UNIVERSITY OF CALIFORNIA
SANTA CRUZ

FEL QUADRUPOLE TUNING VIA BAYESIAN OPTIMIZATION USING PHYSICS-INFORMED GAUSSIAN PROCESS REGRESSION

A dissertation submitted in partial satisfaction of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PHYSICS

by

Dylan Michael Kennedy

December 2021

The Dissertation of Dylan Michael Kennedy
is approved:

Professor Joshua Deutsch, Chair

Professor Bruce Schumm

Professor David Draper

Peter F. Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

FEL Quadrupole Tuning via Bayesian Optimization using Physics-Informed Gaussian Process Regression

by

Dylan Michael Kennedy

Free-Electron Lasers like the one at the SLAC National Accelerator Laboratory are sources of extremely bright X-rays that are useful in a variety of scientific imaging applications. Because there are only a handful of FEL facilities around the world, access to these X-rays is in high demand. Every year, hundreds of hours are spent tuning quadrupole focusing magnets to optimize the X-ray brightness. During this tuning process, the beam typically cannot be used for experiments. In this thesis, I show that by performing Bayesian optimization using a Gaussian process regression model containing prior information derived from an optical model of the accelerator in combination with historical data, we were able to significantly reduce the amount of time spent tuning the quadrupoles in comparison to previous methods.

For Mom and Dad.

Acknowledgments

First and foremost, I would like to thank SLAC for providing me the opportunity to contribute to this work. The support and experienced I received there has been invaluable in my endless journey toward higher education. In particular, it was a special privilege to have worked with Joe Duris and Daniel Ratner, both of whom offered incredible direction and insights throughout my research. I would also like to thank my advisor, Josh, for his candid guidance throughout my time in graduate school, as well as the entire UCSC physics department, and especially the front office, who have all helped make my life easier by accommodating my spontaneous requests countless times without a hint of complaint. Finally, I must thank Sathya Guruswamy from the College of Creative Studies at UCSB, for her decision to personally call me in 2010 and ultimately remove any doubt that I would be a physicist.

Chapter 1

Introduction

Linear particle accelerators have been studied with great interest since their proposal by Gustav Ising in 1924 [1, 2]. In nearly the century that has followed, linear (electron) accelerators have been constructed around the world, with the current generation achieving electron energies in GeV range [1]. One of the most powerful features of these accelerators is that they can be configured as Free-Electron Lasers (FELs). An FEL is a device that uses a relativistic electron-beam to produce extremely intense, coherent synchrotron radiation which can be valuable in a variety of scientific imaging applications [3, 4, 5, 6]. While there are many devices that require fine-tuning to ensure that an FEL is operating at maximal efficiency, a set of devices known as the quadrupole focusing magnets (often referred to as "the quadrupoles", or simply "the quads"), whose magnetic fields are used to focus the electron-beam, require some of the most frequent attention [7].

This thesis is intended to serve as a review for anyone interested in applying

numerical optimization techniques to autonomously tune the quadrupoles of a free-electron laser. It is written with a target audience of first-year grad students in mind. The hope is that an incoming grad student may be able to read this thesis and have a good idea why we did what we did, how we did it, and how one might improve upon our results. To that end, I will begin by briefly discussing the history and uses of free-electron lasers, what it means to "tune" them, and why we care about improving the efficiency of this tuning process. I will then present our reasoning for adopting a Bayesian approach to quadrupole tuning. At the end of this chapter, I will provide a roadmap for the remainder of this thesis.

1.1 The Free-Electron Laser

A free-electron laser is a source of extremely intense radiation that uses electrons traveling at nearly the speed of light as a gain-medium [8]. This is in contrast to a typical laser, which uses atomic or molecular (and generally stationary) material to produce stimulated emission of radiation upon being excited by some external energy source. The first free-electron laser (FEL) was built in 1971 at Stanford University by physicist John Madey [3, 9]. Thirty-eight years later, in 2009, with a yearly operating budget of about $100 million, the Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory achieved its first lasing, becoming the first FEL facility to produce radiation in the X-ray spectrum. To this day, it remains one of the most powerful FEL X-ray sources in the world [3]. After a decade of operation, the X-rays

produced at the Linac Coherent Light Source have proven to be profoundly illuminating, with applications in the fields of materials science, chemistry, plasma physics, life sciences, and beyond [10, 3].

1.2 The Tuning Problem

This thesis addresses the task of tuning the FEL at LCLS using numerical optimization. While there are many devices integral to the operation of an FEL that require frequent adjustment, in the context of this work, "tuning" will refer to the adjustment of the quadrupole magnets. These quadrupoles are devices located periodically along the beamline that act as lenses for the electron-beam, intermittently focusing and refocusing it as it is accelerated over the course of 2 miles to the relativistic speeds necessary to produce X-rays via synchrotron radiation [11].

On a typical day of operation, LCLS will work with numerous research groups, with each group generally requiring an X-ray energy level different from the last [10]. As the energy level of the beam changes, so do the effects of the quadrupole magnets. (We will see a quantitative treatment of these effects in Chapter 2.) As a result, the process of shifting from one energy to another entails changing the settings on these quadrupole magnets so that the beam is properly transported. Thus, tuning the FEL can be understood as an optimization problem. The target that we seek to optimize is the FEL power as a function of the quadrupole magnet settings - i.e. magnetic field gradients - along the beamline. That is, we wish to solve the following, multiple times

per day:

$$\vec{x_*} = \arg\max_{\vec{x}} P(\vec{x}) \tag{1.1}$$

where the components of the input vector $\vec{x}$ are the magnetic field gradients corresponding to the set of quadrupole magnets currently being tuned and the function P is the FEL output power.

Tuning the quadrupoles is necessary to compensate for variables which may change unpredictably over time and are hard to measure quickly [12, 13, 14]. The process of tuning is therefore inevitably guess-and-check. Historically, the job of tuning the quadrupoles was originally performed manually by human operators physically turning knobs. The first advancement to the process of tuning the quadrupoles came when an operator wrapped a rubber band around two of the knobs so that two could be turned with one hand. The next advancement came when an operator put a twist in the rubberband so that the two adjacent knobs would turn opposite to each other if either was adjusted. This second advancement was the result of operators learning something about the underlying function they were optimizing. That is, the operators had observed that adjacent devices seemed to have opposite effects, which, as we will see in Chapter 2, is supported by theory.

While the twisted rubber band was an improvement (and may in fact be little more than LCLS folklore), the biggest advancement to the tuning process came when the human operators were replaced with a numerical optimization algorithm [15, 13, 16]. Early comparisons between automated search and human-driven search found that the

automated process converged faster simply by virtue of being able to adjust the settings faster than humans. The numerical optimization algorithm did not necessarily make better guesses toward the solution than the human operators, but being able to make more guesses and being able to vary more devices (humans, typically possessing only two hands, are severely limited in their ability to adjust many devices simultaneously) in the same amount of time resulted in improved performance.

When it comes to tuning an FEL, it is critical to minimize the time required to reach adequate FEL output power. The longer it takes the operators to tune the beam, the less time the user group will have to perform their experiment. This beam time comes at the cost of tens of thousands of dollars per hour to the Department of Energy, and with the current approach, hundreds of hours a year are spent tuning the quadrupoles. Therefore, any significant systematic improvement to the efficiency of the tuning process will translate directly to more (precious) scientific progress.

Since the first successful demonstration of numerical optimization (using the Nelder-Mead Simplex method) as an approach to quadrupole tuning, efforts have been made to improve the process by incorporating more information about the target function into the search algorithm [15]. The idea was that if operators were able to learn relevant features of the target function, it would be ideal to combine that knowledge with the efficiency of the automated search [17]. To incorporate this information, our group has transitioned in favor of using a Bayesian optimization algorithm, which generates a model of the target function via Gaussian process regression at each iteration of the search [18, 7]. The Gaussian process regression model is a robust fitting tool that

allows us to incorporate prior information about the function that is being optimized in the form of kernel hyperparameters [19]. As we will see in chapter 4, these hyperparameters can be selected to account for devices whose effects are stronger than other devices, correlations between neighboring devices, and even noisy output signals. Being robust to noise is a quality of particular interest to us, as our target function is inherently noisy. Furthermore, Bayesian optimization is especially suited to optimization problems where the target function is extremely expensive to evaluate and whose inputs are multidimensional [20]. While automated search algorithms are able to iterate faster than humans, they are limited by physical constraints (such as magnet settle time) to a maximum sampling rate of about 0.5 Hz. With the cost of owning and operating an FEL like the one at LCLS totaling about $45,000 per hour, this comes out to approximately $25 per point sampled [citation needed]. Bayesian optimization tends to converge to a solution in fewer function evaluations than other optimization algorithms, at the cost of additional compute-power per iteration. That the cost of compute-power pales in comparison to the cost of an additional function evaluation ($25) makes this use case an ideal candidate for Bayesian optimization.

1.3 Roadmap

In the chapters that follow, I catalog our efforts to tailor our Bayesian optimization algorithm to the problem of tuning the FEL quadrupoles at SLAC. In Chapter 2, I will review some essential FEL physics to provide an understanding of the under-

lying mechanism governing our target function. Particular emphasis will be devoted to the fundamental operating mechanism that enables a beam of electrons to act as a free electron laser as well as the role that the quadrupole focusing magnets play in this process. In Chapter 3, I will discuss the fundamentals of Bayesian numerical optimization. In Chapter 4, I will introduce the Gaussian process regression model and illustrate some of the characteristics that make it a good regression tool for our use case. In Chapter 5, I present an analysis into the importance of properly training Gaussian process hyperparameters before performing Bayesian optimization. I then describe in detail our method of training our Gaussian process hyperparameters. I also discuss some computational pitfalls that we encountered, how we defeated them, and I show that our resulting Bayesian optimization algorithm outperforms previous benchmarks. Lastly, in chapter 6, I will summarize our conclusions.

Chapter 2

Physical Background

As discussed in the preceding chapter, the objective of this paper is to improve upon the current approach to numerical optimization for autonomous tuning of the quadrupoles in an FEL device. With this in mind, the first step toward a solution for any numerical optimization problem is to understand as well as possible the nature of the function that is being optimized. In our case, we know that the target function represents the result of a physical process. Specifically, it is governed by the combined physics of a linear particle accelerator, an insertion device called an undulator, and a measurement device called a gas detector. In this chapter, I will cover the basic theoretical mechanics of these three major components.

It should be noted that the goal of this theoretical analysis will not be to develop a perfect (or even particularly accurate) model of the target function. Were it possible to model the function to a high degree of accuracy, a highly strategic approach to numerical optimization would hardly be necessary in the first place. Alas, due to the

complexity of the machine and its many unobservable errors, such a model is perhaps unattainable using theory alone. Not to be discouraged, though, for as we will see, there is still valuable information that can be obtained from the theoretical analysis. However, we should keep in mind that it will only provide us with a rough understanding of how the inputs to our function are expected to affect the output.

I will begin by reviewing some basic FEL theory, focusing mostly on the physics of the undulator, the core component of an FEL, responsible for the generation of the high-intensity X-rays (which make FELs such valuable imaging tools) via synchrotron radiation. In this review, we will learn approximately how the transverse size of the beam affects the FEL output power. I will also discuss the nature of the measurements produced by the gas detector, as the process employed to generate these measurements will add some noise to the target signal. In the latter part of the chapter, I will discuss the physics of transporting an electron beam. There, I will illustrate the focusing/defocusing effects of the quadrupole magnets. It is these quadrupoles that we directly adjust to tune the beam, and as such it is imperative that we understand what function, physically, these magnets perform, and what happens when we change their settings.

2.1 Preliminaries

In this section, I will briefly introduce some of the relativistic quantities that describe the electrons we will be dealing with throughout this chapter. We will use

these results to formulate some approximations that will be used repeatedly in the following sections. Additionally, I will address a standard coordinate system, as well as an important change-of-variable, that is used to describe the motion of charged particles in a linear accelerator.

2.1.1 Useful Approximations

Throughout this chapter, we will be studying electrons whose speed is nearly that of light. Because physics at these speeds is much different than the physics we are accustomed to in our everyday experience, it will be useful to briefly familiarize ourselves with the approximate values of some of the fundamental quantities describing relativistic motion for particles in the energy regime of our interest. A useful computation to remember for the Lorentz factor, γ, which is the ratio of a particle's total energy to its rest energy, is, for an electron [4]:

$$\gamma = \frac{U_e[\text{GeV}]}{mc^2} \approx 1957 U_e[\text{GeV}]. \tag{2.1}$$

For accelerators at modern FEL facilities, the electron energies are typically on the order of one to tens of GeV. At the low end of these energies, the Lorentz factor is on the order of 10^3, but for the higher energy electrons this factor may be upwards of 10^4.

We can use the Lorentz factor to get an estimate for the ratio of the speed of the electrons to the speed of light. This ratio, ubiquitously referred to as β in relativistic texts, is defined by $\beta \equiv v/c$ where v is the speed of the particle and c is the speed of

light. β is related to the Lorentz factor by:

$$\gamma = \frac{1}{\sqrt{1-\beta^2}} \tag{2.2}$$

We can solve for β to arrive at

$$\beta = \sqrt{1-\gamma^2} \approx 1 - \frac{1}{2\gamma^2} \tag{2.3}$$

where the final approximation is the result of Taylor expansion, using the fact that $\gamma^2 \gg 1$. Provided $\gamma \geq 10^3$, the approximated value for beta is accurate to within one part in 10^{12}. We will find this approximation very useful going forward.

It is also interesting to know just how close to the speed of light our electron speeds are reaching. Calculating the fractional difference, again using $\gamma \geq 10^3$, we have

$$\frac{1}{c}(c-v) = 1 - \beta = 1/\gamma^2 \leq 5 \times 10^{-7}. \tag{2.4}$$

That is, the electron speed approaches the speed of light to within a factor on the order of 10^{-7}. In many cases, this means we can approximate the speed of the electrons quite well as simply being equal to the speed of light. However, we must be careful not to do so blindly, because, as we will see in the next section, the fundamental mechanism through which an FEL is able to operate is based on the fact that the electron speed is slightly less than that of light [4].

2.1.2 Longitudinal Description (Change-of-Variable)

In beam physics, there is always a single ideal trajectory for the particles to follow, called the design trajectory. All particle motion is described with respect to this

trajectory. Particles whose motion deviates from the design trajectory are treated as paraxial rays [21]. While these rays are not necessarily parallel to the design trajectory, their angular divergence from the design trajectory must be small in order to keep the beam confined within the tubular transport structures of the accelerator.

When analyzing the particle trajectories, it is most convenient to use a local coordinate system defined with respect to the local design trajectory. That is, we wish to define the z-axis of a local Cartesian coordinate system such that it is aligned with the local design trajectory. Thus, the distance traveled parallel to the design trajectory will be given by the z-coordinate. The paraxial rays, then, may be offset by a small amount in the x- and y-directions, and make some angle θ with respect to the z-axis. We refer to the z-direction as the longitudinal direction, while x and y are referred to as the transverse directions. In keeping with general tradition, the x-direction represents the horizontal offset, while the y-direction represents the vertical offset. We specify the divergence angle θ in 3-dimensions by θ_x and θ_y, where θ_x is the angle made between the projection of the velocity in the x-z plane and the z-axis, and θ_y is the angle made between the projection of the velocity in the y-z plane and the z-axis. Thus, we arrive at the relations

$$\tan\theta_x = v_x/v_z, \tag{2.5}$$

$$\tan\theta_y = v_y/v_z. \tag{2.6}$$

Because the particles move unidirectionally along the positive z-direction, there is a one-to-one correspondence between a particle's z-coordinate and its time coordinate. Thus,

while the fundamental physical laws governing the motion of these particles are typically formulated as differential equations with respect to time, it is possible to perform a change-of-variable, replacing the time coordinate with z. This will be necessary in order to transition to a description of the beam physics that makes use of the principles of geometric optics, as the forces in such optical systems are specified in terms of position rather than time [21, 22]. (As we will see later in this chapter, by employing a few highly accurate approximations, the quadrupole magnets that are the focus of this work can be treated analogously as optical lenses, providing both a familiar interpretation and simplified quantitative framework for computing their effects.) Thus, we will need to replace derivatives with respect to time with derivatives with respect to z. In beam physics texts, derivatives with respect to z are represented using primed notation, while derivatives with respect to time are represented using dot notation [4]. For example, in the case of the horizontal offset coordinate x, we have

$$\dot{x} \equiv \frac{dx}{dt} \equiv v_x, \tag{2.7}$$

$$x' \equiv \frac{dx}{dz}. \tag{2.8}$$

In general, we can transition from one independent variable to another using the chain rule. In this case,

$$\frac{d}{dt} = \frac{dz}{dt}\frac{d}{dz} = v_z \frac{d}{dz}, \tag{2.9}$$

$$\frac{d}{dz} = \frac{1}{v_z}\frac{d}{dt}. \tag{2.10}$$

Applying this to the x coordinate, we get, as expected,

$$x' = \frac{d}{dz}x = \frac{1}{v_z}\frac{d}{dt}x = \frac{v_x}{v_z} = \tan\theta_x. \tag{2.11}$$

The y-direction can be treated similarly. As we have discussed, the velocity in the longitudinal direction approaches that of light. Because the transverse velocities are comparatively small, we can apply the small angle approximation to replace the tangents of the transverse angles by their arguments. By doing so, we arrive at the following,

$$x' \approx \theta_x, \tag{2.12}$$

$$y' \approx \theta_y. \tag{2.13}$$

Thus, the quantities x' and y' are sometimes referred to as the transverse angles, as they are effectively interchangeable under our approximation.

Now that we have established how to change our independent variable from time to longitudinal position, we will be able to understand the forces that we encounter in the following sections in terms of their geometric effects on the particle trajectories.

2.2 FEL Physics

In this section, we will cover the basics of how electrons can be used to produce high-intensity radiation. I will begin with a brief history of synchrotron radiation. I will proceed to discuss the physics of an undulator device, which is an insertion device designed to create oscillations in the motion of charged particles that have been accelerated to relativistic speeds, thereby generating synchrotron radiation [23, 4]. I will then

examine the conditions under which an electron beam passing through an undulator may lead to a phenomenon called self-amplification by spontaneous emission (SASE). When this interaction occurs, the device is operating as what we call a Free-Electron Laser (FEL). We will see that the transverse size of the electron beam is a critical parameter in determining the strength of this interaction. Lastly, I will address the practical consideration of measuring the intensity of the output signal of an FEL using a device called a gas-detector.

2.2.1 Synchrotron Radiation

Synchrotron radiation is generated when a charged particle travels at relativistic speeds along a curved trajectory. This type of radiation can and does occur in nature, for example, when electrons that have been accelerated to ultra-relativistic speeds by a solar flare spiral through a magnetic field. However, it was not until midway through the twentieth century that it was first recognized by humans when it was observed, unexpectedly, by scientists working on the circular accelerator at the General Electric Research Laboratory in Schenectady, New York on April 24, 1947 [24]. Since its discovery, synchrotron radiation has become an invaluable tool in the field of imaging, as it is capable of producing wavelengths that span the entire electromagnetic spectrum [10]. Facilities designed to produce synchrotron radiation, such as the Linac Coherent Light Source at SLAC National Accelerator Laboratory, continue to be constructed around the globe, with each generation capable of producing radiation at higher intensities than the last [3].

At LCLS, the electron beam from the SLAC linear accelerator is used to produce high-powered X-rays via synchrotron radiation. As these electrons approach the end of the accelerator, they are traveling at nearly the speed of light. By passing them though a specially designed magnetic field generated by a device called an undulator, the electrons are driven into rapid transverse oscillations in such a way that causes them to radiate X-rays, and which creates a resonance between the forward-propagating E-M waves and the leading electrons [23, 4]. This resonance leads to a laser-like amplification of the beam of X-rays (hence the name Free-Electron Laser), which can subsequently be used to perform imaging tasks for a wide variety of research purposes. In the sections that follow, I will review the physics of this synchrotron radiation and the amplification process.

2.2.2 The Undulator

An undulator is an insertion device (so-called because it is inserted in the beamline of a particle accelerator) that consists of a series of dipole magnets arranged in a periodic structure with alternating polarities. It is designed to produce synchrotron radiation by forcing the charged particles of a relativistic beam to perform transverse oscillations with respect to their longitudinal position. Depending on the type of undulator, these transverse oscillations may have components in both the x- and y-directions (as is the case in a helical undulator), but we will restrict our discussion to a planar undulator (as this is the kind we have at LCLS), which produces oscillations only in one transverse plane. A schematic of a planar undulator is depicted in Figure 2.1. In

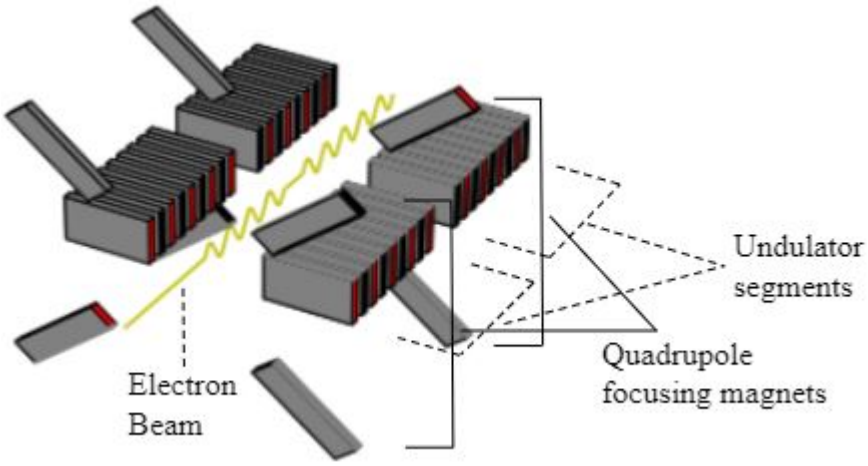


Figure 2.1: Schematic of a Planar Undulator

A schematic of a planar undulator (not to-scale). The undulator shown is composed of a series of dipole magnets (grey) of alternating polarity (north/south depicted by red/black, respectively) situated on each horizontal side of an electron beam (trajectory of beam centroid shown in yellow). Magnetic quadrupoles are configured in between segments of the undulator to keep the beam focused.

this section, we will analyze the motion of an electron passing through such a device, and develop an expression for the fundamental wavelength of the induced synchrotron radiation. This wavelength will be relevant in the following sections when we discuss resonance.

