\mathfrak{H}$ using the geometry of moduli spaces of shtukas. In the case of $G = \mathrm{SL}_n$ with $n > 2$, the example of locally but not globally conjugate Langlands parameters shows that the excursion algebra $\mathfrak{B}$ is strictly larger than the Hecke algebra. This shows that the excursion algebra $\mathfrak{B}$, which was not considered before [Laf18a], plays a fundamental role in the very formulation of the global Langlands correspondence, as soon as we leave the realm of GL_n!

4. Moduli spaces of shtukas

In the function field setting, we have seen that automorphic forms are certain functions on the set $\mathrm{Bun}_G(\mathbb{F}_q)$, that classifies isomorphism classes of G-bundles on the curve X. One goal of this section is to explain that this set has a geometric structure, as the $\mathbb{F}_q$-points of the *moduli stack* of G-bundles on X. We aim to give

[5]This heuristic was originally suggested by Drinfeld and recently made rigorous in [Zhu20]. We discuss this perspective more in § 5.

an idea of the geometry of Bun_G and of the geometry of the related moduli stacks of Hecke modifications.

We then describe much more general geometric objects, that are moduli stacks of *shtukas*[6] on X. The cohomology groups of the moduli stacks of shtukas generalize the spaces of automorphic forms on G and play a fundamental role in the construction of global Langlands parameters in [Laf18a]. We begin to explain why this is the case by discussing a famous lemma of Drinfeld. In this section, we follow the exposition in [Hei18] and [Laf18b].

4.1. **The moduli stack of G-bundles.** It is more natural to consider the set $\mathrm{Bun}_G(\mathbb{F}_q)$ as a *groupoid*, i.e. a category where all the morphisms are invertible. This categorical perspective allows us to keep track of G-bundles on X together with their (finite) groups of automorphisms.

This perspective also leads to the discovery of an algebro-geometric object called Bun_G that parametrizes G-bundles on X. To a scheme S over $\mathbb{F}_q$, Bun_G associates the groupoid $\mathrm{Bun}_G(S)$ of G-bundles on $X \times S$. Roughly, one could think of a G-bundle on $X \times S$ as a family of G-bundles on X parametrized by the points of the scheme S. The geometric object Bun_G is an *Artin stack* defined over $\mathbb{F}_q$. Rather than give the precise definition, we mention that a typical example of an Artin stack is a quotient of an algebraic variety by an algebraic group; one can show that Bun_G looks like this locally.

To recover the groupoid $\mathrm{Bun}_G(\mathbb{F}_q)$ we take the $\mathbb{F}_q$-valued points of Bun_G. Recall that an $\mathbb{F}_q$-valued point is the same as a geometric point that is fixed by the Frobenius automorphism Frob_q, so we could also recover $\mathrm{Bun}_G(\mathbb{F}_q)$ by taking Frobenius fixed points.

We can upgrade the diagram (3.12.1) that is used to define Hecke operators to a diagram of (ind-)stacks

(4.1.1)
$$\begin{array}{ccc} & \mathrm{Hecke}_G & \\ \swarrow & & \searrow \\ \mathrm{Bun}_G & & \mathrm{Bun}_G \times X, \end{array}$$

where the object at the top is the so-called *Hecke stack*. For S a scheme over $\mathbb{F}_q$, the groupoid $\mathrm{Hecke}_G(S)$ is the category of tuples

$$(x, \mathcal{E}, \mathcal{E}', \phi : \ \mathcal{E} - \stackrel{\phi}{-} \to \mathcal{E}' \ )$$

where $x : S \to X$ is a point on the curve, $\mathcal{E}$ and $\mathcal{E}$ are G-bundles on $X \times S$, so objects in $\mathrm{Bun}_G(S)$, and $\phi : \ \mathcal{E} - \stackrel{\phi}{-} \to \mathcal{E}'$ is a modification at x, i.e. an isomorphism away from the graph $\Gamma_x \subset X \times S$ of the point x. The advantage of (4.1.1) over (3.12.1) is that the situation is much more geometric now, and we can allow the point x to move along the curve X.

More generally, we can even consider a version $\mathrm{Hecke}_{G,I}$ of the Hecke stack that parametrizes modifications ϕ at a finite set $(x_i : S \to X)_{i \in I}$ of S-valued points of X. This version lives over a product X^I of I copies of the curve X. As we have done before to define Hecke operators, we can consider substacks of $\mathrm{Hecke}_{G,I}$ where we bound the poles of a modification ϕ at each x_i in terms of an irreducible algebraic representation W_i of $\widehat{G}$. We can still move the points x_i along the curve

[6]The term *shtuka* was introduced by Drinfeld; it means "thing" in Russian.

X, but the advantage of having more than one copy of the curve is that we can also allow different points to collide.

These ideas can be made precise and ultimately lead to the *geometric Satake equivalence*, that relates the geometry of Hecke stacks for G with the representation theory of the Langlands dual group $\widehat{G}$. This equivalence is due to Lusztig, Drinfeld, Ginzburg and Mirkovic-Vilonen [MV07] and plays a fundamental role in the Langlands program over function fields and, in particular, in [Laf18a].

4.2. **Moduli stacks of shtukas.** We now construct some moduli stacks that generalize the groupoid $\mathrm{Bun}_G(\mathbb{F}_q)$, which can be thought of as a discrete, i.e. zero-dimensional stack. These generalizations are moduli stacks of G-shtukas. They were first introduced by Drinfeld for $G = \mathrm{GL}_n$ [Dri80] and then generalized by Varshavsky [Var04] to all reductive groups G, and with an arbitrary number of "legs". The moduli stacks of shtukas combine, in a precise sense, the Hecke stacks we have seen above together with taking Frobenius-fixed points.

Let I be a finite set and, for each $i \in I$, choose an irreducible algebraic representation W_i of $\widehat{G}$ over $\mathbb{Q}_\ell$. We can then form the representation

$$W := \boxtimes_{i\in I} W_i$$

of the product $\widehat{G}^I$ of I copies of $\widehat{G}$. We define $\mathrm{Sht}_{I,W}$ to be the (underlying reduced) stack over X^I whose points over an $\mathbb{F}_q$-scheme S classify G-shtukas. More precisely, the objects of the groupoid $\mathrm{Sht}_{I,W}(S)$ are:

- points $(x_i)_{i\in I} : S \to X^I$ called the legs of the shtuka;
- a G-bundle $\mathcal{E}$ on $X \times S$.
- an isomorphism

$$\phi : \mathcal{E} \mid_{X\times S\setminus \cup_{i\in I}\Gamma_{x_i}} \xrightarrow{\sim} (\mathrm{Id}_X \times \mathrm{Frob}_S)^*\mathcal{E} \mid_{X\times S\setminus \cup_{i\in I}\Gamma_{x_i}},$$

  such that the relative position at x_i of the modification ϕ is bounded in terms of the representation W_i for each $i \in I$.

The moduli stack $\mathrm{Sht}_{I,W}$ naturally lives over X^I, by the map that sends a shtuka to its legs. This fact is the first hint that the cohomology groups of $\mathrm{Sht}_{I,W}$ could provide a link to Langlands parameters.

Remark 4.3. The stack $\mathrm{Sht}_{I,W}$ is a *Deligne–Mumford stack*, which is a particular case of an Artin stack, but is a much nicer geometric object. The typical example of a Deligne–Mumford stack is the quotient of an algebraic variety by a finite étale group scheme. One can show that $\mathrm{Sht}_{I,W}$ looks like this locally. In topology, the corresponding notion is that of an *orbifold*.

Example 4.4. Set $I = \emptyset$ and let $W = \mathbf{1}$ be the trivial representation. We claim that $\mathrm{Sht}_{\emptyset,\mathbf{1}}$ can be identified with the discrete stack $\mathrm{Bun}_G(\mathbb{F}_q)$. To see this, note that the moduli problem for $\mathrm{Sht}_{\emptyset,\mathbf{1}}$ consists of G-bundles $\mathcal{E}$ on $X \times S$ together with an isomorphism

$$\mathcal{E} \xrightarrow{\sim} (\mathrm{Id}_X \times \mathrm{Frob}_S)^*\mathcal{E}.$$

Since G-bundles on $X \times S$ are classified by $\mathrm{Bun}_G(S)$, we see that $\mathrm{Sht}_{\emptyset,\mathbf{1}}$ consists precisely of the Frob_q-fixed points of Bun_G. These are precisely the $\mathbb{F}_q$-valued points.

We will be interested in $H^{\mathrm{cusp}}_{I,W}$, the cuspidal part of the étale cohomology[7] of $\mathrm{Sht}_{I,W}$. We remark that the original condition imposed in [Laf18a] is a technical condition called Hecke-finiteness. This is equivalent to cuspidality by work of Xue [Xue20]. The vector spaces $H^{\mathrm{cusp}}_{I,W}$ are also known to be finite-dimensional $\mathbb{Q}_\ell$-vector spaces by [Xue20].

By Example 4.4, the finite-dimensional vector spaces $H^{\mathrm{cusp}}_{I,W}$ generalize the space of cuspidal automorphic forms on G. However, when I is non-empty, the $H^{\mathrm{cusp}}_{I,W}$ contain much more information: we claim that each $H^{\mathrm{cusp}}_{I,W}$ is a continuous representation of the group Γ_F^I, obtained by taking a product of I copies of the absolute Galois group Γ_F. The fact that there are several copies of the absolute Galois group acting is fundamental and quite subtle; it relies on a famous lemma of Drinfeld, which we discuss in the next subsection.

4.5. **Drinfeld's lemma.** Let $U \subseteq X$ be an open dense subset. For any element i of a finite set I, we can define an ith partial Frobenius morphism Frob_i on the self-product U^I. Explicitly, this is given by

$$\mathrm{Frob}_i : U^I \to U^I, \mathrm{Frob}_i(x_i) = \mathrm{Frob}_U(x_i) \text{ and } \mathrm{Frob}_i(x_j) = x_j \text{ for all } j \neq i.$$

For any scheme $Y \to U^I$, we say that a morphism $F : Y \to Y$ lies "above" Frob_i if the diagram

$$\begin{array}{ccc} Y & \xrightarrow{F} & Y \\ \downarrow & & \downarrow \\ U^I & \xrightarrow{\mathrm{Frob}_i} & U^I \end{array}$$

is commutative.

Lemma 4.5.1 (Drinfeld [Dri80, Laf02, Lau04])**.** *There is an equivalence of categories between:*

- *The category of finite sets equipped with a continuous action of $\pi_1(U, \bar{\eta})^I$.*
- *The category of finite étale covers $Y \to U^I$, equipped with* partial Frobenius morphisms*, i.e. morphisms $F_i : Y \to Y$ above each Frob_i with $i \in I$, that commute with each other and whose composition is equal to Frob_Y.*

Remark 4.6. As we have seen above, a finite étale cover $Y \to U^I$ is equivalent to a finite set equipped with a continuous action of the étale fundamental group $\pi_1(U^I, \bar{\eta}^I)$. Lemma 4.5.1 says that the additional structure needed to upgrade this to a continuous representation of $\pi_1(U, \bar{\eta})^I$ is given by the partial Frobenius morphisms.

We would like to apply Lemma 4.5.1, at least when I is non-empty, to show that the cohomology groups $H^{\mathrm{cusp}}_{I,W}$ are endowed with continuous actions of Γ_F^I. To achieve this, we first introduce certain generalizations $\mathrm{Sht}'_{I,W}$ of $\mathrm{Sht}_{I,W}$, where we factor the modification ϕ as a composition of modifications supported at each of the points $(x_i)_{i\in I}$.

Choose an identification of the finite set I with $\{0, \ldots, n-1\}$ for some positive integer n. We let $\mathrm{Sht}'_{I,W}$ be the (underlying reduced) Deligne–Mumford stack whose

[7]To be precise, we consider the middle degree ℓ-adic intersection cohomology with compact support of the fiber of $\mathrm{Sht}_{I,W}/\Xi$ over a generic geometric point of X^I.

points over an $\mathbb{F}_q$-scheme S classify tuples

$$\left((x_i)_{i\in I},\ \mathcal{E}_0 \overset{\phi_1}{\dashrightarrow} \mathcal{E}_1 \overset{\phi_2}{\dashrightarrow} \ldots \dashrightarrow \mathcal{E}_{n-1} \overset{\phi_n}{\dashrightarrow} \mathcal{E}_n\right), \tag{4.6.1}$$

where

- $x_i : S \to X$ are S-points of X for $i \in I$;
- for $i \in I$, $\mathcal{E}_i$ is a G-bundle on $X \times S$;
- we have $\mathcal{E}_n := (\mathrm{Id}_X \times \mathrm{Frob}_S)^*\mathcal{E}_0$ by definition and, for all $i \in I$, there is an isomorphism

$$\phi_{i+1} : \mathcal{E}_i \mid_{X\times S\setminus\Gamma_{x_i}} \xrightarrow{\sim} \mathcal{E}_{i+1} \mid_{X\times S\setminus\Gamma_{x_i}} .$$

  such that the relative position of $\mathcal{E}_i$ with respect to $\mathcal{E}_{i+1}$ at x_i is bounded in terms of the representation W_i.

There is an obvious morphism

$$\mathrm{Sht}'_{I,W} \to \mathrm{Sht}_{I,W}$$

that forgets the intermediate modifications. It turns out that this induces an isomorphism on the level of the cohomology groups in which we are interested.

The advantage of considering $\mathrm{Sht}'_{I,W}$ over $\mathrm{Sht}_{I,W}$ is that the former can be equipped with partial Frobenius morphisms that are defined moduli-theoretically. Indeed, we can consider the morphism $F_0 : \mathrm{Sht}'_{I,W} \to \mathrm{Sht}'_{I,W}$ that shifts the tuple in (4.6.1) one step to the left. Explicitly, it sends it to the tuple

$$\left((x'_i)_{i\in I},\ \mathcal{E}_1 \overset{\phi_2}{\dashrightarrow} \ldots \dashrightarrow \mathcal{E}_{n-1} \overset{\phi_n}{\dashrightarrow} \mathcal{E}_n \overset{\phi_{n+1}}{\dashrightarrow} \mathcal{E}_{n+1}\right),$$

with $x'_0 := \mathrm{Frob}_X(x_0)$, $x'_i := x_i$ for $i \geq 1$, $\mathcal{E}_{n+1} := (\mathrm{Id}_X \times \mathrm{Frob}_S)^*\mathcal{E}_1$ and $\phi_{n+1} := (\mathrm{Id}_X \times \mathrm{Frob}_S)^*\phi_1$.

Recall our identification of I with the set $\{0, 1, \ldots, n-1\}$. The morphism F_0 lies above the partial Frobenius Frob_0 on X^I. We are allowed to permute the $(x_i)_{i\in I}$, because we can do so over the open subset of X^I where the points are pairwise distinct. This means that we can also construct commuting partial Frobenius morphisms $F_1, \ldots, F_{n-1}$. By a version of Lemma 4.5.1, these partial Frobenius morphisms are precisely the extra structures we need to endow the cohomology groups $H^{\mathrm{cusp}}_{I,W}$ with a continuous action of Γ_F^I.

