Karlsruhe Institute of Technology

# Micro-Bunching Control at Electron Storage Rings with Reinforcement Learning

zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

von der KIT-Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)
angenommene

Dissertation

von

Tobias Boltz

aus Lahr (Schwarzwald)

Tag der mündlichen Prüfung: 12. November 2021
Erster Gutachter: Prof. Dr. Anke-Susanne Müller
Zweiter Gutachter: Prof. Dr. Tamim Asfour

# Contents

# Remarks on Notation

Throughout this thesis, vector-valued quantities are written in bold letters while real and complex numbers as well as scalar functions are not. Vectors are generally assumed to be column vectors unless explicitly written out horizontally. The magnitude of a vector is denoted by the non-bold version of the same letter. Random variables are denoted with capital letters, whereas their instantiations are denoted in lower case. Partly due to the interdisciplinary subject of this thesis, some of the used symbols denote more than one quantity (listed below). In those cases, the meaning should be derived from context.

| | |
|---|---|
| $\doteq$ | equality relationship that is true by definition |
| $\approx$ | approximately equal |
| $\leftarrow$ | assignment |
| | |
| $t$ | continuous time or discrete time step (sometimes $t_i$ is used for clarity) |
| $E$ | particle energy |
| $\boldsymbol{E}$ | electric field (magnitude is written as $\lvert\boldsymbol{E}\rvert$) |
| $v$ | particle velocity ($v \doteq \lvert\boldsymbol{v}\rvert$) or state-value |
| $q$ | particle charge or first generalized longitudinal coordinate or action-value |
| $p$ | particle momentum or second generalized longitudinal coordinate |
| | |
| $Q$ | bunch charge or array estimate of action-value function $q_\pi$ or $q_*$ |
| $I$ | bunch current $I \doteq Q\, f_{\mathrm{rev}}$ |
| $V$ | voltage or array estimate of state-value function $v_\pi$ or $v_*$ |
| $V_{\mathrm{RF}}$ | accelerating voltage (RF voltage) |
| $V_0$ | amplitude of RF voltage |
| $g$ | vacuum gap (vertical distance between plates in the parallel plates model) |
| $h$ | half vacuum gap $h \doteq g/2$ |
| | |
| $\alpha_{\mathrm{c}}$ | momentum compaction factor |
| $\alpha$ | step-size parameter (learning rate) |
| $\beta$ | ratio of particle velocity $v$ and speed of light $c$ |
| $\gamma$ | Lorentz factor or discount factor |
| | |
| $\psi(z, E, t)$ | charge distribution in the longitudinal phase space |
| $\hat{\psi}(z, E, t)$ | normalized charge distribution in phase space $\hat{\psi}(z, E, t) \doteq \psi(z, E, t)/Q$ |
| $\rho(z, t)$ | longitudinal bunch profile |
| $\varrho(z, t)$ | normalized bunch profile $\varrho(z, t) \doteq \rho(z, t)/Q$ |
| $\rho(E, t)$ | energy profile |

| | |
|---|---|
| $\mathcal{F}(x)$ | Fourier transform $\mathcal{F}(x) = \widetilde{x}(\omega) = \int_{\infty}^{\infty} x(t) \mathrm{e}^{-i\omega t}\, \mathrm{d}t$ |
| $\mathcal{F}^{-1}(\widetilde{x})$ | inverse Fourier transform $\mathcal{F}^{-1}(\widetilde{x}) = x(t) = \frac{1}{2\pi} \int_{\infty}^{\infty} \widetilde{x}(\omega) \mathrm{e}^{it\omega}\, \mathrm{d}\omega$ |
| $\Pr\{X = x\}$ | probability that a random variable $X$ takes on the value $x$ |
| $X \sim p$ | random variable $X$ distributed according to $p(x) \doteq \Pr\{X = x\}$ |
| $\mathbb{E}[X]$ | expectation of a random variable $X$, i.e., $\mathbb{E}[X] \doteq \sum_x p(x)x$ |
| $\mathrm{avg}[X]$ | sample average of a random variable $X$, i.e., $\mathrm{avg}[X] \doteq \frac{1}{n} \sum_{i=1}^{n} x_i$ |
| $A_t$ | action at time $t$ (random variable) |
| $a$ | a particular action (instantiation) |
| $\mathcal{A}$ | short for $\mathcal{A}(s)$, set of all actions available in state $s$ |
| $\pi(s)$ | action taken in state $s$ under deterministic policy $\pi$ |
| $\pi(a\|s)$ | probability of taking action $a$ in state $s$ under stochastic policy $\pi$ |
| $\boldsymbol{w}, \boldsymbol{w}_t$ | $d$-dimensional vector of weights underlying an approximate value function |
| $\hat{v}(s, \boldsymbol{w})$ | approximate value of state $s$ given weight vector $\boldsymbol{w}$ |
| $\mu(s)$ | on-policy distribution over states |
| $\boldsymbol{x}(s)$ | vector of features visible when in state $s$ |
| $\boldsymbol{w}^\top \boldsymbol{x}$ | inner product of vectors, $\boldsymbol{w}^\top \boldsymbol{x} \doteq \sum_i w_i x_i$ |
| $\boldsymbol{\theta}, \boldsymbol{\theta}_t$ | $d'$-dimensional parameter vector of target policy |
| $\pi(a\|s, \boldsymbol{\theta})$ | probability of taking action $a$ in state $s$ given parameter vector $\boldsymbol{\theta}$ |
| $\pi_{\boldsymbol{\theta}}$ | policy corresponding to parameter $\boldsymbol{\theta}$ |
| $J(\boldsymbol{\theta})$ | performance measure for the policy $\pi_{\boldsymbol{\theta}}$ |
| $\mathcal{A}_\pi(s, a)$ | advantage of taking action $a$ in state $s$ under policy $\pi$ |
| $\{x\}$ | set of possible $x$-values $\{x\} \doteq \mathcal{X} \doteq \{x \mid \text{for } x \in \mathcal{X}\}$ |
| $\{x\} \times \{y\}$ | Cartesian product $\{x\} \times \{y\} \doteq \mathcal{X} \times \mathcal{Y} \doteq \{(x, y) \mid \text{for } x \in \mathcal{X} \text{ and } y \in \mathcal{Y}\}$ |

# 1. Introduction

At the time this thesis is written, the world finds itself amidst and partly in the process of recovering from the COVID-19 pandemic caused by the SARS-Cov-2 virus. One major contribution to the worldwide efforts of bringing this pandemic to an end are the vaccines developed by different research teams all around the globe. Produced in a remarkably short time frame, a crucial first step for the discovery of these vaccines was mapping out the atomic structure of the proteins making up the virus and their interactions. Due to the bright X-rays required in the process, synchrotron light sources play an active role in the ongoing efforts of accomplishing that goal [1]. Synchrotron light sources are particle accelerators that are capable of providing intense electromagnetic radiation by accelerating packages of electrons, called bunches, and forcing them on curved trajectories. Besides the support of research on the SARS-Cov-2 virus, the remarkable properties of synchrotron radiation lead to a multitude of applications in a variety of scientific fields such as materials science, geology, biology and medicine. As a special form of synchrotron radiation, this thesis is concerned with the coherent synchrotron radiation (CSR) generated by short electron bunches in a storage ring. At wavelengths larger than the size of the emitting electron structure, the particles within a bunch radiate coherently. This coherent emission of synchrotron radiation scales with the number of involved particles and can thus enhance the intensity of the emitted radiation by several orders of magnitude. As a consequence, modern synchrotron light sources, such as the Karlsruhe Research Accelerator (KARA) at the Karlsruhe Institute of Technology (KIT), are deliberately operating with short bunch lengths to extend the radiated CSR spectrum to higher frequencies and to increase the intensity of the emitted radiation. Yet, the continuous reduction of the bunch length at high beam intensities eventually leads to complex longitudinal dynamics caused by the self-interaction of the electron bunches with their own emitted CSR. This phenomenon, generally referred to as micro-bunching or micro-wave instability, can lead to the formation of dynamically changing micro-structures within the charge distribution of the electron bunches and thus to a fluctuating emission of CSR. Moreover, it can cause oscillations of the bunch length and the energy spread, which can be detrimental to the operation of a synchrotron light source. On the other hand, as electron structures smaller than the full electron bunch, the micro-structures created by the instability lead to an increased emission of CSR at frequencies up to the THz frequency range. The instability can thus also be beneficial for a variety of applications that rely on intense radiation in that particular frequency range.