First, we must construct a valid model for the magnetic field of a planar undulator. To do this, we will make use of the magnetic scalar potential, which we denote ϕ. Recall that in free space (such as the space between the magnets of the undulator through which our electron beam will pass), this scalar potential must satisfy

Laplace's equation:

$$\nabla^2 \phi = 0. \tag{2.14}$$

This condition restricts our set of physically allowed potentials. By constructing a potential that satisfies Laplace's equation, we guarantee that the associated magnetic field satisfies Maxwell's equations in vacuum. Specifically, the resulting field will have identically zero divergence and zero curl. Furthermore, we must ensure that our function ϕ describes a magnetic field that displays transverse oscillations with respect to the z-direction in order to induce transverse oscillations in the charged particles traveling longitudinally through the field. The magnetic field can be obtained from ϕ using the following relation:

$$\vec{B} = \nabla \phi. \tag{2.15}$$

To satisfy the above conditions, we consider the field given by the scalar potential [4]

$$\phi = -\frac{B_0}{k_u} \sinh(k_u y) \sin(k_u z). \tag{2.16}$$

Here, B_0 is the maximum magnetic field strength (amplitude), and k_u is the spatial frequency (wavenumber), determined by the spacing of the dipole magnets. This potential satisfies Laplace's equation, and results in a magnetic field given by

$$\vec{B} = \nabla \phi \tag{2.17}$$

$$= -B_0 \cosh(k_u y) \sin(k_u z) \hat{y} - B_0 \sinh(k_u y) \cos(k_u z) \hat{z}. \tag{2.18}$$

For sufficiently small transverse offsets from the z-axis, the above magnetic field can be approximated as

$$\vec{B} = -B_0 \sin(k_u z)\hat{y}, \tag{2.19}$$

where we have discarded the component of the field along the z-direction for simplicity (0th order approximation). In other words, the above is the resulting magnetic field along the z-axis.

We can now compute the trajectory of an electron passing through this undulator field. To do this, we will assume that the electron initially has no transverse offset and no transverse velocity. That is, it is initially directed along the z-axis. The force on the electron from the undulator field is given by the Lorentz force:

$$\vec{F} = -e(\vec{E} - \vec{v} \times \vec{B}). \tag{2.20}$$

Assuming the Lorentz force resulting from the interaction between the undulator field and the electron is much larger than the force resulting from the induced synchrotron radiation, the net force on the electron will be approximately equal to the Lorentz force. Writing the net force as the change per time of the relativistic momentum, and assuming zero electric field, we have

$$\vec{F} = \frac{d\vec{p}}{dt} = \frac{d}{dt}(\gamma m_0 c\vec{v}) = -e(\vec{v} \times \vec{B}) \tag{2.21}$$

$$= -e(-v_z B_y \hat{x} + v_x B_y \hat{z}), \tag{2.22}$$

where m_0 is the electron's rest mass and $\vec{p} = \gamma m_0 c\vec{v}$ is the electron's relativistic momentum. Note that under our assumptions, the energy of the electron is conserved,

and the velocity along y remains zero. To understand the resulting trajectory, we first look only at the transverse component of this equation, and use eq (2.9) to replace our derivative with respect to time with a derivative with respect to z. Cancelling out the z-component of the velocity that subsequently appears on both sides, we get

$$\frac{d}{dz}(\gamma m_0 c v_x) = -eB_0 \sin(k_u z). \tag{2.23}$$

Integrating over z yields the result

$$v_x = \frac{eB_0}{\gamma m_0 k_u} \cos(k_u z) \tag{2.24}$$

$$= \frac{Kc}{\gamma} \cos(k_u z) \tag{2.25}$$

where in the final step we have made a notational simplification by defining the undulator deflection parameter as $K = \frac{eB_0}{m_0 c k_u}$. We see that the transverse velocity oscillates sinusoidally with respect to its longitudinal position.

Recalling, now, that our assumptions have implied a conservation of energy for our electron, we can infer from the above results how the longitudinal velocity evolves. Squaring both sides of eq (2.3) and using the fact that $v_y = 0$ under our assumptions, we have

$$1 - 1/\gamma^2 = \beta^2 = \frac{1}{c^2}(v_x^2 + v_z^2), \tag{2.26}$$

Solving for v_z and substituting the above result for v_x yields

$$v_z = c\sqrt{1 - 1/\gamma^2 - v_x^2/c^2} \tag{2.27}$$

$$= c\sqrt{1 - 1/\gamma^2 - \frac{K^2}{\gamma^2}\cos^2(k_u z)} \tag{2.28}$$

$$\approx c[1 - \frac{1}{2\gamma^2} - \frac{K^2}{2\gamma^2}\cos^2(k_u z)]. \tag{2.29}$$

From this, we can see that the longitudinal velocity has an oscillatory term that is the result of the energy transfer to the transverse direction x. Alternatively, we can think of this oscillatory term as being the result of the electron maintaining a constant speed but experiencing a path length increase (with respect to a straight-line trajectory along z) due to its transverse excursions in the horizontal plane. Note that the $\cos^2(k_u z)$ term in the above expression for the longitudinal velocity v_z oscillates with exactly twice the frequency of both the undulator field and the transverse velocity v_x. We can thereby easily observe that the average longitudinal velocity over one full period of the electron motion is given by

$$\overline{v_z} = c(1 - \frac{1 + K^2/2}{2\gamma^2}). \tag{2.30}$$

We are now able to determine the fundamental wavelength of the radiation produced by the oscillatory motion of the electron described above. For simplicity, we will assume the observer is situated on the z-axis. The phase of the observed synchrotron radiation will correspond directly to the phase of the transverse oscillations, which have been analyzed in the stationary reference frame. The wavelength of the radiation, which we will denote λ, measured by a stationary observer on the z-axis will therefore

be compressed according to the classical Doppler effect. In other words, the observed wavelength will be equal to the distance that light propagates ahead of the electron after one period of oscillation (which is of course less than the distance that light would propagate away from a stationary electron given the same amount of time). Thus, if we denote the wavelength of the transverse oscillations as $\lambda_u = 2\pi/k_u$, we have

$$\lambda = (c - \overline{v_z})\Delta t, \tag{2.31}$$

where $\Delta t = \lambda_u/\overline{v_z}$ is the period of one oscillation. Substituting this and our result for $\overline{v_z}$ into the above, we get

$$\lambda = (c/\overline{v_z} - 1)\lambda_u \tag{2.32}$$

$$= ([1 - \frac{1 + K^2/2}{2\gamma^2}]^{-1} - 1)\lambda_u \tag{2.33}$$

$$\approx \frac{1 + K^2/2}{2\gamma^2}\lambda_u. \tag{2.34}$$

This is the fundamental wavelength of undulator radiation.

2.2.3 Amplification by Spontaneous Emission

In the previous section, we computed the trajectory of an individual electron passing through an undulator and derived an expression for the wavelength of the resulting radiation. In this section, we will use those results to see how a beam of electrons passing through an undulator can produce an interaction, between the forward-propagating radiation field and the electrons themselves, that leads to amplification of the emitted radiation signal.

As discussed in the previous section, the radiation field emitted by an individual electron propagates ahead of the electron because the electron travels slower than light. However, if we consider a beam of electrons distributed longitudinally, the leading electrons will experience an interaction with the forward-propagating radiation field from the trailing electrons. If this interaction causes the affected electrons to decelerate, conservation of energy dictates that the kinetic energy lost will be gained as additional radiation. Under the proper conditions, it is possible to sustain this process throughout many undulator periods. When this occurs, the process is called self-amplification by spontaneous emission, or SASE, and it can be employed to amplify the strength of the undulator radiation by many orders of magnitude [4]. This interaction is the basis for an FEL device, and the resulting amplification is referred to as FEL gain.

To understand a bit about this process, we will analyze the interaction of the radiated EM field and the oscillating electrons by modeling the radiation as a forward-propagating plane wave:

$$\vec{E}(z,t) = E_0 \cos(kz - \omega t + \phi)\hat{x} \tag{2.35}$$

$$\vec{B}(z,t) = \frac{E_0}{c} \cos(kz - \omega t + \phi)\hat{y} \tag{2.36}$$

Recalling that the magnetic field does no work, the power delivered to an electron interacting with this electromagnetic wave at the position z and time t is given by:

$$P = \vec{F} \cdot \vec{v} = -e\vec{E} \cdot \vec{v} \tag{2.37}$$

$$= \frac{-eE_0Kc}{\gamma} \cos\left(kz - \omega t + \phi\right) \cos\left(k_u z\right) \tag{2.38}$$

where we have used the result for the x-component of the electron velocity derived in eq (2.25) from the previous section. If the power delivered to the electron is negative, the electron is decelerated, and the kinetic energy lost will be radiated into the EM-field, increasing the field strength. Therefore, in order to achieve appreciable gain, there must be a significant decrease in the electron kinetic energy, which requires that it sustains a net deceleration over multiple undulator periods. To establish how this is possible, we first use the trig identity $\cos u \cos v = \frac{1}{2}[\cos(u+v) + \cos(u-v)]$ to replace the product of the trig functions in the above expression with a sum:

$$P = \frac{-eE_0Kc}{2\gamma}[\cos((k+k_u)z - \omega t + \phi) + \cos((k-k_u)z - \omega t + \phi)]. \tag{2.39}$$

Now we can see that there can only be a significant net energy exchange if at least one of the two terms in the above sum does not average to zero. To accomplish this, we will require that at least one of the cosine terms has an approximately constant argument.

Let us define the phases ψ_+ and ψ_- to be the arguments of the above cosine terms. That is, let

$$\psi_+ = (k+k_u)z - \omega t + \phi \tag{2.40}$$

$$\psi_- = (k-k_u)z - \omega t + \phi \tag{2.41}$$

To see how these phases evolve throughout the undulator, we will take the derivative of each with respect to longitudinal position. To do this, let us look specifically at the phase ψ_+, after which we will apply the same procedure to ψ_-. We must keep in mind

that the time t is the time at which the electron arrives at position z. Thus we have

$$\frac{d\psi_+}{dz} = (k + k_u) - \omega\frac{dt}{dz} = (k + k_u) - \omega/v_z \tag{2.42}$$

$$= (k + k_u) - \frac{\omega}{c}[1 - 1/\gamma^2 - \frac{K^2}{\gamma^2}\cos^2 k_u z)]^{-\frac{1}{2}} \tag{2.43}$$

$$\approx (k + k_u) - k[1 + \frac{1}{2\gamma^2} + \frac{K^2}{2\gamma^2}\cos(k_u z)], \tag{2.44}$$

where we have used the fact that $\omega = ck$ and also substituted our result from eq for v_z and Taylor expanded. If the average rate of change of this phase throughout the undulator is zero, it is possible to have a sustained exchange of energy. Averaging the above expression over one undulator period, and including the similar result for ψ_-, we have

$$\overline{\frac{d\psi_+}{dz}} = k_u - k\frac{1 + K^2}{2\gamma^2}, \tag{2.45}$$

$$\overline{\frac{d\psi_-}{dz}} = -k_u - k\frac{1 + K^2}{2\gamma^2}. \tag{2.46}$$

From these results, we can see that over each undulator period, the phase ψ_- decreases by more than 2π. Such a rapid, and indeed monotonic, evolution of the phase ψ_- means that the second cosine term in equation (2.39) will oscillate rapidly throughout the undulator, and therefore cannot result in a sustained interaction between the field and the electron. Ignoring terms like this one that oscillate rapidly throughout the undulator is called *wiggle averaging* [4].

Contrarily, there exists what we call a *resonance condition* for the phase ψ_+. That is, if

$$k_u/k = \frac{1 + K^2}{2\gamma^2}, \tag{2.47}$$

then

$$\overline{\frac{d\psi_+}{dz}} = 0, \tag{2.48}$$

indicating that the phase ψ_+ will be, on average, constant throughout the undulator, and the interaction described by the first cosine term in eq (2.39) will result in a sustained exchange of energy. Notice that this condition is exactly satisfied by the fundamental wavelength of the undulator radiation given by eq (2.34)! Thus, we can see that there is a natural resonance that is created between the undulating electrons and the radiation field produced by their very motion.

The resonance discussed here is the foundational mechanism that enables an undulator to operate as an FEL device. Next, we will take a look at the equations that govern the output power of the FEL radiation beam, and see how they depend on the physical beam characteristics, in order to understand how to optimize this output.

2.2.4 Dependencies of the FEL Output

In the previous section, we studied the theoretical basics of SASE, the fundamental mechanism that enables an undulator to operate as a laser-like light amplification device. Of course, our analysis thus far has provided little more than a conceptual understanding, as we have only studied the interaction of a single electron with a simplified model of the undulator radiation. In regard to this work, we ultimately would like to determine what physical parameters affect the strength of the amplification interaction within a beam of electrons. In particular, we would like to establish the mechanism

by which the quadrupole magnets affect the FEL output power. However, while it has been important to establish a conceptual understanding of the underlying mechanism governing our target function, the analysis that leads to the quantitative understanding of a high-gain FEL device that we ultimately desire, even in its simplest form, becomes considerably more complex, and therefore exceeds the boundaries of this text. Thus, for a more complete treatment of FEL physics, we refer the reader to [8, 4]. For our purposes, we will simply summarize and collect the relevant results of said analysis, here.

Conveniently, most of the properties of a high-gain FEL can be characterized by a single, dimensionless quantity called the Pierce parameter, defined as [4]

$$\rho = [\frac{I}{8\pi I_A}(\frac{K[JJ]}{1+K^2/2})^2\frac{\gamma\lambda^2}{2\pi s^2}]^{1/3}. \tag{2.49}$$

Each of the terms in the above expression is defined in Table 2.1 below.

With ρ defined, we can compactly write down what are, to us, the most important results of this FEL theory. Firstly, we have that as the beam passes through the undulator, the SASE FEL power (initially) increases exponentially as

$$P \sim P_0 \exp(z/L_G). \tag{2.50}$$

Here, P_0 is the starting power (for an FEL starting from shot noise, $P_0 \sim$ kW) and L_G is the FEL gain length, which scales inversely with the Pierce parameter,

$$L_g \sim \rho^{-1}. \tag{2.51}$$

Parameter	Description
I	Beam current.
$I_A = 4\pi mc^3/e \approx 17kA$	Alfven current.
K	Undulator focusing parameter.
$[JJ] = J_0(\frac{K^2}{4+2K^2}) - J_1(\frac{K^2}{4+2K^2})$	Bessel function adjustment factor. J_n is the n^{th} Bessel function.
γ	Electron Lorentz factor.
λ	Fundamental wavelength of undulator radiation.
s	Cross-sectional width of the beam through the undulator.

Table 2.1: Definition of Quantities in Eq (2.49)

As can be seen in the above expression for the power, P, the gain length is effectively the longitudinal distance over which the power of emitted radiation is amplified by a factor of the natural number, e. Unfortunately, the exponential gain is not sustainable indefinitely. Recall that in our analysis of the undulator radiation, we assumed that the electron energy was conserved. Of course, if the undulator resonance is exploited to produce high-powered x-rays, the electrons will eventually lose significant kinetic energy. While the FEL power can continue to be amplified in this case by

using a tapered undulator, at a certain point, the marginal increase in power begins to scale linearly with respect to longitudinal position rather than exponentially [8]. This transition is called *saturation*. The power level at which the beam saturates, called the *saturation power*, or P_{sat}, is also related to the Pierce parameter [4]:

$$P_{sat} \sim \rho P_{beam}, \tag{2.52}$$

where

$$P_{beam} = \frac{I}{e}\gamma mc^2 \tag{2.53}$$

is the kinetic power of the beam. Thus, the Pierce parameter can be thought of as approximately the efficiency at which the FEL device converts the kinetic energy of the electron beam to radiation.

Given these relationships, we can see that by maximizing the Pierce parameter, we maximize the saturation power while minimizing the gain length. This means that not only will the exponential gain result in greater amplification, but the subsequent linear gain will also be greater because there are more remaining undulator periods over which it can be sustained. Thus, the problem of maximizing the FEL output radiation power is essentially equivalent to maximizing the Pierce parameter.

So now the question becomes, "How do the quadrupoles affect the Pierce parameter?" As mentioned in the introduction, the quadrupole magnets are used to focus the beam. While this focusing is necessary simply to transport the beam to the undulator, we can now see by examining eq (2.49) that the Pierce parameter is inversely

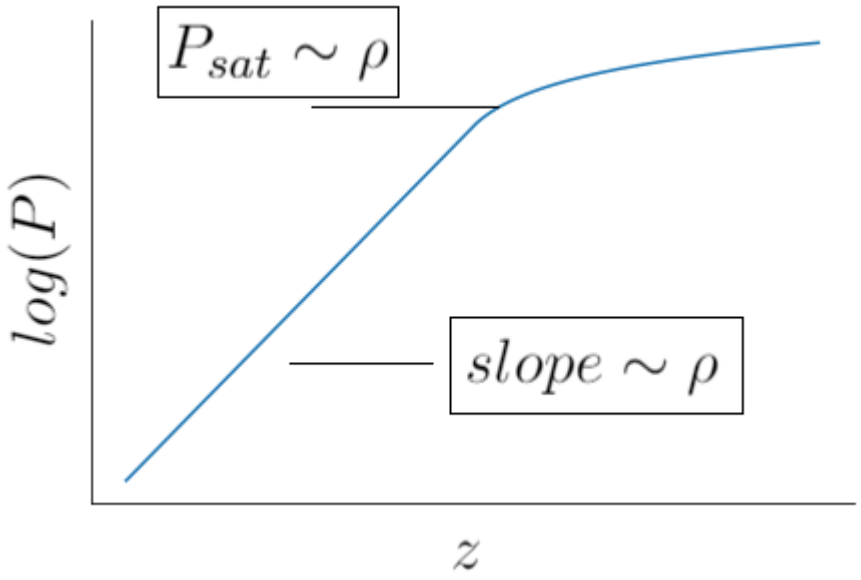


Figure 2.2: FEL Power Growth through Undulator

A basic illustration of the FEL X-ray power growth as a function of longitudinal position z in the undulator. The power P grows exponentially before reaching the saturation power P_{sat}, after which the growth is approximately linear. The Pierce parameter $\rho \propto \frac{1}{s^{2/3}}$ where s is the cross-sectional width of the beam in the undulator.

related to the cross-sectional width of the beam[1]. Thus, it is especially important that the beam is kept as tight as possible while it is passing through the undulator, where it produces the amplification discussed here [25]. To accomplish this, quadrupole magnets are placed periodically within the undulator to focus and refocus the beam as it performs its oscillations. In the upcoming sections, we will examine the mechanics of how these quadrupole magnets can be used as lenses to focus a charged particle beam,

[1]In the FEL theory discussed here, the beam is assumed to be symmetric in the two transverse directions, so only a single linear parameter (the beam radius) is needed to describe the cross-sectional width. In the more general case of an elliptical beam, s is replaced with $\sqrt{A/\pi}$ where A is the cross-sectional area of the beam.

from which we will begin to gain an understanding of the functional dependence of the FEL output power on the quadrupole settings.

2.2.5 Measuring the FEL Output

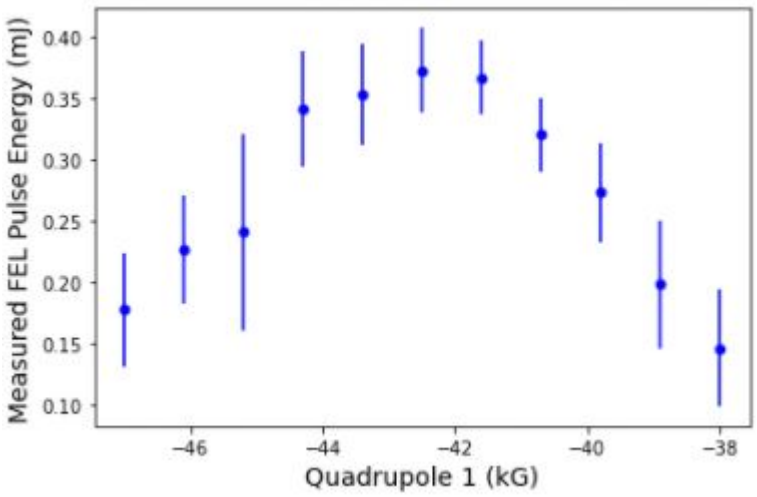


Figure 2.3: Measurements of the FEL X-ray Pulse Energy

Some measurements from a gas detector of the X-ray pulse energy at LCLS for different settings of a particular matching quadrupole. 120 pulses are measured for each setting. Mean energies are shown by the dots, with error bars representing one standard deviation.

In order to optimize the intensity of the FEL radiation, we must of course have a way to observe the target function. The X-ray pulse energy of the FEL at LCLS is measured by a device called a *gas detector*. The gas detector measures the intensity of the radiation by correlating it to the UV response of a population of N_2 molecules through which the radiation beam is passed [26]. While the the amount of UV radiation measured by the gas detector for a given radiation beam is correlated to

the X-ray intensity, it is also affected by the statistical properties of the gas, as well as background noise (although measures have of course been taken to minimize the effect of the latter). The resulting output signal from the gas detector is consequently noisy. An example of some measurements from the gas detector is shown in Fig 2.3. Presence of noise in the function observations is a critical factor to consider when it comes to selection of an optimization algorithm. We will see in Chapter 3 that in the Bayesian approach to optimization, uncertainty in the observations is explicitly factored into the decision-making process.

2.3 Electron Beam Transport Physics

In the previous sections, we saw how electrons traveling at nearly the speed of light can be used to generate extremely intense X-rays useful for imaging. This, of course, requires that we have access to an accelerated beam of electrons. To accelerate electrons up to these relativistic speeds requires longitudinal forces to be applied on the particles over distances on the order of miles. Over sufficiently long distances such as this, the electron beam will diverge significantly unless subjected to focusing forces. Thus, quadrupole focusing magnets are used throughout the beamline to maintain small transverse cross-sections of the beam, keeping the beam far away from apertures to avoid scraping and wakefields that may cause unpredictable and undesired behavior. Furthermore, as we saw in the previous section, it is particularly critical that we minimize the transverse size of the beam as it passes through the undulator in order to maximize

the FEL output power. In this section, we will examine the mechanics of transporting an electron beam and conditioning it to maximize the strength of the amplification interaction throughout the undulator.