5. Excursion operators and Galois representations

The goal of this section is to explain how to go from the cohomology of the moduli spaces of shtukas discussed in § 4 to Galois representations. It turns out that we can take these cohomology groups as a black box for this part of the argument. We formalize what we know so far into a system of functors that satisfy certain compatibilities, and we explain heuristically how these functors give rise to a semi-simple Galois representation. In addition to [Laf18a] and [Laf18b], we also follow the perspective developed in [Zhu20].

Let $\widehat{G}$ be a split reductive group over $\mathbb{F}_q$. Let $\mathrm{Rep}\,\widehat{G}$ denote the category of finite-dimensional algebraic representations of $\widehat{G}$ on $\mathbb{Q}_\ell$-vector spaces. Let $\mathrm{Rep}\,\Gamma_F$ denote the category of continuous, finite-dimensional representations of Γ_F on $\mathbb{Q}_\ell$-vector spaces. Assume that we have a system of $\mathbb{Q}_\ell$-linear functors

$$I : \mathrm{Rep}\,\widehat{G}^I \to \mathrm{Rep}\,\Gamma_F^I, \quad W \mapsto H_I(W),$$

where I runs over the category of (possibly empty) finite sets. Concretely, this means that, for every $\widehat{G}^I$-equivariant morphism

$$u : W \to W',$$

there exists a Γ_I^F-equivariant morphism

$$H_I(u) : H_I(W) \to H_I(W').$$

Assume also that the functors H_I satisfy certain compatibilities. More precisely, every map of finite sets $\zeta : I \to J$ induces a diagonal morphism

$$\widehat{G}^J \to \widehat{G}^I, (g_j)_{j\in J} \mapsto (g_{\zeta(i)})_{i\in I}$$

which in turn induces a restriction functor on the level of representations $\mathrm{Rep}\ \widehat{G}^I \to \mathrm{Rep}\ \widehat{G}^J$, $W \mapsto W^\zeta$. Then the system of functors H_I should be equipped with isomorphisms

$$\chi_\zeta : H_I(W) \xrightarrow{\sim} H_J(W^\zeta)$$

which satisfy the following properties:

- these isomorphisms are functorial in W;
- they are Γ_F^J-equivariant, where the action of Γ_F^J on $H_I(W)$ factors through the diagonal morphism $\Gamma_F^J \to \Gamma_F^I$, $(\gamma_i)_{i\in J} \mapsto (\gamma_{\zeta(i)})_{i\in I}$;
- they are compatible with composition.

Example 5.1. If W is an irreducible representation of $\widehat{G}^I$, we could take

$$H_I(W) := H^{\mathrm{cusp}}_{I,W},$$

the cuspidal part of the cohomology of the moduli stack of shtukas with legs indexed by the finite set I and modifications bounded by W, as described in § 4. In particular, this gives

$$H_\emptyset(\mathbf{1}) = \mathfrak{H} = C_c^{\mathrm{cusp}}(\mathrm{Bun}_G(\mathbb{F}_q)/\Xi, \mathbb{Q}_\ell),$$

the space of cuspidal automorphic forms on G. Associated to the map $\emptyset \to \{0\}$ there is an isomorphism

$$H_\emptyset(\mathbf{1}) \xrightarrow{\sim} H_{\{0\}}(\mathbf{1}),$$

where the LHS parametrizes shtukas with no legs and the RHS parametrizes shtukas with an inactive leg.

More generally, the geometric Satake equivalence allows us to define each H_I as a functor $\mathrm{Rep}\ \widehat{G}^I \to \mathrm{Rep}\ \Gamma_F^I$ and to prove that these functors satisfy the desired additional compatibilities.

Given such a system of functors, satisfying all these compatibilities, V. Lafforgue proves Theorem 3.16 in three steps.

(1) He constructs an action of a certain excursion algebra $\mathfrak{B}$ on $H_\emptyset(\mathbf{1})$.
(2) A $\widehat{G}$-valued *pseudo-representation* is a notion that generalizes the characteristic polynomial of a representation when $\widehat{G} = \mathrm{GL}_n$. He extracts the data of a $\widehat{G}$-valued pseudo-representation of Γ_F from the action of the excursion algebra $\mathfrak{B}$.
(3) Finally, he goes from $\widehat{G}$-valued pseudo-representations to $\widehat{G}$-valued semi-simple representations.

We first recall the explicit construction of excursion operators on $H_\emptyset(\mathbf{1})$, then we give a more conceptual explanation of this action in terms of the moduli stack of $\widehat{G}(\overline{\mathbb{Q}}_\ell)$-valued representations of Γ_F.

Let I be a non-empty finite set. Let $(\gamma_i)_{i\in I} \in \Gamma_F^I$. We let $\zeta : I \to \{*\}$ be the unique map. For any $W \in \mathrm{Rep}\, \widehat{G}^I$, we write $W^\vee$ for the $\mathbb{Q}_\ell$-linear dual of W. Choose $x \in W^{\Delta(\widehat{G})}$, $\xi \in (W^\vee)^{\Delta(\widehat{G})}$; by definition, these give rise to $\widehat{G}$-equivariant morphisms $x : \mathbf{1} \to W^\zeta$ and $\xi : W^\zeta \to \mathbf{1}$. We can define an excursion operator as the following composition:

$$\begin{array}{ccccccc} H_\emptyset(\mathbf{1}) & \xrightarrow{\sim} & H_{\{*\}}(\mathbf{1}) & \xrightarrow{H_{\{*\}}(x)} & H_{\{*\}}(W) & \xrightarrow{\chi_\zeta^{-1}}_{\sim} & H_I(W) \\ & & & & & & \downarrow{\scriptstyle (\gamma_i)_{i\in I}} \\ H_\emptyset(\mathbf{1}) & \xleftarrow{\sim} & H_{\{*\}}(\mathbf{1}) & \xleftarrow[H_{\{*\}}(\xi)]{} & H_{\{*\}}(W) & \xleftarrow[\chi_\zeta]{\sim} & H_I(W). \end{array} \tag{5.1.1}$$

The excursion operator consists of three steps:

- a *creation* step induced by x, which creates I legs over the generic point of the curve X;
- the action of $(\gamma_i)_{i\in I}$, which moves the I legs shtuka independently and brings them back to the same generic point;
- an *annihilator* operator induced by ξ, which annihilates the I legs.

According to [Laf18b], this is called an *excursion operator* because it moves around the legs of the shtuka. Using the properties of the system of functors $(H_I)_I$, one can prove that the excursion operators satisfy certain important compatibilities. A crucial one is the following fact, which will make the link to pseudo-representations.

Fact 5.1.1. *The excursion operator defined in* (5.1.1) *only depends on* $(\gamma_i) \in \Gamma_F^I$ *and on the function*

$$f : \widehat{G}^I \to \mathbb{Q}_\ell, (g_i) \mapsto \langle \xi, (g_i)x \rangle.$$

We therefore denote the excursion operator by $S_{I,f,(\gamma_i)}$.

As I, f *and* (γ_i) *vary, the* $S_{I,f,(\gamma_i)}$ *generate a commutative algebra* $\mathfrak{B}$.

Note that the function f is invariant under the diagonal action of $\widehat{G}$ on both the left and the right; denote these by $\Delta(a_l(\widehat{G}))$ and $\Delta(a_r(\widehat{G}))$ respectively. If we rewrite our non-empty finite set as $I \sqcup \{*\}$, we have excursion operators defined in terms of

$$\mathbb{Q}_\ell[\widehat{G}^{I\cup\{*\}}]^{\Delta(a_l(\widehat{G}))\times\Delta(a_r(\widehat{G}))} \simeq \mathbb{Q}_\ell[\widehat{G}^I]^{\mathrm{Ad}\widehat{G}}.$$

The algebra on the RHS is by definition the algebra of regular functions on the geometric invariant theory (GIT) quotient $\widehat{G}^I//\widehat{G}$, where the action of $\widehat{G}$ on $\widehat{G}^I$ is the adjoint action, given by conjugation.

The heuristic is now the following. We should think of the I-tuple $(\gamma_i) \in \Gamma_F^I$ as a homomorphism from the free group on I generators to the absolute Galois group Γ_F, giving a way to "probe" Γ_F by a free finitely generated group. If we forgot about the topology on Γ_F, and viewed it merely in the category of groups, we could recover it as the direct limit of its finitely generated subgroups. This means that we could reconstruct Γ_F by probing it with finite sets I.

The stacky quotient $\widehat{G}^I/\widehat{G}$ is a moduli of $\widehat{G}$-valued representations of the free group on I generators, up to $\widehat{G}$-conjugacy. The associated GIT quotient $\widehat{G}^I//\widehat{G}$

parametrizes $\widehat{G}$-valued pseudo-representations. As I ranges over all finite sets, the direct limit of the algebras $\mathbb{Q}_\ell[\widehat{G}^I]^{\mathrm{Ad}\widehat{G}}$ should recover the algebra of regular functions on the GIT quotient of $\widehat{G}$-valued representations of Γ_F, up to $\widehat{G}$-valued conjugacy. On the other hand, the direct limit of the algebras $\mathbb{Q}_\ell[\widehat{G}^I]^{\mathrm{Ad}\widehat{G}}$ over all finite sets I is precisely the algebra $\mathfrak{B}$ of all excursion operators. Therefore, a maximal ideal of $\mathfrak{B}$ should be the same data as a $\widehat{G}(\overline{\mathbb{Q}}_\ell)$-valued pseudo-representation of Γ_F. In turn, the latter is equivalent to a semi-simple $\widehat{G}(\overline{\mathbb{Q}}_\ell)$-valued representation.

Example 5.2. Let Γ be the free group on one element and $\widehat{G} = \mathrm{GL}_n/\mathbb{Q}_\ell$. We have a natural map

$$\mathrm{Tr} : \mathrm{Hom}(\Gamma, \mathrm{GL}_n)/\mathrm{GL}_n \to \mathrm{Hom}(\Gamma, \mathrm{GL}_n)//\mathrm{GL}_n$$

from the stacky moduli of GL_n-valued representations of Γ, up to conjugacy, to the moduli of GL_n-valued pseudo-representations. $\mathrm{Hom}(\Gamma, \mathrm{GL}_n)//\mathrm{GL}_n$ is an open subset of n-dimensional affine space $\mathbb{A}^n_{\mathbb{Q}_\ell}$ and Tr is the map that sends a matrix up to conjugacy to its characteristic polynomial. In turn, the $\overline{\mathbb{Q}}_\ell$-valued points of $\mathrm{Hom}(\Gamma, \mathrm{GL}_n)//\mathrm{GL}_n$ are in bijection with semi-simple conjugacy classes in $\mathrm{GL}_n(\overline{\mathbb{Q}}_\ell)$.

The above heuristic is made rigorous through the following theorem proved in [Laf18a].

Theorem 5.3. *For each $\overline{\mathbb{Q}}_\ell$-valued maximal ideal ν of $\mathfrak{B}$ there exists a unique global Langlands parameter ρ, such that the following equality holds for all I, f and $(\gamma_i)_{i\in I}$:*

$$\nu(S_{I,f,(\gamma_i)_{i\in I}}) = f((\rho(\gamma_i))_{i\in I}). \tag{5.3.1}$$

The key ingredient in the proof of this theorem is a result of Richardson [Ric88], which identifies the points of the GIT quotient $\widehat{G}^I//\widehat{G}$ with *semi-simple* conjugacy classes: conjugacy classes of tuples $(g_i)_{i\in I} \in \widehat{G}$ such that the Zariski closure of the subgroup of $\widehat{G}$ generated by the g_i is itself a reductive group. This uses the fact that there is a bijection between the points of the quotient $\widehat{G}^I//\widehat{G}$ and the closed orbits of the conjugation action of $\widehat{G}$ on $\widehat{G}^I$.

Finally, in order to deduce Theorem 3.16 from Theorem 5.3, one needs to express Hecke operators as particular instances of excursion operators. Let v be a place of X and W be an irreducible algebraic representation of $\widehat{G}$. It turns out that the Hecke operator $T_{v,W}$ can be recovered as the excursion operator $S_{\{1,2\},f,(\mathrm{Frob}_v,1)}$, where $f \in \overline{\mathbb{Q}}_\ell[\widehat{G}\backslash\widehat{G}^2/\widehat{G}]$ is defined by $f((g_1,g_2)) := \mathrm{tr}_W(g_1 g_2^{-1})$. This compatibility between Hecke operators and excursion operators is not at all obvious. Observe that it follows from (5.3.1) that, for each maximal ideal ν of $\mathfrak{B}$ with associated Langlands parameter ρ, we have the equality

$$\nu(S_{\{1,2\},f,(\mathrm{Frob}_v,1)}) = \mathrm{tr}(W \circ \rho)(\mathrm{Frob}_v),$$

which is reminiscent of the Galois side of Eichler–Shimura relations such as (3.14.1) and (3.14.2). The compatibility between Hecke operators and excursion operators is proved in [Laf18a] using a geometric argument and represents a vast generalization of the Eichler–Shimura relations.

References

[BHKT19] Gebhard Böckle, Michael Harris, Chandrashekhar Khare, and Jack A. Thorne, *$\hat{G}$-local systems on smooth projective curves are potentially automorphic*, Acta Math. **223** (2019), no. 1, 1–111.

[Bla94] Don Blasius, *On multiplicities for* SL(n), Israel J. Math. **88** (1994), no. 1-3, 237–251.

[Del74] Pierre Deligne, *La conjecture de Weil. I*, Inst. Hautes Études Sci. Publ. Math. (1974), no. 43, 273–307.

[Dri80] V. G. Drinfeld, *Langlands' conjecture for* GL(2) *over functional fields*, Proceedings of the International Congress of Mathematicians (Helsinki, 1978), Acad. Sci. Fennica, Helsinki, 1980, pp. 565–574.

[Dri87a] ———, *Cohomology of compactified moduli varieties of F-sheaves of rank* 2, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **162** (1987), no. Avtomorfn. Funkts. i Teor. Chisel. III, 107–158, 189.

[Dri87b] ———, *Moduli varieties of F-sheaves*, Funktsional. Anal. i Prilozhen. **21** (1987), no. 2, 23–41.

[Dri88] ———, *Proof of the Petersson conjecture for* GL(2) *over a global field of characteristic p*, Funktsional. Anal. i Prilozhen. **22** (1988), no. 1, 34–54, 96.

[Eme13] Matthew Emerton, *Galoisian sets of prime numbers*, MathOverflow, 2013, URL: https://mathoverflow.net/q/12382 (version: 2013-06-30).

[Eme20] Matthew Emerton, *Langlands reciprocity: L-functions, automorphic forms, and Diophantine equations*, to appear in "The Genesis of the Langlands program" (2020).

[FS20] Laurent Fargues and Peter Scholze, *Geometrization of the local langlands correspondence*, in preparation, 2020.

[GL17] Alain Genestier and Vincent Lafforgue, *Chtoucas restreints pour les groupes réductifs et paramétrisation de Langlands locale*, arXiv e-prints (2017), arXiv:1709.00978.

[Hei18] Jochen Heinloth, *Langlands parameterization over function fields following V. Lafforgue*, Acta Math. Vietnam. **43** (2018), no. 1, 45–66.

[Laf02] Laurent Lafforgue, *Chtoucas de Drinfeld et correspondance de Langlands*, Invent. Math. **147** (2002), no. 1, 1–241.

[Laf18a] Vincent Lafforgue, *Chtoucas pour les groupes réductifs et paramétrisation de Langlands globale*, J. Amer. Math. Soc. **31** (2018), no. 3, 719–891.