Over the past years, the micro-bunching instability has been extensively studied at the KIT storage ring KARA and other synchrotron light sources. Facilitated by the development of novel diagnostics and simulation tools, the instability and the underlying longitudinal beam dynamics were observed and analyzed in great detail and across a large range of machine parameters. Building upon the gained insights and experience with the instability,

the work summarized in this thesis takes these efforts one step further by approaching the topic of control over the occurring micro-bunching dynamics. In a careful analysis of the perturbation generated by the CSR self-interaction, an effective method of influencing the formation of micro-structures is identified and the resultant opportunities of exerting control over these dynamics are pursued. As indicated above, the benefits of extensive control over the micro-bunching instability are twofold. A practical method of mitigating the CSR-induced perturbation at an electron storage ring would extend the regime of stable operation to shorter bunch lengths and higher bunch currents. As illustrated in the context of the COVID-19 pandemic, particle accelerators in general and synchrotron light sources in particular are instruments that facilitate basic scientific research in various domains. An extension of the sustainable beam properties that can be provided to external experiments is thus a major benefit. Additionally, successful mitigation of the micro-bunching instability would expand the capabilities to optimize for related beam properties, at existing facilities, but also for future synchrotron light sources. On the other hand, a deliberate and controlled excitation of the micro-structures can amplify the intensity of the CSR emitted in the frequency range corresponding to the spatial extent of the structure and could thus be used to tailor the emission of CSR to dedicated experiments. In an attempt to support these complementary objectives, the presented work is mainly concerned with finding direct ways of interacting with the micro-structure formation process in order to influence the beam dynamics in either direction. For the objective of mitigating the micro-bunching dynamics, which turns out to be the more challenging task, the necessity for dynamic adjustments of the applied control signal naturally motivates the use of reinforcement learning (RL) methods. The general task is thus formalized as an RL problem and different state-of-the-art algorithms are applied to solve the underlying control problem. The pursued approach towards micro-bunching control at electron storage rings is developed and tested on the basis of simulation data and its feasibility verified in first experiments at KARA.

After this introduction, the content of this thesis is divided into eight further chapters. While chapter 2 covers the required fundamentals of accelerator physics, the concept of coherent synchrotron radiation and the micro-bunching instability are introduced in chapter 3. Beyond a description of the driving mechanism underlying the instability and its characteristic features, the chapter also contains a brief summary of existing mitigation techniques and prior efforts to influence the micro-bunching dynamics. As a substantial part of the developed approach to micro-bunching control relies on the use of reinforcement learning methods, chapter 4 provides an introduction to the general subject and covers a selection of modern RL algorithms. In an analysis of the longitudinal dynamics underlying the micro-bunching instability and the relation between individual particle trajectories and collective motion, the theoretical basis for the control pursued in the thesis is derived in chapter 5. As the necessity of dynamic control directly motivates the use of reinforcement learning methods, chapter 6 formalizes the general task as a reinforcement learning problem. The main results obtained by applying the developed methods in simulations are presented in chapter 7. While these results verify the general feasibility of extensive micro-bunching control, further challenges regarding the stability and generalization of the achieved RL-based control are discussed in the final section. Although the full implementation of the developed methods at KARA was beyond the

scope of this thesis, chapter 8 presents a range of first experiments to verify the feasibility in practice. The thesis finally concludes with a brief summary and outlook towards future work in chapter 9.

# 2. Fundamentals of Accelerator Physics

*Nothing happens until something moves.*

— Albert Einstein

The main objective of the work summarized in this thesis was to identify and pursue an avenue towards control of the micro-bunching dynamics which occur in electron storage rings under specific operating conditions. As modern storage rings tend to be quite complex systems, this chapter aims to provide some of the most relevant fundamentals of accelerator physics. For a more exhaustive and detailed introduction it is referred to the existing textbooks, e.g. [2, 3].

## 2.1. Charged Particles in Electromagnetic Fields

In order to accelerate particles, or more precisely to increase their kinetic energy, they have to be subjected to an external force. Of the four fundamental forces: gravity, the weak and the strong force, and the electromagnetic force, only the latter is suitable for the technical requirements of a typical particle accelerator. A particle with charge $q$ which is moving with velocity $\boldsymbol{v}$ in external electromagnetic fields is subject to the Lorentz force

$$\boldsymbol{F}_\mathrm{L} = q\ (\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B})\ , \tag{2.1}$$

where $\boldsymbol{E}$ and $\boldsymbol{B}$ denote the electric and the magnetic field, respectively. During its motion from position $\boldsymbol{r}_1$ to $\boldsymbol{r}_2$, the particle gains the energy

$$\Delta E = \int_{\boldsymbol{r}_1}^{\boldsymbol{r}_2} \boldsymbol{F}_\mathrm{L}\, \mathrm{d}\boldsymbol{r} = q \int_{\boldsymbol{r}_1}^{\boldsymbol{r}_2} (\boldsymbol{v} \times \boldsymbol{B} + \boldsymbol{E})\, \mathrm{d}\boldsymbol{r}\ . \tag{2.2}$$

As $\mathrm{d}\boldsymbol{r}$ and $\boldsymbol{v}$ are parallel, the term $(\boldsymbol{v} \times \boldsymbol{B})\, \mathrm{d}\boldsymbol{r}$ vanishes and Eq. (2.2) can be simplified to

$$\Delta E = q \int_{\boldsymbol{r}_1}^{\boldsymbol{r}_2} \boldsymbol{E}\, \mathrm{d}\boldsymbol{r} = q\, \Delta V\ , \tag{2.3}$$

where $\Delta V$ is the potential difference induced by the electric field. The energy gain of the particle is thus independent of the magnetic field, the desired acceleration has to be achieved solely through the use of electric fields. Nonetheless, magnetic fields are still essential for the operation of particle accelerators as they are used to deflect and focus the particle's trajectory. As the force induced by a constant magnetic field is perpendicular to

the particle's velocity, it bends the path of motion and, in the absence of further effects, leads to a circular trajectory. In that case, the Lorentz force acts as the centripetal force

$$F_\mathrm{c} = F_\mathrm{L} \quad \Rightarrow \quad \frac{mv^2}{R} = qvB\,, \tag{2.4}$$

where $m$ denotes the particle's mass and $R$ is the radius of the trajectory. For relativistic particles, where the velocity is close to the speed of light $v \approx c$, the force induced by a magnetic field is typically much stronger than what can be realized via electric fields. Hence, for storage rings operating at relativistic energies, electric fields are used to increase the particle's energy whereas bending and focusing of the beam is achieved via magnetic fields.