2.3.1 Particle Beam Evolution in Free Space

In the absence of external forces, Newton's first law of motion tells us that the trajectory of an electron will be a straight line. Because of the high speeds of the particles in our beams (and vacuum chambers), we are able to ignore the weak acceleration due to Earth's gravitational field (the time it takes for a particle to be transported through any given section of the linac is so small that gravity is unable to accelerate it appreciably). Thus, if the particles are not subjected to any electromagnetic forces, their paths will be linear. The segments of the linac that do not contain electromagnetic fields are therefore approximated as free space and are referred to as *drift* spaces. We can describe the trajectory of an electron through such space as a function of its longitudinal coordinate:

$$x = x_0 + x'\Delta z \tag{2.54}$$

$$x' = x'_0 \tag{2.55}$$

Of course, in the case of a particle beam, we have a collection of particles with various transverse positions and velocities. To understand how a beam evolves, we can take a statistical approach of examining how the RMS of the particle coordinates evolve with

respect to z. That is, the evolution of the mean square of the x coordinates is given by

$$< x_f^2 > =< x_i^2 + 2x_i x' \Delta z + x'^2 \Delta z^2 > \tag{2.56}$$

$$=< x_i^2 > +2 < x_i x' > \Delta z + < x'^2 > \Delta z^2. \tag{2.57}$$

If we assume that the initial transverse particle coordinates are uncorrelated to the transverse velocities, then $< x_i x' >= 0$ and the middle term vanishes. We can compute the final RMS of the particle x-coordinates as

$$x_{rms,f} = \sqrt{< x_f^2 >} = \sqrt{< x_i^2 > + < x'^2 > \Delta z^2} \tag{2.58}$$

$$= \sqrt{x_{rms,i}^2 + {x'_{rms}}^2 \Delta z^2} \tag{2.59}$$

$$= x_{rms,i} \sqrt{1 + \frac{x'_{rms}}{x_{rms,i}^2} \Delta z^2}, \tag{2.60}$$

where $x_{rms,i}$ and x'_{rms} are the RMS of the initial positions and angles (in the x-direction), respectively. From this result, we can see that over very short distances Δz, the RMS x-coordinate increases quadratically, while over sufficiently long distances, the evolution of the RMS x-coordinate approaches a linearly increasing function with respect to z. In fact, for the latter conclusion, which is the more important for our discussion anyway, we need not make the assumption that the initial transverse positions and velocities are uncorrelated. Additionally, we can see that the latter conclusion holds for sufficiently large positive and negative values of Δz. The physical interpretation of this result is that any particle beam traveling (exclusively) through free space was at one point converging, and will eventually diverge. These results are illustrated in Figure (2.3), which depicts a beam consisting of many particles with various positions and velocities.

The RMS transverse positions are illustrated as a *beam envelope* which converges to a minimum size at what is referred to as the *beam waist*, before diverging. (While so far we have only considered the horizontal transverse direction, an identical argument can be made for the vertical direction.)

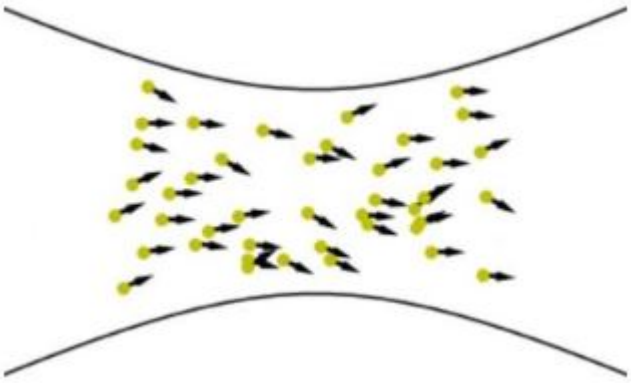


Figure 2.4: Particle Beam Evolution in Free-Space

A side-view of a beam of particles (yellow) with small transverse velocities (black arrows) traveling through free-space. The beam envelope, represented by the solid black curves above and below, converges to a minimum cross-sectional width at the *beam waist* before diverging indefinitely.

The fact that any beam traveling through free space will eventually diverge leads to the necessity of focusing devices. These devices will be responsible for maintaining small deviations from the design trajectory throughout the linac. While there are a number of ways to achieve focusing, the modern approach is to use quadrupole magnetic fields, which we will study in the following section.

2.3.2 Single Particle Motion in Quadrupole Magnetic Field

Transporting a beam containing trillions of relativistic electrons with varied positions and momenta requires that it periodically be subjected to focusing forces. Perhaps the most obvious way of focusing a charged particle beam would be to subject it to an electric field gradient. However, because the magnetic component of the Lorentz force exhibits a linear dependence on the particle speed, in the case of a beam of electrons traveling at relativistic speeds approaching that of light, it is possible to obtain a much stronger focusing effect by using a magnetic field. Achieving a lens-like focusing effect using a magnetic field is not quite as straightforward as in the case of an electric field, so in this section I will cover the physics of this so-called *strong focusing*.

The modern approach to focusing in accelerator physics employs quadrupole magnets that generate transverse fields whose strength increases linearly with the distance from the z-axis [21]:

$$B_y = -gx, \tag{2.61}$$

$$B_x = -gy. \tag{2.62}$$

Here, g is the magnitude of the quadrupole's magnetic field gradient, which can be manipulated by adjusting the currents through the electromagnets composing the quadrupole. For an electron traveling along the z-direction with transverse positions x and y, this

magnetic field results in a Lorentz force with components

$$F_x = ev_z B_y = -ev_z gx, \tag{2.63}$$

$$F_y = -ev_z B_x = ev_z gy. \tag{2.64}$$

By design, each component of the resultant force depends only on the electron's position along the respective direction. That is, the transverse forces are decoupled. Furthermore, the components scale linearly with their respective transverse coordinate. This means that electrons whose positions are further from the design trajectory (z-axis) will experience a greater force. Notice, however, that for positive values of g, while the x-component of the force is directed toward the z-axis, the y-component of the force is directed away (and the opposite is true if the sign of g is reversed). Thus, such a quadrupole device will be focusing only in one transverse direction, and necessarily defocusing in the other.

Substituting the x-component of the above force into the relativistic Newton's 2nd law, and recalling that the magnetic field does no work (γ remains constant), we have:

$$F_x = \frac{dp_x}{dt} \tag{2.65}$$

$$-ev_z gx = \gamma m_0 \frac{d^2}{dt^2} x \tag{2.66}$$

$$-ev_z gx = \gamma m_0 v_z^2 x''. \tag{2.67}$$

Where in the last step above we have replaced the derivative with respect to time on the right by a derivative with respect to longitudinal position using our change-of-

variable rules from Section 2.1.2. Rearranging, and following a similar procedure for the y-component, we arrive at the (familiar) differential equations:

$$x'' + k^2 x = 0, \tag{2.68}$$

$$y'' - k^2 y = 0, \tag{2.69}$$

where we have introduced the focusing strength parameter k such that $k^2 = \frac{eg}{\gamma m_0 v_z} = \frac{eg}{p}$, in which p is the electron momentum. The x-equation, of course, results in simple harmonic motion, while the solutions in y are hyperbolic:

$$x = x_0 \cos\left[k(z - z_0)\right] + \frac{x_0'}{k} \sin\left[k(z - z_0)\right], \tag{2.70}$$

$$y = y_0 \cosh\left[k(z - z_0)\right] + \frac{y_0'}{k} \sinh\left[k(z - z_0)\right]. \tag{2.71}$$

These are the equations describing the evolution of a single electron traveling a distance $(z - z_0)$ through a magnetic quadrupole with constant gradient in the orientation described by eqs (2.61) and (2.62) above. If the quadrupole is rotated 90° about the z-axis, the equations of motion for x and y will be swapped, resulting in simple harmonic motion in y and hyperbolic motion in x.

2.3.3 Linear Optics and Transport Matrix Formalism

Having just solved for the trajectory of our electron passing through the quadrupole magnetic field, we will now rewrite the results in the formalism of linear algebra. At a given longitudinal position z, the state of the electron trajectory is defined by its position x, y and angle x', y' relative to the ideal trajectory (z-axis). That

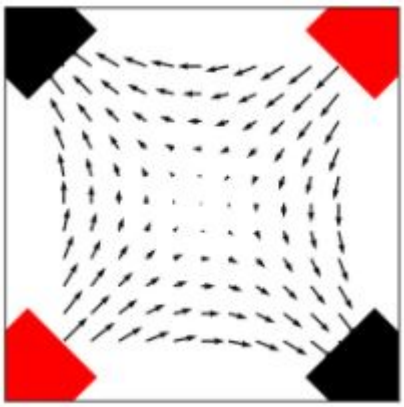


Figure 2.5: Quadrupole Focusing Magnet

An illustration of the magnetic field of a quadrupole focusing magnet as seen looking along the optical axis (into page). Magnetic field lines are depicted by black arrows. The red and black objects in the corners represent magnetic north and south poles, respectively, of electromagnets with variable strengths.

is, we can describe the transverse state of our electron using two vectors:

$$\vec{x} = \begin{pmatrix} x \\ x' \end{pmatrix}, \vec{y} = \begin{pmatrix} y \\ y' \end{pmatrix}. \tag{2.72}$$

Including the derivatives of eqs (2.70) and (2.71) to describe the evolution of the angles x', y' results in a system of linear equations that can be written in matrix notation as

$$\begin{pmatrix} x \\ x' \end{pmatrix} = \begin{pmatrix} \cos\left[k(z-z_0)\right] & \frac{1}{k}\sin\left[k(z-z_0)\right] \\ -k\sin\left[k(z-z_0)\right] & \cos\left[k(z-z_0)\right] \end{pmatrix} \begin{pmatrix} x_0 \\ x_0' \end{pmatrix}, \tag{2.73}$$

$$\begin{pmatrix} y \\ y' \end{pmatrix} = \begin{pmatrix} \cosh\left[k(z-z_0)\right] & \frac{1}{k}\sinh\left[k(z-z_0)\right] \\ -k\sinh\left[k(z-z_0)\right] & \cosh\left[k(z-z_0)\right] \end{pmatrix} \begin{pmatrix} y_0 \\ y_0' \end{pmatrix}. \tag{2.74}$$

More compactly, we can write the transformation as

$$\vec{x} = \boldsymbol{M_f} \cdot \vec{x_0}, \tag{2.75}$$

$$\vec{y} = \boldsymbol{M_d} \cdot \vec{y_0}, \tag{2.76}$$

wherein we have defined the transport matrices for a quadrupole of length $L = z - z_0$, which focuses in x and defocuses in y, as

$$\boldsymbol{M_f} = \begin{pmatrix} \cos\left[k(z-z_0)\right] & \frac{1}{k}\sin\left[k(z-z_0)\right] \\ -k\sin\left[k(z-z_0)\right] & \cos\left[k(z-z_0)\right] \end{pmatrix}, \tag{2.77}$$

$$\boldsymbol{M_d} = \begin{pmatrix} \cosh\left[k(z-z_0)\right] & \frac{1}{k}\sinh\left[k(z-z_0)\right] \\ -k\sinh\left[k(z-z_0)\right] & \cosh\left[k(z-z_0)\right] \end{pmatrix}. \tag{2.78}$$

If the quadrupole were rotated 90° about the z-axis, it would focus in y and defocus in x, and the matrices $\boldsymbol{M_f}, \boldsymbol{M_d}$ in eqs (2.75) and (2.76) would be swapped. One can verify that in the limit that the focusing strength k goes to zero, we recover the equations for straight line motion discussed in Section (2.3.1). That is, the transport matrix for a drift space of length l can be written

$$\boldsymbol{M_0} = \begin{pmatrix} 1 & l \\ 0 & 1 \end{pmatrix}. \tag{2.79}$$

A useful approximation can be made by applying the so-called thin-lens limit, where we assume the length of the quadrupole L approaches zero, but the product k^2L remains constant and finite. In this limit, the position of the electron is unaltered (as is required by continuity), with only the angles x', y' being affected. In other words, this approximation describes a lens that is sufficiently short such that the beam size does not vary

appreciably as the beam passes through it. The transport matrices for a quadrupole in the thin-lens limit become:

$$\boldsymbol{M_f} = \begin{pmatrix} 1 & 0 \\ -1/f & 1 \end{pmatrix}, \tag{2.80}$$

$$\boldsymbol{M_d} = \begin{pmatrix} 1 & 0 \\ 1/f & 1 \end{pmatrix}, \tag{2.81}$$

wherein we have defined the focal length $f = (k^2 L)^{-1}$. One may recognize that in the thin-lens limit, we have recovered the exact transformation rules for thin optical lenses. That is, the quadrupole acts as a simple converging lens in the x direction, and a diverging lens in the y direction.

Having written the evolution of the electron state in the formalism of linear transport, it becomes quite simple (by means of recursion) to compute the effect of a series of optical elements. For a segment of the accelerator consisting of N elements, we have for the state of the electron in the x-direction:

$$\vec{x} = \boldsymbol{M_N M_{N-1} ... M_2 M_1} \vec{x_0}, \tag{2.82}$$

where $\boldsymbol{M_i}$ is the transport matrix of the ith element (whether it be focusing quad, defocusing quad, or drift space), indexed such that $i = 0$ corresponds to the element the electron passes through first and $i = N$ corresponds to the element the electron passes through last. Notice that the total transport matrix, $\boldsymbol{M_{tot}}$, which transforms the electron from its initial state $\vec{x_0}$ to its final state $\vec{x}$, is equal to the product of the

transport matrices, multiplied in reverse-chronological order. That is,

$$\boldsymbol{M_{tot}} = \prod_{i=0}^{N-1} \boldsymbol{M_{N-i}} = \boldsymbol{M_N M_{N-1} ... M_2 M_1}. \tag{2.83}$$

2.3.4 Beam Envelope and Twiss Parameters

So far, we have seen how we can use transport matrix formalism to study the evolution of a single electron. Of course, as we saw in Section (2.2.4), we are ultimately concerned with the evolution of the shape, specifically the transverse cross-sectional width, of a beam consisting of what is typically billions of electrons. Therefore, we now turn to a statistical approach to understand how the shape of the beam evolves.

As we discussed previously, the state of a single electron is defined by its phase-space coordinates $\vec{x} = \begin{pmatrix} x \\ x' \end{pmatrix}$ and $\vec{y} = \begin{pmatrix} y \\ y' \end{pmatrix}$. In the case of a beam, at a particular longitudinal position, we have many electrons with varied phase-space coordinates. Conveniently, the distribution of phase-space coordinates can typically be well-approximated by two separate 2-dimensional Gaussians, one for the x and x' phase-space coordinates, and one for the y and y' phase-space coordinates [22]. Under such an approximation, curves of constant phase-space density are ellipses. Thus, to define the state of the beam, we can construct an ellipse in each of the $\vec{x}$- and $\vec{y}$- phase-spaces that envelopes a certain percentage of the electrons, as shown in Figure (2.5). An important feature of ellipses is that they remain ellipses under linear transformations. Furthermore, if the matrix representing the linear transformation has unit determinant, it will preserve the area of the transformed ellipse. Notice that the transport matrices for quadrupoles (both focusing and defocusing), as well as drift spaces, indeed satisfy this condition. Therefore,

if the electrons pass through any number of such elements (e.g. quad-drift-quad-drift etc.) the characterizing ellipse will simply be transformed into another ellipse whose area in phase space is unchanged. Here, we will derive the transformation rule for the parameters defining these characterizing ellipses. For simplicity, we will look only at the x-direction, but the treatment will generalize to the y-direction, as well.

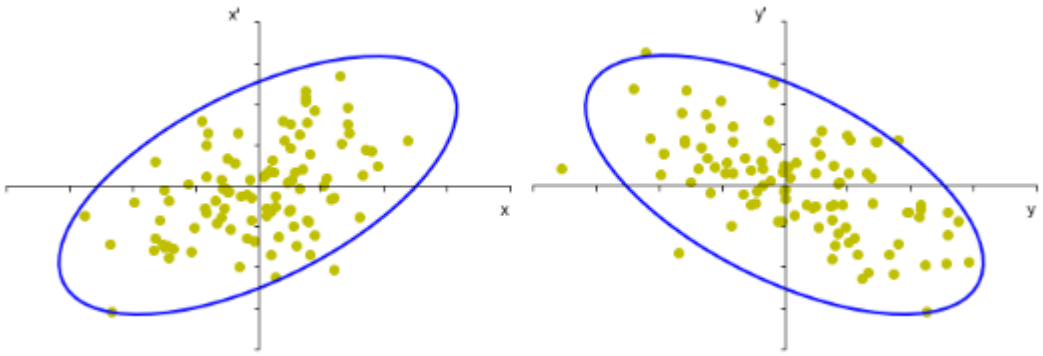


Figure 2.6: Particle Beam Representation in Phase-Space

Example of a beam of particles (yellow) Gaussianly distributed in the $\vec{x}$ and $\vec{y}$ phase-spaces. The two characterizing ellipses (blue) are generally independent, with each described by its own set of Twiss parameters, which are manipulated via the quadrupole focusing magnets.

Suppose that, at a longitudinal position z, the distribution of the electrons in the x and x' phase-space is described by the zero-mean, 2-dimensional (correlated), normalized Gaussian probability density function:

$$F_0(x, x') = \frac{1}{2\pi\epsilon_x} \exp\left[-\frac{1}{2\epsilon_x}(\gamma_x x^2 + \beta_x x'^2 + 2\alpha_x x x')\right] \tag{2.84}$$

To interpret the quantities $\epsilon_x, \gamma_x, \beta_x, \alpha_x$, we compare this to the general zero-mean,

2-dimensional normalized Gaussian probability density function given by:

$$f(\vec{x}) = \frac{1}{2\pi|\mathbf{\Sigma}_x|^{\frac{1}{2}}} \exp\,[-\frac{1}{2}\vec{x}^{\top}\mathbf{\Sigma}_x^{-1}\vec{x}], \tag{2.85}$$

where $\vec{x} = \binom{x}{x'}$, and

$$\mathbf{\Sigma}_x = \begin{pmatrix} \sigma_x^2 & \sigma_{xx'} \\ \sigma_{xx'} & \sigma_{x'}^2 \end{pmatrix} = \begin{pmatrix} < x >^2 & < xx' > \\ < xx' > & < x' >^2 \end{pmatrix} \tag{2.86}$$

is the non-singular covariance matrix of second-order moments.

Looking at the normalization factor in front of the exponential, we can immediately see that

$$\epsilon_x = |\mathbf{\Sigma}_x|^{\frac{1}{2}} = \sqrt{\sigma_x^2\sigma_{x'}^2 - \sigma_{xx'}^2}. \tag{2.87}$$

Because the matrix $\mathbf{\Sigma}$ is non-singular, we can invert it using the rule for any invertible 2x2 matrix $\boldsymbol{L}$. That is, if we have some 2x2 matrix $\boldsymbol{L}$ such that $\boldsymbol{L} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ and $|\boldsymbol{L}| \neq 0$, then

$$\boldsymbol{L}^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{|\boldsymbol{L}|}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix}. \tag{2.88}$$

Applying this to the above, we get

$$\mathbf{\Sigma}_x^{-1} = \frac{1}{|\mathbf{\Sigma}_x|}\begin{pmatrix} \sigma_{x'}^2 & -\sigma_{xx'} \\ -\sigma_{xx'} & \sigma_x^2 \end{pmatrix} = \frac{1}{\epsilon_x^2}\begin{pmatrix} \sigma_{x'}^2 & -\sigma_{xx'} \\ -\sigma_{xx'} & \sigma_x^2 \end{pmatrix}. \tag{2.89}$$

Substituting this expression for $\mathbf{\Sigma}_x^{-1}$ into eq (2.85) and comparing the result with eq (2.84), we can see that the parameters $\beta_x, \gamma_x, \alpha_x$, are related to the second-order mo-

ments by:

$$\sigma_x^2 =< x^2 >= \epsilon_x \beta_x, \tag{2.90}$$

$$\sigma_{x'}^2 =< x'^2 >= \epsilon_x \gamma_x, \tag{2.91}$$

$$\sigma_{xx'} =< xx' >= -\epsilon_x \alpha_x. \tag{2.92}$$

Notice that if we use the above expressions to rewrite eq (2.87) in terms of $\beta_x, \gamma_x, \alpha_x$, we can square both sides and cancel out the factors ϵ_x^2, leaving us with the useful identity:

$$1 = \beta_x \gamma_x - \alpha_x^2. \tag{2.93}$$

Now suppose this distribution of electrons is linearly transported to the longitudinal position z_f (some distance downstream of z), such that an electron initially found at the coordinates $\vec{x} = \binom{x}{x'}$ ends up at the final coordinates $\vec{x_f} = \binom{x_f}{x'_f}$ according to:

$$\begin{pmatrix} x_f \\ x'_f \end{pmatrix} = \boldsymbol{M} \cdot \begin{pmatrix} x \\ x' \end{pmatrix} \tag{2.94}$$

where $\boldsymbol{M} = \begin{pmatrix} C & S \\ C' & S' \end{pmatrix}$ and $|\boldsymbol{M}| = 1$. More formally, suppose the probability mass initially contained within a square infinitesimal area element at the coordinates $\binom{x}{x'}$ is mapped to an infinitesimal parallelogram at $\binom{x_f}{x'_f}$ according to the above transformation. Because the transport matrix has unit determinant, this transformation is area-preserving. Therefore, to conserve probability mass, the transformed probability density at $\binom{x_f}{x'_f}$ must be equal to the initial probability density at $\binom{x}{x'}$. That is, if we

call the final probability density F, then we have

$$F(x_f, x'_f) = F_0(x, x'). \tag{2.95}$$

To understand how the shape of the final probability density function F has changed from F_0, we can compute the inverse of the transport matrix (by once again using the rule from eq (2.88) for 2x2 invertible matrices) and use it to rewrite the coordinates $\binom{x}{x'}$ in terms of $\binom{x_f}{x'_f}$:

$$\boldsymbol{M}^{-1} = \begin{pmatrix} C & S \\ C' & S' \end{pmatrix}^{-1} = \frac{1}{|\boldsymbol{M}|} \begin{pmatrix} S' & -S \\ -C' & C \end{pmatrix} = \begin{pmatrix} S' & -S \\ -C' & C \end{pmatrix} \tag{2.96}$$

$$\begin{pmatrix} x \\ x' \end{pmatrix} = \boldsymbol{M}^{-1} \cdot \begin{pmatrix} x_f \\ x'_f \end{pmatrix} = \begin{pmatrix} S' & -S \\ -C' & C \end{pmatrix} \cdot \begin{pmatrix} x_f \\ x'_f \end{pmatrix}. \tag{2.97}$$