[Laf18b] ———, *Shtukas for reductive groups and Langlands correspondence for function fields*, Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. I. Plenary lectures, World Sci. Publ., Hackensack, NJ, 2018, pp. 635–668.

[Lap99] Erez M. Lapid, *Some results on multiplicities for* SL(n), Israel J. Math. **112** (1999), 157–186.

[Lar94] Michael Larsen, *On the conjugacy of element-conjugate homomorphisms*, Israel J. Math. **88** (1994), no. 1-3, 253–277.

[Lar96] ———, *On the conjugacy of element-conjugate homomorphisms. II*, Quart. J. Math. Oxford Ser. (2) **47** (1996), no. 185, 73–85.

[Lau04] Eike Sören Lau, *On generalised D-shtukas*, Bonner Mathematische Schriften [Bonn Mathematical Publications], vol. 369, Universität Bonn, Mathematisches Institut, Bonn, 2004, Dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, 2004.

[LMF20] The LMFDB Collaboration, *The L-functions and modular forms database*, `http://www.lmfdb.org`, 2020, [Online; accessed 8 December 2020].

[LZ18] Vincent Lafforgue and Xinwen Zhu, *Décomposition au-dessus des paramètres de Langlands elliptiques*, arXiv e-prints (2018), arXiv:1811.07976.

[MV07] I. Mirković and K. Vilonen, *Geometric Langlands duality and representations of algebraic groups over commutative rings*, Ann. of Math. (2) **166** (2007), no. 1, 95–143.

[Ric88] R. W. Richardson, *Conjugacy classes of n-tuples in Lie algebras and algebraic groups*, Duke Math. J. **57** (1988), no. 1, 1–35.

[Sch18] Peter Scholze, *p-adic geometry*, Proceedings of the International Congress of Mathematicians—Rio de Janeiro 2018. Vol. I. Plenary lectures, World Sci. Publ., Hackensack, NJ, 2018, pp. 899–933.

[Str17] Benoît Stroh, *La paramétrisation de Langlands globale sur les corps de fonctions*, no. 390, 2017, Séminaire Bourbaki. Vol. 2015/2016. Exposés 1104–1119, pp. Exp. No. 1110, 169–197.

[TW95] Richard Taylor and Andrew Wiles, *Ring-theoretic properties of certain Hecke algebras*, Ann. of Math. (2) **141** (1995), no. 3, 553–572.

[Var04] Yakov Varshavsky, *Moduli spaces of principal F-bundles*, Selecta Math. (N.S.) **10** (2004), no. 1, 131–166.

[Wei49] André Weil, *Numbers of solutions of equations in finite fields*, Bull. Amer. Math. Soc. **55** (1949), 497–508.

[Wei16] Jared Weinstein, *Reciprocity laws and Galois representations: recent breakthroughs*, Bull. Amer. Math. Soc. (N.S.) **53** (2016), no. 1, 1–39.

[Wil95] Andrew Wiles, *Modular elliptic curves and Fermat's last theorem*, Ann. of Math. (2) **141** (1995), no. 3, 443–551.

[Wym72] B. F. Wyman, *What is a reciprocity law?*, Amer. Math. Monthly **79** (1972), 571–586; correction, ibid. 80 (1973), 281.

[Xue20] Cong Xue, *Cuspidal cohomology of stacks of shtukas*, Compos. Math. **156** (2020), no. 6, 1079–1151.

[XZ17] Liang Xiao and Xinwen Zhu, *Cycles on Shimura varieties via geometric Satake*, arXiv e-prints (2017), arXiv:1707.05700.

[Zhu20] Xinwen Zhu, *Coherent sheaves on the stack of Langlands parameters*, arXiv e-prints (2020), arXiv:2008.02998.

Email address: a.caraiani@imperial.ac.uk

Imperial College London, 180 Queen's Gate, Kensington, London SW7 2AZ

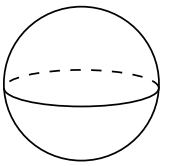
GETTING A HANDLE ON THE CONWAY KNOT

JENNIFER HOM

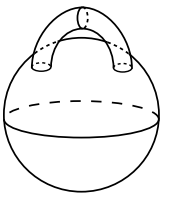
Abstract. A knot is said to be slice if it bounds a smooth disk in the 4-ball. For 50 years, it was unknown whether a certain 11 crossing knot, called the Conway knot, was slice or not, and until recently, this was the only one of the thousands of knots with fewer than 13 crossings whose slice-status remained a mystery. We will describe Lisa Piccirillo's proof that the Conway knot is not slice. The main idea of her proof is given in the title of this talk.

1. Introduction

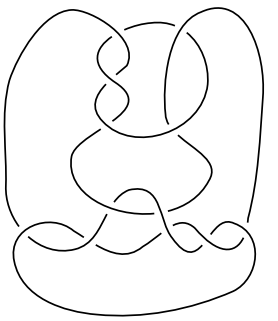
Here is a 3-ball:



and here is a 3-ball with a handle attached:



This is the Conway knot:

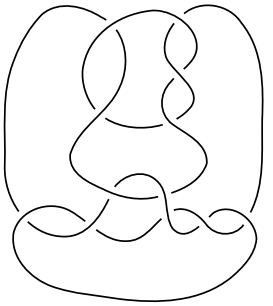


2020 *Mathematics Subject Classification.* Primary 57K10.
The author was partially supported by NSF grant DMS-1552285.

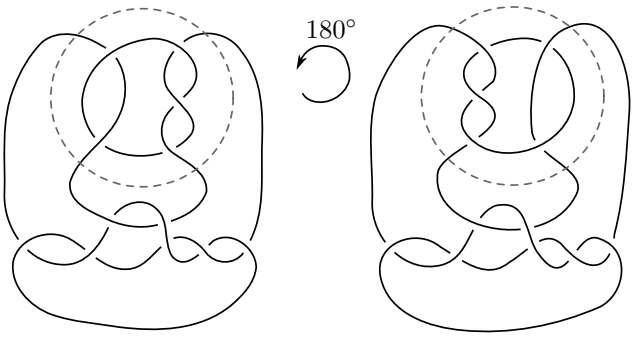
Our knots will live in the 3-sphere S^3, which is the boundary of the 4-ball B^4. A knot is *slice* if it bounds a smooth disk in the 4-ball. The term slice comes from the fact that such knots are cross sections (i.e., slices) of higher dimensional knots.

Main Theorem (Piccirillo [Pic20])**.** *The Conway knot is not slice.*

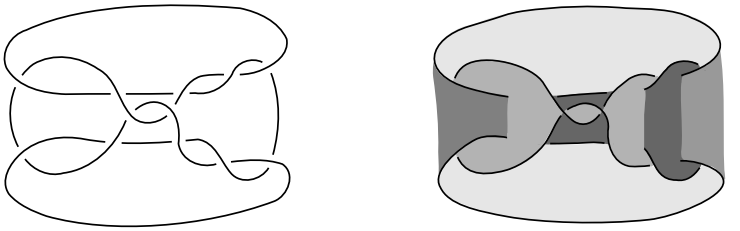
This knot is not the Conway knot:



It is called the Kinoshita-Terasaka knot, and it is related to the Conway knot by *mutation*, that is, we cut out a ball containing part of the knot, rotate it 180°, and glue it back in.



The Kinoshita-Terasaka knot is slice. Here is a slightly different diagram of the Kinoshita-Terasaka knot. As we can see, it bounds an immersed disk in S^3:



Thinking of this immersed disk as sitting in the S^3 boundary of the 4-ball, we can push the surface into the 4-ball and eliminate the arcs of self-intersection by pushing one sheet of the surface near the arc deeper into the 4-ball, giving us an embedded disk in the 4-ball.

One way to study knots is to use a knot invariant. A knot invariant is a mathematical object (like a number, a polynomial, or a group) that we assign to a knot.

Knot invariants can be used to distinguish knots. Certain knot invariants obstruct a knot from being slice. One such invariant is Rasmussen's s-invariant, which to a knot K assigns an integer $s(K)$. If $s(K) \neq 0$, then K is not slice.

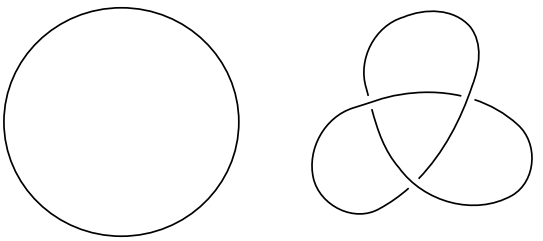
Since the Conway knot and the Kinoshita-Terasaka knots are mutants, they have a lot in common. For example, the s-invariant of both knots is zero. In fact, all known knot invariants that obstruct sliceness vanish for the Conway knot. That leads one to wonder: how did Piccirillo show that the Conway knot is not slice? Her key idea was to find some other knot K' such that the Conway knot is slice if and only if K' is slice, and to obstruct K' from being slice. The goal of these notes is to give some context for her result and sketch the main ideas of her proof.

2. Telling knots apart

The fundamental group is one of the first algebraic invariants encountered in a topology class. A knot is homeomorphic to S^1, so its fundamental group is always isomorphic to the integers. However, instead of studying the knot, we can study the space around the knot. That is, we consider the *knot complement*, consisting of the 3-sphere minus a neighborhood of the knot. The *knot group* is the fundamental group of the knot complement.

Typically, one studies knots up to *ambient isotopy*. Intuitively, this means that we can wiggle and stretch our knot, but we cannot cut it nor let it pass through itself. Since isotopic knots have homeomorphic complements and homeomorphic spaces have isomorphic fundamental groups, the knot group is an invariant of the isotopy class of a knot.

Here are two knots, the unknot and the trefoil:



Example 2.1. The knot group of the unknot is $\mathbb{Z}$.

Example 2.2. The knot group of the trefoil is $\langle x, y \mid x^2 = y^3 \rangle$. This group is non-abelian, since it surjects onto the symmetric group S_3. Therefore, the trefoil and the unknot are different.

Since it can often be difficult to tell if two group presentations describe isomorphic groups, it can be convenient to pass to more tractable invariants. One example is the Alexander polynomial, denoted $\Delta(t)$, which Fox [Fox53] showed can be algorithmically computed from a group presentation for the knot complement.

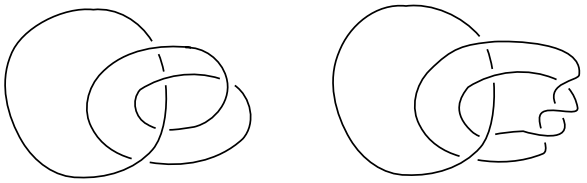
Example 2.3. The Alexander polynomial of the unknot is 1.

Example 2.4. The Alexander polynomial of the trefoil is $t^2 - t + 1$.

Example 2.5. The Conway knot and the Kinoshita-Terasaka knot both have Alexander polynomial 1.

The Alexander polynomial is invariant under mutation, which explains why the Conway knot and the Kinoshita-Terasaka knot have the same Alexander polynomial. There are several other polynomial knots invariants, such as the Jones, HOMFLY-PT, and Kauffman polynomials, all of which are also invariant under mutation. Knot Floer homology [OS04] and Khovanov homology [Kho00] categorify the Alexander and Jones polynomials; that is, to a knot, they assign a graded vector space whose graded Euler characteristic is the desired polynomial. A certain version of knot Floer homology is invariant under mutation [Zib19], as are versions of Khovanov homology [Blo10, Weh10]. Moreover, Rasmussen's s-invariant is invariant under mutation [KWZ19]; this gives a quick way to determine that the s-invariant of the Conway knot is zero, since it is the mutant of a slice knot.

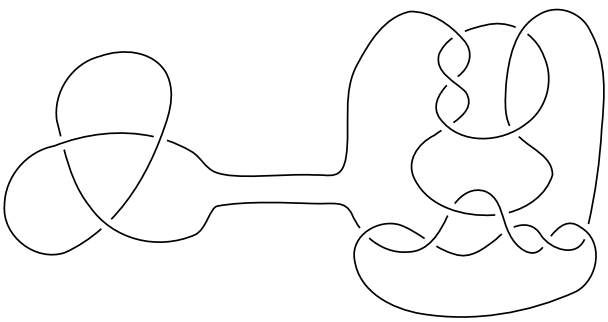
As we already observed, isotopic knots have homeomorphic complements. What about the converse? If two knots have homeomorphic complements, then are they isotopic? This question was answered in the affirmative in 1989 by Cameron Gordon and John Luecke [GL89], who proved that knots are determined by their complements. This is in contrast to links. For example, the two links below have homeomorphic complements, but are not isotopic, since in the first, both components are unknots, while in the second, one component is the trefoil.



3. Measuring the complexity of a knot

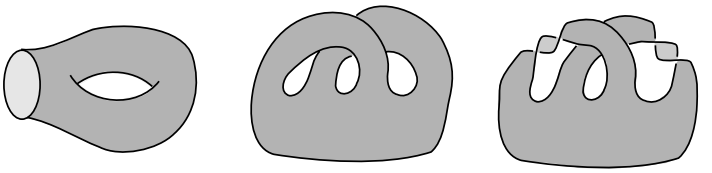
How can we measure the complexity of a knot K? One such measure is the *unknotting number*, denoted $u(K)$, which is the minimal number of times a knot must be passed through itself to untie it. Both the Conway knot and the Kinoshita-Terasaka knot can be unknotted by changing a single crossing, hence the unknotting number is one for both of them. Note that a knot has unknotting number zero if and only if it is the unknot.

There is a natural way to add together two knots K_1 and K_2, called the *connected sum*, denoted $K_1 \# K_2$. Here is the connected sum of the trefoil and the Conway knot:



What is the unknotting number of $K_1\#K_2$? A natural guess is that $u(K_1\#K_2) = u(K_1)+u(K_2)$. One can readily check that $u(K_1\#K_2) \leq u(K_1)+u(K_2)$. However, whether or not the reverse inequality holds remains an open question!

Here is another measure of complexity. Every knot in the 3-sphere bounds a compact, oriented, connected surface. Such surface is called a *Seifert surface* for the knot. Recall that compact, oriented surfaces with connected boundary are characterized up to homeomorphism by their genus. The surfaces below are all have genus one:



The boundary of each of the first two surfaces is the unknot. The boundary of the last surface is the trefoil.

The *genus* of a knot K is the minimal genus of a Seifert surface for K. The unknot is the only knot that bounds a disk. In other words, a knot had genus zero if and only if is the unknot. In contrast to unknotting number, we know how genus behaves under connected sum; Schubert [Sch49] showed that genus is additive under connected sum, that is, $g(K_1\#K_2) = g(K_1) + g(K_2)$.

Example 3.1 ([Gab86])**.** The Conway knot has genus three. The Kinoshita-Terasaka knot has genus two.

The unknot is the only knot with unknotting number zero, and it's also the only knot with genus zero. What about a measure of complexity where there are nontrivial knots that are also simple? Enter the *slice genus*.

Recall that S^3 is the boundary of the 4-ball, and that a knot K in S^3 is *slice* if it bounds a smooth disk in the 4-ball. Such a disk is a called a *slice disk* for K. Not every knot K bounds a smooth disk in the 4-ball, but every knot does bound a smooth compact, oriented, connected surface in the 4-ball. (One way to obtain such a surface is by pushing a Seifert surface for K into the 4-ball.) The minimal genus of such surface is called *slice genus* of K. Slice knots are precisely those knots with slice genus zero. Of course the unknot is slice, but there are also infinitely many nontrivial knots which are slice. For example, the Kinoshita-Terasaka knot is slice. Unlike the ordinary genus of a knot, slice genus is not additive under connected sum.