## 2.2. Electron Storage Rings

Around the 1920s, the first machines to accelerate particles generally used a static electric field induced between two electrodes. Driven by the desire for higher particle energies, several different designs were conceived in order to increase the maximum achievable voltage. Ultimately, though, this approach is limited by an effect known as corona discharge, which describes a spark-over between the electrodes and leads to the collapse of the high voltage. In later machines, and in most modern particle accelerators, this limitation is overcome by the use of an alternating voltage. As the electric field is varying, this brings with it the necessity to carefully control the timing of the particle's passage through the accelerating section. Typically, one aims to expose the particle to the sinusoidal voltage

$$V_\mathrm{RF}(t) = V_0 \sin(2\pi f_\mathrm{RF} t + \varphi_0)\,, \tag{2.5}$$

at a designated phase $\varphi_\mathrm{s}$, which increases the particle energy by

$$\Delta E = qV_0 \sin(\varphi_\mathrm{s})\,. \tag{2.6}$$

Here, the accelerating voltage is defined by the amplitude $V_0$, the radio frequency (RF) $f_\mathrm{RF}$ and the initial phase $\varphi_0$. The required electric field is usually realized in an RF cavity, a special type of metallic resonator which can generate a standing wave at its resonant frequencies. In a linear accelerator, several of these accelerating structures are arranged along a straight line, where each additional passage further increases the particle's energy by the same amount. In principle, one can achieve arbitrarily high energies with this approach. In practice, however, the growing size and the involved costs quickly restrict such efforts. A more efficient approach is to use magnetic fields to force the particles on a circular orbit on which they repeatedly pass through the same accelerating structure. Yet, to keep the accelerated particles in a fixed aperture, one has to consider the dependency of the trajectory on the particle energy. For relativistic particles, Eq. (2.4) yields the bending radius

$$R = \frac{E}{qcB}\,, \tag{2.7}$$

which means, in a constant magnetic field, particles with higher energy will travel on a trajectory with a larger radius. This can be addressed in different ways, yielding several

types of circular particle accelerators. One such machine is the synchrotron, which achieves a constant radius by ramping up the magnetic field strength synchronously to the rising particle energy. In order to maintain a fixed phase for the acceleration process, the radio-frequency $f_{\mathrm{RF}}$ in Eq. (2.5) has to be an integer multiple of the revolution frequency

$$f_{\mathrm{RF}} = h\, f_{\mathrm{rev}} \,, \tag{2.8}$$

where $h$ is the so-called harmonic number. A particle which passes the accelerating structure at a phase that significantly deviates from the design phase $\varphi_{\mathrm{s}}$, e.g. because of perturbations along its trajectory, will be decelerated and eventually hit the vacuum chamber wall. Thus, the particle beam in a synchrotron cannot be continuous. Instead, particles can only be accumulated in small packages, the so-called bunches, around the valid phases. Nonetheless, it is common to refer to the total amount of particles or charge in the machine as the beam current

$$I_{\mathrm{beam}} \doteq Q_{\mathrm{total}}\, f_{\mathrm{rev}} \,. \tag{2.9}$$

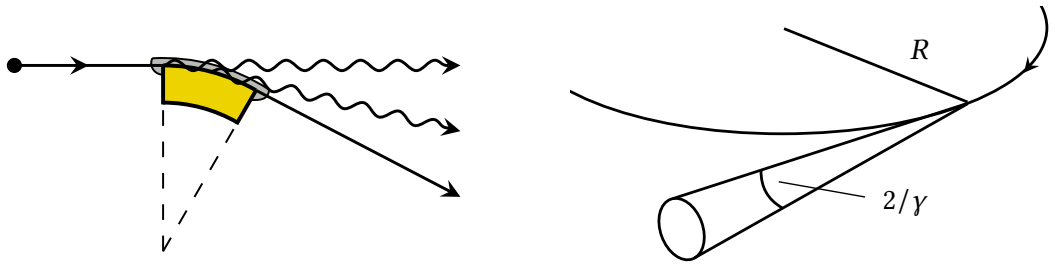
Analogously, one defines the bunch current

$$I \doteq Q_{\mathrm{bunch}}\, f_{\mathrm{rev}} \,. \tag{2.10}$$

As discovered during the 1950s, one major benefit of the synchrotron is that it can be operated as a storage ring, in which particles are stored at a constant energy. This facilitates the accumulation of high beam currents and is particularly convenient for a range of use cases. Most notably, it enables particle colliders to achieve high collision rates which is of the utmost importance in particle physics experiments, e.g. the search for new elementary particles. The usage of electrons in these experiments leads, due to their point particle nature, to clean collisions and thus allows for precise measurements. Yet, it was quickly discovered that the accelerated electrons lose a part of their energy during every revolution by emitting electromagnetic waves. As this phenomenon was first observed at the General Electric 70 MeV synchrotron [4], the detected radiation was named synchrotron radiation.

## 2.3. Synchrotron Radiation

In the years after its discovery, synchrotron radiation was treated as a by-product of high energy particle accelerators and its characteristics were studied parasitically at these machines. Over time though, the exceptional properties of this type of radiation were realized and in 1968 the first dedicated storage ring for the production of synchrotron radiation began its operation [5]. The first generation of synchrotron light sources made use of the radiation that is emitted in the main bending magnets. This is also the case for the second generation of machines where the transverse beam size was minimized to improve the spatial resolution in experiments. Nowadays, in machines of the third generation, so-called insertion devices are installed into the machines to produce additional synchrotron radiation with improved horizontal focusing and higher photon flux.




(a) Emission of synchrotron radiation in a bending magnet

(b) Synchrotron radiation cone

Figure 2.1.: (a) During their deflection in a bending magnet, electrons emit synchrotron radiation tangential to their path of motion. (b) The radiation is emitted in a narrow cone with an opening angle that is determined by the particle energy. Figure adapted from [2].

The synchrotron radiation, due to the deflecting force of a bending magnet, is emitted in a narrow cone tangential to the trajectory of the electron beam, as illustrated in Fig. 2.1. For relativistic energies, the cone's half opening angle is approximately given by

$$\Theta_{\mathrm{cone}} \approx 1/\gamma \,, \tag{2.11}$$

with the relativistic Lorentz factor $\gamma = 1/\sqrt{1-\beta^2}$ and $\beta = v/c$. In case of the KIT storage ring KARA with electron energies between 0.5 GeV and 2.5 GeV, the full opening angle is between 0.41 mrad and 2.04 mrad. The strong focusing of the emitted photon beam is one of the properties that distinguishes synchrotron light sources from more conventional ways to produce electromagnetic radiation. The instantaneous power emitted by a single electron passing through a bending magnet is given by [2]

$$P_{\mathrm{s}} = \frac{e^2 c}{6\pi\varepsilon_0} \frac{1}{(m_{\mathrm{e}}c^2)^4} \frac{E^4}{R^2} \,, \tag{2.12}$$

with the elementary charge $e$, the electron rest mass $m_{\mathrm{e}}$, and the dielectric constant of vacuum $\varepsilon_0$. During a full revolution, the electron typically passes several of these bending magnets and loses the energy

$$\Delta E = \oint P_{\mathrm{s}} \,\mathrm{d}t = \frac{1}{f_{\mathrm{rev}}} P_{\mathrm{s}} = \frac{2\pi R}{c} P_{\mathrm{s}} \,, \tag{2.13}$$

which has to be compensated by the RF system of the storage ring. Owing to the narrow angle of emission, the radiation is observed in short electromagnetic pulses and the radiated power spectrum depicted in Fig. 2.2 covers a broad range of frequencies. It can be described by [6]

$$\mathcal{P}(\omega) = \frac{P_{\mathrm{s}}}{\omega_{\mathrm{c}}} S\left(\frac{\omega}{\omega_{\mathrm{c}}}\right) \,, \tag{2.14}$$