Substituting for $\binom{x}{x'}$ in equation (2.84), we have

$$F(x_f, x'_f) = F_0(S'x_f - Sx'_f, -C'x_f + Cx'_f) \tag{2.98}$$

$$= \frac{1}{2\pi\epsilon_x} \exp\left[-\frac{1}{2\epsilon_x}(\gamma_x(S'x_f - Sx'_F)^2 + \beta_x(-C'x_f + Cx'_f)^2\right. \tag{2.99}$$

$$\left. +2\alpha_x(S'x_f - Sx'_f)(-C'x_f + Cx'_f))\right] \tag{2.100}$$

$$= \frac{1}{2\pi\epsilon_{x_f}} \exp\left[-\frac{1}{2\epsilon_{x_f}}(\gamma_{x_f}x_f^2 + \beta_{x_f}{x'_f}^2 + 2\alpha_{x_f}x_f x'_f)\right], \tag{2.101}$$

where

$$\epsilon_{x_f} = \epsilon_x, \tag{2.102}$$

$$\beta_{x_f} = C^2\beta_x - 2SC\alpha_x + S^2\gamma_x, \tag{2.103}$$

$$\alpha_{x_f} = -CC'\beta_x + (SC' + S'C)\alpha_x - SS'\gamma_x, \tag{2.104}$$

$$\gamma_{x_f} = C'^2\beta_x - 2S'C'\alpha_x + S'^2\gamma_x. \tag{2.105}$$

Now, the final probability density is written in the same form as the initial, except that the parameters $\beta_{x_f}, \alpha_{x_f}, \gamma_{x_f}$ have been linearly transformed from the initial parameters $\beta_x, \alpha_x, \gamma_x$ according to

$$\begin{pmatrix} \beta_{x_f} \\ \alpha_{x_f} \\ \gamma_{x_f} \end{pmatrix} = \begin{pmatrix} C^2 & -2SC & S^2 \\ -CC' & (SC' + S'C) & -SS' \\ C'^2 & -2S'C' & S'^2 \end{pmatrix} \cdot \begin{pmatrix} \beta_x \\ \alpha_x \\ \gamma_x \end{pmatrix}. \tag{2.106}$$

Thus, we arrive at the transformation rule for the parameters of a Gaussian-distributed beam (or any elliptically distributed beam, for that matter). The parameters $\beta_x, \alpha_x, \gamma_x$ are called the *Twiss parameters*, and are the primary descriptors of the transverse state of the beam in accelerator optics. Note that the normalization factor, ϵ_x, which we call the *geometric emittance*, remains unchanged as a consequence of the area-preserving linear transformation [27]. The geometric emittance is so-called because it has the geometric interpretation that it is $1/\pi$ times the area of the ellipse defined by the set of points that are unit *Mahalanobis distance* away from the center of the distribution [28]:

$$d_M^2 \equiv \vec{x}^\top \boldsymbol{\Sigma}^{-1} \vec{x} = 1. \tag{2.107}$$

(To convince ourselves of this, we could show that the linear transformation given by

$$\vec{x} = \frac{1}{\sqrt{\epsilon}} \begin{pmatrix} \sigma_{x'} & \frac{\sigma_{xx'}}{\sigma_{x'}} \\ 0 & \sqrt{\sigma_x^2 - \frac{\sigma_{xx'}^2}{\sigma_{x'}^2}} \end{pmatrix} \vec{v} \tag{2.108}$$

is an area-preserving transformation that maps the circle $\vec{v}^{\top}\vec{v} = \epsilon$ onto the ellipse above.) In this regard, the geometric emittance is an invariant (under our assumptions) measure of the phase-space area of the beam, with larger emittances describing wider, flatter distributions.

2.3.5 Alternating-gradient Focusing Lattice

We have seen in eqs (2.68) and (2.69) that a transverse quadrupole magnetic field acting on a charged particle beam produces forces that scale linearly with the transverse displacements. In one transverse dimension, the force produces a focusing effect (pushing electrons toward the optical axis), while in the other it produces a defocusing effect (pushing electrons away from the optical axis). From these results, it is clear that if we are to achieve a net focusing effect in both transverse directions, we will require more than one quadrupole. The obvious solution is to use one quadrupole to focus the beam in the x-direction, and another (rotated 90° about the z-axis with respect to the first) to focus the beam in the y-direction. This arrangement is referred to as an alternating-gradient focusing lattice [11]. Still, it is not immediately obvious that such a configuration is capable of achieving net focusing in both directions.

To convince ourselves that it is, indeed, possible to achieve net focusing, let us consider an alternating-gradient focusing lattice consisting of two identical quadrupoles

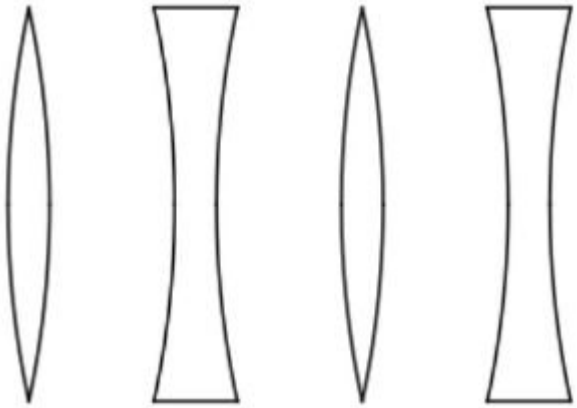


Figure 2.7: Alternating-Gradient Focusing Lattice

A periodic series of converging and diverging lenses (quadrupoles) in alternating order. (Note: This optical schematic can only describe the behavior of the quadrupoles in one transverse direction. The schematic for the other transverse direction will have the positions of the converging and diverging lenses swapped.)

separated by some distance L and rotated 90° about the z-axis with respect to one another. It is clear from the thin-lens approximation that in order to realize any net effect, it is necessary to include a drift space between (and after) the two quads. (Without a drift space between the quads, their effects would cancel: $\boldsymbol{M_f \cdot M_d = M_d \cdot M_f = I}$.) For simplicity, let us examine the effect that such a configuration will have on an individual electron. Let $\vec{x_0} = \binom{x_0}{x_0'}, \vec{y_0} = \binom{y_0}{y_0'}$ be the initial phase-space coordinates of the electron just before entering the first quad, $\vec{x_1}, \vec{y_1}$ be the coordinates just before the second quad, and $\vec{x_2}, \vec{y_2}$ be the coordinates just after the second quad. Without loss of generality, we assume that the quadrupole the electron passes through first has a focusing effect in the x-direction, and a defocusing effect in the y-direction (and thus vice-versa for the

second quadrupole). Using the transport rules from Section 2.3.3, and working in the thin-lens approximation, the transformation of the electron phase-space coordinates is given by:

$$\vec{x_1} = \begin{pmatrix} x_1 \\ \\ x_1' \end{pmatrix} = \boldsymbol{M_0} \cdot \boldsymbol{M_f} \cdot \vec{x_0} = \begin{pmatrix} 1 & L \\ \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ \\ -1/f & 1 \end{pmatrix} \cdot \begin{pmatrix} x_0 \\ \\ x_0' \end{pmatrix}, \tag{2.109}$$

$$\vec{x_2} = \begin{pmatrix} x_2 \\ \\ x_2' \end{pmatrix} = \boldsymbol{M_d} \cdot \vec{x_1} = \begin{pmatrix} 1 & 0 \\ \\ 1/f & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ \\ x_1' \end{pmatrix}. \tag{2.110}$$

By writing the transformation in terms of the intermediate phase-space coordinates $\vec{x_1}$, we are more readily able to see how the quadrupoles' effects depend on the electron positions. Specifically, we are interested in seeing the net effect on the angle(s) after the electron has passed through both quadrupoles:

$$\Delta x' = x_2' - x_0' = \frac{1}{f}x_1 + x_1' - x_0' = \frac{1}{f}x_1 + (-\frac{1}{f}x_0 + x_0') - x_0' \tag{2.111}$$

$$= \frac{1}{f}(x_1 - x_0). \tag{2.112}$$

A similar result can be obtained for the net deflection in the y-direction:

$$\Delta y' = y_2' - y_0' = -\frac{1}{f}(y_1 - y_0). \tag{2.113}$$

From the above equations, we can see that the electron will be subjected to a net restoring impulse (i.e. be deflected toward the optical axis) in both the x- and y-directions, provided $x_1 < x_0$ and $y_1 > y_0$. Recall that the first quadrupole provides a deflection toward the optical axis in x, and away from the axis in y. Therefore, if the electron is initially diverging in both x and y, these conditions can be met with the appropriate choice of focal length.

That net focusing is achievable by an alternating-gradient focusing lattice is indeed the result of the fact that the impulse delivered by each quadrupole lens scales linearly with the respective displacement from the optical axis. In the arrangement discussed above, the first quadrupole provides an impulse that tends to reduce the displacement in the x-direction and tends to increase the displacement in the y-direction. The second quadrupole will then deliver an impulse that is opposite to the first in both directions, but its defocusing effect in the x-direction will be smaller in magnitude than the focusing effect of the first quad, and its focusing effect in the y-direction will be larger in magnitude than the defocusing effect of the first quad, yielding a net impulse whose x- and y- components are both directed toward the optical axis. In other words, the focusing effect of each quad will be stronger than the defocusing effect of the other.

Focusing the electron beam into a small transverse cross section before it passes through the undulator is a critical step in the operation of a high-energy X-ray FEL, and is made possible by the *strong focusing* afforded by these alternating-gradient focusing lattices [25]. Next, we will see how a series of identical, symmetric alternating-gradient focusing cells like the one we have examined here can be used to maintain minimal deviations about some target beam size.

2.3.6 The Matched Beam: Beam Size

As we saw in Section (2.2.4), the FEL gain-length is inversely related to the cross-sectional size of the beam. Therefore, not only must we focus the beam down before it reaches the undulator, we must also ensure the beam remains as tight as possible

as it passes through the undulator in order to maximize the final intensity of the generated radiation. To accomplish this, a series of symmetric alternating-gradient focusing cells (just like the one we examined in the previous section) are situated throughout the undulator. Here, we will show that if the beam enters such a lattice with the proper Twiss parameters, the ensuing motion will be periodic. This allows for the construction of a focusing lattice that is able to maintain minimal deviations from some specified target beam size (determined by the strength of and distance between the quadrupoles). The set of Twiss parameters that results in this periodic motion is referred to as the *match* and the process of manipulating the Twiss parameters at the beginning of the undulator via upstream quadrupoles to achieve the match is called *matching the beam.*

With this knowledge, finally, we are beginning to form a quantifiable understanding of the process of quadrupole tuning. In fact, in the context of this paper, "matching the beam" is essentially equivalent to "quadrupole tuning." However, it should be noted that in general, this is not the case. Due to the length of the linac, there are close to 100 quadrupole focusing magnets, many of which are simply used to prevent beam-loss while transporting the beam to the undulator hall. Typically, however, only 24 of these quadrupoles are used for fine-tuning of the Twiss parameters, with 4 quadrupoles in particular being the most commonly used in the process of matching the beam. In this work, we have chosen to focus on this smaller subset of quadrupoles known as the *matching quads.* These are the quadrupoles immediately upstream of the undulator, and the ones responsible for the final and finest adjustments to the Twiss parameters before the beam enters the focusing channel of the undulator.

Now, let us calculate the matched Twiss parameters for a given alternating-gradient focusing cell. The easiest way to do this is to imagine that there are two identical alternating-gradient focusing cells placed one after the other (with congruent drift spaces between each quadrupole), and to then calculate the Twiss parameters at a position halfway through the first quadrupole that result in the same Twiss parameters after it has been transported through exactly one full period of the system. That is, we wish to impose the periodic conditions:

$$\begin{pmatrix} \beta_x \\ \alpha_x \\ \gamma_x \end{pmatrix} = \boldsymbol{\Lambda_x} \cdot \begin{pmatrix} \beta_x \\ \alpha_x \\ \gamma_x \end{pmatrix}, \begin{pmatrix} \beta_y \\ \alpha_y \\ \gamma_y \end{pmatrix} = \boldsymbol{\Lambda_y} \cdot \begin{pmatrix} \beta_y \\ \alpha_y \\ \gamma_y \end{pmatrix}, \tag{2.114}$$

where $\boldsymbol{\Lambda_x}, \boldsymbol{\Lambda_y}$ are the transformation matrices for the Twiss parameters, which we compute using eq (2.106), representing transport through one full period of the system. Note that we are not imposing a periodic condition on any individual particle, but for the beam envelope. It is therefore important not to confuse the transport rules for the particle coordinates given in eq (2.82) with the transport rules for the Twiss parameters. That said, to compute the matrices $\boldsymbol{\Lambda_x}, \boldsymbol{\Lambda_y}$, we will first compute the matrices representing transport through one period for the individual particles (using our results from Section 2.3.3), and then use eq (2.106) to express the elements of $\boldsymbol{\Lambda_x}, \boldsymbol{\Lambda_y}$ in terms of those results. The single-particle transport matrices for our complete

focusing cell are given by:

$$\boldsymbol{M_{tot,x}} = \boldsymbol{M_{f,\frac{1}{2}}} \cdot \boldsymbol{M_0} \cdot \boldsymbol{M_d} \cdot \boldsymbol{M_0} \cdot \boldsymbol{M_{f,\frac{1}{2}}} \tag{2.115}$$

$$= \begin{pmatrix} 1 & 0 \\ -\frac{1}{2f} & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & L \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ \frac{1}{f} & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & L \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ -\frac{1}{2f} & 0 \end{pmatrix} \tag{2.116}$$

$$= \begin{pmatrix} 1 - \frac{l^2}{2f^2} & 2l(1 + \frac{l}{2f}) \\ -\frac{l}{2f^2}(1 - \frac{l}{2f}) & 1 - \frac{l^2}{2f^2} \end{pmatrix} \tag{2.117}$$

$$\boldsymbol{M_{tot,y}} = \boldsymbol{M_{d,\frac{1}{2}}} \cdot \boldsymbol{M_0} \cdot \boldsymbol{M_f} \cdot \boldsymbol{M_0} \cdot \boldsymbol{M_{d,\frac{1}{2}}} \tag{2.118}$$

$$= \begin{pmatrix} 1 & 0 \\ \frac{1}{2f} & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & L \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ -\frac{1}{f} & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & L \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ \frac{1}{2f} & 0 \end{pmatrix} \tag{2.119}$$

$$= \begin{pmatrix} 1 - \frac{l^2}{2f^2} & 2l(1 - \frac{l}{2f}) \\ -\frac{l}{2f^2}(1 + \frac{l}{2f}) & 1 - \frac{l^2}{2f^2} \end{pmatrix} \tag{2.120}$$

Substituting these results into eq (2.106), we have for our matrices $\boldsymbol{\Lambda_x}, \boldsymbol{\Lambda_y}$:

$$\boldsymbol{\Lambda_x} = \begin{pmatrix} (1 - \frac{l^2}{2f^2})^2 & -4l(1 + \frac{l}{2f})(1 - \frac{l^2}{2f^2}) & 4l^2(1 + \frac{l}{2f})^2 \\ \frac{l}{2f^2}(1 - \frac{l^2}{2f^2})(1 - \frac{l}{2f}) & -\frac{l^2}{f^2}(1 - \frac{l^2}{4f^2}) + (1 - \frac{l^2}{2f^2})^2 & -2l(1 + \frac{l}{2f})(1 - \frac{l^2}{2f^2}) \\ \frac{l^2}{4f^4}(1 - \frac{l}{2f})^2 & \frac{1}{f^2}(1 - \frac{l^2}{2f^2})(1 - \frac{l}{2f}) & (1 - \frac{l^2}{2f^2})^2 \end{pmatrix} \tag{2.121}$$

$$\boldsymbol{\Lambda_y} = \begin{pmatrix} (1 - \frac{l^2}{2f^2})^2 & -4l(1 - \frac{l}{2f})(1 - \frac{l^2}{2f^2}) & 4l^2(1 - \frac{l}{2f})^2 \\ \frac{l}{2f^2}(1 - \frac{l^2}{2f^2})(1 + \frac{l}{2f}) & -\frac{l^2}{f^2}(1 - \frac{l^2}{4f^2}) + (1 - \frac{l^2}{2f^2})^2 & -2l(1 - \frac{l}{2f})(1 - \frac{l^2}{2f^2}) \\ \frac{l^2}{4f^4}(1 + \frac{l}{2f})^2 & \frac{1}{f^2}(1 - \frac{l^2}{2f^2})(1 + \frac{l}{2f}) & (1 - \frac{l^2}{2f^2})^2 \end{pmatrix} \tag{2.122}$$

While an argument of symmetry can be made as to why the parameters α_x, α_y must be zero to satisfy the conditions in eqs (2.114), I will simply show that by setting them to zero, we can easily solve for the remaining parameters. First, let us recall the identity from eq (2.93). With $\alpha_x = \alpha_y = 0$, this gives us

$$\gamma_x = 1/\beta_x, \tag{2.123}$$

$$\gamma_y = 1/\beta_y. \tag{2.124}$$

The conditions for periodicity therefore become

$$\beta_x = (1 - \frac{l^2}{2f^2})^2\beta_x + 4l^2(1 + \frac{l}{2f})^2(1/\beta_x), \tag{2.125}$$

$$\beta_y = (1 - \frac{l^2}{2f^2})^2\beta_y + 4l^2(1 - \frac{l}{2f})^2(1/\beta_y). \tag{2.126}$$

Above, I have only kept the results from the first row of each matrix product, as the other equations would be redundant. Solving for β_x, β_y, we get, provided $f > l/2$,

$$\beta_x = 2f\sqrt{\frac{f + \frac{l}{2}}{f - \frac{l}{2}}}, \tag{2.127}$$

$$\beta_y = 2f\sqrt{\frac{f - \frac{l}{2}}{f + \frac{l}{2}}}. \tag{2.128}$$

Thus, we arrive at the matched Twiss parameters for the alternating-gradient focusing cell discussed here. As we can see from our results for the matched parameters, $\beta_x > \beta_y$, which makes sense because the first quadrupole in our focusing cell will focus the beam in the x-direction and defocus the beam in the y-direction, causing β_y to grow and β_x to shrink. When the beam reaches the second quadrupole, the initial values of β_x and β_y will have been swapped. The second quadrupole will then have the opposite effect

of the first, returning β_x and β_y to their original values as the beam completes its pass through one full period of the focusing channel. The propagation of a matched beam through such a focusing channel is shown in Figure 2.8.

Collecting our results, we have for an alternating-gradient focusing cell composed of quadrupoles of focal length $f = (k^2 L)^{-1}$ separated by drift spaces of length l, whose first quadrupole focuses in x and defocuses in y, that the matched Twiss parameters (at a longitudinal position halfway through the first quad) are given by:

$$\begin{pmatrix} \beta_x \\ \alpha_x \\ \gamma_x \end{pmatrix} = \begin{pmatrix} 2f\delta \\ 0 \\ \frac{1}{2f\delta} \end{pmatrix}, \begin{pmatrix} \beta_y \\ \alpha_y \\ \gamma_y \end{pmatrix} = \begin{pmatrix} \frac{2f}{\delta} \\ 0 \\ \frac{\delta}{2f} \end{pmatrix}, \tag{2.129}$$

where I have defined $\delta = \sqrt{\frac{f+l/2}{f-l/2}}$.

Recalling from eq (2.90) that the values β_x, β_y describe the widths of Gaussian distributions of the transverse positions of the particles in the beam, we calculate the approximate cross-sectional beam size as

$$s = \sqrt{\sigma_x \sigma_y} = [\epsilon_x \epsilon_y \beta_x \beta_y]^{1/4}. \tag{2.130}$$

We can see in Figure 2.8 that when the beam is matched, the beam size is held constant at a value determined by the matched Twiss parameters β_x, β_y. As we discussed in Section (2.2.4), the output power of an FEL device is inversely related to the cross-sectional size of the beam as it passes through the undulator. By designing this focusing channel with small values of β_x, β_y, it is possible to squeeze the beam very tight, leading to the extremely intense amplification for which FELs are so valued.