The Alexander polynomial can obstruct sliceness: if K is slice, then $\Delta_K(t)$ is of the form $t^n f(t) f(t^{-1})$ for some polynomial f and some natural number n.

Example 3.2. The trefoil is not slice, since its Alexander polynomial $t^2 - t + 1$ is irreducible.

Closely related to the notion of sliceness is the following equivalence relation: two knots K_0 and K_1 are *concordant* if they cobound an annulus A in $S^3 \times [0, 1]$,

where the boundary of A is $K_0 \subset S^3 \times \{0\}$ and $K_1 \subset S^3 \times \{1\}$. One can check that a knot is slice if and only if it concordant to the unknot.

Note that we required our surfaces to be smoothly embedded. What would happen if we just asked for topologically embedded disks in B^4? It turns out that every knot bounds a topologically embedded disk in B^4. Recall that the *cone* of a space X is $\mathrm{Cone}(X) = (X \times [0,1])/(X \times \{0\})$. Since $\mathrm{Cone}(S^3, K) = (B^4, B^2)$, every knot K in S^3 bounds a topological disk in B^4, but the disk is not smoothly embedded, because of the cone point. Rather than requiring smoothness, one can instead require that the disk be locally flat; a knot that bounds a locally flat disk is called *topologically slice*. Freedman [Fre83] proved that any knot with Alexander polynomial one is topologically slice; in particular, the Conway knot is topologically slice. Work of Donaldson [Don83] implies that there are topologically slice knots that are not slice. Many slice obstructions actually obstruct topological sliceness, which is part of the reason why showing the Conway knot is not slice is so difficult.

4. An equivalent condition for sliceness

There are many invariants that obstruct sliceness, such as the aforementioned factoring of the Alexander polynomial, integer-valued invariants τ and ν coming from knot Floer homology [OS03, OS11], and Rasmussen's integer-valued invariant s coming from Lee's perturbation of Khovanov homology [Ras10, Lee05]. These invariants (and many more!) all vanish for the Conway knot. (In my PhD thesis, I defined a new slice obstruction. One of the first questions people asked me was what its value was on the Conway knot; sadly, the obstruction vanishes for the Conway knot.)

Recall that in Section 2, starting from a knot K in S^3, we built a 3-manifold, the knot complement. Piccirillo's strategy for showing that the Conway knot is not slice relies on building a 4-manifold, called the *knot trace*, from a knot K in S^3. We will denote the trace of K by $X(K)$. The following folklore result (see [FM66]) is a key ingredient in Piccirillo's proof:

Trace Embedding Lemma. *A knot K is slice if and only if its trace $X(K)$ smoothly embeds in S^4.*

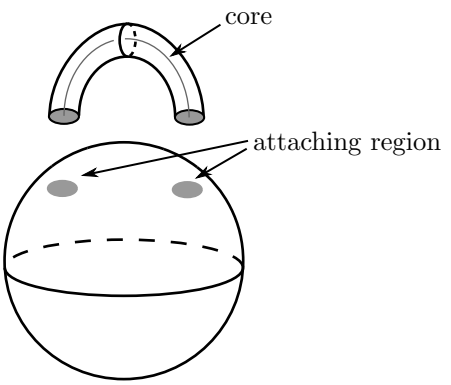
In contrast to the fact that knots are determined by their complements, knots are not determined by their traces. That is, there exist non-isotopic knots K and K' with the same (i.e., diffeomorphic) traces [Akb77]. Allison Miller and Lisa Piccirillo [MP18] proved something even stronger: they showed that there exist knots K and K' with the same trace such that K and K' are not even concordant. This disproved a conjecture of Abe [Abe16]. Miller and Piccirillo's result implies that it's possible to have knots K and K' with the same trace, but for, say, $s(K)$ to be zero while $s(K')$ is nonzero.

We are slowing uncovering Piccirillo's strategy for proving the Conway knot is not slice: find a knot K' with the same trace as the Conway knot, and show that K' is not slice. Then the Trace Embedding Lemma implies that the Conway knot is not slice either.

5. Handles and traces

Let B^n denote the n-ball. Recall the 3-ball with a handle attached from beginning of these notes. More specifically, the handle consists of $B^1 \times B^2$ attached

to $S^2 = \partial B^3$ along $S^0 \times B^3 = \partial B^1 \times B^2$. This handle is called a 3-dimensional 1-handle.



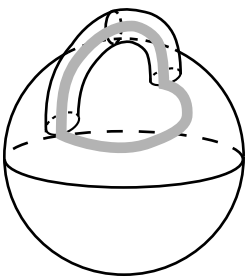
More generally, we consider an n-dimensional k-handle $B^k \times B^{n-k}$. Such a handle can be attached to an n-manifold M with boundary by identifying a submanifold $S^{k-1} \times B^{n-k} \subset \partial M$ with $S^{k-1} \times B^{n-k} = \partial B^k \times B^{n-k}$. The submanifold $S^{k-1} \times B^{n-k} \subset \partial M$ is called the *attaching region* of the handle. The *core* of the handle is $B^k \times \{0\}$, where we think of B^k as the unit ball in $\mathbb{R}^k$.

To build the knot trace, we will consider a 4-dimensional 2-handle $B^2 \times B^2$ attached to $S^3 = \partial B^4$. We need to specify the attaching region $S^1 \times B^2 \subset S^3$. This is just a tubular neighborhood of a knot. (The careful reader will note that we need to specify a parametrization of the neighborhood with $S^1 \times B^2$; this is called the *framing* of the knot. For ease of exposition, we will largely suppress this key point from our discussion.) The *trace* of a knot K is the result of attaching a (0-framed) 2-handle to $S^3 = \partial B^4$ along K. This is just a higher dimensional analog of the 1-handle attached to the 3-ball above.

6. Knots with the same trace

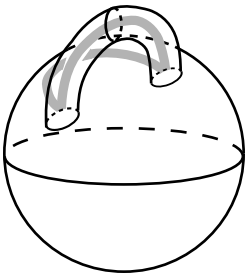
In order to understand Piccirillo's construction of a knot with the same trace as the Conway knot, it will be helpful to consider an analogy one dimension lower, in 3-dimensions, where we can more easily visualize things.

Consider the 3-ball with a 1-handle attached. Recall that a (3-dimensional) 2-handle is just a thickened disk $B^2 \times B^1$, which we attached along an annulus $S^1 \times B^1$. Suppose we attached a 2-handle along the grey annulus:



Observe that the resulting manifold M_1 is homeomorphic (in fact, diffeomorphic, after smoothing corners) to B^3!

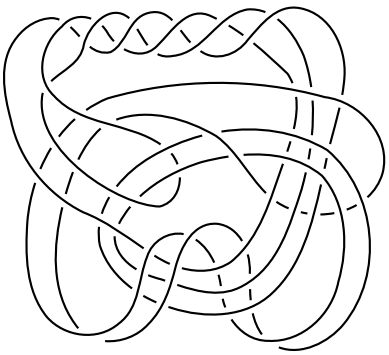
We could instead attach a 2-handle along the following grey thickened curve:



This would yield a manifold, M_2, which is again homeomorphic to B^3.

If we attached 2-handles to both of the grey curves, we obtain a manifold M that is homeomorphic to B^3 with a 2-handle attached. Note that M is built from a 3-ball, one 1-handle, and two 2-handles. We could view M as $M_1 \cong B^3$ with a 2-handle attached or we could view M as $M_2 \cong B^3$ with a 2-handle attached. Notice that the attaching regions for these 2-handles are just (thickened) embedded circles in $S^2 = \partial B^3$. Of course, embedded circles in S^2 are not especially interesting. But what happens when we bump things up a dimension?

Now consider the trace $X(C)$ of the Conway knot C. Piccirillo found a clever way to build $X(C)$ as a 4-ball, a 1-handle, and two 2-handles. (All of the handles here are 4-dimensional.) If you take the 4-ball, the 1-handle, and the first 2-handle, you get a 4-ball, and the second 2-handle is attached along the Conway knot C in S^3 (the boundary of the 4-ball). On the other hand, if you take the 4-ball, the 1-handle, and the second 2-handle, you still get a 4-ball, and the remaining 2-handle is attached along some different knot K'. This means that C and K' have the same trace! Here is Piccirillo's knot K' that has the same trace as the Conway knot:

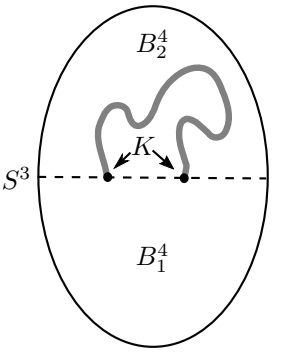


7. Proof of the Trace Embedding Lemma

Now that we have seen handles and traces, we will sketch the proof of the Trace Embedding Lemma.

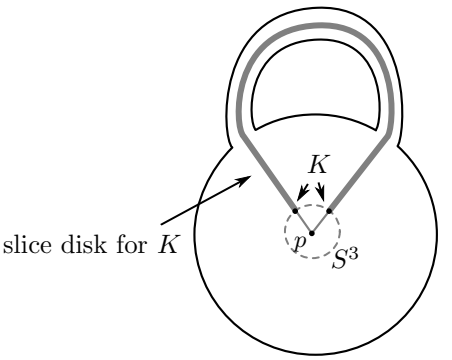
Suppose that K is slice. This means that K bounds a smooth disk in the 4-ball. Recall that S^4 is the union of two 4-balls, say B^4_1 and B^4_2. Think of K as sitting in the common S^3 boundary of these two 4-balls. Since K is slice, it bounds a slice disk D in say B^4_2. Recall that a 4-dimensional 2-handle is just $D^2 \times D^2$. Then B^4_1

together with a closed neighborhood of D is the trace of K, smoothly embedded in S^4. A schematic of S^4 as the union of two 4-balls is shown below:



The slice disk is represented by the thick grey curve. The trace of K consists of B_1^4 together with a neighborhood of the slice disk for K.

Now suppose that $X(K)$ embeds in S^4. Consider the piecewise linear embedded S^2 in $X(K)$ consisting of the core of the 2-handle together with the cone of K. Smoothly embed $X(K)$ in S^4; composition gives a piecewise linear embedding of S^2 in S^4, which is smooth away from the cone point p. Now take a small neighborhood around p in S^4. The complement of this neighborhood is a 4-ball B. Consider the piecewise linear embedding of S^2 intersected with B; we've cut out the cone point, so this gives a slice disk in B for K in ∂B. A schematic of the trace embedded in S^4 is shown below:



The 4-ball B is everything outside of the S^3 dotted circle, and the thick grey curve shows the slice disk for K.

8. Showing that K' is not slice

The goal is now to find a way to show that K', the knot that shares a trace with the Conway knot, is not slice. It turns out that some slice obstructions, such as the invariant ν coming from knot Floer homology, are actually trace invariants: if two knots K_1 and K_2 have the same trace, then $\nu(K_1) = \nu(K_2)$ [HMP19].

Luckily, the same is not true for Rasmussen's s-invariant. Using a computer program and some simple algebraic observations, Piccirillo shows that $s(K') = 2$,

implying that K' is not slice. Since K' and the Conway knot have the same trace, the Trace Embedding Lemma implies that the Conway knot is not slice.

9. What's next?

Now that we know exactly which knots with fewer than 13 crossings are slice, what's next? Of course, one could try to determine exactly which knots with fewer than 14 or 15 crossings are slice. But why not try to apply some of our tools to other open problems?

The smooth 4-dimensional Poincaré conjecture posits that a smooth 4-manifold that is homeomorphic to S^4 is actually diffeomorphic to S^4. To disprove the conjecture, one wants to find an *exotic* S^4, that is, a smooth 4-manifold that is homeomorphic but not diffeomorphic to S^4. One possible approach (outlined in [FGMW10]) to disprove the smooth 4-dimensional Poincaré conjecture relies on Rasmussen's s-invariant, as follows.

There are many constructions of potentially exotic 4-spheres Σ (see, for example [CS76]; note that certain infinite subfamilies of these are known to be standard by [Akb10, Gom10, MZ19]). By removing a neighborhood of a point in Σ, one can instead study potentially exotic 4-balls β. The difficult part is now determining whether or not β is exotic, or if it is in fact just the standard B^4.

While slice obstructions like ν actually obstruct a knot from being slice in an exotic 4-ball, it remains possible that the s-invariant only obstructs a knot from being slice in the standard 4-ball. The game is then to try to find a knot K that is slice in a potentially exotic 4-ball β. If $s(K)$ is non-zero, then K is not slice in the standard 4-ball, thereby implying that β must be exotic.

Both of the key steps in this approach (constructing the potentially exotic 4-ball and computing s) seem difficult. But maybe there is some other way to get handle on the problem in order to trace a solution. I look forward to hearing a Current Events Bulletin talk about such a result!

Acknowledgements

I would like to thank JungHwan Park and Lisa Piccirillo for helpful comments on an earlier draft of these notes.

References

[Abe16] Tetsuya Abe, *On annulus twists*, RIMS Kôkyûroku (2016), 2004:108–114.

[Akb77] Selman Akbulut, *On 2-dimensional homology classes of 4-manifolds*, Math. Proc. Cambridge Philos. Soc. **82** (1977), no. 1, 99–106.

[Akb10] ———, *Cappell-Shaneson homotopy spheres are standard*, Ann. of Math. (2) **171** (2010), no. 3, 2171–2175.

[Blo10] Jonathan M. Bloom, *Odd Khovanov homology is mutation invariant*, Math. Res. Lett. **17** (2010), no. 1, 1–10.

[CS76] Sylvain E. Cappell and Julius L. Shaneson, *There exist inequivalent knots with the same complement*, Ann. of Math. (2) **103** (1976), no. 2, 349–353.

[Don83] S. K. Donaldson, *An application of gauge theory to four-dimensional topology*, J. Differential Geom. **18** (1983), no. 2, 279–315.

[FGMW10] Michael Freedman, Robert Gompf, Scott Morrison, and Kevin Walker, *Man and machine thinking about the smooth 4-dimensional Poincaré conjecture*, Quantum Topol. **1** (2010), no. 2, 171–208.

[FM66] Ralph H. Fox and John W. Milnor, *Singularities of 2-spheres in 4-space and cobordism of knots*, Osaka Math. J. **3** (1966), 257–267.

[Fox53] Ralph H. Fox, *Free differential calculus. I. Derivation in the free group ring*, Ann. of Math. (2) **57** (1953), 547–560.

[Fre83] Michael Freedman, *The disk theorem for four-dimensional manifolds*, Proceedings of the ICM (1983), 647–663.

[Gab86] David Gabai, *Genera of the arborescent links*, Mem. Amer. Math. Soc. **59** (1986), no. 339, i–viii and 1–98.

[GL89] C. McA. Gordon and J. Luecke, *Knots are determined by their complements*, J. Amer. Math. Soc. **2** (1989), no. 2, 371–415.

[Gom10] Robert E. Gompf, *More Cappell-Shaneson spheres are standard*, Algebr. Geom. Topol. **10** (2010), no. 3, 1665–1681.