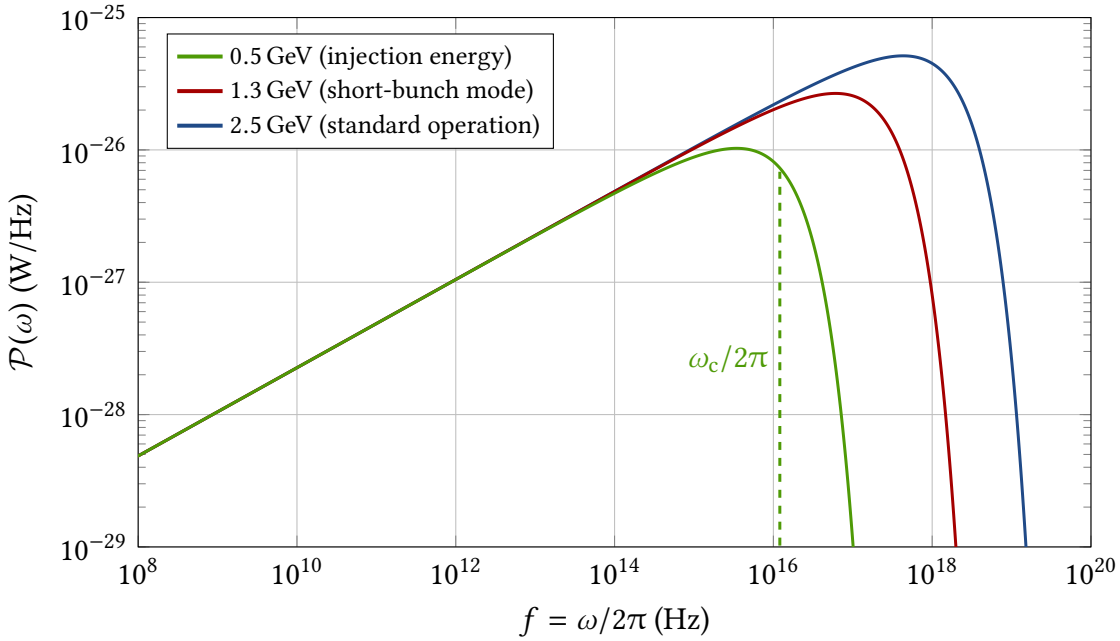



Figure 2.2.: Radiated power spectrum $\mathcal{P}(\omega)$ of a single electron passing through a bending magnet ($R = 5.559\,\mathrm{m}$) for the typical beam energies at KARA. The green dashed line marks the critical frequency $\omega_c$ at injection energy, it divides the power spectrum into two sections of equal integrated radiation power.

where $\omega_c = 3c\gamma^3/2R$ denotes the critical frequency, which divides the power spectrum into two sections of equal integrated radiation power

$$\int_0^{\omega_c} \mathcal{P}(\omega)\,\mathrm{d}\omega = \int_{\omega_c}^{\infty} \mathcal{P}(\omega)\,\mathrm{d}\omega = \frac{1}{2}P_s\,. \tag{2.15}$$

The spectral function $S$ is given by

$$S(\xi) = \frac{9\sqrt{3}}{8\pi}\xi \int_{\xi}^{\infty} K_{5/3}(\xi)\,\mathrm{d}\xi\,, \tag{2.16}$$

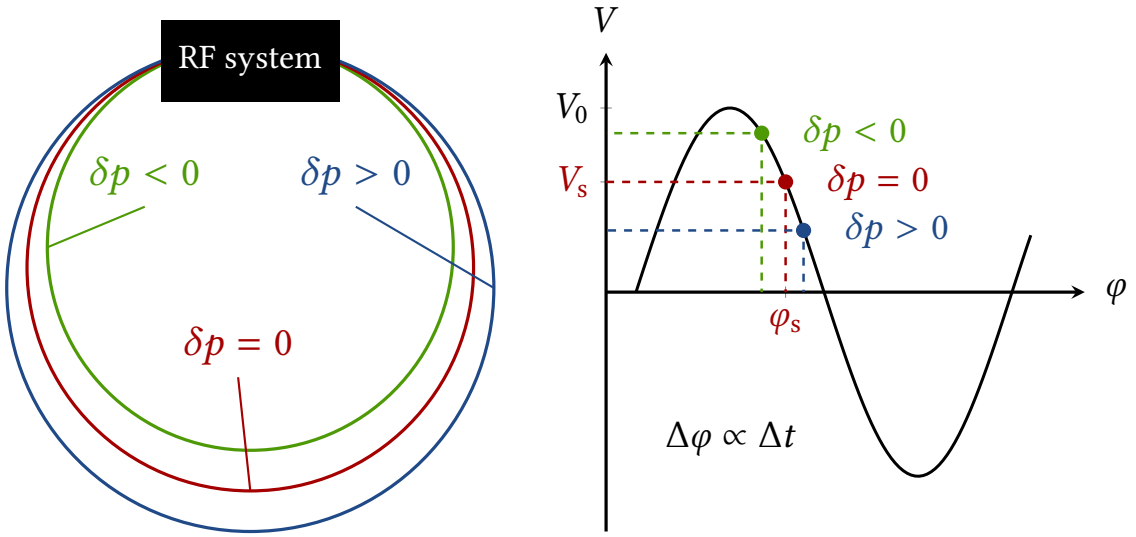
where $K_{5/3}$ denotes the modified Bessel function. Finally, the combined, incoherent emission of all electrons within a bunch yields the incoherent power spectrum

$$\mathcal{P}_{\mathrm{ISR}}(\omega) = N_e\,\mathcal{P}(\omega)\,, \tag{2.17}$$

where $N_e$ is the number of electrons, which is typically around $10^9$ at KARA.

## 2.4. Longitudinal Beam Dynamics

With each passage through a bending magnet, and with each revolution in a storage ring, the accelerated electrons lose a part of their energy in the form of synchrotron radiation. In order to maintain a constant beam energy and a fixed trajectory, this energy loss has




(a) Orbits for deviating momenta

(b) RF voltage at different phases

Figure 2.3.: Principle of phase focusing. (a) Compared to the orbit of the synchronous particle (red line), particles with a lower momentum travel on a shorter orbit (green line), whereas a higher momentum increases the path length (blue line). (b) Depending on their arrival time in the accelerating structure, particles are exposed to a different voltage and thereby gain different amounts of energy. This results in a focusing effect towards the momentum and phase of the synchronous particle. Figure adapted from [2].

to be compensated by the accelerating RF system. A particle, which gains exactly that amount of energy in the accelerating section that is radiated off during one full revolution in the storage ring, is called synchronous particle. Corresponding to the design energy, it holds the reference momentum $p_s$ and passes the RF system precisely at the synchronous phase $\varphi_s$. Any particle with a non-zero momentum deviation

$$\delta p \doteq (p - p_s)/p_s \,, \tag{2.18}$$

is subject to a different deflection in the bending magnets and therefore travels on a deviating orbit. According to Eq. (2.7), a lower momentum leads to a smaller bending radius and a shorter orbit. Meanwhile, the change in velocity is negligible for the ultra-relativistic particle energies considered in this thesis. A particle with $\delta p < 0$ thus arrives earlier in the accelerating structure, is exposed to a higher voltage and gains more energy. Vice versa, a particle with higher momentum has a longer path, arrives later and gains less energy. Combined, this results in a focusing effect which restores the deviating particles towards the synchronous phase. The overall concept is known as phase focusing and illustrated in Fig. 2.3. The relation between the deviation in orbit length

$$\delta L \doteq (L - L_s)/L_s \,, \tag{2.19}$$

and the momentum deviation is defined as the momentum compaction factor

$$\alpha_c \doteq \frac{\delta L}{\delta p} \,. \tag{2.20}$$

As a smaller momentum compaction factor leads to a stronger longitudinal compression around the synchronous particle, it has a direct impact on the length of the electron bunches in a storage ring. For the short-bunch operation mode at KARA, the magnet optics are thus adjusted such that the momentum compaction factor is decreased by more than a full order of magnitude, which is why it is also referred to as low-alpha mode.