2.3.7 The Mismatched Beam: Device Correlations

When the Twiss parameters of the beam are matched into the focusing channel of the undulator, the beam will experience minimal deviations from the target beam size. However, in reality, the Twiss parameters are never perfectly matched, and it is important to understand what effect this mismatch has on the FEL. Using the optical transport physics we have developed thus far, we can easily see what happens to the beam as it propagates through the periodic focusing channel of the undulator in the case where the Twiss parameters are mismatched. To understand these effects in quad-space, first, we construct an optical model of the matching quads and the undulator quads using some reasonable values from real operation. Assuming the beam was matched at these

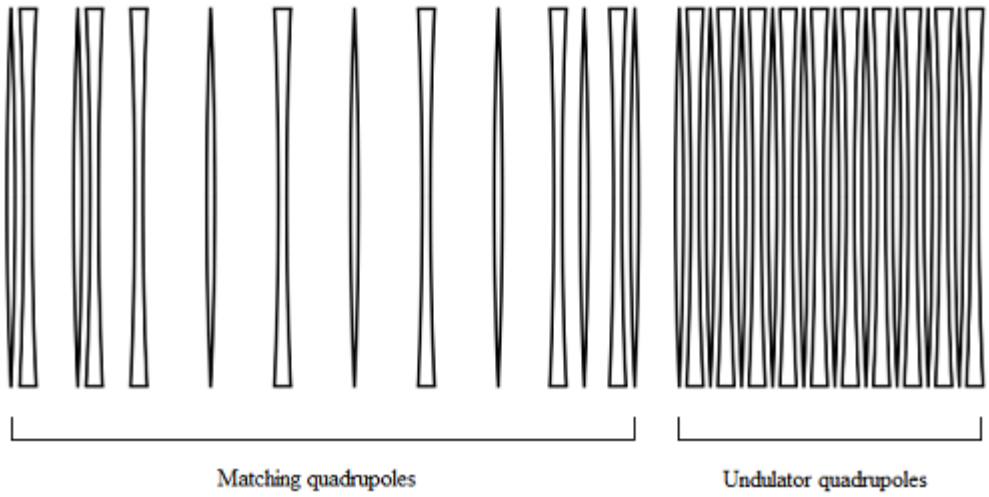


Figure 2.8: Schematic of our Optical Model

A schematic of our optical model of the matching quadrupoles and undulator quadrupoles in one transverse direction. The undulator quadrupoles form a periodic alternating-gradient focusing lattice.

settings, we then propagate the matched Twiss parameters at the undulator backwards upstream to the beginning of the optical lattice. (Note that since we calculated the matched Twiss parameters at a longitudinal position halfway through the first undulator quad, we must remember to include the first half of the this undulator quad in the back propagation.) Then, if we vary any of the matching quadrupoles and propagate the beam forward, the beam will no longer be matched when it reaches the undulator. The results of detuning (that is, perturbing from their matched settings) two adjacent quadrupoles as just described are included in figure 2.8. Whereas when the Twiss parameters are matched, the β parameters, which describe the transverse coss-sectional beam widths, oscillate uniformly, when the Twiss parameters are *mismatched*, the β parameters oscillate much more wildly. A consequence of these larger and less uniform oscillations is a larger beam size averaged throughout the undulator. According to our FEL model, this will lead to a smaller Pierce parameter, and decreased FEL output power.

Using the aforementioned average beam size as a measure of match quality, we can perform scans in quad-space to visualize the effects of simultaneously adjusting multiple devices. A 2-dimensional raster scan performed about the matched settings for 2 adjacent matching quadrupoles is shown on the left in Figure 2.9. There are two important results in this figure. The first, which could have been guessed, is that the further either quadrupole is perturbed from its matched setting, the more the average beam size increases. Second, the the results shown are not isotropic in quadrupole space. Indeed, there is a striking correlation between the two devices. That is, the beam size

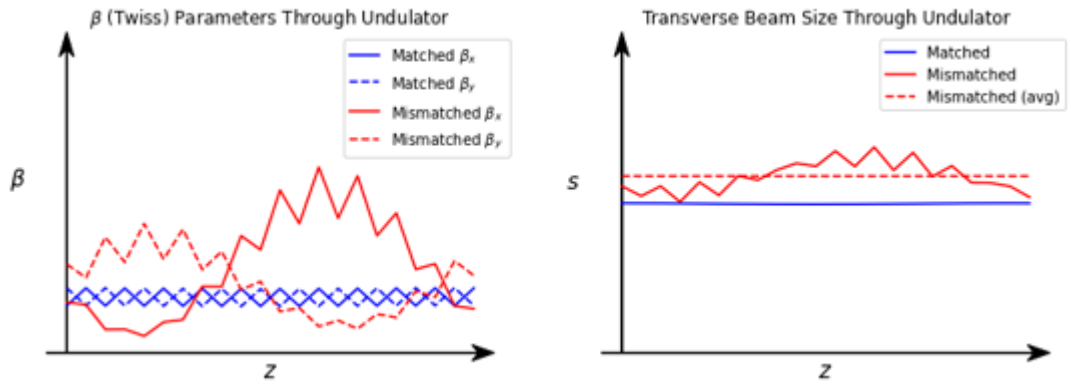
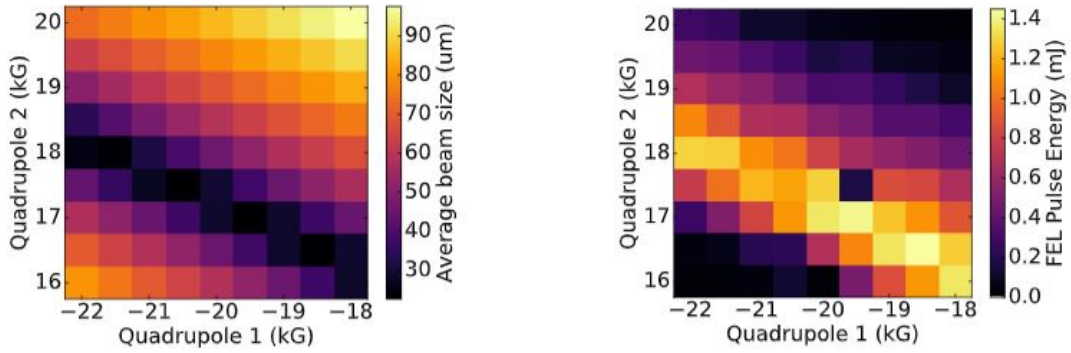


Figure 2.9: Optical Model: Beam Propagation

(Left) Evolution of the β (Twiss) parameters as a function of longitudinal position z in the undulator in a case that the Twiss parameters are matched (blue) and one in that they are mismatched (red) into the periodic focusing channel of undulator quadrupoles. **(Right)** Evolution of the beam size, s, as a function of longitudinal position z in the undulator for the same cases (blue and red) shown to the left. The dotted red line shows the average beam size in the case that the beam is mismatched.

grows much more rapidly when the quadrupoles are both detuned in the same direction than it does when they are detuned oppositely to each other. Note that this is exactly the behavior that the operators had observed on the machine when they were in charge of tuning it manually (recall the twisted rubber band). In fact, raster scans in quad space like the one above have been performed on the real machine to investigate the effect of adjusting the quadrupoles on the FEL output power, and the results confirm the presence of these anti-correlations between neighboring matching quads. An example of such a scan is presented on the right in Figure 2.9, below.

From an optimization standpoint, these results provide valuable insights. The



(a) Predicted Average Beam Size

(b) Measured FEL Pulse Energy

Figure 2.10: Device Correlations

(a) The average beam size through the undulator predicted by our optical model as a function of the field strengths of two adjacent matching quadrupoles for a given assumption of the matched settings. **(b)** Real observations of the measured FEL Pulse Energy at LCLS as a function of the same two quadrupoles and over the same range of values for which we examined the results of our optical beam size model in (a). The field strengths that produced the highest FEL output power in this scan were used as the assumed match in the aforementioned model. Figure appears in [7].

first is that the principal directions of the target function are not aligned with the device axes, which is information that we will want incorporated into our model. The second is that these principal directions appear to be well-predicted by the optical model of the beam size. As we will see in chapter 5, the ability to predict these correlations will be very useful when it comes to learning a model from data which is otherwise too sparse to illuminate the full structure of the underlying function. Lastly, cross-sections of the target function appear convex. While our Bayesian approach to optimization is

theoretically a global optimization strategy, it is convenient that our target function does not appear to contain local minima.

2.4 Summary

A Free-Electron Laser is a device that converts the kinetic energy of a beam of relativistic electrons to high-powered X-rays via synchrotron radiation. The synchrotron radiation is induced by an arrangement of magnets called an undulator. 1-d FEL theory predicts that the output power of the FEL X-rays will scale as $\sim \frac{1}{s^{2/3}}$ where s is the transverse cross-sectional beam size through the undulator. The beam size in the undulator is kept small with periodic quadrupole focusing magnets. The evolution of the beam envelope through the undulator depends on the Twiss parameters of the beam as it enters the periodic focusing lattice. Matching quadrupoles, immediately upstream of the periodic focusing lattice, are used to manipulate the Twiss parameters at the entrance of the undulator, thereby affecting the beam size through the undulator and subsequently the FEL output power. Beam conditions can change unpredictably and in ways that are hard to measure, making frequent tuning of the quadrupoles necessary to maintain a beam that is close to the match.

Tuning the matching quadrupoles (adjusting the strengths of their magnetic field gradients in order to improve the X-ray brightness) is essentially an optimization problem. In an effort to understand the target function in this optimization, measurements of the response of the FEL output power with respect to changes in the matching

quadrupole strengths have been performed. These measurements have repeatedly confirmed the presence of strong correlations between neighboring focusing elements, which are predicted by an optical beam size model. While a complete and reliable model of the FEL response to the quadrupoles has not been demonstrated, the ability to predict correlations between devices with a simple optical model will be valuable when it comes to numerical optimization. We must also keep in mind that the FEL output signal monitor (our target function for optimization) is noisy as an inevitable consequence of the non-intrusive mechanism employed by the gas detector to measure the X-ray intensity.

Chapter 3

Bayesian Optimization

When discussing numerical optimization, one of the first algorithms that comes to mind is likely gradient descent. Invented by Cauchy in 1847, it is an algorithm so fundamental that it predates the computer by nearly 100 years [29]. Its descendants are among the most commonly used algorithms in numerical optimization today. The principle of using the gradient, or an approximation thereof, and higher derivatives to model the local curvature of the target function has been demonstrated to be extremely powerful by modern optimization algorithms such as Momentum, AdaGrad, Adam, BFGS, etc. Of course, gradient-based algorithms perform best when they have access to the exact derivatives of the target function, which is a luxury we are not afforded when it comes to quadrupole tuning. In cases where the exact derivatives are not available, some optimization algorithms will calculate approximations using a finite difference approach. However, this adds numerous function evaluations per iteration, and with a target function like ours, becomes prohibitively time-consuming.

With gradients and higher derivatives unknown to us, we are left with a far shorter list of optimization algorithms from which to choose. Then, of course, there is the additional fact that our target function is noisy, which constrains our list of suitors further, still. One algorithm in particular that has proven to be fairly robust to noise and that does not require derivative information is the Nelder-Mead Simplex Method [30, 15]. Indeed, this algorithm has been used effectively to optimize the FEL at SLAC. In fact, the performance of this algorithm on our tuning problem is rather remarkable considering the ease of its implementation. The Nelder-Mead Simplex Method requires little training, and, as shown by previous studies, outperforms most human operators. With the promising results from the Nelder-Mead Simplex Method serving as proof that numerical optimization is a viable approach to quadrupole tuning, our group at LCLS was inspired to achieve further improvement by adopting a Bayesian approach to the numerical optimization.

Bayesian optimization is a computationally intensive (per iteration, in comparison to the other common optimization algorithms discussed previously) global search strategy that is useful in cases where the target function is extremely expensive to evaluate and where derivatives of the function are not readily available [31, 32]. Additionally, the ability to incorporate Bayesian treatment of uncertainty in the function outputs makes Bayesian optimization robust to signal noise. While Bayesian optimization is especially suited for functions with multidimensional inputs, its computationally intensive nature typically prohibits it scaling to functions of excessively high dimensions.

Bayesian optimization is typically used in situations where the evaluation of

the target function requires the execution of some time-intensive simulation or, as in our case, a physical process. For example, it has been used in search and rescue to find missing persons, where each "function evaluation" might amount to real-time observation of a specified location via satellite or drone [33]. Similarly, it has been used in the search for natural resources as a tool to help engineers decide where would be most profitable to more closely prospect [34, 35]. In such cases, the penalty for trying a bad guess for the solution is extremely high (potentially life or death, in the case of search and rescue), and so it is very important under such circumstances that every guess made by the search algorithm be as effective as possible. To accomplish this, Bayesian optimization algorithms attempt to utilize all the available information regarding the target function at every step of the optimization. In this regard, it is unique from local optimization algorithms, which only consider the curvature in the vicinity of the previously tested point to advance to a position that is marginally better than the last. Bayesian optimization algorithms will require more computation per iteration than other common algorithms, but the result is that the optimizer tends to converge to a solution with fewer evaluations of the target function, and one which is capable of circumnavigating local optima. Essentially, when performing Bayesian optimization, at each iteration, we think very hard about what point to sample next because thinking (computing) is much cheaper than evaluating the target function.

Here, I will describe the general approach to optimizing a function using Bayesian optimization. While the implementation of such algorithms can vary, the fundamental components are always present. First, a Bayesian regression tool is re-

quired to model the target function. Second, something called an acquisition function must be defined to quantify the value associated with each prospective point in the sample space. The acquisition function serves as the metric by which we are able to compare prospective points, allowing us to select the point that will provide us with the greatest return. The general form of a Bayesian optimization algorithm is outlined below.

Algorithm 1 Bayesian Optimization.

$i \leftarrow 1$

while $i \leq maxiter$ **do**

  if $i \neq 1$ **then**

    Numerically solve: $\vec{x^*} = argmin(\mathcal{A}(\vec{x}))$

  else

    $\vec{x^*} \leftarrow \vec{x}_{start}$

  end if

  Evaluate $y^* = y(\vec{x^*})$

  Update Bayesian regression model with new data point $(\vec{x^*}, y^*)$.

  $i \leftarrow i + 1$

end while

At each iteration, a Bayesian regression model yields an uncertain prediction for every possible point in the space. The acquisition function uses these predictions to assign a scalar value to each input. An optimization is performed on the acquisition function to find the input with greatest acquisition value. Indeed, at every iteration

of the Bayesian optimization algorithm, another optimization is performed to choose which point to sample from the target function next. Once the point has been sampled, the Bayesian regression model is updated and the cycle is repeated, in our case until the process is manually terminated. A few iterations from a simple 1-d example of Bayesian optimization are shown in figure 3.1.

It is important to note that during the optimization of the acquisition function, the target function need not be sampled, only the Bayesian regression model. The acquisition function is therefore cheap to evaluate, but derivative information may not be available, and, perhaps more troublingly, the function often possesses multiple local maxima, so care must be taken to ensure that the acquisition function is properly optimized (see Ch. 5.3 for more details on optimizing the acquisition function) [36].

In the following sections, I will provide an illustrative example of Bayesian regression and further discuss the role of the acquisition function.

3.1 Bayesian Regression vs Ordinary Least Squares (OLS)

At every iteration of the Bayesian optimization algorithm outlined above, a Bayesian regression model must be fit to the available data. Bayesian regression differs from Ordinary Least-Squares (OLS) regression, which is the approach with which the reader is probably most familiar, in that rather than simply providing point estimates as predictions, it also provides the associated uncertainty in those predictions. Additionally, Bayesian modeling requires definitions of certain *prior* beliefs about the target

function from which the data is presumed to have been drawn. As an illustrative example, let us compare 1-dimensional Bayesian linear regression to Ordinary Least-Squares linear regression.

Suppose we would like to fit a line to a collection of n data points $\{x_i, y_i\}_{i=1}^n$ drawn from a 1-d linear function with uniform Gaussian-distributed noise, ϵ, such that the i^{th} output y_i is related to the i^{th} input x_i by:

$$y_i = f(x_i) + \epsilon_i, \tag{3.1}$$

$$f(x_i) = mx_i + b, \tag{3.2}$$

$$\epsilon_i \sim N(0, \sigma_0^2). \tag{3.3}$$

In the Ordinary Least-Squares approach to linear regression, we attempt to find the parameters m^*, b^* that minimize the sum of the squared residuals (SSR), where the residuals are the differences between the observed values $\{y_i\}_{i=1}^n$ and the values predicted by the regression model $\{y_i^*\}_{i=1}^n$:

$$y_i^* = m^* x_i + b^*, \tag{3.4}$$

$$SSR = \sum_{i=1}^{n} [y_i - y_i^*]^2 \tag{3.5}$$

$$= \sum_{i=1}^{n} [y_i - (m^* x_i + b^*)]^2. \tag{3.6}$$

In the case of the linear regression model, we can see that the SSR is quadratic with respect to the parameters m^*, b^*, and as such the values of these parameters that minimize

the SSR can be calculated analytically:

$$m^* = \frac{n\sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}, \tag{3.7}$$

$$b^* = \frac{\sum_{i=1}^n y_i - m^* \sum_{i=1}^n x_i}{n} = \overline{y} - m^* \overline{x}. \tag{3.8}$$

An example of such a linear regression performed on some sample data is shown on the left in Figure 3.2. Notice that the prediction $y^*(x) = m^* x + b^*$ for a given input value x is a point estimate, and we have made no assumptions on the possible values of m^*, b^*.

Now suppose we would like to fit the same data using a Bayesian approach to linear regression. In this case, the objective is to solve for the parameters of the linear model that produce the greatest model *likelihood*. Indeed, the most fundamental difference between the Bayesian approach to regression and OLS is that in Bayesian regression, uncertainty is assumed in the model. For our example, we will assume a linear model with additive, zero-mean, Gaussian-distributed noise of variance σ^2:

$$y^*(x) = m^* x + b^* + \epsilon, \tag{3.9}$$

$$\epsilon \sim N(0, \sigma^2). \tag{3.10}$$

The definition of our model in combination with the data gives rise to the *likelihood* of the *data given the model*. If we denote the sets $\mathcal{X} = \{x_i\}_{i=1}^n, \mathcal{Y} = \{y_i\}_{i=1}^n$, we have:

$$p(\mathcal{Y}|\mathcal{X}, m^*, b^*) = \prod_{i=1}^n p(y_i|x_i, m^*, b^*) \tag{3.11}$$

$$= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{[y_i - (m^* x_i + b^*)]^2}{2\sigma^2}\} \tag{3.12}$$

$$= \frac{1}{[2\pi\sigma^2]^{\frac{n}{2}}} \exp\{-\frac{\sum_{i=1}^n [y_i - (m^* x_i + b^*)]^2}{2\sigma^2}\}. \tag{3.13}$$

We now want to use Baye's theorem to write the probability of the model parameters m^*, b^* in terms of the likelihood of the data given the model. The resulting probability distribution for the model parameters is called the *posterior* distribution. To do this, we first need to define *prior* probability distributions for the parameters m^*, b^*. The inclusion of prior probability distributions for the model parameters in Bayesian regression is another key difference from OLS. These prior distributions contain information about our prior, or *a priori*, knowledge regarding the possible values of the model parameters. The prior probability distributions may be inspired by domain knowledge, but often they are chosen simply for calculational convenience. In our case, we will assume that the parameters m^*, b^* are independently normally distributed with zero mean and respective variances σ_m^2, σ_b^2:

$$m^* \sim N(0, \sigma_m^2), \tag{3.14}$$

$$b^* \sim N(0, \sigma_b^2). \tag{3.15}$$

Recalling Baye's theorem $p(A|B)p(B) = p(B|A)p(A)$, we have

$$p(m^*, b^*|\mathcal{Y}, \mathcal{X})p(\mathcal{Y}|\mathcal{X})p(\mathcal{X}) = p(\mathcal{Y}|\mathcal{X}, m^*, b^*)p(m^*)p(b^*). \tag{3.16}$$

Dropping the normalization constants that do not depend on the parameters m^*, b^*, we get

$$p(m^*, b^*|\mathcal{Y}, \mathcal{X}) \propto p(\mathcal{Y}|\mathcal{X}, m^*, b^*)p(m^*)p(b^*) \tag{3.17}$$

$$\propto \exp\{-\frac{\sum_{i=1}^{n}[y_i - (m^* x_i + b^*)]^2}{2\sigma^2}\} \exp\{-\frac{m^{*2}}{2\sigma_m^2}\} \exp\{-\frac{b^{*2}}{2\sigma_b^2}\} \tag{3.18}$$

where we have replaced the equality with a proportionality. We can combine the exponentials and complete the square to arrive at a posterior distribution that is also normally distributed, allowing us to recover the constants of proportionality. (I haven't included the detailed results here because this example is intended to be illustrative only of the general process of Bayesian reasoning for the purpose of regression. We do not use Bayesian linear regression at all for tuning the quadrupoles. However, a nice feature of Bayesian linear regression with Gaussian priors and Gaussian noise is that these expressions can indeed be evaluated analytically.)

Finally, with the posterior distribution for the parameters m^*, b^*, we can form our Bayesian prediction. To do this, we take the weighted average of the results of all possible models by integrating over all possible values of the parameters m^*, b^*. That is, for some new input x, we have for the predicted output y^*

$$p(y^*|x, \mathcal{X}, \mathcal{Y}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} p(y^*|x, m^*, b^*) p(m^*, b^*|\mathcal{X}, \mathcal{Y}) dm^* db^*. \tag{3.19}$$

The above integral can be approximated using Markov-Chain Monte-Carlo sampling to produce an approximate predictive distribution for possible values of the output y^*. The prediction from the Bayesian linear regression model, evaluated using the same data as in the OLS regression example earlier, is included on the right in Figure 3.2.

3.2 Acquisition Functions

At each iteration, a Bayesian optimization algorithm, like all numerical optimization algorithms, must somehow decide where to take the next step in its search for

the optimum. Using the predictive posterior of an updated Bayesian regression model, we can construct what is called an *acquisition function*, which assigns a scalar value to each prospective point in the search space, typically by evaluating a weighted average over all possible values of the predicted output. (In the case of a maximization problem, we want the acquisition function to be greater in locations where the modeled function outputs are likely to be greater than the previously observed values.) With a single scalar value assigned to each prospective input, we can perform a well-defined comparison. To that end, a numerical optimizer is used to solve for the input vector $\vec{x_*}$ which maximizes the acquisition function $\mathcal{A}(\vec{x})$:

$$\vec{x_*} = \arg\max_{\vec{x}} \mathcal{A}(\vec{x}). \tag{3.20}$$

The solution $\vec{x_*}$ is then used as the next input for our target function. That is, we *acquire* the new point $\{\vec{x_*}, y_*\}$, update our Bayesian regression model to incorporate this new data, and then reiterate this process. (In practice, we construct our Bayesian prediction using a Gaussian process regression model that maps points in multi-dimensional quadrupole-space to a scalar measure of the FEL output power. We seek the location in the input space that produces the largest possible FEL output power.)

Our flexibility when it comes to defining the acquisition function allows us to tailor the behavior of our optimization algorithm. For example, depending on the selection of the acquisition function, the optimizer may either be compelled to take exploratory steps, to places where the model predicts the possibility of vast improvements in the target signal (at the risk of moving away from the optimum), or to take very

conservative steps, to places where the model predicts that the target signal is almost certain to improve (but at the cost of convergence rate) [37]. This spectrum of behavior is commonly referred to as *exploration vs exploitation*. In the remainder of this chapter, I will introduce some of the most common acquisition functions, including the ones used in our studies.