[HMP19] Kyle Hayden, Thomas E. Mark, and Lisa Piccirillo, *Exotic Mazur manifolds and knot trace invariants*, 2019, arXiv:1908.05269.

[Kho00] Mikhail Khovanov, *A categorification of the Jones polynomial*, Duke Math. J. **101** (2000), no. 3, 359–426.

[KWZ19] Artem Kotelskiy, Liam Watson, and Claudius Zibrowius, *On symmetries of peculiar modules; or, δ-graded link Floer homology is mutation invariant*, 2019, arXiv:1910.14584.

[Lee05] Eun Soo Lee, *An endomorphism of the Khovanov invariant*, Adv. Math. **197** (2005), no. 2, 554–586.

[MP18] Allison N. Miller and Lisa Piccirillo, *Knot traces and concordance*, J. Topol. **11** (2018), no. 1, 201–220.

[MZ19] Jeffrey Meier and Alexander Zupan, *Generalized square knots and homotopy 4-spheres*, 2019, arXiv:1904.08527.

[OS03] Peter Ozsváth and Zoltán Szabó, *Knot Floer homology and the four-ball genus*, Geom. Topol. **7** (2003), 615–639.

[OS04] ———, *Holomorphic disks and knot invariants*, Adv. Math. **186** (2004), no. 1, 58–116.

[OS11] Peter S. Ozsváth and Zoltán Szabó, *Knot Floer homology and rational surgeries*, Algebr. Geom. Topol. **11** (2011), no. 1, 1–68.

[Pic20] Lisa Piccirillo, *The Conway knot is not slice*, Ann. of Math. (2) **191** (2020), no. 2, 581–591.

[Ras10] Jacob Rasmussen, *Khovanov homology and the slice genus*, Invent. Math. **182** (2010), no. 2, 419–447.

[Sch49] Horst Schubert, *Die eindeutige Zerlegbarkeit eines Knotens in Primknoten*, S.-B. Heidelberger Akad. Wiss. Math.-Nat. Kl. **1949** (1949), no. 3, 57–104.

[Weh10] Stephan M. Wehrli, *Mutation invariance of Khovanov homology over $\mathbb{F}_2$*, Quantum Topol. **1** (2010), no. 2, 111–128.

[Zib19] Claudius Zibrowius, *On symmetries of peculiar modules; or, δ-graded link Floer homology is mutation invariant*, 2019, arXiv:1909.04267.

School of Mathematics, Georgia Institute of Technology, Atlanta, GA, USA
Email address: hom@math.gatech.edu

Rectangles, Curves, and Klein Bottles

Richard Evan Schwartz *

October 20, 2020

1 Introduction

This article starts with the question of picking out four special points on a curve in the plane and ends with a discussion of Shevchishin's theorem that you cannot embed a Klein bottle in $\boldsymbol{R}^4$, four dimensional Euclidean space, if it is Lagrangian. I will explain below what this means.

The notorious Toeplitz Conjecture, which goes all the way back to 1911, asks whether any Jordan curve contains 4 points which make the vertices of a square. (The edges of the square might intersect the curve in a messy way.) Such a collection of points is called an *inscribed square*. Figure 1 shows an example of a red square inscribed in a hexagon.

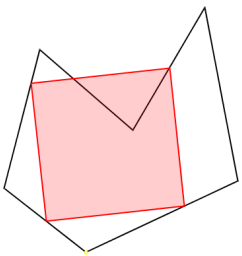


Figure 1: A square inscribed in a hexagon.

It is important to remember that a Jordan curve is any *continuous* loop in the plane. If you put reasonable constraints on the curve, then the result

* Supported by N.S.F. Research Grant DMS-1204471

has been known for a long time. In 1913, A. Emch [**E**] proved the result for convex curves. In 1929, L. G. Shnirlmann [**Shn**] proved the result for sufficiently smooth curves. In general, the problem is that the curve could be unreasonable. Perhaps you start with a polygon, and then make little changes to the curve on small scales. Then you go in with a microscope and make even smaller changes, and so on.

The square peg problem has a long and sprawling history. See, for instance, [**AA**], [**ACFSST**], [**CDM**], [**E**], [**FG**], [**H1**], [**H2**], [**Jer**], [**Mak1**], [**Mak2**], [**Ma1**], [**Ma2**], [**M**], [**N**], [**NW**], [**S1**], [**Shn**], [**St**], [**Ta**], [**Tv**], [**Va**]. B. Matschke's paper [**Ma**] gives a survey of what had been known up to 2014, and I. Pak's book [**Pa**] has an even more recent survey. The state of the art is the recent result [**FG**] that a locally 1-lipschitz Jordan curve has an inscribed square. What this condition means is that locally the curve is parametrized by a distance-non-increasing map from the line into the plane.

One can relax the question and ask about inscribed rectangles. The first general result along these lines, due to H. Vaughan, is that every Jordan curve (no matter how wild) has an inscribed rectangle. The reference for this is a bit hard to track down. Vaughan gave the proof in a lecture at U.I.U.C. in the 1977. My own involvement in this business is that I proved [**S2**] that any Jordan curve really has a lot of inscribed rectangles: All but at most four points of any Jordan curve are vertices of inscribed rectangles.

As an aside, M. Meyerson [**M**] proved in 1980 that all but at most 2 points of any Jordan curve are vertices of inscribed equilateral triangles. This kind of result is not known for any other shape of triangle – e.g., right-angled isosceles – though M. Neilson [**N**] shows that a dense set of points in any Jordan curve are vertices of inscribed triangles of any desired shape.

Just like triangles, a rectangle has a shape to it, namely its *aspect ratio*, the ratio of its length to its width. One can ask whether every Jordan curve has an inscribed rectangle of any given aspect ratio. In 2018, C. Hugelmeyer made the first progress on this problem, showing in [**H1**] that every smooth Jordan curve has an inscribed rectangle of aspect ratio $\sqrt{3}$. He later showed the following result [**H2**]: For any smooth Jordan curve, at least one third of the aspect ratios (as measured in a natural way) arise as aspect ratios of inscribed rectangles. This result involved a clever conversion of the problem into a question about certain Moebius bands intersecting in $\boldsymbol{R}^4$. Roughly speaking, Hugelmeyer constructs a continuous 1-parameter family of embedded Moebius bands, all having the same boundary. Using topological methods and a bit of measure theory, he then shows that at least one third

of the pairs have to intersect away from their boundary.

This year (as of this writing, 2020), Josh Greene and Andrew Lobb [**GL**] made a breakthrough on the aspect ratio problem for inscribed rectangles. They proved that any smooth Jordan curve has inscribed rectangles of every aspect ratio. Their proof builds in Hugelmeyer's idea, and considers a related 1-parameter family of Moebius bands embedded in $\boldsymbol{R}^4$. The added twist is that they use the additional structure of $\boldsymbol{R}^4$ coming from its identification with $\boldsymbol{C}^2$, the space of pairs of complex numbers, and this allows them to bring in tools from symplectic geometry. They then use symplectic methods to show that *every pair* of Moebius bands must intersect each other away from the common boundary. This breakthrough was the subject of a recent article in Quanta magazine [**Q**].

Where do the Klein bottles come from? Well, if two Moebius bands meet along a common boundary then their union is a Klein bottle with a kind of seam along the boundary. Suitably smoothing out this seam, you wind up with a Klein bottle. As I will explain, Greene and Lobb arrange for both the embeddings and the smoothing to be compatible with symplectic geometry, and the result is that the Klein bottle has the special property of being Lagrangian.

In this article, I will give an account of some of my favorite results in this area, and then focus on the Greene-Lobb result. More honestly, I will give an account of the results whose proofs I actually understand well enough to give a nice explanation. The reader should know that my taste is partly dictated by my ignorance of the wider field. I am probably omitting a lot of beautiful material just by accident.

Here is an outline of the paper. In the brief §2, I will say a few words about Jordan curves. In §3 I will sketch proofs of Meyerson's Theorem and of the square peg result for Jordan curves which are locally graphs of functions, as well as a few other related results. The material in §3 is not needed for the Greene-Lobb result. In §4 I will explain the ideas behind the Greene-Lobb result, using some of the symplectic geometry as a black box.

This article is a companion to my (upcoming) talk at the Current Events section of the 2021 J.M.M. meetings. I would like to thank David Eisenbud for inviting me to speak on this topic. I would like to thank Dan Cristofaro-Gardiner and Josh Greene for helpful conversations. I would also like to thank the Simons Foundation for their support, in the form of a Simons Sabbatical Fellowship, and also the Institute for Advanced Study, in the form of a 1-year membership funded by a grant from the Ambrose Monell Foundation.

2 Jordan Curves

2.1 Basic Definition

A *Jordan curve* is the image $J = f(S^1)$ of a continuous and one-to-one map $f : S^1 \to \boldsymbol{R}^2$. Here S^1 is the unit circle. The famous *Jordan Curve Theorem* says that $\boldsymbol{R}^2 - J$ has two components, one bounded and one unbounded. The bounded one is called the *inside* and the unbounded one is called the *outside*. There are many proofs of the Jordan Curve Theorem. See e.g. [**T**].

The case for polygons is fairly elementary: Color the points of $\boldsymbol{R}^2 - J$ black or white according as to whether a generic ray emanating from the point intersects J an odd or an even number of times. (The argument given in §3.4 below justifies the claim that this parity does not depend on the line.) These black and white regions turn out to be the inside and the outside regions.

If you want to avoid using the Jordan Curve Theorem, which in general is rather tricky to prove, let me suggest an alternate definition. Say that a *special Jordan curve* is the image $h(S^1)$ where $h : \boldsymbol{R}^2 \to \boldsymbol{R}^2$ is a homeomorphism – i.e., a bijection which is continuous and whose inverse is continuous. In this case, J automatically inherits the topological properties of S^1, such as having an inside and an outside. The Jordan Curve Theorem and the 2-dimensional Shoenflies Theorem together say that every Jordan Curve is special.

2.2 Polygonal Approximation

A Jordan curve is *approximable* by a sequence $\{J_n\}$ of polygons such that $d(J_n, J) \to 0$ as $n \to \infty$. Here, $d(J_n, J)$ is the infimal value of ϵ so that every point of J is within ϵ of J_n and *vice versa*. This metric is called the *Hausdorff metric*.

Every Jordan curve J is approximable by polygons. The cheapest approach to proving this is just to "connect the dots". Take a finite sequence of points going around J and then connect these points in their cyclic order. This will produce a polygon that approximates J, and the approximation gets better the more points we take. In general these polygons need not be embedded, and so you have to do some careful pruning to make this work.

One case where the connect-the-dots approach works cleanly is when J is a *local graph*. What this means is that there is a finite covering $N_1, ..., N_k$ of J by rectangles, and a corresponding collection of rotations $\rho_1, ..., \rho_n$ such

that $\rho_j(J \cap N_j)$ is the graph of a function in N_j for each $j = 1, ..., k$.

The connect-the-dots approach produces a sequence of polygonal approximations $\{J_n\}$. Being polygons, these approximations are automatically local graphs. However, the method does better. The approximations can be made to be local graphs in a *uniform* way, in the sense that the above finite list of coverings and rotations works for all members of the approximating family.

Here is a sketch of Tverberg's proof in the general case.

Lemma 2.1 *Every Jordan curve is approximable by polygons.*

Proof: Impose a grid of mesh size $1/n$ on the plane and (for the sake of cleanliness) adjust the grid so that none of its vertices belong to J. Let Q be a some square in the grid. As we traverse S^1 there is a smallest arc $A_Q \subset S^1$ such that $f(S^1 - A_Q)$ is disjoint from Q. Say that Q-*surgery* is the operation of replacing $f(S^1 - A_Q)$ by the line segment connecting the endpoints of $f(S^1 - A_Q)$ and keeping the rest of J the same. The resulting loop you get from Q-surgery need not be embedded, though it intersects Q in a line segment whose endpoints belong to J.

The curve J intersects finitely many grid squares, say $Q_1, ..., Q_m$. (Order them in some way.) Starting with $J_0 = J$, let J_1 be the result Q_1-surgery on J. Let J_2 be the result of Q_2-surgery on J_1. And so on. The curve J_n is an embedded polygon whose vertices all lie in J. Though J_n may intersect fewer grid squares than J, it still must be a good approximation for the following reason: If J_n fails to intersect some grid square that J intersects, it means that there was a nearby surgery that wiped out this intersection, and so J_n contains a point near the missing square. ♠

There is one additional case where we use this kind of polygonal approximation. Following Meyerson, say that a *triod* is the union of 3 continuous arcs joined at a single point, like the letter Y, and otherwise disjoint from each other. The triod is *polygonal* if it is a finite union of line segments. The same kind of polygonal approximation shows that an arbitrary triod can is approximable by polygonal triods.

3 Some Results about Inscribed Shapes

3.1 Triangles with an Arbitrary Shape

Let Δ be a triangle. Say that another triangle T has the same *shape* as Δ if there is an orientation preserving similarity which maps Δ to T. Such a map is the composition of a rotation, a dilation, and a translation. In this section I will prove that every point of every differentiable Jordan curve is the vertex of an inscribed triangle of any given shape. This result is, in a sense, the triangular analogue of the Greene-Lobb result.

Let J be a differentiable curve and let p_0 be a point on J. Let p_t be a parametrization of the J so that as t ranges from 0 to 1 the point p_t moves all the way around J, say counterclockwise. For each choice of $t \in (0,1)$ there is a unique point q_t so that the points (p_0, p_t, q_t) are vertices of a triangle T_t which has the same shape as Δ. In Figure 2, I have drawn T_t in red when t is near 0 and in blue when t is near 1.

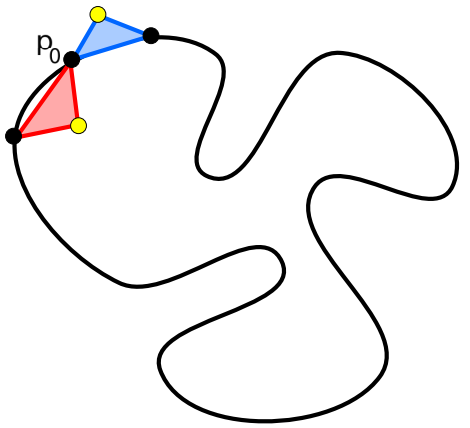


Figure 2: A family of triangles

Notice the in Figure 2 the point q_t, in yellow, starts out on the inside of J when t is near 0 and switches to the outside when t is near 1. Since q_t varies continuously, there must be some time $u \in (0,1)$ for which $p_u \in J$. But then all the vertices of T_u lie in J. This completes the proof.

The proof gives a bit more: Any point of differentiability on any Jordan curve is the vertex of triangles of arbitrary shape.

3.2 Meyerson's Theorem

Meyerson's Theorem says that all but at most 2 points of an arbitrary Jordan curve J are vertices of inscribed equilateral triangle. In this section I will give a proof that is complete modulo details of polygonal approximation. The proof here is somewhat like Meyerson's proof, and of course is based on his ideas, but it relies more heavily on polygonal approximation to make the analysis simpler.