In consequence of the restoring effect of phase focusing, particles with deviating momenta perform a longitudinal oscillation around the synchronous particle, the so-called synchrotron motion. For small deviations from the synchronous phase

$$\phi \doteq \varphi - \varphi_\mathrm{s} \,, \tag{2.21}$$

it can be described by the equation of motion [3]

$$\ddot{\phi} + \frac{2}{\tau_\mathrm{d}} \dot{\phi} + \omega_{\mathrm{s},0}^2 \, \phi = 0 \,, \tag{2.22}$$

with the longitudinal damping time $\tau_\mathrm{d}$ and the nominal synchrotron frequency

$$f_{\mathrm{s},0} = \frac{\omega_{\mathrm{s},0}}{2\pi} = f_\mathrm{rev} \sqrt{\frac{eV_0 h \cos(\varphi_\mathrm{s})}{2\pi\beta^2 E} \left( \frac{1}{\gamma^2} - \alpha_\mathrm{c} \right)} \,. \tag{2.23}$$

The damping term in Eq. (2.22) is related to the energy dependent emission of synchrotron radiation, which counteracts large deviations from the synchronous particle. As the damping time is typically much larger than the oscillation period, it has only a small effect on the phase oscillation and may be neglected. In that case, particles which deviate from the synchronous phase are subject to a linear restoring force generated by the RF system and perform perfectly harmonic oscillations (illustrated in Fig. 2.4a). While deviations in phase and energy are damped on average, the quantized emission of photons by individual electrons is a statistical process which simultaneously introduces as spread in the energy distribution of the bunch. In the equilibrium between quantum excitation and radiation damping, the energy distribution assumes a Gaussian shape with a standard deviation referred to as the natural energy spread [3]

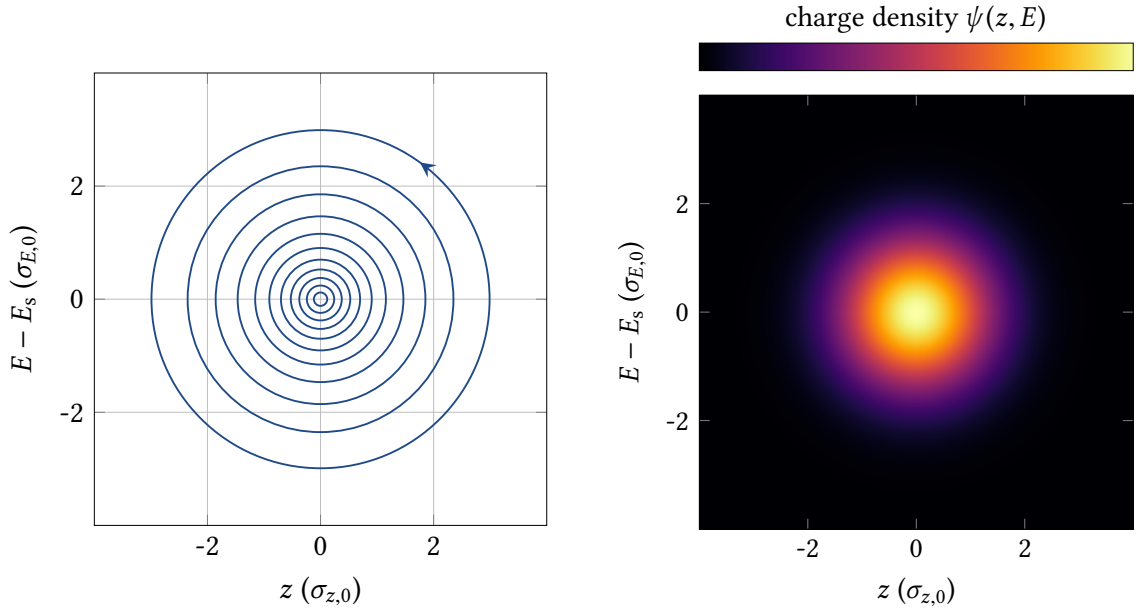
$$\sigma_{\delta,0} = \frac{\sigma_{E,0}}{E} = \sqrt{\frac{55\,\hbar c}{32\sqrt{3}\, m_\mathrm{e} c^2} \frac{\gamma^2 R}{\mathcal{J}_z}} \,, \tag{2.24}$$

with the longitudinal damping partition number $\mathcal{J}_z \approx 2$. This spread in the energy distribution is, via the momentum compaction factor, also transfered to a spread along the longitudinal coordinate

$$z \doteq -\frac{\beta c}{2\pi h f_\mathrm{rev}} \phi \,, \tag{2.25}$$

and is analogously referred to as the natural bunch length

$$\sigma_{z,0} = \left| \frac{c}{2\pi\beta f_{\mathrm{s},0}} \left( \frac{1}{\gamma^2} - \alpha_\mathrm{c} \right) \right| \sigma_{\delta,0} \,. \tag{2.26}$$




(a) Harmonic particle trajectories

(b) Equilibrium charge distribution

Figure 2.4.: (a) Driven by the restoring force of the RF system, off-momentum particles perform oscillations around the position and energy of the synchronous particle. (b) Radiation damping and quantum excitation lead to a Gaussian equilibrium distribution of particles in the longitudinal phase space.

The equilibrium charge distribution in the longitudinal phase space, spanned by the particle energy and longitudinal position relative to the synchronous particle, is thus a two-dimensional Gaussian as shown in Fig. 2.4b. The integral of the phase space density $\psi(z, E)$ over the longitudinal coordinate yields the energy distribution or energy profile

$$\rho(E) = \int_{-\infty}^{\infty} \psi(z, E)\, \mathrm{d}z \,, \tag{2.27}$$

with the standard deviation given by Eq. (2.24). Analogously, by integrating over the energy one obtains the longitudinal bunch profile

$$\rho(z) = \int_{-\infty}^{\infty} \psi(z, E)\, \mathrm{d}E \,. \tag{2.28}$$

While the synchrotron motion is nearly harmonic and the equilibrium distribution is Gaussian, there are many sources of perturbations and deviations from this idealized case in a real particle accelerator. These lead to a more complex synchrotron motion and deformations of the charge distribution in the longitudinal phase space. Moreover, in non-equilibrium cases, the charge distribution may also vary over time.

# 3. Collective Effects

> *In the case of all things which have several parts and in which the totality is not, as it were, a mere heap, but the whole is something beside the parts, there is a cause; […]*
>
> — Aristotle, *Metaphysics*

A description of the motion of a single particle under the influence of external fields, as outlined in chapter 2, is a first step to understand the dynamics in a particle accelerator. Yet, in order to accurately describe the longitudinal dynamics of even a single bunch of particles, it is often not sufficient to regard the beam as a collection of individual, non-interacting particles. Particularly at high intensities, the interaction of the beam with its immediate surroundings generates non-negligible electromagnetic fields. These so-called wake fields act back on the beam and perturb its motion. The micro-bunching instability is the result of such a perturbation, generated collectively by the electrons within a bunch. This chapter introduces a formalism to describe the temporal evolution of the charge distribution under the influence of additional wake fields. Furthermore, it provides a thorough description of the micro-bunching instability and its characteristics. A more general treatment of the subject of collective beam instabilities can be found in [7].

## 3.1. Vlasov-Fokker-Planck Equation

Given the large number of particles within a bunch, it is convenient to describe collective effects by modeling the charge distribution with a normalized density function

$$\hat{\psi}(z, E, t) \doteq \psi(z, E, t)/Q \quad \text{with} \quad Q = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \psi(z, E, t)\,\mathrm{d}z\mathrm{d}E\,, \tag{3.1}$$

instead of considering individual particles. By introducing the generalized coordinates

$$q \doteq z/\sigma_{z,0} \quad \text{and} \quad p \doteq (E - E_\mathrm{s})/\sigma_{E,0}\,, \tag{3.2}$$

the charge distribution can be described in the dimensionless longitudinal phase space spanned by $q$ and $p$, where the origin marks the position of the synchronous particle. Assuming a linear momentum compaction factor and a linear accelerating voltage, the Hamiltonian of the unperturbed system is given by

$$\mathcal{H}_0 = \frac{1}{2}\left(q^2 + p^2\right)\,, \tag{3.3}$$

which describes a one-dimensional oscillator and corresponds to the perfectly harmonic particle oscillations at the synchrotron frequency illustrated in Fig. 2.4a. If the system

is conservative, that is, in the absence of damping and diffusion effects, the temporal evolution of the charge distribution can be described by the Vlasov equation [7]

$$\frac{\partial\hat{\psi}}{\partial\Theta} + \frac{\partial\mathcal{H}}{\partial p}\frac{\partial\hat{\psi}}{\partial q} - \frac{\partial\mathcal{H}}{\partial q}\frac{\partial\hat{\psi}}{\partial p} = 0\,, \tag{3.4}$$

where $\Theta = f_{\mathrm{s},0}t$ denotes the time in multiples of the nominal synchrotron period. Following the notation in [8], the Vlasov-Fokker-Planck (VFP) equation

$$\frac{\partial\hat{\psi}}{\partial\Theta} + \frac{\partial\mathcal{H}}{\partial p}\frac{\partial\hat{\psi}}{\partial q} - \frac{\partial\mathcal{H}}{\partial q}\frac{\partial\hat{\psi}}{\partial p} = \frac{1}{f_{\mathrm{s},0}\tau_{\mathrm{d}}}\frac{\partial}{\partial p}\left(p\hat{\psi} + \frac{\partial\hat{\psi}}{\partial p}\right)\,, \tag{3.5}$$

introduces additional terms on the right-hand side to account for the effects of radiation damping and quantum excitation. Collective effects can be included as a perturbation to the Hamiltonian