3.2.1 Probability of Improvement

Given a predictive Bayesian posterior distribution, $p(y|\vec{x})$, for our target function, $f(\vec{x})$, the most introductory example of an acquisition function that we can define is the *probability of improvement*. As the name suggests, this function uses the predictive posterior distribution to calculate, as a function of the input vector $\vec{x}$, the probability (according to our current model) that the value returned by the target function is better than the best value previously seen by the optimization algorithm. That is, in the case of a maximization problem, we define the probability of improvement, $\mathcal{PI}(\vec{x})$, as

$$\mathcal{PI}(\vec{x}) = \int_{y_{best}}^{\infty} p(y|\vec{x})dy. \tag{3.21}$$

Note the lower bound on the integral. Only predicted values that are above the best seen value contribute to the result. The integral in eq (3.21) is the complementary cumulative distribution, sometimes called the Q-function, of the posterior distribution $p(y|\vec{x})$ evaluated at y_{best}. That is, for the probability of improvement $\mathcal{A}(\vec{x})$, we have

$$\mathcal{PI}(\vec{x}) = 1 - \Phi(y_{best}|\vec{x}), \tag{3.22}$$

where Φ is the cumulative distribution function of $p(y|\vec{x})$:

$$\Phi(y_{best}|\vec{x}) = \int_{-\infty}^{y_{best}} p(y|\vec{x})dy. \tag{3.23}$$

If $p(y|\vec{x})$ is normally distributed, as was the case in our Bayesian linear regression example, and as is the case with the Gaussian process regression model we will eventually use, these integrals unfortunately cannot be evaluated analytically. However, cumulative density functions for common distributions like this have been extensively studied, and numerical approximations for the function Φ in the case of the normal distribution can easily be looked up. A 1-d example of the probability of improvement, given a Bayesian regression model (GP) and some data, is plotted in Figure 3.3.

3.2.2 Expected Improvement

While the probability of improvement provides a nice introductory example to defining a Bayesian metric by which to choose the next point to acquire in our Bayesian optimization algorithm, it may lead to optimizer behavior that is undesirably conservative in its exploration. For example, consider the case where the model predicts a 95% probability of improvement at position $\vec{x_1}$ and a 90% improvement at $\vec{x_2}$, but where the uncertainty in the prediction at $\vec{x_2}$ is much higher than at $\vec{x_1}$. In this case, while the model predicts that the probability of improving by moving to position $\vec{x_2}$ is slightly lower, it also predicts that there is a *possibility* of seeing a much larger improvement as compared to $\vec{x_1}$. If the goal of our Bayesian optimization algorithm is to converge to a global maximum in the fewest possible steps, it may be advantageous to

assign weights to the magnitudes of these possible improvements that our model predicts for each input. To do this, we define the *expected improvement*, $\mathcal{EI}(\vec{x})$, as [20, 38]

$$\mathcal{EI}(\vec{x}) = \int_{y_{best}}^{\infty} (y - y_{best}) p(y|\vec{x}) dy, \tag{3.24}$$

where once again $p(y|\vec{x})$ is the predictive posterior distribution from our Bayesian regression model at the given iteration, and y_{best} is the highest value of the target function observed by the optimizer thus far.

When the posterior $p(y|\vec{x})$ is normally distributed, while we cannot evaluate the exact result of this entire integral analytically, we can evaluate half of it, and we can write the result of the other half in terms of the cumulative density function of the standard normal distribution, Φ. That is, if

$$p(y|\vec{x}) \sim N(\mu(\vec{x}), \sigma^2(\vec{x})), \tag{3.25}$$

then

$$\mathcal{EI}(\vec{x}) = (\mu(\vec{x}) - y_{best})\Phi(Z) + \sigma(\vec{x})\phi(Z), \tag{3.26}$$

where ϕ is the standard normal distribution, and $Z = \frac{\mu(\vec{x}) - y_{best}}{\sigma(\vec{x})}$ is the *Z-score*, or *standard score*, of the difference between the predictive mean, $\mu(\vec{x})$, and the best observed target function value, y_{best}, defined in terms of the standard deviation $\sigma(\vec{x})$ of the predictive distribution. We can see from the expression above for the expected improvement that the first term in the sum will assign greater value to positions $\vec{x}$ where the predictive mean $\mu(\vec{x})$ is much higher than the previously best seen value y_{best}, and where the uncertainty $\sigma(\vec{x})$ is low. In contrast, the second term in the sum will assign

greater value to positions $\vec{x}$ where the model is more uncertain. Combined, these two terms simultaneously weigh the value of exploitation and exploration. A 1-dimensional example of the expected improvement, given the same Bayesian regression model and data as in the earlier example for probability of improvement, is included in Figure 3.3. The expected improvement acquisition function, made attractive by its analytical transparency and reasonable balance between exploration and exploitation, was used in many of the optimization studies we performed, which we will discuss more in chapter 5.

3.2.3 Other Acquisition Functions

As we have seen in the previous two examples, the definition of the acquisition function will affect the behavior of our optimization algorithm. Further modifications to the exploration/exploitation behavior could be made by either adding tune-able parameters to the acquisition functions already discussed, or by defining entirely unique acquisition functions. For example, another common acquisition function, called *Upper Confidence Bound*, is designed such that the optimizer attempts to minimize a quantity called the *regret* [39]. Tuning the parameters of such acquisition functions poses an additional problem to the task of tailoring an optimization algorithm to a particular application. Because Bayesian optimization is generally reserved for cases where the target function is extremely expensive to evaluate, tuning these parameters will likely need to be achieved through Monte-Carlo testing on some surrogate model of the target function. In our studies, we defaulted to using the expected improvement for its effec-

tiveness, interpretability, and ease of implementation. Extensive experimentation with different acquisition functions is beyond the scope of this work.

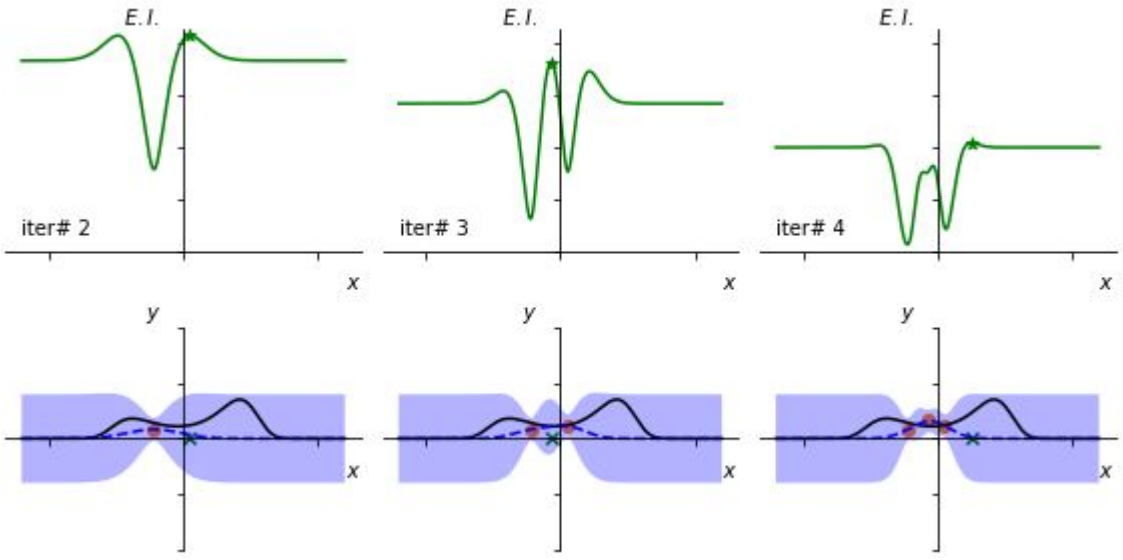


Figure 3.1: Iterations of an Example Bayesian Optimization Algorithm

(Top) The Expected Improvement acquisition function evaluated at the three consecutive states of our Gaussian process regression model shown below. The green star represents the acquisition function maximum. **(Bottom)** Three consecutive iterations of a Bayesian optimization algorithm performed on a target function shown by the solid black curve. Gaussian process regression (zero-mean) was used to form our Bayesian prediction whose maximum-likelihood estimator is shown by the dashed blue curve and 95% confidence interval shown by the blue shaded region. The set of (noisy) observations at each iteration is shown in orange. The x-position of the next point to acquire is shown by the green x-mark.

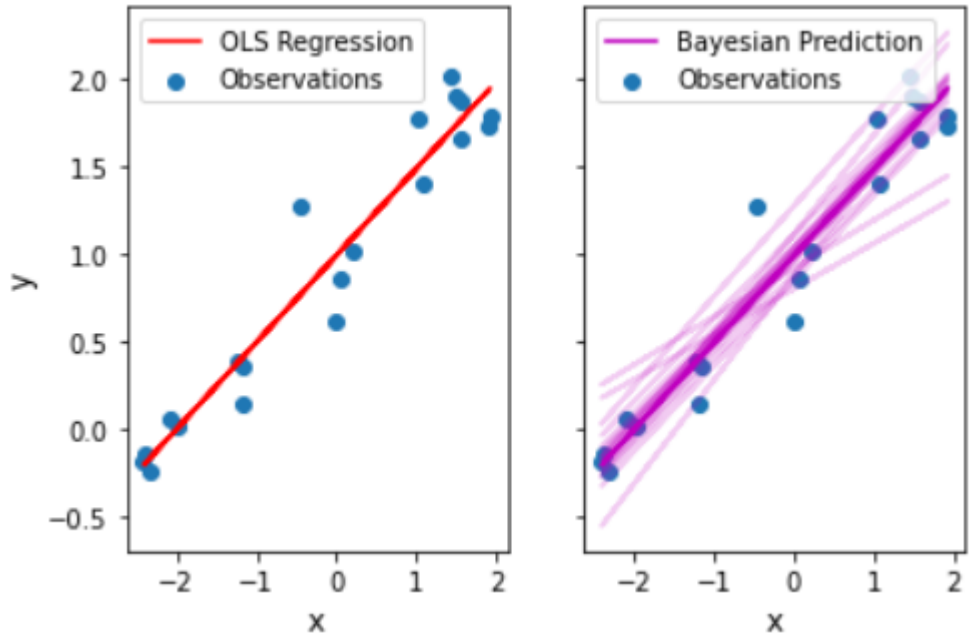


Figure 3.2: OLS vs Bayesian Linear Regression

(Left) Ordinary Least-Squares (OLS) linear regression (shown in red) on some noisy observations shown in blue. **(Right)** Samples from a Bayesian posterior distribution over linear models (magenta) on the same set of data (blue) as used in the regression to the left.

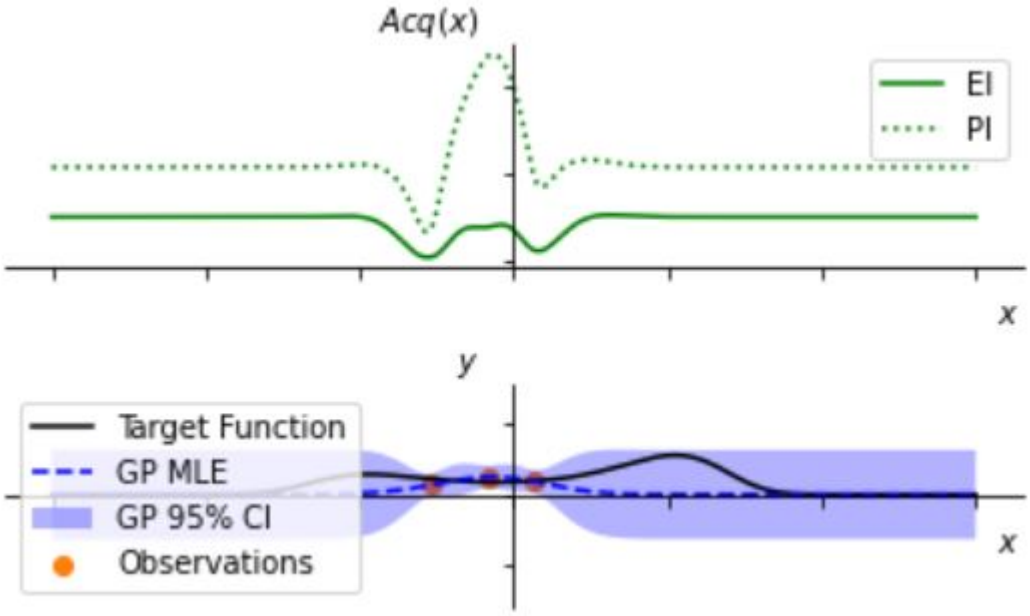


Figure 3.3: Common Acquisition Functions

(Top) The Expected Improvement (EI) and Probability of Improvement (PI) acquisition functions evaluated using the Bayesian regression model shown below. **(Bottom)** Gaussian process regression (blue) performed on noisy observations (orange) sampled from an underlying target function (black).

Chapter 4

Gaussian Processes

A *Gaussian process* is a *stochastic process*, or collection of random variables, that has the property that any finite subset of its variables is described by a joint Gaussian distribution [19]. Gaussian processes are commonly used as regression models in Bayesian optimization. Recall how in the case of Bayesian linear regression discussed in the preceding chapter, our predictions arose from some assumptions about the functional dependence of the observed outputs on the inputs, and the prior probability distributions for the parameters of that dependence. That is, the prior distribution of possible functions was defined in *parameter space*. In Gaussian process regression, we arrive at our predictions by assuming a prior probability distribution directly over the set of all possible function observations in what we call *function space*. (In this paper, we use an implementation of the GP using Python's Scikit-learn package [40].)

4.1 Gaussian Process Regression

To elaborate, suppose we have n observations of some function $f(\vec{x})$ evaluated at the positions $\{x_i\}_{i=1}^n$ and we want to make a prediction for the function output evaluated at some point x_*, where x_* could be anywhere in the input space. In order to model the function $f(\vec{x})$ using a Gaussian process, we define the vector $\vec{f}$ such that

$$\vec{f} = (f(\vec{x_1}), f(\vec{x_2}), ..., f(\vec{x_n}), f(\vec{x_*}))^\top, \tag{4.1}$$

and assume that this vector represents the result of a single sample from a multivariate Gaussian distribution (or joint normal distribution) of dimension $n + 1$:

$$\vec{f} \sim N(\vec{\mu}, \mathbf{\Sigma}). \tag{4.2}$$

Here $\vec{\mu} = (\mu(\vec{x_1}), \mu(\vec{x_2}), ..., \mu(\vec{x_n}), \mu(\vec{x_*}))^\top$ is the center, or *mean*, of the distribution, determined by the function $\mu(\vec{x})$, and $\mathbf{\Sigma}$ is the distribution's *covariance matrix*, which defines the distribution's shape. – In this context, the shape of the distribution will describe how similar our model believes the function outputs to be to each other, with the element $\mathbf{\Sigma}_{ij}$ measuring the similarity between the function outputs f_i and f_j. – To complete our assumptions, we need only to specify the function $\mu(\vec{x})$ as well as the elements of the covariance matrix $\mathbf{\Sigma}$. (Prescriptions for these specifications will be provided shortly.) We then simply condition the probability of the output vector $\vec{f}$ on its observed components $\{f_i\}_{i=1}^n$ to obtain the *conditional probability* of the unobserved component $f_{n+1} = f(\vec{x_*})$. The result of conditioning a multivariate Gaussian in this

fashion can readily be looked up, yielding our prediction [19]:

$$f(\vec{x_*})|\{f(\vec{x_i})\}_{i=1}^n = f_{n+1}|\{f_i\}_{i=1}^n \tag{4.3}$$

$$\sim N(\overline{\mu}, \overline{\sigma^2}). \tag{4.4}$$

Here, $\overline{\mu}$ and $\overline{\sigma^2}$ are the scalar values which describe the mean and variance, respectively, of the predictive 1-dimensional Gaussian distribution, given by

$$\overline{\mu} = \mu(\vec{x_*}) + \mathbf{\Sigma}_{1\times n}\mathbf{\Sigma}_{n\times n}^{-1}(\vec{f_n} - \vec{\mu_n}), \tag{4.5}$$

$$\overline{\sigma^2} = \mathbf{\Sigma}_{1\times 1} - \mathbf{\Sigma}_{1\times n}\mathbf{\Sigma}_{n\times n}^{-1}\mathbf{\Sigma}_{n\times 1}, \tag{4.6}$$

where the vectors $\vec{f_n}$ and $\vec{\mu_n}$ are the n-dimensional sub-vectors of the $(n+1)$-dimensional vectors $\vec{f}$ and $\vec{\mu}$, respectively, containing their first n elements, and $\mathbf{\Sigma}_{n\times n}$, $\mathbf{\Sigma}_{1\times n}$, $\mathbf{\Sigma}_{n\times 1}$, and $\mathbf{\Sigma}_{1\times 1}$ are the $n \times n$, $1 \times n$, $n \times 1$, and 1×1, resepctively, sub-matrices of the $(n+1) \times (n+1)$ covariance matrix $\mathbf{\Sigma}$ defined as

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{n\times n} & \mathbf{\Sigma}_{1\times n} \\ \mathbf{\Sigma}_{n\times 1} & \mathbf{\Sigma}_{1\times 1} \end{pmatrix}. \tag{4.7}$$

To fully define our prediction, as briefly mentioned above, we need only to specify the function $\mu(\vec{x})$, which we call the *prior mean*, and our matrix $\mathbf{\Sigma}$, which we call the *model covariance*. The prior mean function can simply be a prior regression for our function $f(x)$, if one is available, but often a zero-mean prior, $\mu(\vec{x}) \equiv 0$, is assumed. To express the model covariance, we select a valid (positive semi-definite) *covariance function*, or *kernel*, $K(\vec{x}, \vec{x'})$ that we will use to populate the elements of $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma}_{ij} = K(\vec{x_i}, \vec{x_j}). \tag{4.8}$$

Notice that in doing so, we are defining the similarity $\mathbf{\Sigma}_{ij}$ between the function outputs f_i and $f_j$in terms of the corresponding inputs $\vec{x_i}$, $\vec{x_j}$. The selection of our kernel, which we will discuss momentarily, is instrumental in determining the fitting behavior of our Gaussian process regression model. First, however, let us make one modification to the treatment we have just described.

Recall that one of our foundational assumptions in our above model was that we were given a set of some observed function values $\{f(\vec{x_i})\}_{i=1}^n$. If, instead, we assume that we are given some *noisy* observations of the function, $\{y_i\}_{i=1}^n$, such that $y_i = f(\vec{x_i}) + \epsilon$, where the noise ϵ is independent of the inputs and distributed as $\epsilon \sim N(0, \sigma^2)$, then we have for our expressions above,

$$\mathbf{\Sigma}_{n\times n} \to \mathbf{\Sigma}'_{n\times n}, \tag{4.9}$$

$$\vec{f_n} \to \vec{y_n}, \tag{4.10}$$

where

$$\mathbf{\Sigma}'_{n\times n} = \mathbf{\Sigma}_{n\times n} + \sigma_n^2 I, \tag{4.11}$$

$$\vec{y_n} = (\, y_1 \; y_2 \; \cdots \; y_n \,)^\top, \tag{4.12}$$

in which I is the n-dimensional identity matrix. That is, under our modified assumptions, our predictive equations given some set of noisy observations become [19]:

$$f(\vec{x_*})|\{y_i\}_{i=1}^n \sim N(\overline{\mu'}, \overline{\sigma'^2}), \tag{4.13}$$

$$\overline{\mu'} = \mu(\vec{x_*}) + \mathbf{\Sigma}_{1\times n}\mathbf{\Sigma}'^{-1}_{n\times n}(\vec{y_n} - \vec{\mu_n}), \tag{4.14}$$

$$\overline{\sigma'^2} = \mathbf{\Sigma}_{1\times 1} - \mathbf{\Sigma}_{1\times n}\mathbf{\Sigma}'^{-1}_{n\times n}\mathbf{\Sigma}_{n\times 1}. \tag{4.15}$$

(Note in both the noise-free and the noisy case, our predictive equations involve the inversion of an $n \times n$ matrix, where n is the number of function observations on which we are conditioning our model. Because the computational complexity of such matrix inversion goes as $\sim \mathcal{O}(n^3)$, the Gaussian process regression model will not typically scale well to "big data" applications. Fortunately, in Bayesian optimization, this tends not to be a problem, because the number of function observations is intentionally minimized.)

4.2 The Covariance Function and Model Hyperparameters

The Gaussian process regression model is a type of function approximator that generates a prediction for a novel test point by taking a weighted average of the training observations, where the weights in that average are determined by the respective distance of each training example to the test point as measured by what we call a *covariance function*, or *kernel*. More concretely, examining eq (4.14), we can see that our predictive mean is a linear combination of the training data, with weights determined by the submatrices of $\mathbf{\Sigma}$. We therefore select our kernel function $K(\vec{x}, \vec{x'})$ to populate the elements of the matrix $\mathbf{\Sigma}$ in a way that describes our prior beliefs about the correlation between the possible function outputs in terms of their inputs. Selection of the kernel function therefore has a major effect on the fitting behavior of our Gaussian process, determining qualities such as the smoothness and characteristic *length scales* (a term which we will define momentarily) of the functions that compose our model. To

understand the role of the kernel function more concretely, we introduce the *Squared-Exponential* kernel, which is the kernel function we used in our studies, and discuss its properties.

4.2.1 The Squared-Exponential Kernel

The *Squared-Exponential* (or SE) kernel is one of the most popular choices of covariance functions in kernel-based learning [19]. While it has a number of properties that make it an attractive choice, our primary reasons for its adoption were its empirical efficacy in modeling smooth functions and its analytical interpretability.

For a function of 1-dimensional input, the SE kernel defines the similarity between two function values $f(x)$, $f(x')$ in terms of the euclidean distance between their inputs according to

$$K(x, x') = \sigma^2 \exp\{-\frac{1}{2}\frac{|x - x'|^2}{l^2}\}, \tag{4.16}$$

where σ^2 and l are positively-valued scalars, called *hyperparameters*, that must be learned in some way. Selection of these hyperparameters will affect the fitting behavior of our Gaussian process. The hyperparameter σ^2 is effectively the *prior variance*, while the hyperparameter l can be thought of as the characteristic *length scale* – the distance in the input space over which we may begin to see a significant variation in our modeled output. We can clearly see by this definition that the assigned similarity between two function values will be maximized when their inputs are very close together, meaning that observations nearer to the point at which we would like to make a prediction will

receive a greater weight.