Lemma 3.1 *Suppose that $p_0 \in J$ is some point, and there exist two other points p_1', p_2' in the region bounded by J such that $p_0 p_1' p_2'$ is an equilateral triangle. Then J has an inscribed equilateral triangle with vertex p_0.*

Proof: Let B and U denote the bounded and unbounded components of $\boldsymbol{R}^2 - J$. Let ρ be the 60 degree rotation about p_0 such that $R(p_1') = p_2'$. Consider extending the rays $p_0 p_1'$ and $p_0 p_2'$ outward until they first hit J at points p_1, p_2. Without loss of generality $p_0 p_1$ is not longer than $p_0 p_2$. Hence $R(p_1) \in J \cup B$. Let q_1 be a point of J maximally far from p_0. We have $R(q_1) \in J \cup U$. So, by continuity there is some $r_1 \in J$ such that $R(r_1) \in J$. Our equilateral triangle has vertices $p_0, r_1, R(r_1)$. ♠

Recall that a triod is a continuous version of the letter Y. Call the triod *good* if there is an equilateral triangle inscribed in the triod having one end of the triod as vertex. Otherwise call it *bad*. Call such triangles *end-inscribed triangles*. The key observation is that any 3 vertices of J are the endpoints of a triod that stays entirely in the region bounded by J. This is easy to see if J is a special Jordan curve. Just take one for the round disk and map it over.

Suppose for the moment that all triods are good. Choose any $a, b, c \in J$ and take a triod staying entirely inside J and having a, b, c as endpoints. Since this triod is good, there is an equilateral triangle inscribed in it having one of a, b, c as vertex, say a. But then the previous lemma applies to this triangle and shows that J has an inscribed equilateral triangle with a as vertex. So, to prove Meyerson's Theorem we just have to show that all triods are good.

We will prove that all triods are good in three steps: polygonal triods, end-straight triods, general triods. The polygonal case really shows the meat of the argument. The other cases just amount to fooling around with approximations and limits.

Lemma 3.2 *A polygonal triod is good.*

Proof: Assume not, for the sake of contradiction. Let A denote the union of the first two legs of T. Let a be the endpoint of T not in A. For any $x \in T$ let A_x denote the result of rotating A by 60 degrees clockwise about x. When $x \in T - \partial A$, the we have $\partial A \cap A_x = \emptyset$ and $A \cap \partial A_x = \emptyset$. Otherwise we'd get the desired triangle. This means that the mod 2 intersection number I_x between A and A_x is well-defined and constant for all $x \in T - \partial A$.

Let b be an endpoint of A. The two arcs A and A_b meet only at b and make a 60 degree angle. So, by compactness, A_x and A cross exactly once, at x, for x sufficiently close to b. Hence $I_x = 1$ for all $x \in T - \partial A$. In particular, $I_a = 1$. But then we have an inscribed equilateral triangle with vertex a. ♠

A triod is *end straight* the triad is polygonal sufficiently near the ends.

Lemma 3.3 *An end straight triod is good.*

Proof: Let T be end-straight. We can approximate T by a sequence $\{T_n\}$ of polygonal triods having the same final segments. By the previous lemma, T_n has an end-inscribed equilateral triangle Δ_n. Not all points of Δ_n can be on the same final segment of T_n. Note also that $T_n \to T$ and T is embedded. Combining these two observations, we see that there is a uniform positive lower bound to the size of Δ_n. Hence we can take a limit and find the end-inscribed equilateral triangle on T. ♠

Lemma 3.4 *An arbitrary triod is good.*

Proof: Now let T be an arbitrary triod, with ends a, b, c. For any large integer n, move out along the triple point of T until you reach the first point that is exactly $1/n$ from a. Call this point a'. Likewise define b', c'. Let T_n be the triod obtained by adding the segments aa', bb', cc' and erasing the arcs of T which join a to a', etc. If n is large enough, all points of $T' - (aa')$ are further than $1/n$ from a. Etc.

By construction T_n is end-straight. Let Δ_n be an end-inscribed triangle on T_n. Note that Δ_n, being equilateral, cannot have a as a vertex, and another vertex on aa'. So, either Δ_n is inscribed in T, and we're done, or else (after relabeling) Δ_n has a as a vertex and one point in bb'. Letting $n \to \infty$ we get an inscribed equilateral triangle with both a and b as vertices. ♠

3.3 Squares Inscribed in Local Graphs

In this section I will show that any local graph has an inscribed square.

If some J has an inscribed square Q, then the vertices of Q inherit two cyclic orders, one from the inclusion in Q and one from their inclusion in J. We call Q *gracefully inscribed* in J if these two orders coincide. Below I will sketch a proof that every polygon has a gracefully inscribed square.

Let $\{J_n\}$ be a sequence of polygons approximating a local graph J and having the uniformity property discussed in connection with this approximation. Let Q_n be a square gracefully inscribed in J_n. After passing to a subsequence we reduce to two cases. Either there is a positive lower bound δ to the diameters of Q_n or else there is a single point $p \in J$ such that $Q_n \to p$. In the first case we can take a limit on a subsequence and find a square of sidelength at least δ inscribed in J. Let us rule out the second case.

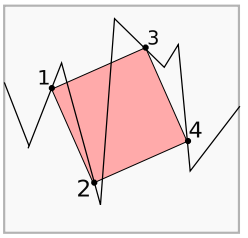


Figure 3: A contradiction at a small scale

Rotating and scaling, we can assume that the limit point is the origin, and that J and J_n intersect $[-1,1]^2$ in sets which are graphs of functions. This is meant to hold for all n. The cyclic order on the vertices of Q_n imposed by J_n goes from left to right, as indicated in Figure 3. However, this order cannot coincide with the cyclic on the vertices imposed by Q_n. This is a contradiction. So, the second case cannot occur.

This proof suggests a stronger version of the square peg conjecture that has been discussed quite often in connection with this problem. Say that a polygon P is *wide* if the bounded component of $\boldsymbol{R}^2 - P$ contains a disk of radius 1.

Conjecture 3.5 (Big Peg) *There is some $\epsilon_0 > 0$ with the following property. Every wide polygon has an inscribed square of sidelenght at least ϵ_0.*

The Big Peg Conjecture and polygonal approximation immediately imply the original Square Peg Conjecture. The Big Peg Conjecture is quite seductive because it only involves polygons.

3.4 A Warmup Problem

Before getting to the existence of gracefully inscribed squares, let's consider a warmup problem that captures many features of the argument we give below. The argument we give is one of the steps in the proof of the polygonal Jordan curve theorem.

Let's prove that a generic horizontal line intersects a generic polygon X an even number of times. Here, a *generic polygon* means one having no pair of vertices on the same horizontal line. A *generic horizontal line* (with respect to X) is one which does not contain a vertex of X. Let L_1 be a generic line. Start with a line L_0 lying entirely below X. Let L_t be the family of horizontal lines which sweeps upward. Call a parameter t *critical* if L_t contains a vertex of X and otherwise *ordinary*.

At the ordinary parameters, the intersection points vary continously and so their number does not change. There are only finitely many critical parameters, and at each critical parameter there is only one intersection point that lies at a vertex. As we wiggle the line up or down near a critical parameter, we see that near the critical parameter there are only three things that can happen. Figure 4 shows two of them, and the third possibility is like the first one but turned upside-down.

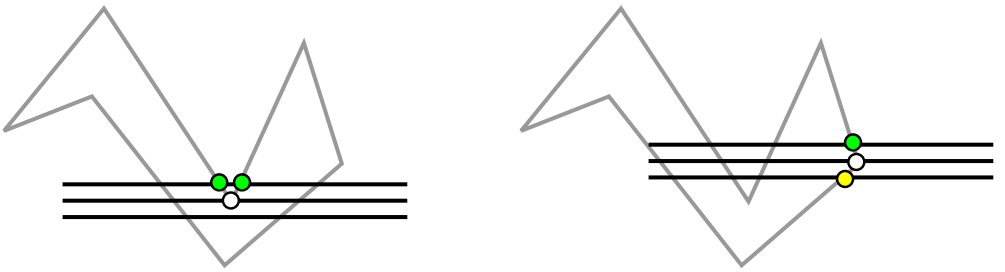


Figure 4: The behavior of intersection points.

In all cases, the parity of the number of intersections does not change. So, L_1 intersects X an even number of times. An examination of the argument shows that all we really needed was that X does not have any horizontal sides. We made X even more generic just so as to deal with the critical intersections one at a time.

3.5 Existence of Gracefully Inscribed Squares

We show that a generic polygon has an odd number of gracefully inscribed squares. Here *generic* means that no two sides of the polygon are parallel, no three sides lie in lines having a triple intersection, no three vertices lie in a square, and so on. The reasons for using generic polygons are similar to the reasons in the warmup problem. For instance, if a polygon has two long parallel sides close together it will have infinitely many inscribed squares.

There are a variety of proofs that a generic polygon has an odd number of inscribed squares. See [**Shn**], [**St**] or [**P**, Theorem 23.11]. These arguments do not specifically ask for gracefully inscribed squares, but the variational proof – at least the one I sketch below – works when we restrict our attention to gracefully inscribed squares.

Suppose P_1 is a generic polygon. Start with some easy-to-understand polygon P_0 having the same number of sides as P_1 and having an odd number of inscribed squares. For instance, P_0 could be a slight perturbation of a subdivision of an obtuse triangle. Now consider a continuous family P_t of polygons that interpolates between between P_0 and P_1. You cannot necessarily make all the polygons in the family *completely generic*. For instance, you may not be able to avoid some edges becoming parallel along the way. However, if the edges of P_t are very short, then a square inscribed in P_t can have at most 2 vertices inscribed in this union of parallel edges. Also, you can make coincidences like parallel edges happen one at a time.

Say that a vertex of an inscribed square is *critical* if it is a vertex of P_t, and otherwise ordinary. Call the square critical if it has a critical vertex, and call the parameter t critical if there is an associated critical square. We can make the family generic enough so that there are only finitely many critical parameters, and at each critical parameter there is only one critical square, and this critical square has only one critical vertex. Morever, we can make all the ordinary vertices vary continuously with the parameter. The continuity property is a "local" one: it only involves a statement about how a polygon interacts with at most 4 lines, and it can be settled by a direct algebraic calculation as in [**S1**] or [**S2**]. So, the square count changes only when we pass through a critical parameter.

What happens when we pass through a critical parameter? Notice that the question is also local: at most 5 lines are involved. Figure 5 shows a typical picture. The black edges at the intersection point are edges of the polygon and the red and blue segments are meant to depict the lines

containing these edges.

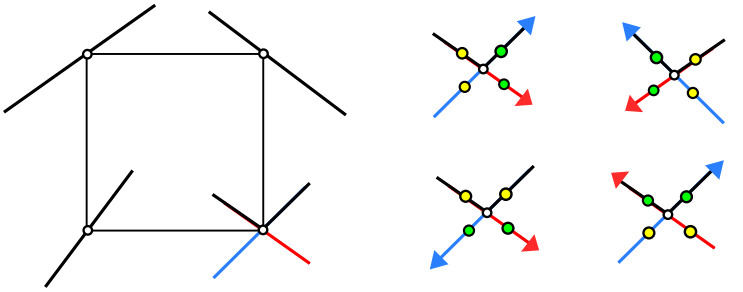


Figure 5: The behavior of intersection points.

We can think of the square in question as two copies of the same square, one red and one blue. The red (respectively blue) square is inscribed in the union of the black and red (respectively blue) lines. As the parameter varies, the red square and the blue squares separate from each other in a continuous way. We can translate the picture so that the red and blue lines always intersect at the origin. If the family is generic enough, the critical vertices, both from the red square and the blue square, vary monotonically through the origin.

We have illustrated this with the figures at the right. The yellow intersection points are the locations of the critical vertices just before we reach the critical parameter and the green points indicate the positions just after. Figure 5 shows the four possibilities for the directions that the points can move as a function of the varying parameter. The directions indicate the motion as a function of the parameter. The red and blue squares may or may not be inscribed in the polygon; it depends on whether the critical vertices lie on the segments of the polygon or on the red or blue segments extending them. In the bottom left case at right in Figure 5, both squares are inscribed in the polygon before the critical parameter and neither are inscribed after. In the top left case, one polygon is inscribed in the polygon before and one is inscribed after. The other two cases have similar treatments. Thus, the parity of the number of inscribed squares does not change as we pass through a critical point.

Notice that if the squares are gracefully inscribed before, they are gracefully inscribed after. So, the parity of the number of graceful squares does not change either. Finally, the square inscribed in a (subdivided, perturbed) obtuse triangle is gracefully inscribed. Our initial polygon has an odd number of gracefully inscribed squares and therefore so does the final one.

3.6 Existence of an inscribed Rectangle

In the next chapter I will explain Vaughan's proof that every Jordan curve has an inscribed rectangle. For the reader who really wants an elementary explanation, Vaughan's proof leaves a bit to be desired: It requires some graduate level algebraic topology. Let me sketch a different proof based on the material in my papers [**S1**] and [**S2**]. This proof is more complicated but avoids algebraic topology.

It turns out that generically the space of rectangles inscribed in 4 lines is a 1-dimensional manifold. Rather beautifully, the set of centers of these rectangles generically forms a hyperbola. See [**S1**]. This fact, coupled with the kind of analysis done in the previous section, shows that the space of rectangles inscribed in a generic polygon is a 1-dimensional manifold. Some of the components are loops and some of the other components are arcs. The endpoints of the arc components correspond to rectangles of aspect ratio 0 or ∞. (Working with labeled rectangles, we can tell the difference.) It turns out that a component of the manifold in question contains an even number of gracefully inscribed squares unless it has one of two properties:

- It is a loop connecting a gracefuly inscribed square $ABCD$ to the same inscribed square $BCDA$ with its vertices rotated. Call this a *rotator*. Every vertex of the polygon is the vertex of some rectangle in the rotator.

- It is an arc whose one end corresponds to rectangles of aspect ratio near 0 and whose other end corresponds to rectangles of aspect ratio near ∞. Call this a *sweepout*. All but at most 4 vertices of the polygon are vertices of a rectangle in a sweepout.

Given that there are an odd number of graceful squares, the generic polygon has an odd number of rotators and sweepouts combined; hence at very least it has one or the other. The existence of a sweepout would establish that the polygon has inscribed rectangles of every aspect ratio, but I could not rule out the existence of rotators. But, in either case, if we have sequence of polygons $\{J_n\}$ approximating a Jordan curve, we can extract from either a sweepout or a rotator a uniformly large rectangle R_n (in the sense of its minimum side length) inscribed in R_n. (I'll say more about this below.) The limit $\lim R_n$ will be a nontrivial rectangle inscribed in J. In [**S2**] I soup up this argument to show that all but at most 4 points of J are vertices of inscribed rectangles.

Now I will say more about finding the big rectangle R_n. Consider a set S_n of (say) 100 disjoint points on S_n. We choose so that $\{S_n\}$ converges to a set S of 100 distinct points on J. Each rectangle R inscribed in J_n cuts J_n off into 4 arcs, A, B, C, D going in order. Let $I(R)$ denote the number of points in $A \cup C$ minus the number of points in $B \cup D$. By a rough form of continuity, we can always find some rectangle R_n, either in a rotator or a sweepout, having $|I(R_n)| < 10$. Then some pair of adjacent arcs cut off by R_n, say A_n, B_n, are such that that $S_n \cap A_n$ and $S_n \cap B_n$ each have at least 10 points. By construction, no side of R_n can shrink to a point. I have deliberately used more points than strictly necessary so as to avoid needing a careful count.