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_{\mathrm{c}} \quad \text{with} \quad \mathcal{H}_{\mathrm{c}} = \frac{e f_{\mathrm{rev}}}{\sigma_{E,0} f_{\mathrm{s},0}}\int_q^\infty V_{\mathrm{c}}(q',t)\,\mathrm{d}q'\,, \tag{3.6}$$

where $V_{\mathrm{c}}$ denotes the potential induced by the collective effect during a full revolution in the storage ring and scales with the charge involved in the interaction.

## 3.2. Coherent Synchrotron Radiation

The incoherent synchrotron radiation power in Eq. (2.17) is calculated as the sum of the single particle emission of all contributing electrons. Besides the mere number of particles, the emitted radiation also depends on their distribution within the bunch. If the size of the emitting structure, that is, the length of the electron bunch, is in the order of the emitted wavelength ($\sigma_z/\lambda \approx 1$) or even smaller, the particles radiate coherently. This effect can greatly enhance the intensity of the emitted radiation at the low frequency end of the spectrum as shown in Fig. 3.1. The radiated power spectrum of the coherent synchrotron radiation (CSR) is given by [3]

$$\mathcal{P}_{\mathrm{CSR}}(\omega) = N_{\mathrm{e}}(N_{\mathrm{e}}-1)|\widetilde{\varrho}(\omega)|^2\,\mathcal{P}(\omega)\,, \tag{3.7}$$

with the Fourier transform of the normalized bunch profile

$$\widetilde{\varrho}(\omega) = \mathcal{F}(\varrho(t)) = \frac{1}{Q}\int_{-\infty}^{\infty}\rho(t)\mathrm{e}^{-i\omega t}\,\mathrm{d}t\,. \tag{3.8}$$

The combination of the coherent and incoherent emission finally yields the total radiated power spectrum

$$\mathcal{P}_{\mathrm{tot}}(\omega) = \mathcal{P}_{\mathrm{ISR}}(\omega) + \mathcal{P}_{\mathrm{CSR}}(\omega) = N_{\mathrm{e}}\left[1 + (N_{\mathrm{e}}-1)|\widetilde{\varrho}(\omega)|^2\right]\,\mathcal{P}(\omega)\,. \tag{3.9}$$

Because of the quadratic dependency of Eq. (3.7) on the number of electrons within the bunch, the coherent emission of synchrotron radiation can increase the intensity of

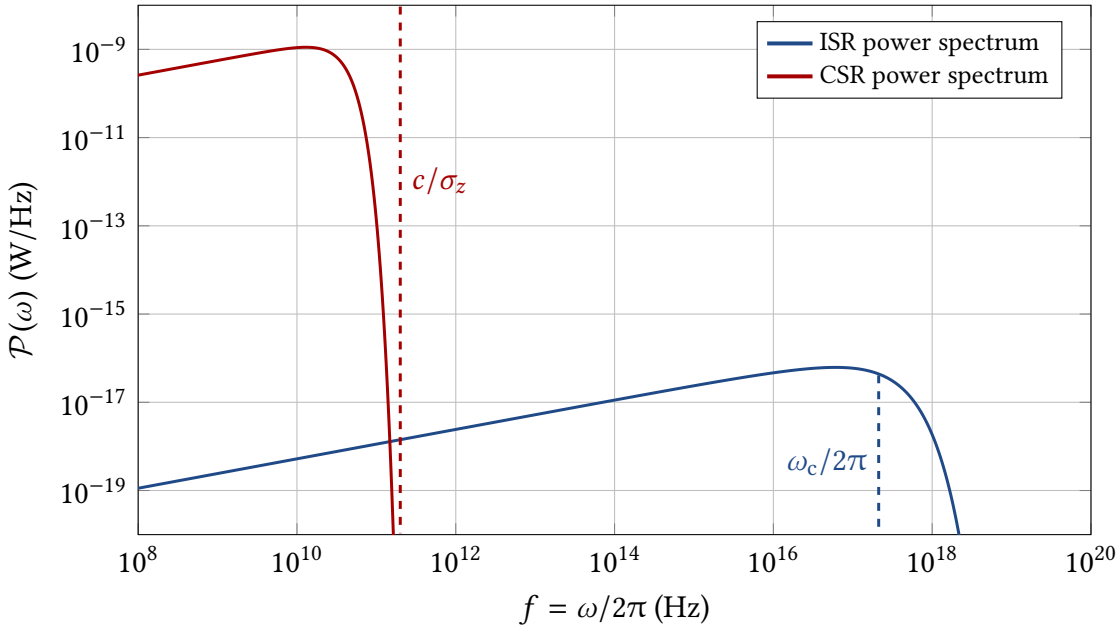



Figure 3.1.: The coherent emission at wavelengths longer than the bunch length, that is, at frequencies below $c/\sigma_z$, leads to a radiated power that exceeds the incoherent synchrotron radiation power by several orders of magnitude. However, the emitted CSR power drops off rapidly at higher frequencies. The shown power spectra are calculated for an exemplary Gaussian-shaped electron bunch with $\sigma_z/c = 5\,\mathrm{ps}$ and $I_{\mathrm{bunch}} = 1\,\mathrm{mA}$ in the short-bunch operation mode of KARA ($E = 1.3\,\mathrm{GeV}$, $R = 5.559\,\mathrm{m}$).

the emitted radiation by several orders of magnitude. As the frequency range covered by the CSR power spectrum directly depends on the bunch length, this creates a clear incentive to push for shorter bunch lengths in the operation of modern synchrotron light sources. However, the increased spatial compression eventually gives rise to a strong self-interaction of the electron bunch with its own emitted CSR, which causes complex longitudinal dynamics and limits the minimal achievable bunch length. The interaction of the electron bunch with the self-generated CSR is possible because of the marginally shorter path of a photon propagating on a straight line compared to the curved path of the electron bunch in a bending magnet, as illustrated in Fig. 3.2. In this way, the coherent

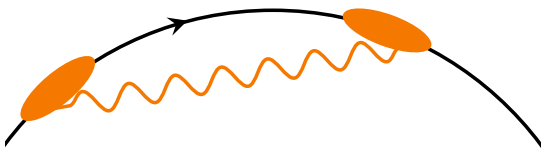


Figure 3.2.: The curved trajectory of an electron bunch in a bending magnet enables the self-interaction of the bunch with its own emitted CSR. By traveling on a straight line, the emitted photons can catch up with the electrons at the head of the bunch.