The Squared-Exponential kernel can be generalized to functions of multi-dimensional inputs. In this case, we have

$$K(\vec{x}, \vec{x'}) = \sigma^2 \exp\{-\frac{1}{2}(\vec{x} - \vec{x'})^\top \mathbf{\Sigma}_{SE}(\vec{x} - \vec{x'})\}, \tag{4.17}$$

where σ^2 is once again a scalar value describing the prior variance, and instead of a scalar length scale hyperparameter, we have the covariance matrix $\mathbf{\Sigma}_{SE}$, which we will call the *matrix of covariance hyperparameters*, or simply the *covariance hyperparameters*. (Note that we must be careful not to confuse this matrix of hyperparameters with the covariance matrix discussed in section (4.1).) While the matrix $\mathbf{\Sigma}_{SE}$ is not quite as easily digested as the scalar length scale hyperparameter from the 1-dimensional case, rest assured that it still contains information about the characteristic length scales (now plural) of our Gaussian process regression model.

In addition to the hyperparameters σ^2, and $\mathbf{\Sigma}_{SE}$, we will also consider as a hyperparameter the assumed noise variance σ_n^2 that we introduced at the end of section (4.1). We will denote the set of hyperparameters for a Gaussian process as $\mathbf{\Theta}$. **Together with our choices of the prior mean function $\mu(\vec{x})$, and the functional form of our Squared-Exponential kernel given by eq (4.17), the choices for the values of these hyperparameters $\mathbf{\Theta} = \{\sigma^2, \mathbf{\Sigma}_{SE}, \sigma_n^2\}$ fully define our prior model, allowing us to evaluate the predictive equations (4.14) and (4.15), given some data.** When sufficient training data is available, the hyperparameter values are typically optimized to produce a model with the highest likelihood via numerical

optimization (in Scikit-learn, this is done with a built-in function) [19, 40]. We will talk more about how we selected the hyperparameter values for our problem in the next chapter. In the remainder of this chapter, I will attempt to illustrate the general behavior of our Gaussian process regression model, as well as present some useful ways of formulating the matrix $\mathbf{\Sigma}_{SE}$.

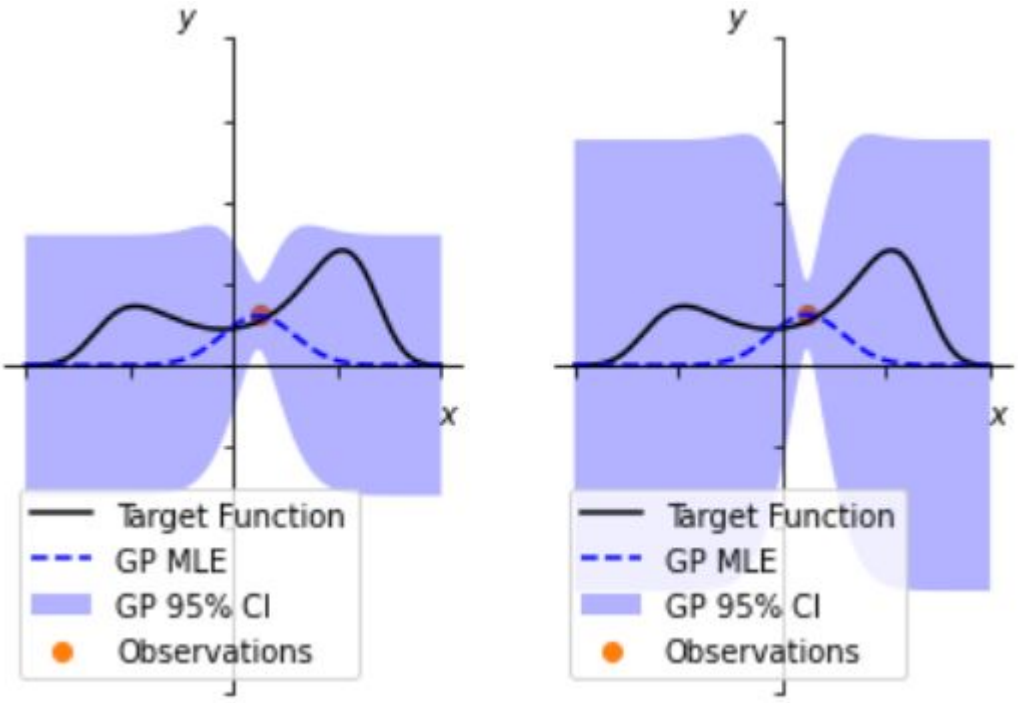


Figure 4.1: Variance Hyperparameter

(Left) Gaussian process regression (blue) performed on a single noisy sample (orange) from an underlying target function (black). **(Right)** The same Gaussian process regression model (blue) conditioned on the same sample (orange) as shown on the left, but with kernel variance hyperparameter increased by a factor of ~ 2.

4.2.2 Prior Mean and Variance

The prior mean $\mu(\vec{x})$, and variance σ^2, both of which are scalars, simply designate the mean and the variance, respectively, of the predictive distribution for our function output $f(\vec{x})$ prior to any observations. After conditioning the model on some observations, the resulting predictive distributions will approach the prior distributions for the function outputs $f(\vec{x})$ at positions where the input $\vec{x}$ is very far (many multiples of the length scale, measured in the input space) away from the inputs of the observations. See illustration in Figure 4.1.

4.2.3 Noise

While the variance hyperparameter defines the uncertainty in our predictions very far from our observations, the noise hyperparameter σ_n^2 will describe the persistent uncertainty in our model in places very *near* to our observations. This effect is illustrated in Figure 4.2. We used two different values for the noise hyperparameter to construct two otherwise identical Gaussian process regression models that were then given identical sets of noisy observations. When the models are given an observation at a particular point, they remains uncertain in their predictions at that position. If the noise parameter is overestimated, the GP will be overly reluctant to infer structure from the data, whereas when the noise parameter is more appropriately selected, the GP does a better job of fitting the curvature.

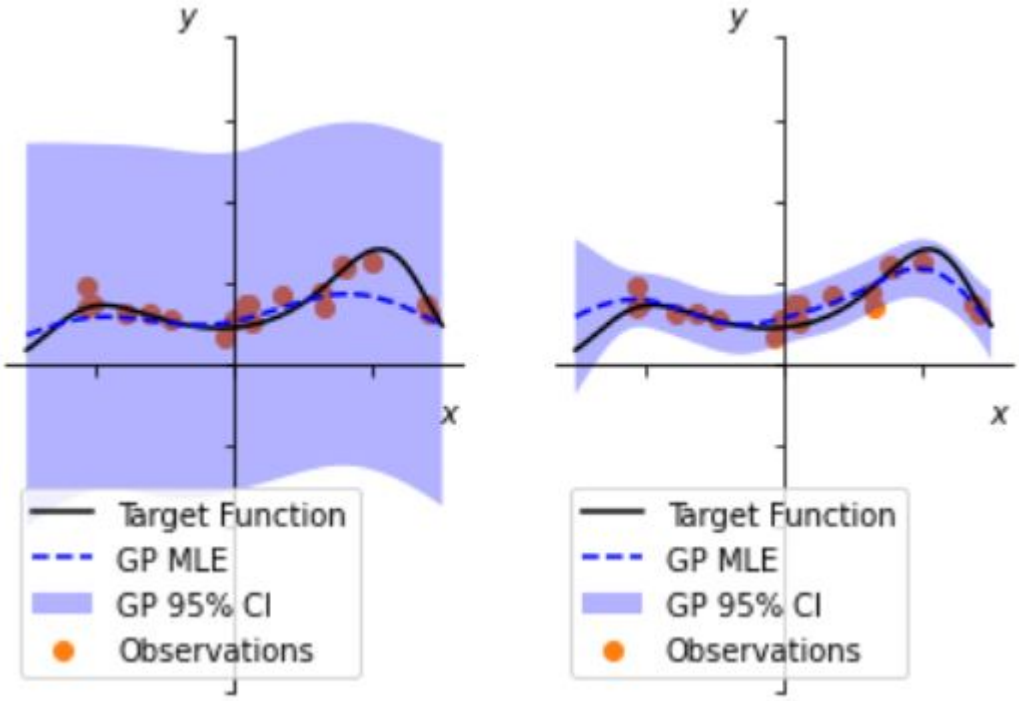


Figure 4.2: Noise Hyperparameter

(Left) Gaussian process regression (blue) performed on some noisy samples (orange) from an underlying target function (black). **(Right)** The same Gaussian process regression model (blue) conditioned on the same samples (orange) as shown on the left, but with a much smaller and more appropriate noise hyperparameter (reduced by a factor > 50).

4.2.4 Length scales in 1-Dimension

The effect of the length scale hyperparameter in 1-dimension is illustrated in Figure 4.3. We used two different values for the length scale hyperparameter to construct two otherwise identical Gaussian process regression models that were then given identical sets of observations drawn from a smooth underlying function with additive Gaussian noise. The smaller length scale used in the GP whose results are plotted on

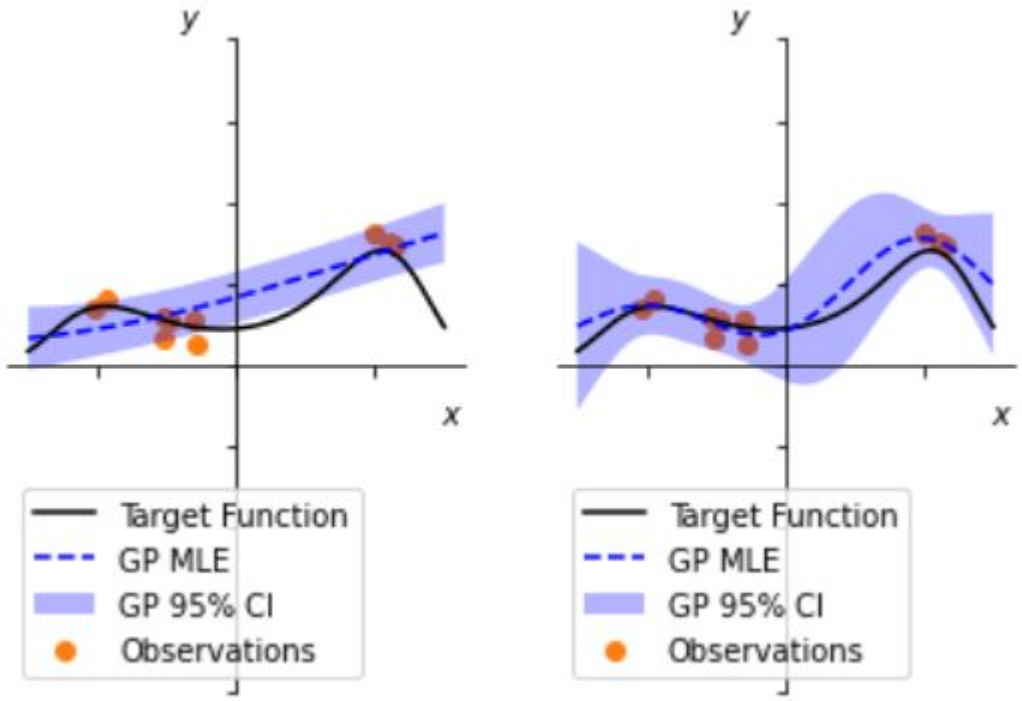


Figure 4.3: Length scale Hyperparameter: 1-d

(Left) Gaussian process regression (blue) performed on some noisy samples (orange) from an underlying target function (black). **(Right)** The same Gaussian process regression model (blue) conditioned on the same samples (orange) as shown on the left, but with a shorter and more appropriate length scale hyperparameter (reduced by a factor of ~ 6).

the right in Figure 4.3 causes the model to (accurately, in this case) predict greater curvature in the space between the samples.

4.2.5 Length scales in Higher Dimensions: Principal Directions (Correlations)

In the case where we are attempting to model a function of multi-dimensional inputs, it is possible that the target function may exhibit output values that vary more rapidly with respect to changes in the inputs along certain directions in the input space than along others. That is, different cross sections of the function may be best described by multiple different characteristic length scales. Conveniently, if we are aware of such behavior, we can incorporate this knowledge into our covariance matrix $\mathbf{\Sigma}_{SE}$ of our squared-exponential kernel. For example, if it is simply the case that we know that certain components of our input vectors affect our function output more than others, this can easily be treated by constructing the covariance matrix $\mathbf{\Sigma}_{SE}$ as the diagonal matrix

$$\mathbf{\Sigma}_{SE} = \begin{pmatrix} 1/l_1^2 & & \\ & \ddots & \\ & & 1/l_n^2 \end{pmatrix}, \tag{4.18}$$

where n is the dimension of the input vectors to our function, and l_i is the characteristic length scale of the target function along the direction given by the i^{th} Cartesian unit vector. However, generally speaking, the directions of maximum curvature for a target function may not be aligned with the Cartesian coordinate axes. To handle the more general case, where the input components are correlated, we must include off-diagonal components in our covariance matrix. Specifically, if we are aware that our n-dimensional function decomposes well into a set of some *principal directions* given

by the n-dimensional unit vectors $\{\hat{v_i}\}_{i=1}^n$, along which the function varies with the respective (positive) characteristic length scales $\{l_i\}_{i=1}^n$, we can construct the matrix $\mathbf{\Sigma}_{SE}$ according to the matrix product [19]

$$\mathbf{\Sigma}_{SE} = \mathbf{\Lambda}\mathbf{\Lambda}^\top, \tag{4.19}$$

where the i^{th} column of the matrix $\mathbf{\Lambda}$ is given by the vector $\vec{v_i} = \frac{1}{l_i}\hat{v_i}$. (Note that this is equivalent to performing a change-of-basis on our inputs that diagonalizes the matrix $\mathbf{\Sigma}_{SE}$.)

As we will discuss in the next chapter, a case that is of particular interest to us is one in which we are aware of the principal directions but not the associated length scales. Under such circumstances, to select our hyperparameter values, we simply fix the vectors $\{\hat{v_i}\}_{i=1}^n$ in the above formulation of the matrix $\mathbf{\Sigma}_{SE}$, and only optimize over the possible length scale hyperparameter values $\{l_i\}_{i=1}^n$ to produce maximum likelihood estimates of the length scales. By taking this approach we are effectively adding prior information to our model. In cases where training data is sparse, this may lead to a model with superior predictive accuracy. Alternatively, we can think of this procedure as reducing the number of hyperparameters that must be learned, which is desirable when data is sparse, because having too many model parameters can lead to over-fitting.

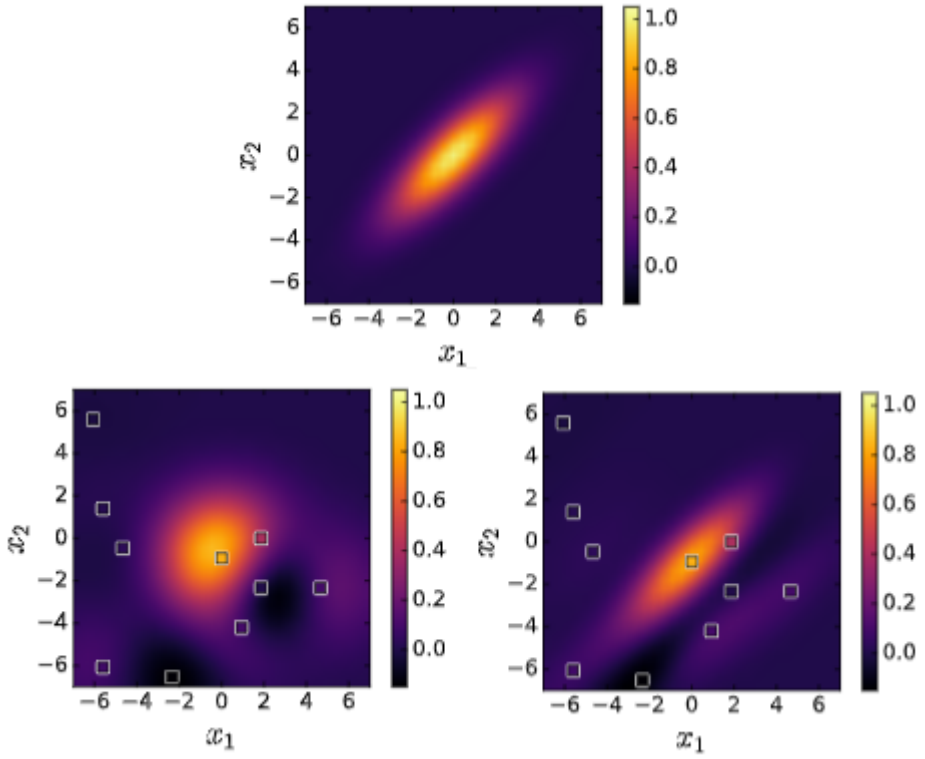


Figure 4.4: Length scale Hyperparameters: 2-d (Correlations)

(Top) A heatmap showing a 2-d, correlated Gaussian target function as a function of the inputs x_1 and x_2. **(Left)** A heatmap showing the maximum likelihood estimator of a Gaussian process regression model with diagonal (uncorrelated) matrix of length scale hyperparameters conditioned on a set of noisy observations of the target function shown above. The coordinates of the observations are highlighted by white boxes. **(Right)** A heatmap showing the maximum likelihood estimator of a Gaussian process regression model with properly optimized general (correlated) matrix of length scale hyperparameters conditioned on the same set of observations (white boxes) as used in the regression to the left. Figure appears in [7].

Chapter 5

Optimizing the Optimizer

In this chapter I will discuss the steps we took to tailor our Bayesian optimization algorithm to the problem of tuning the quadrupoles at LCLS. Before I jump into the details of training our regression model for use on the FEL, I will present an investigation into the effects of a suboptimal regression model on the performance of our Bayesian optimization algorithm in a toy environment. As we will see, the algorithm is fairly robust, but as the dimensionality of the search space increases, the inclusion of accurate correlation hyperparameters (in the Squared Exponential kernel) becomes increasingly beneficial.

After we have established expectations regarding the performance of our algorithm using our results from the toy environment, I will discuss our approach to estimating the optimal hyperparameters, in particular the length scale hyperparameter matrix $\mathbf{\Sigma}_{SE}$, for our real target function as accurately as possible. I will then address the steps we took to effectively maximize the acquisition function, which can be rid-

dled with troublesome local maxima, at every iteration of our algorithm. Finally, I will present the results of deploying our Bayesian optimization algorithm to the problem of tuning the matching quadrupoles at LCLS.

5.1 Effects of Correlation Hyperparameters on Optimizer Performance

In the previous chapter, we saw how selection of model hyperparameters such as length scales and noise made drastic differences in the GP's ability to accurately fit observed data. While in cases with low-dimensional input space it is easy to visually assess the quality of the fit, and quality of fit can always be measured by the model likelihood, ultimately, the quantity we truly care about is the amount of time it takes our Bayesian Optimization algorithm to find a sufficiently high signal so that researchers are able to use the FEL for their experiments. It stands to reason that a model with higher likelihood, which produces a better fit to observed data, will require fewer function evaluations (and therefore less time) to converge on a solution. Nevertheless, we should still like to observe these effects empirically in order to quantify these benefits.

We are particularly interested to see how the inclusion of appropriate correlation hyperparameters, or off-diagonal terms in the matrix $\mathbf{\Sigma}_{SE}$, affects the performance of our algorithm. To fully explicate why we expect this to be so important, let us once again consider the measured FEL response shown in Figure 2.9. To cast these results in GP terminology, the target function exhibits vastly different characteristic length

scales. Indeed, the optimizer mostly needs to closely search along the direction with the largest rate of change, because its position along the other principal direction affects the output much less. As we saw in Figure 4.4, when the GP model has appropriate length scale correlation hyperparameters, the prediction for points along the principal direction with the largest length scale varies, as designed, less rapidly from the nearby observed function values. Additionally, the uncertainty in our predictions near some data will increase less rapidly along this principal direction. Recalling the definitions and behaviors of the acquisition functions discussed in Chapter 3, we can convince ourselves that this will result in the optimizer being compelled to take larger steps along this direction in search of improvement. While this conclusion may seem obvious, its implications are not to be underappreciated. That is, in the case of a strongly correlated and convex target function with finite widths (as we appear to have), far fewer observations will be required to effectively map the function along the directions with the largest length scales than will be required to map the function along the directions with the shortest length scales, but only if our GP model is informed with appropriate length scale and correlation hyperparameters.

Alternatively, we can think of the correlation hyperparameters as effectively decoupling the optimization problem. Recall from Section 4.2.5 that using a GP with correlation hyperparameters is equivalent to first performing a change of basis on our input vectors and then using a GP with a diagonal matrix of length scales. When properly chosen, the change-of-basis rotates the target function into a coordinate-space where it no longer exhibits correlations between inputs. In the case of a correlated,

convex target function like ours, this means that the optimizer should effectively be able to replace an n-dimensional optimization problem with n 1-dimensional optimization problems. Indeed, the optimizer only needs to find the maximum of the function along each, now independent, input component. This translates to a significant amount of space, which should increase with the dimensionality of the inputs, that can effectively be disregarded by the optimizer without loss of performance.

To see the effects of the correlation hyperparameters on our Bayesian optimization algorithm, we used as a simple synthetic target function a correlated n-dimensional Gaussian function with additive Gaussian noise. We then selected 100 random starting positions, each one Mahalanobis distance away from the target function maximum (meaning every optimizer started at the same underlying function value). From each of these starting positions, we deployed two versions of our Bayesian optimization algorithm: one using a GP with diagonal matrix of length scale hyperparameters, and one using a GP with properly optimized matrix of length scale hyperparameters. All other GP hyperparameters were identical between the two versions, and Expected Improvement was used as the acquisition function. The optimizers were declared to have converged when they reached a point in the input space corresponding to an underlying target function value of at least 95% of the maximum. The results are plotted in Figure 5.1. The number of steps to converge grows apparently exponentially with the dimensionality of the input space in the case that the target function is correlated and the GP length scale hyperparameter matrix is not. In contrast, when the correlation hyperparameters are optimized, the number of steps to converge increases only linearly

with the dimensionality of the inputs. In other words, the higher the dimensionality of the optimization problem, the more important it is to include accurate correlation information in our GP model (provided the target function is strongly correlated).