4 Existence of Inscribed Rectangles

4.1 Vaughan's Theorem

Let me first explain Vaughan's argument that every Jordan curve has an inscribed rectangle. This result was the inspiration for the work of Hugelmeyer and Greene-Lobb. The argument relies on the topological fact that there are no continuous embeddings K of a Klein bottle into $\boldsymbol{R}^3$. Working with homology and cohomology, we have $H_1(K) = \boldsymbol{Z} \oplus \boldsymbol{Z}/2$ and Alexander Duality gives $H^1(\boldsymbol{R}^3 - \boldsymbol{K}) = \boldsymbol{Z} \oplus \boldsymbol{Z}/2$ but this last group must be torsion free, by the Universal Coefficient Theorem.

Given a Jordan curve J, let S denote the set of unordered and unequal pairs of points in J. The space S is a Moebius band. There are various ways to see this. This topological statement works the same for any Jordan curve, so we might as well consider the unit circle, and we also might as well consider it as a subset of the real projective plane. Every unordered pair of points in the circle determines a unique point in $\boldsymbol{RP}^2$: Take the two tangent lines to the circle at these points and intersect them. (If the points coincide it is natural to take this intersection point to be equal to the point itself rather than the entire line of intersections.) This map identifies the space S with the complement of the closed unit disk in the projective plane, and this is a Moebius band.

Vaughan defines a map $\phi : S \to \boldsymbol{R}^3$ by the formula

$$\phi(a,b) = \Big(\frac{a+b}{2}, |a-b|\Big). \tag{1}$$

Geometrically ϕ maps the ordered pair to a point encoding the midpoint of the segment $\overline{ab}$ and its length. If $\phi(a_1,b_1) = \phi(a_2,b_2)$ it means that the corresponding segments have the same length and meet at their midpoint. This gives an inscribed rectangle. So, we just have to prove that ϕ is not one-to-one.

We argue by contradiction. Notice that the image $\phi(S)$ lies in the upper half space, and $\phi(\partial S)$ lies in the XY-plane. Let ρ denote reflection in the XY-plane. The union

$$K = \phi(S) \cup \phi(\partial S) \cup \rho \circ \phi(S)$$

consists of 2 Moebius bands meeting along their boundary and thus is a Klein bottle. The 3 pieces separately are disjoint, and if ϕ is one-to-one then K is embedded. This contradicts the non-existence of an embedded Klein bottle.

4.2 Symplectic Geometry

The standard symplectic structure on $\boldsymbol{R}^4$ can be described entirely in terms of real numbers but is nice to describe it in terms of complex numbers. We can naturally identity $\boldsymbol{R}^4$ with the space $\boldsymbol{C}^2$. The map is given by

$$(x_1, y_1, x_2, y_2) \to (z_1, z_2), \qquad z_j = x_j + iy_j.$$

On $\boldsymbol{C}^2$ there is a natural operation on pairs of vectors $V, W \in \boldsymbol{C}^2$. Writing $V = (V_1, V_2)$ and $W = (W_1, W_2)$, we define

$$\langle V, W \rangle = V_1 \overline{W}_1 + V_2 \overline{W}_2. \tag{2}$$

Here $\overline{z}$ denotes the complex conjugate of z. This is known as a *Hermitian inner product*. It is linear in each argument and also $\langle V, W \rangle = \overline{\langle W, V \rangle}$.

It is instructive to write this in real coordinates. Let $V = (a_1 + ia_2, a_3 + ia_4)$ and $W = (b_1 + ib_2, b_3 + ib_4)$. Then

$$\langle V, W \rangle = (a_1b_1 + a_2b_2 + a_3b_3 + a_4b_4) + i(a_1b_2 - a_2b_1 + a_3b_4 - a_4b_3).$$

The real part of this expression is the dot product, and the imaginary part is (up to a "rotation") the standard symplectic form on $\boldsymbol{R}^4$. We will write the imaginary part as ω. So,

$$\omega(V, W) = \mathrm{Im}\langle V, W \rangle. \tag{3}$$

The complex structure gives a nice map on $\boldsymbol{C}^2$, namely $iV = (iV_1, iV_2)$. This operation extends to an operation on planes in $\boldsymbol{C}^2$ of indeed on any other set. We just multiply every vector in the set by i. A 2-plane $\Pi \subset \boldsymbol{R}^2$ is called *totally real* if $i\Pi$ and Π are perpendicular. This condition is the same as requiring that ω, when restricted to the plane, is identically 0. The plane $\boldsymbol{R}^2$ sitting inside $\boldsymbol{C}^2$ is the prototypical totally real plane. Any plane spanned by $(1, 0)$ and $(0, u)$, for unit complex u, is totally real. The reason is that the map $(z, w) \to (z, uw)$ preserves the Hermitian inner product and maps $\boldsymbol{R}^2$ to the plane we are considering.

A 2-plane Π is called *complex* if $i\Pi$ and Π are parallel. The plane $\boldsymbol{C}^1 \subset \boldsymbol{C}^2$ is the prototypical complex plane. The general plane Π in $\boldsymbol{C}^2$ is some kind of interpolation between these two cases: Π and $i(\Pi)$ will make some angle that ranges between 0 and $\pi/2$. This angle is sometimes called the *angle of holomorphy*.

4.3 Lagrangian Surfaces

Let U be an open set in $\boldsymbol{R}^4$. A *diffeomorphism* from U to $\boldsymbol{R}^4$ is a bijection $f : U \to \boldsymbol{R}^4$ which is non-singular and smooth. This is to say that the differential map df (the matrix of partial derivatives) is invertible, and f has partial derivatives of all orders. The Inverse Function Theorem says that f^{-1} will have these same properties.

A *smooth embedded surface* $\Sigma \subset \boldsymbol{R}^4$ is a set with the following property. For each point $p \in \Sigma$ there is an open disk Δ_p and a smooth diffeomorphism $\phi : \Delta_p \to \boldsymbol{R}^4$ which maps $\Sigma \cap \Delta_p$ to a $\boldsymbol{R}^2$. In other words, up to diffeomorphism, Σ looks locally just like a plane sitting in $\boldsymbol{R}^4$. Each point $p \in \Sigma$ has a tangent plane T_p. This is the plane which the differential of our diffeomorphism ϕ maps to $\boldsymbol{R}^2$. There seem to be two senses of what it means for Σ to be Lagrangian. The weak definition is that there is no point p for which T_p is complex. The strong definition is that T_p is totally real for all $p \in \Sigma$. Shevchishin's Theorem about Klein bottles works for the weak definition (and, of course, for the strong definition as well). The construction by Greene and Lobb gives a Klein bottle which, if embedded, would be Lagrangian in the strong sense.

It is well known that one can embed the Klein bottle as a surface in $\boldsymbol{R}^4$. The classic approach is to almost embed it in $\boldsymbol{R}^3$ as one of those famous glass blown models, and then fix it up. These glass blown models do not quite work, because one of the necks of the bottle crashes through the surface and makes a seam. This is shown in Figure 7.

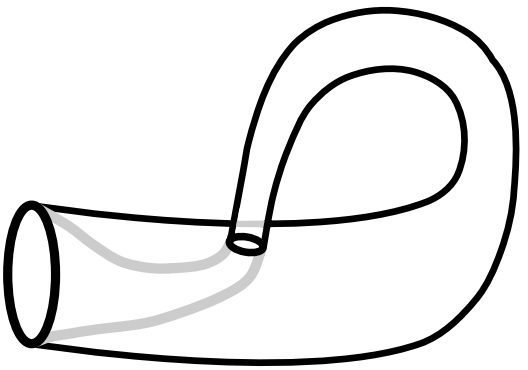


Figure 7: A Klein bottle in $\boldsymbol{R}^3$ with a seam.

To embed this surface in $\boldsymbol{R}^4$ we add a fourth coordinate to separate out the two parts of the surface which crash into each other. Imagine traveling

on the outside of the bottle, starting at the base and moving around. Make the 4th coordinate zero near the base and then (initially moving left to right) gradually increase it as you move around. By the time you wind around to the seam, the neck has a large positive 4th coordinate and is disjoint from the seam. Now gradually decrease the 4th coordinate so that the neck can rejoin with the base in $\boldsymbol{R}^4$.

Shevchishin's Theorem says, in particular, that some tangent plane along the resulting surface must be complex.

Theorem 4.1 (Shevchishin) *There does not exist a smooth Lagrangian Klein bottle in* $\boldsymbol{C}^2$.

Here is a lower dimensional analogy. Let X be the plane with the origin removed. Each point $p \in X$ has a special tangent line through it, namely the line through p that is perpendicular to the ray $\overline{0p}$. It is possible to embed the circle in X in many ways, but it is impossible to do so in such a way that its tangent line is never special. The tangent line will be special at each point of the embedded circle that is maximally far from the origin.

There are several proofs of Shevchishin's result, though from talking to symplectic geometry folks I have the impression that perhaps some of the proofs initially had gaps in them. All the proofs rely on some symplectic geometry machinery that is beyond the scope of the article. One of the proofs is in the paper [**N**] by S. Nemirovski.

I won't pretend to understand a proof of Shevchishin's Theorem, but let me give you a sense of the depth of the result. I will repeat, as filtered through my own understanding, the sketch of a proof that Helmut Hofer mentioned when I asked him about it after Josh Greene's talk at IAS. Any errors in this account are due to my misunderstanding. The space K has a smooth foliation by closed loops. This is easy! Each of these loops can be filled in by a pseudo-holomorphic curve. This is harder. The Lagrangian condition forces these holomorphic curves to be transverse to K, and then certain technical conditions force them to vary continuously and to be disjoint.

Consider the union of these disks. If we remove one of them, then topologically we have the product of a disk and a segment. When we put in the missing one we are gluing the ends of this solid tube together. There are essentially two choices for the gluing: orientation preserving and orientation reversing. Now for the punchline: A pseudo-holomorphic curve has a canonical orientation coming from the (almost) complex structure involved in its

definition. So, the gluing must be orientation preserving and therefore the union must be a solid torus rather than a twisted disk bundle. But then the boundary of a solid torus cannot be a Klein bottle, and this is a contradiction.

I got the impression that the details of this approach have not been worked out. Nemirovski's proof goes through something called Luttinger surgery and a result of Gromov-McDuff on the classification of open symplectic manifolds which are standard (i.e. look like symplectic $\boldsymbol{R}^4$) at infinity. The Gromov-McDuff result, in turn, involves the kind of pseudo-holomorphic disks mentioned above.

4.4 Lagrangian Smoothing

The classic *bump function* is a smooth function f such that $f(x) = 0$ for $|x| \geq 2$ and $f(x) = 1$ for $|x| \leq 1$. This function is used all over the place in the theory of smooth manifolds. It is also used in the construction below, which is called *Lagrangian smoothing*.

Let's consider a very simple situation first. In $\boldsymbol{C}^2$ we consider two totally real planes which intersect along a line. The first plane is $\Pi_1 = \boldsymbol{R}^2$. The second plane also goes through the origin and is spanned by $(1, 0)$ and $(0, i)$. Both planes contain the vector $(1, 0)$ and so intersect along a line. Let

$$Y = \Pi_1 \cup \Pi_2.$$

Note that Y is the union of 2 totally real planes. Put another way, Y is the union of 2 Lagrangian surfaces which meet transversely along a curve which happens to be a line.

Now consider a family of planes X_t that is perpendicular to Π_1 and Π_2. The plane X_t contains the point $(t, 0)$ and is spanned by the vectors $(0, 1)$ and $(0, i)$. These planes are all complex planes, parellel to the second copy of $\boldsymbol{C}$, namely $\{0\} \times \boldsymbol{C}$. Each plane X_t intersects Y in the union Y_t of two perpendicular lines, as shown at left in Figure 8. We can produce a surface by replacing each Y_t by a union Z_t of two smooth curves, as shown at right in Figure 8. We can make Y_t and Z_t agree outside, say, the unit disk. This construction makes use of a bump function.

Figure 8: A local model for Lagrangian smoothing

The union $Z = \bigcup Z_t$ is a Lagrangian surface which agrees with Y outside a small neighborhood of the line. The reason why Z is Lagrangian is that the tangent planes at each point are spanned by vectors of the form $(0, u)$ and $(1, 0)$ for some unit complex u. As we have already mentioned, such planes are totally real.

This operation is a local model for Lagrangian smoothing Suppose we have 2 Lagrangian embedded surfaces which meet along a closed curve. Using a suitable change of coordinates, which comes from a variant of the so-called Darboux Theorem, one can arrange that the local picture is just like in the simple model above. One then performs the local surgery described above and produces a union of 2 disjoint Lagrangian surfaces which agrees with the original union outside a small-as-you-like neighborhood of the original curve of intersection. The details of this coordinate change are worked out in [**GL**].

4.5 Sketch of the Proof

Inspired by Hugelmeyer's papers, [**H1**] and [**H2**], Greene and Lobb use a construction that is similar to the one given in the proof of Vaughan's Theorem. They also consider the space S of unordered distinct points on J. They define the map $\phi : S \to \boldsymbol{C}^2$ using the map

$$f(a, b) = \Big(\frac{a+b}{2}, \frac{(a-b)^2}{2\sqrt{2}|a-b|}\Big). \tag{4}$$

One can extract from $f(a, b)$ the location of the center of $\overline{ab}$, the length of $\overline{ab}$, and twice the angle that it makes with the X-axis. They also introduce the map

$$R_\phi(z, w) = (z, e^{i\phi}(w)). \tag{5}$$

If $f(a_1, b_1) = R_{2\phi} \circ f(a_2, b_2)$ it means that the segments $\overline{a_1 b_1}$ and $\overline{a_2 b_2}$ have the same midpoint and the same length. Also, one of them is a rotation of the other through an angle of ϕ. Therefore, an intersection like this corresponds to an inscribed rectangle whose diagonals make an angle of ϕ. Let

$$M_\phi = R_{2\phi} \circ f(S).$$

The set M_0 is an embedded Moebius band because one can recover a and b from $f(a, b)$. The set M_ϕ is just a rotation of M_0. These two Moebius bands limit on a common boundary, namely $f(\partial S)$. The union

$$K_\phi = M_0 \cup f(\partial S) \cup M_\phi$$

is a Klein bottle, and it is embedded unless M_0 and M_ϕ intersect away from their common boundary.

The upshot of the discussion above is that if K_ϕ is not embedded, then J has an inscribed rectangle whose diagonals make an angle of ϕ with each other. To prove that J has an inscribed rectangle of every aspect ratio it suffices to prove that K_ϕ is never embedded.

Greene and Lobb have cooked up their map so that M_0 is Lagrangian in the strong sense. There are two main points to the proof. First, f extends to a map from $\boldsymbol{C}^2$ to $\boldsymbol{C}^2$ which preserves the symplectic form ω. So, f maps Lagrangian surfaces to Lagrangian curves. Second, the set of ordered distinct pairs of points in J is a Lagrangian surface inside $\boldsymbol{C}^2$. Indeed, the tangent plane at each point of J is spanned by vectors of the form $(z, 0)$ and $(0, z)$.

The union K_ϕ has a seam along $f(\partial S)$ that looks locally like one quarter of the left side of Figure 8, except that the angle between the two surfaces is 2ϕ rather than $\pi/2$. The left side of Figure 9 shows what we mean.

Figure 9: Another Lagrangian smoothing construction.