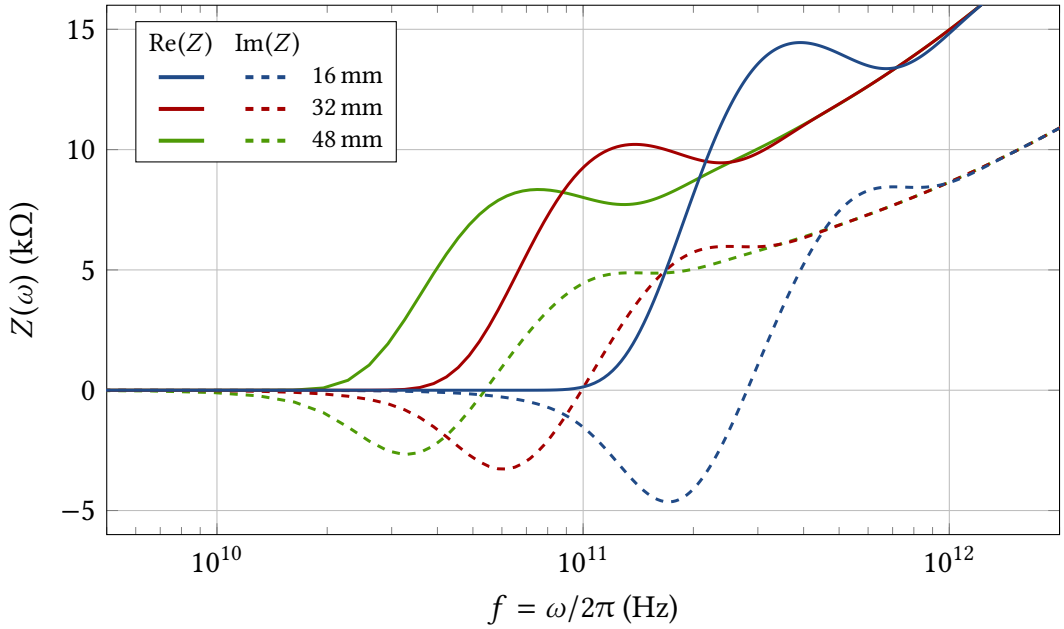



Figure 3.3.: CSR parallel plates impedance for different vacuum gaps. It describes the self-interaction of an electron bunch with its own emitted CSR under the shielding effect of the surrounding vacuum chamber walls. A reduced vacuum gap leads to a stronger shielding, which shifts the cut-off frequency due to the beam pipe to higher frequencies. The beam pipe installed at KARA has a total height of $g = 32\,\mathrm{mm}$ (red curves).

synchrotron radiation emitted by the tail of the bunch catches up with the electrons at the head and causes an energy transfer. This self-interaction is conveniently described by an impedance $Z_{\mathrm{CSR}}(\omega)$ in the frequency domain. Due to the enormous complexity of an exhaustive analytical description, one resorts to approximations of the beam's immediate surroundings. As implied by its name, the free space model [9] neglects any shielding effects by the beam pipe and assumes a circular motion of the electron bunch in vacuum. However, it is crucial to consider the interaction with the vacuum chamber walls as, in this way, the generated wake field may also effect the tail, not just the head of the bunch. To account for the effects of the surrounding vacuum chamber, the parallel plates model introduces two indefinitely extending, perfectly conducting horizontal plates. An approximation for the resulting CSR parallel plates impedance illustrated in Fig. 3.3 can be found in [10, 11]

$$Z_{\mathrm{CSR}}^{\mathrm{PP}}(\omega) \approx \frac{4\pi^2 2^{1/3} R}{\varepsilon_0 c g} \left(\frac{\omega}{\omega_{\mathrm{rev}}}\right)^{-1/3} \sum_p \left[\mathrm{Ai}'(u_p)\mathrm{Ci}'(u_p) + u_p \mathrm{Ai}(u_p)\mathrm{Ci}(u_p)\right] , \tag{3.10}$$

with the Airy functions Ai and Bi, $\mathrm{Ci} = \mathrm{Ai} - i\,\mathrm{Bi}$, the vacuum gap $g$ between the parallel plates and

$$u_p = \frac{\pi^2 (2p+1)^2 R^2}{2^{2/3} g^2} \left(\frac{\omega}{\omega_{\mathrm{rev}}}\right)^{-4/3} . \tag{3.11}$$

While the parallel plates model is a simplistic description of the beam pipe's shielding effect, and additional sources of impedances contribute to the actual beam dynamics in a storage ring, Eq. (3.10) provides a solid basis for describing the influence of the CSR wake field on the longitudinal dynamics of an electron bunch. Given an expression for the impedance, one obtains the CSR-induced wake potential via

$$V_{\mathrm{CSR}}(q,t) = \frac{Q}{2\pi}\int_{-\infty}^{\infty} \widetilde{\varrho}(\omega,t) Z_{\mathrm{CSR}}(\omega) \mathrm{e}^{i\omega q}\,\mathrm{d}\omega\ . \tag{3.12}$$

This additional potential continuously acts back on the electron bunch and causes a perturbation of the simple longitudinal dynamics discussed in the previous chapter. At high bunch currents, this effect eventually gives rise to the micro-bunching instability introduced in section 3.4.

## 3.3. Vlasov-Fokker-Planck Solver

While there is no analytic solution to the VFP equation for the full Hamiltonian defined by Eq. (3.6) and Eq. (3.12), it can be solved numerically on a discretized grid. To that end, the charge density function $\hat{\psi}(q,p)$ is modeled as a two-dimensional discretized distribution in the longitudinal phase space. Given an initial distribution, the VFP equation is solved iteratively in small time steps to simulate the temporal evolution of the charge distribution under the given boundary conditions. At each time step, the longitudinal bunch profile can be calculated by integrating over the generalized energy coordinate. Following Eq. (3.7) one obtains the CSR power spectrum and via Eq. (3.12) the CSR wake potential, which determines the evolution of the charge distribution in the subsequent time step.

Based on the approach in [12], the simulation code Inovesa [11] is a relatively new, massively parallelized implementation of such a VFP solver. Its fast runtime allows for extensive simulation studies using merely standard desktop PCs. With the approximation of the shielding effect via the CSR parallel plates impedance defined in Eq. (3.10), the generated simulation data has shown high qualitative and good quantitative agreement with measurements at KARA [11, 13, 14]. This is an essential finding for the work summarized in this thesis as it renders Inovesa a tool which can be used for two major purposes: First and foremost, it enables dedicated studies to advance the understanding of the longitudinal dynamics underlying the micro-bunching instability. Secondly, considering the overarching objective of extensive control over the micro-bunching dynamics, it can be used to test different interactions with the beam and allows for the development of a feedback system in a well-defined, low-noise environment.

## 3.4. Micro-Bunching Instability

Electron storage rings which provide a short-bunch mode or operate at high bunch currents typically observe a threshold current above which the energy spread of the beam increases abruptly and the emitted radiation starts to fluctuate. At bunch currents clearly above the threshold, the emitted CSR power and other observed beam properties like the bunch length

and the energy spread display a characteristic bursting behavior. Owing to these sawtooth-shaped bursts, the phenomenon was initially referred to as sawtooth instability [15] or bursting CSR [16]. Over the last two decades it has been observed at a wide range of facilities [16–29]. It was quickly realized that the observed bursts were accompanied by a fast beam instability and a model to explain these effects was proposed in [30]. Nowadays, it is generally accepted that the observed behavior is caused by the formation of micro-structures within the bunch arising from the self-interaction of the beam with its own emitted CSR [31]. For typical beam pipes with a diameter of several millimeters, the interaction with the beam, as described by Eq. (3.12), happens primarily in the microwave frequency range. In more recent work, the dynamics are thus referred to as a microwave instability (crucial part of the impedance) or as micro-bunching instability (emphasizing the dynamics within the bunch). Throughout this thesis, the term micro-bunching instability is used.