5.2 Learning the Optimal GP Hyperparameters

In this section I will describe our approach to learning the optimal hyperparameters for our Gaussian process regression model. In chapter 4, we saw that given sufficient data, we were able to adjust the model hyperparameters to maximize the likelihood of the model given the data. By doing so, the Gaussian process regression model was able to achieve an excellent fit in our toy examples. However, in those examples, the number of input dimensions was at most two. While LCLS does a good job of archiving historical data from routine quadrupole tuning sessions, it is uncommon that fewer than 4 quadrupoles are adjusted simultaneously during these optimization scans. This means that the data we have available to us on which to train our Gaussian process regression model belongs to a higher dimensional and therefore much larger space than the previous examples. While our ability to learn the noise and variance hyperparameters does not suffer greatly from this higher dimensionality, the number of hyperparameters in our matrix $\mathbf{\Sigma}_{SE}$ grows quadratically with increasing dimension. We found that attempting to learn a complete matrix of length scale hyperparameters on a scan containing ~ 100 noisy observations in 4+ dimensional space results in over-fitting. Of course, in trying to learn some length scale hyperparameters for our model, we are making the assump-

tion that there is some underlying structure to our target function that is stable from one scan to the next. This assumption is well-supported, as 2-dimensional raster scans in quad-space performed months apart on the same 2 quadrupoles consistently produce results displaying the same correlated structure discussed in section 2.3.7. While these raster scans are prohibitively time-consuming to be performed repeatedly in a way that would fully characterize the underlying function, the knowledge that there is some stability to its structure suggests that data from the routine optimization scans should be able to be combined in some way.

Regrettably, combining scan data to learn a complete matrix of length scale hyperparameters is not as easy as it sounds. Unobservable errors in the real machine may change unpredictably from day to day, which, along with variation in beam energy and upstream lattice conditions, makes aggregating data into larger sets tricky, because the data do not necessarily represent the same underlying function. First, if we want to learn some stable underlying structure, we need to allow for different length scales for different beam energies, because electrons of different energies are affected differently by the quadrupoles. Thus, the training data must somehow be split according to beam energy. This is straightforward, and still results in plenty of data for each division. The changing upstream lattice conditions and unobservables, on the other hand, effectively displace our function in quad-space, which means we cannot simply stack data sets from different days on top of each other. Because the Twiss parameters are not routinely measured, we do not know exactly in what way the upstream optical conditions have changed, so the nature of the displacement is uknown. To attempt to combine the data in an

intelligent way, we took the best values from each of our various optimization scans and assumed that, for each scan, these quadrupole settings corresponded to the maximum of the underlying function. We then zeroed all the input data in each scan relative to each scan's optimal point before combining. Extracting length scale information from the combined set of data did not produce reliable results, presumably because of combined errors from our inevitably inaccurate estimation of the true optimum for each scan and from the noise in the observations.

Of course, it is possible to forego a GP model with complete length scale information in favor of one with a only diagonal matrix of kernel hyperparameters. The number of length scale hyperparameters then scales linearly with dimensionality instead of quadratically. Estimates for these hyperparameters can therefore be produced much more confidently from the available scan data, without needing to worry about combination. However, as we have seen, our Bayesian optimization algorithm should perform significantly better if we are able to include accurate off-diagonal information in our hyperparameter matrix to describe the correlations between neighboring quadrupoles that our optical model predicts and observations (for select cases) have confirmed. Indeed, as has been hinted at in the previous chapters, it is our ability to predict the principal directions of curvature of our target function using our optical model that comes to our rescue.

The optical model described in Chapter 2 is simply composed of a series of matrix products. As such, it is very fast to evaluate. The principal directions of curvature of the resulting beam size function, in quad-space (see Figure 2.9), can be computed

numerically by calculating the eigenvectors of the Hessian matrix of second order derivatives evaluated at the function minimum. These principal direction vectors can then be used to formulate our matrix of length scale hyperparameters $\mathbf{\Sigma}_{SE}$ as described in eq (4.19). The number of hyperparameters in our resulting GP model scales linearly with dimensionality, and can therefore be trained using much less data. Results from an experimental FEL quadrupole tuning session at LCLS using Bayesian optimization via GP regression with principal direction information supplied by our optical model and corresponding length scale magnitudes (and other hyperparameters) learned from data are shown in Figure 5.3.

5.3 Optimizing the Acquisition Function

The results in Figures 5.1 and 5.3 emphasize the importance of proper model selection in maximizing the efficiency of our experimental optimization algorithm, and accordingly we have spent the vast majority of this thesis discussing how to most effectively model our target function. However, modeling the target function well is not the only important step in a Bayesian optimization algorithm. There is another step that is just as integral to all Bayesian optimization algorithms that must be addressed, as the specifics of its implementation can make the difference between an optimizer that converges with excellent efficiency and one that does not converge at all. This step, of course, is the optimization of the acquisition function.

As we saw in chapter 3, the acquisition function is the metric by which we

compare candidate points in the prospective search space. At each iteration of the Bayesian optimization algorithm, the acquisition function must be searched to find the input vector which maximizes its value. It is through this optimization that we ensure that the next point the optimizer acquires is truly the best (according to our model) of all possible candidates. Despite its critical role, proper optimization of the acquisition function is a topic that can easily be glossed over in discussion of Bayesian optimization. Nevertheless, it is a somewhat non-trivial problem, as the acquisition function can in fact have even more complex structure than the target function, making it a tricky function to optimize. Even when the target function is convex, the acquisition function at a given iteration may suffer from multiple local maxima. And of course, the acquisition function will have as many dimensions as the target function, meaning it, too, may suffer from the curse of dimensionality.

Originally, we had designed to optimize the acquisition function using a stock Nelder-Mead simplex method optimizer, but found in our Monte-Carlo investigations discussed in section 5.1 that as we increased the dimensionality of our toy target function, the optimizer would sometimes get lost – that is, completely fail to converge. Because our target function in these Monte-Carlo tests was convex, and the starting point for the optimizer was always reasonably close to the optimum, *and* because Bayesian optimization is renowned as a global optimization strategy, we concluded that perhaps there was an error in our implementation. Sure enough, close examination (via raster-scans in 2-d examples) of our results for the optimization of the acquisition function confirmed that we were not successfully selecting the input vector that produced the

function's global maximum.

Fortunately, our acquisition function is cheap enough to evaluate that these issues with its optimization are able to be solved using brute-force. Specifically, our solution was to deploy many Nelder-Mead simplex method optimizers in parallel starting from different positions scattered randomly about hyper-spheres centered on the best 3 points our optimizer had observed thus far. (The radii of the hyper-spheres determined the spacing of the starting positions, and therefore had to be reflective of the widths of possible local maxima in which our optimizers could get trapped. In our case, we used a fraction of one of the length scale magnitudes.) By selecting the best result of the parallelized optimizations as our next-acquired point, our aforementioned convergence problems in higher dimensions were apparently solved. While this method of optimizing the acquisition function is somewhat computationally expensive, it is quick enough to be evaluated using parallel processing without becoming a limiting factor in our overall Bayesian optimization algorithm, since the target function can only be evaluated at a rate of 0.5 Hz, regardless. This swarm-like approach to optimizing the acquisition function was used in the Monte-Carlo study shown in Figure 5.3 as well as the experimental tuning session whose results are shown in Figure 5.3.

5.4 Results

To test our completed optimization algorithm on the FEL at LCLS, we began with a beam that had been previously matched as well as possible. The beam was then

detuned by selecting a random 4-vector (in the quad-space corresponding to the devices we would eventually seek to retune) and stepping along that direction until the target signal was reduced by about 90%. This position was saved as the starting point for the optimization algorithms that we compared. With our Bayesian optimization algorithm having already been demonstrated as an effective approach to quadrupole tuning using a GP with diagonal matrix of length scale hyperparameters, we sought to evaluate the benefit of adding correlation information from the optical model. The results of the comparison are shown on the left in Figure 5.3. The Bayesian optimization algorithm whose GP includes the correlation information from the optical model significantly outperformed the less-informed Bayesian optimization algorithm in 4-dimensions. These results are consistent with results produced from applying our Bayesian optimization algorithms to a synthetic (correlated Gaussian) target function with correlations equal to those predicted by the beam size response of our optical model in a configuration representing the FEL at time of our experiment, shown on the right in Figure 5.3.

As briefly mentioned in Chapter 2, we have restricted our focus to the quadrupoles at the very end of the linac, just before the undulator. The simple optical model presented in Chapter 2 fails to accurately model the FEL response to quadrupoles much further upstream, due to more complicated optical effects of other devices in between. (By looking only at the *matching quadrupoles* located just before the undulator, we are able to ignore these more complicated effects, while maintaining sufficiently many degrees of freedom to match the Twiss parameters into the undulator focusing lattice.) Relevantly, Bayesian optimization without supplemental information from the optical

model should still be considered a viable approach to quadrupole tuning.

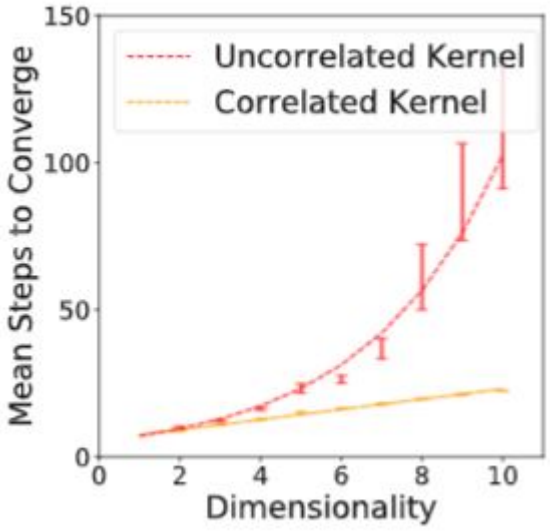


Figure 5.1: Effects of Correlation Hyperparameters in Increasing Dimensions

Comparison of Bayesian optimization convergence time on a correlated target function. The results plotted in yellow represent a GP with optimized length scale correlation hyperparameters, while the results plotted in red represent a GP with diagonal (uncorrelated) length scale hyperparameter matrix. Each bar shows the standard error about the mean for 100 trials. The correlated GP kernel (yellow linear fit) performs as well as optimization of an isotropic target function with an isotropic GP kernel, growing linearly with the number of dimensions. Steps to converge with mismatched kernel grows approximately exponentially (red exponential fit). Figure appears in [7].

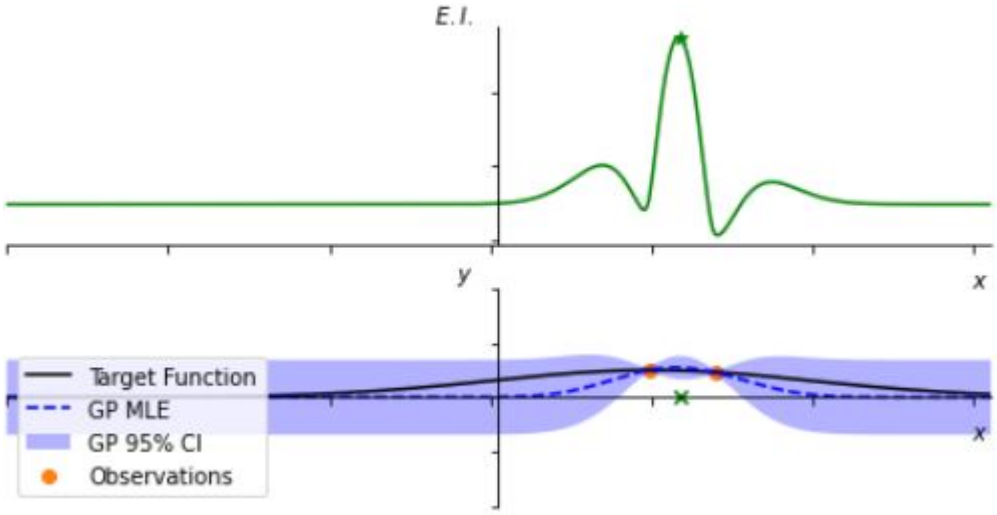


Figure 5.2: Acquisition Function: Local Maxima

(Top) Expected Improvement (E.I.) acquisition function evaluated using GP model shown below. The function contains two local maxima. (Bottom) A GP model (blue) conditioned on data (orange) sampled from a convex target function (black).

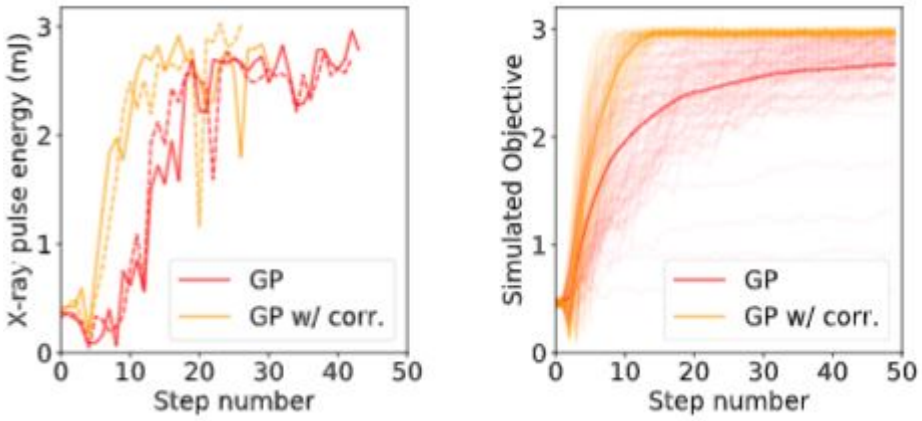


Figure 5.3: Bayesian Optimization Results at LCLS

(Left) Optimization results at LCLS for 4 matching quadrupoles: GP without correlation hyperparameters (red) vs GP with correlations (yellow). Each scan was performed twice with identical starting conditions, shown with different linestyles. Each step takes approximately 3 to 4 seconds. **(Right)** Simulations using an optical model with quadrupoles configured as they were during the live tests shown on the left. 100 individual scans for each method, with means shown by thick lines, are consistent with measurements. Figure appears in [7].

Chapter 6

Conclusion

At FEL X-Ray sources like LCLS, frequent tuning of the quadrupole magnets, especially the matching quadrupoles, is necessary to achieve optimal X-ray intensities under ever-changing electron-beam conditions. Numerical approaches to tuning the quadrupoles have been demonstrated as viable in previous studies. In this work we showed that by informing a GP with prior information about correlations between quadrupoles derived from an optical model, our Bayesian optimization algorithm was able to beat previous benchmarks. Marginal improvements to the tuning process like this could, over the course of years, help make available tens or hundreds of hours of additional beam time to research groups – time that would have otherwise been lost to the tuning process. Further improvements may be possible in the Bayesian optimization approach by incorporation of an appropriately constructed prior mean function, as well as through further experimentation with acquisition functions.

Additionally, it should be noted that more could have been done to model the

FEL response to the quadrupoles than was presented here. In Chapter 2, we did not go further than establishing an approximate relationship between the beam size and the FEL output power via the Pierce parameter. While attempts were made to predict the FEL output power using a combination of the optical model and simple 1-d FEL physics, the results did not exhibit the same scaling as the measured response. Thus, the results were no more useful in training our GP than the results presented in Chapter 2 derived from the optical model alone. That is, although the correlations between the quadrupoles were well-predicted, the length scales were not. It may be possible to accurately model our target function length scales via FEL simulators like GENESIS, but this approach is computationally expensive and cannot be executed sufficiently quickly to be useful in practice [41]. The attraction of the simple optical beam size model is that it can be evaluated quickly enough to be used to compute predicted correlations for a given quadrupole lattice configuration on-demand. With these practical concsiderations in mind, further attempts could be made at constructing a complete FEL model (which combines the effects of the quadrupoles and the undulator) that is fast to evaluate.

Bibliography

[1] T.P. Wangler. *RF Linear Accelerators*. Physics textbook. Wiley, 2008.

[2] G Ising. Prinzip einer Methode zur Herstellung von Kanalstrahlen hoher Voltzahl. *Ark. Mat. Astron. Fys.*, 18:1–4, 1924.

[3] Claudio Pellegrini. The history of x-ray free-electron lasers. *The European Physical Journal H*, 37(5):659–708, 2012.

[4] Kwang-Je Kim, Zhirong Huang, and Ryan Lindberg. *Synchrotron radiation and free-electron lasers*. Cambridge university press, 2017.

[5] Paul Emma, R Akre, J Arthur, R Bionta, C Bostedt, J Bozek, A Brachmann, P Bucksbaum, Ryan Coffee, F-J Decker, et al. First lasing and operation of an ångstrom-wavelength free-electron laser. *nature photonics*, 4(9):641–647, 2010.

[6] Lcls overview. https://lcls.slac.stanford.edu/overview. Accessed: 2021-12-01.

[7] Joseph Duris, Dylan Kennedy, Adi Hanuka, Jane Shtalenkova, Auralee Edelen, P Baxevanis, Adam Egger, T Cope, M McIntire, S Ermon, et al. Bayesian optimization of a free-electron laser. *Physical review letters*, 124(12):124801, 2020.

[8] Zhirong Huang and Kwang-Je Kim. Review of x-ray free-electron laser theory. *Physical Review Special Topics-Accelerators and Beams*, 10(3):034801, 2007.

[9] D. A. G. Deacon, L. R. Elias, J. M. J. Madey, G. J. Ramian, H. A. Schwettman, and T. I. Smith. First operation of a free-electron laser. *Phys. Rev. Lett.*, 38:892–894, Apr 1977.

[10] Lcls facts and infographics. https://lcls.slac.stanford.edu/fact-sheets-and-infographics. Accessed: 2021-12-01.

[11] Ernest D Courant and Hartland S Snyder. Theory of the alternating-gradient synchrotron. *Annals of physics*, 3(1):1–48, 1958.

[12] Alexander Scheinker, Dorian Bohler, Sergey Tomin, Raimund Kammering, Igor Zagorodnov, Holger Schlarb, Matthias Scholz, Bolko Beutner, and Winfried Decking. Model-independent tuning for maximizing free electron laser pulse energy. *Physical Review Accelerators and Beams*, 22(8):082802, 2019.

[13] Sergey Tomin, G Geloni, I Zagorodnov, A Egger, W Colocho, A Valentinov, Y Fomin, I Agapov, T Cope, D Ratner, et al. Progress in automatic software-based optimization of accelerator performance. *Proc. IPAC'16*, pages 3064–3066, 2016.

[14] Xiaobiao Huang and James Safranek. Online optimization of storage ring nonlinear beam dynamics. *Physical Review Special Topics-Accelerators and Beams*, 18(8):084001, 2015.

[15] Xiaobiao Huang. Robust simplex algorithm for online optimization. *Physical Review Accelerators and Beams*, 21(10):104601, 2018.

[16] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

[17] Roger C Conant and W Ross Ashby. Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2):89–97, 1970.

[18] Mitchell McIntire, Daniel Ratner, and Stefano Ermon. Sparse gaussian processes for bayesian optimization. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, UAI'16, page 517–526, Arlington, Virginia, USA, 2016. AUAI Press.

[19] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer school on machine learning*, pages 63–71. Springer, 2003.

[20] J Mockus. The bayes methods for seeking the extremal point. *Kybernetes*, 1974.

[21] J Rossbach and Peter Schmueser. Basic course on accelerator optics. Technical report, P00011673, 1993.

[22] Jamie Rosenzweig. *Fundamentals of beam physics*. Oxford University Press Oxford, 2003.

[23] MR Howells and Brian M Kincaid. The properties of undulator radiation. Technical report, P00021955, 1993.

[24] F. R. Elder, A. M. Gurewitsch, R. V. Langmuir, and H. C. Pollock. Radiation from electrons in a synchrotron. *Phys. Rev.*, 71:829–830, Jun 1947.

[25] HP Freund and PJM Van der Slot. Studies of a terawatt x-ray free-electron laser. *New journal of physics*, 20(7):073017, 2018.

[26] Stefan Hau-Riege. Gas detector lcls engineering specifications document. Technical report, Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), 2007.

[27] Helmut Wiedemann. *Particle accelerator physics*. Springer Nature, 2015.

[28] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55, 1936.

[29] Claude Lemaréchal. Cauchy and the gradient method, 2012.

[30] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[31] Apoorv Agnihotri and Nipun Batra. Exploring bayesian optimization. *Distill*, 2020. https://distill.pub/2020/bayesian-optimization.

[32] Auralee Edelen, Nicole Neveu, C Mayes, C Emma, and D Ratner. Machine learning models for optimization and control of x-ray free electron lasers. In *NeurIPS Machine Learning for the Physical Sciences Workshop*, 2019.

[33] Thomas Kratzke, Lawrence Stone, and J.R. Frost. Search and rescue optimal planning system. pages 1 – 8, 08 2010.

[34] Noel Cressie. The origins of kriging. *Mathematical geology*, 22(3):239–252, 1990.

[35] Jean Pierre Delhomme. Kriging in the hydrosciences. *Advances in water resources*, 1(5):251–266, 1978.

[36] Johannes Kirschner, Mojmir Mutny, Nicole Hiller, Rasmus Ischebeck, and Andreas Krause. Adaptive and safe bayesian optimization in high dimensions via one-dimensional subspaces. In *International Conference on Machine Learning*, pages 3429–3438. PMLR, 2019.

[37] Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions with application to active user modeling and hierarchical reinforcement learning. 2009.

[38] Jonas Močkus. On bayesian methods for seeking the extremum. In *Optimization techniques IFIP technical conference*, pages 400–404. Springer, 1975.

[39] Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

[40] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu

Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

[41] S. Reiche. GENESIS 1.3: a fully 3D time-dependent FEL simulation code. *Nuclear Instruments and Methods in Physics Research A*, 429(1-3):243–248, June 1999.