Simplifying things a bit, what Greene and Lobb do is smooth out the seam by doing "half" of the Lagrangian smoothing discussed above. The result would be a smooth embedded Lagrangian Klein bottle, which is a

contradiction. This proves that M_0 and M_ϕ indeed intersect away from their common boundary.

Let me say a few more words about what it means to do "half" the Lagrangian smoothing. What they do is pass to a double cover, writing f as a composition $f = \sigma \circ \widehat{f}$, where $\widehat{f}$ is a map defined in a way very similar to f and σ is a 2-fold branched covering map from $\widehat{K}_\phi$ to K_ϕ. Here $\widehat{K}_\phi$ is the object like K_ϕ that is constructed using $\widehat{f}$ in place of f. They perform the smoothing of $\widehat{K}_\phi$ in a way that is equivariant with respect to σ, and then they push down the image via σ. Effectively, this does the smoothing as indicated in Figure 9.

5 References

[**AA**] A. Akopyan and S Avvakumov, *Any cyclic quadrilateral can be inscribed in any closed convex smooth curve*. Forum of Mathematics, Sigma Vol. 6, E7, 2018

[**ACFSST**] J. Aslam, S. Chen, F. Frick, S. Saloff-Coste, L. Setiabrata, H. Thomas, *Splitting Loops and necklaces: Variants of the Square Peg Problem*, Forum of Mathematics, Sigma Vol. 8, E5, 2020

[**CDM**], J. Cantarella, E. Denne, and J. McCleary, *transversality in Configuration Spaces and the Square Peg Problem*, arXiv 1402.6174 (2014).

[**E**] A. Emch, *Some properties of closed convex curves in the plane*, Amer. J. Math. **35** (1913) pp 407-412.

[**FG**] *Non-orientable slice surfaces and inscribed rectangles* arXiv 2003.01590v1 (2020).

[**H1**] C. Hugelmeyer, *Every Smooth Jordan Curve has an inscribed rectangle with aspect ratio equal to* $\sqrt{3}$. arXiv 1803:07417 (2018)

[**H2**] C. Hugelmeyer, *Inscribed Rectangles in a smooth Jordan curve attain at least one third of all aspect ratios*, arXiv 1911:07336 (2019)

[**Jer**]. R. Jerrard, *Inscribed squares in plane curves*, T.A.M.S. **98** pp 234-241

(1961)

[**Mak1**] V. Makeev, *On quadrangles inscribed in a closed curve*, Math. Notes **57(1-2)** (1995) pp. 91-93

[**Mak2**] V. Makeev, *On quadrangles inscribed in a closed curve ande vertices of the curve*, J. Math. Sci. **131**(1) (2005) pp 5395-5400

[**Ma1**] B. Matschke, *A survey on the Square Peg Problem*, Notices of the A.M.S. **Vol 61.4**, April 2014, pp 346-351.

[**Ma2**] B. Matschke, *Quadrilaterals inscribed in convex curves*, arXiv 1801:01945v2

[**M**] M. Meyerson *Equilateral Triangles and Continuous Curves*, Fundamenta Mathematicae, 110.1, 1980, pp 1-9.

[**N**] M. Neilson, *Triangles Inscribed in Simple Closed Curves*, Geometriae Dedicata (1991)

[**Ne**], S. Nemirovski, *Homology Class of a Lagrangian Klein Bottle*, arXiv: 0106122v4 (2008)

[**NW**] M. Neilson and S. E. Wright, *Rectangles inscribed in symmetric continua*, Geometriae Dedicata **56(3)** (1995) pp. 285-297

[**P**] I. Pak, *Lectures on Discrete and Polyhedral Geometry*, online book. https://www.math.ucla.edu/ pak/book.htm

[**Q**] K. Hartnett, *New Geometric Perspective Cracks an Old Problem about Rectangles*, https://www.quantamagazine.org

[**S1**] R. E. Schwartz, *Four lines and a rectangle*, J. Experimental Math. (to appear) 2020

[**S2**] R. E. Schwartz, *A Trichotomy for Rectangles Inscribed in a Jordan Curve*, Geometriae Dedicata, 2018.

[**Shn**], L. G. Shnirelman, *On certain geometric properties of closed curves* (in Russian), Uspehi Matem. Nauk **10** (1944) pp 34-44;
available at http://tinyurl.com/28gsys.

[**St**], W. Stromquist, *Inscribed squares and square-like quadrilaterals in closed curves*, Mathematika **36** (1989) pp 187-197

[**Ta**], T. Tao, *An integration approach to the Toeplitz square peg conjecture* Foum of Mathematics, Sigma, 5 (2017)

[**Tv**], H. Tverberg, *A Proof of the Jordan Curve Theorem*,
Bulletin of the London Math Society, 1980, pp 34-38.

[**Va**], H. Vaughan, *Rectangles and simple closed curves*, Lecture, Univ. of Illinois at Urbana-Champagne.

CURRENT EVENTS BULLETIN

Previous speakers and titles

For PDF files of talks, and links to Bulletin of the AMS articles, see http://www.ams.org/ams/current-events-bulletin.html.

January 17, 2020 (Denver, CO)

Jordan S. Ellenberg, University of Wisconsin-Madison
Geometry, Inference, and Democracy

Bjorn Poonen, Massachusetts Institute of Technology
A p-adic approach to rational points on curves

Suncica Canic, University of California, Berkeley
Recent Progress on Moving Boundary Problems

Vlad C. Vicol, Courant Institute of Mathematical Sciences, New York University
Convex integration and fluid turbulence

January 18, 2019 (Baltimore, MD)

Bhargav Bhatt, University of Michigan
Perfectoid geometry and its applications

Thomas Vidick, California Institute of Technology
Verifying quantum computations at scale: a cryptographic leash on quantum devices

Stephanie van Willigenburg, University of British Columbia
The shuffle conjecture

Robert Lazarsfeld, Stony Brook University
Tangent Developable Surfaces and the Equations Defining Algebraic Curves

January 12, 2018 (San Diego, CA)

Richard D. James, University of Minnesota
Materials from mathematics

Craig L. Huneke, University of Virginia
How complicated are polynomials in many variables?

Isabelle Gallagher, Université Paris Diderot
From Newton to Navier-Stokes, or how to connect fluid mechanics equations from microscopic to macroscopic scales

Joshua A. Grochow, University of Colorado, Boulder
The Cap Set Conjecture, the polynomial method, and applications (after Croot-Lev-Pach, Ellenberg-Gijswijt, and others)

January 6, 2017 (Atlanta, GA)

Lydia Bieri, University of Michigan
Black hole formation and stability: a mathematical investigation.

Matt Baker, Georgia Tech
Hodge Theory in Combinatorics.

Kannan Soundararajan, Stanford University
Tao's work on the Erdos Discrepancy Problem.

Susan Holmes, Stanford University
Statistical proof and the problem of irreproducibility.

January 8, 2016 (Seattle, WA)

Carina Curto, Pennsylvania State University
What can topology tell us about the neural code?

Lionel Levine, Cornell University and *Yuval Peres, Microsoft Research and University of California, Berkeley
Laplacian growth, sandpiles and scaling limits.

Timothy Gowers, Cambridge University
Probabilistic combinatorics and the recent work of Peter Keevash.

Amie Wilkinson, University of Chicago
What are Lyapunov exponents, and why are they interesting?

January 12, 2015 (San Antonio, TX)

Jared S. Weinstein, Boston University
Exploring the Galois group of the rational numbers: Recent breakthroughs.

Andrea R. Nahmod, University of Massachusetts, Amherst
The nonlinear Schrödinger equation on tori: Integrating harmonic analysis, geometry, and probability.

Mina Aganagic, University of California, Berkeley
String theory and math: Why this marriage may last.

Alex Wright, Stanford University
From rational billiards to dynamics on moduli spaces.

January 17, 2014 (Baltimore, MD)

Daniel Rothman, Massachusetts Institute of Technology
Earth's Carbon Cycle: A Mathematical Perspective

Karen Vogtmann, Cornell University
The geometry of Outer space

Yakov Eliashberg, Stanford University
Recent advances in symplectic flexibility

Andrew Granville, Université de Montréal
Infinitely many pairs of primes differ by no more than 70 million (and the bound's getting smaller every day)

January 11, 2013 (San Diego, CA)

Wei Ho, Columbia University
How many rational points does a random curve have?

Sam Payne, Yale University
Topology of nonarchimedean analytic spaces

Mladen Bestvina, University of Utah
Geometric group theory and 3-manifolds hand in hand: the fulfillment of Thurston's vision for three-manifolds

Lauren Williams, University of California, Berkeley
Cluster algebras

January 6, 2012 (Boston, MA)

Jeffrey Brock, Brown University
Assembling surfaces from random pants: the surface-subgroup and Ehrenpreis conjectures

Daniel Freed, University of Texas at Austin
The cobordism hypothesis: quantum field theory + homotopy invariance = higher algebra

Gigliola Staffilani, Massachusetts Institute of Technology
Dispersive equations and their role beyond PDE

Umesh Vazirani, University of California, Berkeley
How does quantum mechanics scale?

January 6, 2011 (New Orleans, LA)

Luca Trevisan, Stanford University
Khot's unique games conjecture: its consequences and the evidence for and against it

Thomas Scanlon, University of California, Berkeley
Counting special points: logic, Diophantine geometry and transcendence theory

Ulrike Tillmann, Oxford University
Spaces of graphs and surfaces

David Nadler, Northwestern University
The geometric nature of the Fundamental Lemma

January 15, 2010 (San Francisco, CA)

Ben Green, University of Cambridge
Approximate groups and their applications: work of Bourgain, Gamburd, Helfgott and Sarnak

David Wagner, University of Waterloo
Multivariate stable polynomials: theory and applications

Laura DeMarco, University of Illinois at Chicago
The conformal geometry of billiards

Michael Hopkins, Harvard University
On the Kervaire Invariant Problem

January 7, 2009 (Washington, DC)

Matthew James Emerton, Northwestern University
Topology, representation theory and arithmetic: Three-manifolds and the Langlands program

Olga Holtz, University of California, Berkeley
Compressive sensing: A paradigm shift in signal processing

Michael Hutchings, University of California, Berkeley
From Seiberg-Witten theory to closed orbits of vector fields: Taubes's proof of the Weinstein conjecture

Frank Sottile, Texas A & M University
Frontiers of reality in Schubert calculus

January 8, 2008 (San Diego, California)

Günther Uhlmann, University of Washington
Invisibility

Antonella Grassi, University of Pennsylvania
Birational Geometry: Old and New

Gregory F. Lawler, University of Chicago
Conformal Invariance and 2-d Statistical Physics

Terence C. Tao, University of California, Los Angeles
Why are Solitons Stable?

January 7, 2007 (New Orleans, Louisiana)

Robert Ghrist, University of Illinois, Urbana-Champaign
Barcodes: The persistent topology of data

Akshay Venkatesh, Courant Institute, New York University
Flows on the space of lattices: work of Einsiedler, Katok and Lindenstrauss

Izabella Laba, University of British Columbia
From harmonic analysis to arithmetic combinatorics

Barry Mazur, Harvard University
The structure of error terms in number theory and an introduction to the Sato-Tate Conjecture

January 14, 2006 (San Antonio, Texas)

Lauren Ancel Myers, University of Texas at Austin
Contact network epidemiology: Bond percolation applied to infectious disease prediction and control

Kannan Soundararajan, University of Michigan, Ann Arbor
Small gaps between prime numbers

Madhu Sudan, MIT
Probabilistically checkable proofs

Martin Golubitsky, University of Houston
Symmetry in neuroscience

January 7, 2005 (Atlanta, Georgia)

Bryna Kra, Northwestern University
The Green-Tao Theorem on primes in arithmetic progression: A dynamical point of view

Robert McEliece, California Institute of Technology
Achieving the Shannon Limit: A progress report

Dusa McDuff, SUNY at Stony Brook
Floer theory and low dimensional topology

Jerrold Marsden, Shane Ross, California Institute of Technology
New methods in celestial mechanics and mission design

László Lovász, Microsoft Corporation
Graph minors and the proof of Wagner's Conjecture

January 9, 2004 (Phoenix, Arizona)

Margaret H. Wright, Courant Institute of Mathematical Sciences, New York University
The interior-point revolution in optimization: History, recent developments and lasting consequences

Thomas C. Hales, University of Pittsburgh
What is motivic integration?

Andrew Granville, Université de Montréal
It is easy to determine whether or not a given integer is prime

John W. Morgan, Columbia University
Perelman's recent work on the classification of 3-manifolds

January 17, 2003 (Baltimore, Maryland)

Michael J. Hopkins, MIT
Homotopy theory of schemes

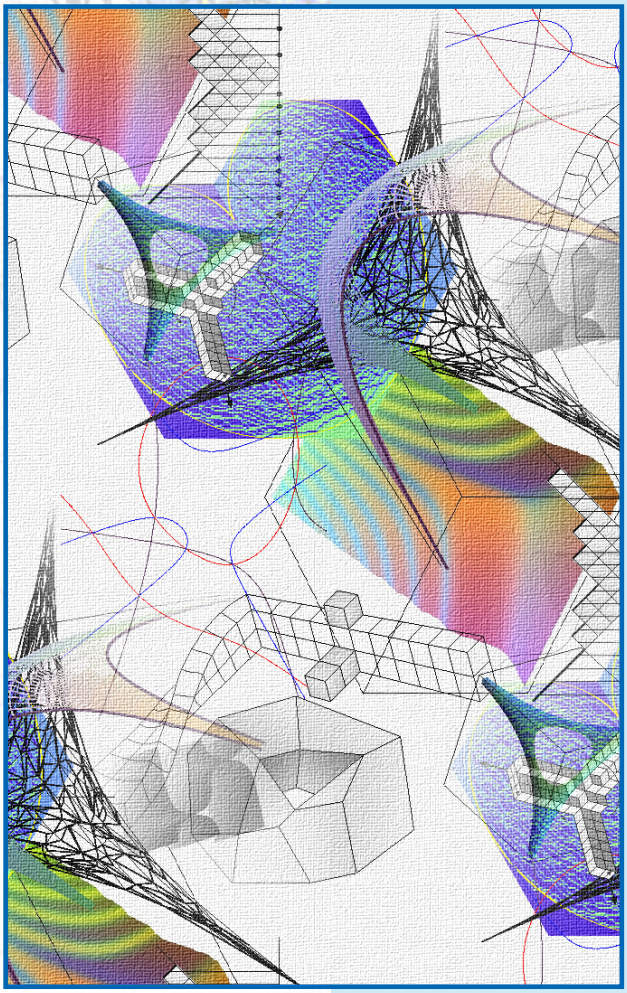
Ingrid Daubechies, Princeton University
Sublinear algorithms for sparse approximations with excellent odds

Edward Frenkel, University of California, Berkeley
Recent advances in the Langlands Program

Daniel Tataru, University of California, Berkeley
The wave maps equation

2021 CURRENT EVENTS BULLETIN

Committee



The back cover graphic is reprinted courtesy of Andrei Okounkov.

Cover graphic associated with Ana Caraiani's talk courtesy of Ana Caraiani.

Cover graphic associated with Jennifer Hom's talk, Chalk art by Catherine Owens, image courtesy of Lisa Piccirillo.

Cover graphic associated with Richard Evan Schwartz' talk courtesy of Richard Evan Schwartz.