### 3.4.1. Driving Mechanism

As the CSR-induced wake potential defined in Eq. (3.12) scales with the total charge involved in the interaction, its effect on the longitudinal beam dynamics is heavily dependent on the bunch current. While, at low currents, the strength of the wake potential is small compared to the accelerating RF potential, it grows with increasing current and eventually builds up to a significant perturbation. The sum of the approximately linear RF potential and the CSR wake potential yields the effective potential which the electrons are exposed to during a full revolution in the storage ring

$$V_{\mathrm{eff}}(q,t) \doteq V_{\mathrm{RF}}(q) + V_{\mathrm{CSR}}(q,t)\ . \tag{3.13}$$

Although the interaction with the respective fields occurs at different locations in the storage ring, that is, the RF cavities and the bending magnets, averaging over one revolution is reasonable as the synchrotron motion happens at a much slower time scale than a single passage of the storage ring ($f_{\mathrm{s},0} \ll f_{\mathrm{rev}}$). For bunch currents below the instability threshold, the influence of the CSR wake potential leads to a deformation of the longitudinal charge distribution. Yet, for small bunch currents, an equilibrium between the charge distribution and the resultant CSR wake potential is reached, leading to a stationary charge distribution and wake potential

$$\psi(z,E,t) = \psi(z,E) \quad \text{and} \quad V_{\mathrm{CSR}}(z,t) = V_{\mathrm{CSR}}(z)\ . \tag{3.14}$$

Above the threshold current, however, this is no longer the case. Instead, the continuous CSR self-interaction loop illustrated in Fig. 3.4 leads to a permanently varying charge distribution with dynamically forming and evolving micro-structures in the longitudinal phase space. As the CSR wake potential is derived from the bunch profile, the resultant perturbation of the effective potential is continuously varying as well. At large currents, this volatile interplay between charge distribution and CSR wake potential leads to a self-amplification of the micro-structures forming within the bunch. As the small longitudinal structures represent a high-frequency contribution to the Fourier transformed bunch profile, they overlap with higher values of the CSR impedance and may thereby yield

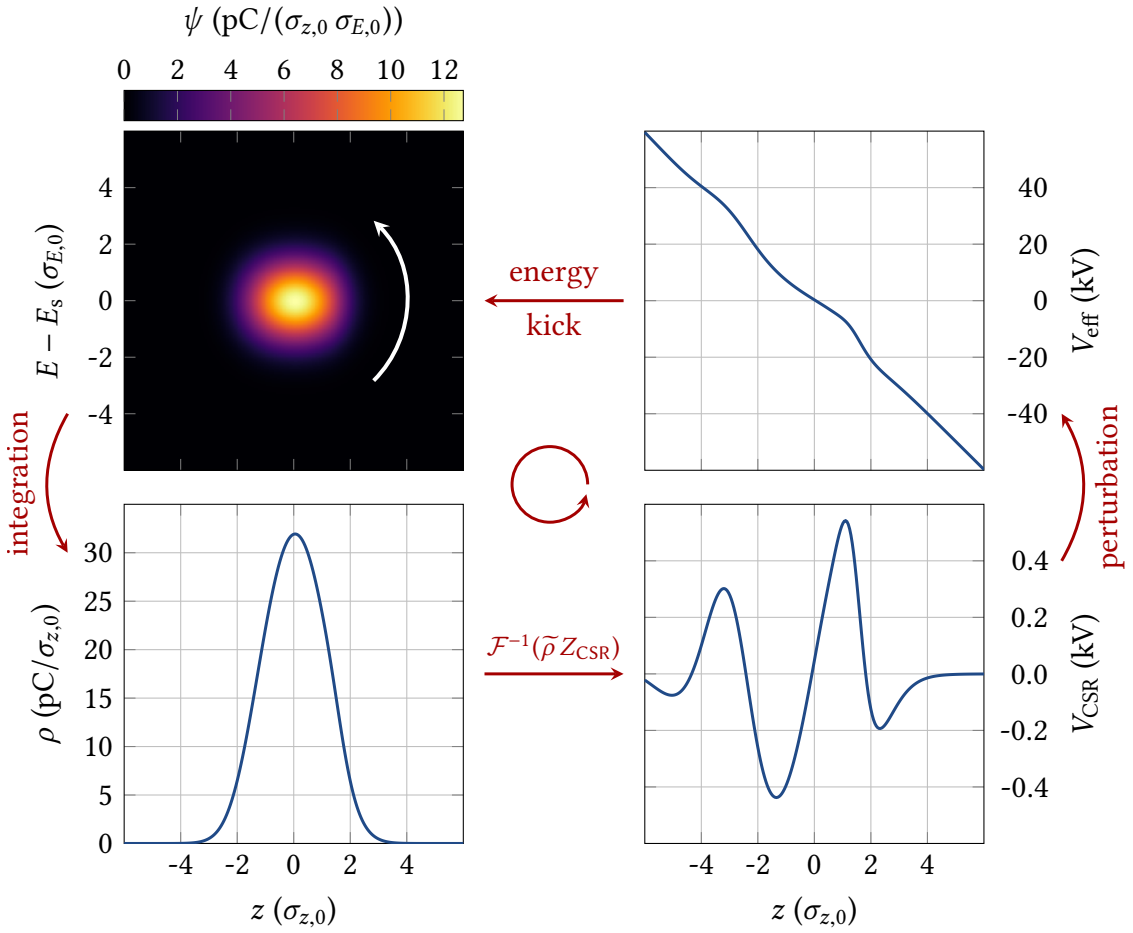



Figure 3.4.: Principle of CSR self-interaction. Given a charge distribution in the longitudinal phase space (upper left), one obtains the bunch profile (lower left) by integrating over the energy. A Fourier transformation, multiplication with the CSR impedance and subsequent inverse Fourier transformation yield the CSR wake potential (lower right). The additional wake potential causes a perturbation of the effective potential (upper right), which affects the particles' energy gain during a full revolution in the storage ring. In this way, the CSR wake potential acts back on the bunch and alters the charge distribution in phase space.

relatively large contributions to the CSR wake potential. This additional part of the wake potential can drive the formation of new micro-structures or enhance the already existing structures in the longitudinal phase space. The synchrotron radiation emitted by a bunch undergoing such dynamics is also continuously varying. As the CSR spectrum in Eq. (3.7) is directly dependent on the bunch profile, spontaneous growth of micro-structures in the longitudinal phase space can lead to a burst of radiation at photon frequencies corresponding to the spatial extent of the structure. As a consequence of the amplification gained by coherent emission this generally also increases the total power radiated by the electron bunch.

The volatility of the driving mechanism that generates the micro-bunching instability is one of the reasons why control of these dynamics is a challenging and delicate endeavor.

### 3.4.2. Characteristic Features

The nature of the CSR self-interaction illustrated in Fig. 3.4 and the scaling of the CSR wake potential with bunch current lead to a few characteristic features of the micro-bunching instability. The current dependency is thereby not limited to a mere instability threshold, but creates rich longitudinal dynamics at higher currents. A concise but distinctive depiction of the current dependency is given by the exemplary CSR power spectrogram shown in Fig. 3.6. Here, the CSR power time signal

$$P_{\mathrm{CSR}}(t) = \int_{\omega_1}^{\omega_2} \mathcal{P}_{\mathrm{CSR}}(\omega, t)\, \mathrm{d}\omega \,, \tag{3.15}$$

shown for an exemplary bunch current in Fig. 3.5, is calculated for each current and the magnitude of its Fourier transform is displayed as a horizontal line to create the overall spectrogram. The color code is defined by the spectral intensity and thus indicates the most dominant fluctuation frequencies. The CSR power signal in Eq. (3.15) is directly correlated to the evolution of the micro-structures within the bunch as, at each point in time, the CSR power spectrum in Eq. (3.7) is determined by the longitudinal bunch profile. Any shift of the fluctuation frequencies thus corresponds to a change in the temporal evolution of the charge distribution in the longitudinal phase space. The most notable differences, as pointed out below, mark transitions into different regimes of the instability, which may serve as a useful categorization of the occurring micro-bunching dynamics. Although these dynamics are dependent on different machine parameters like the beam energy, the accelerating voltage or the momentum compaction factor, the discussed qualitative

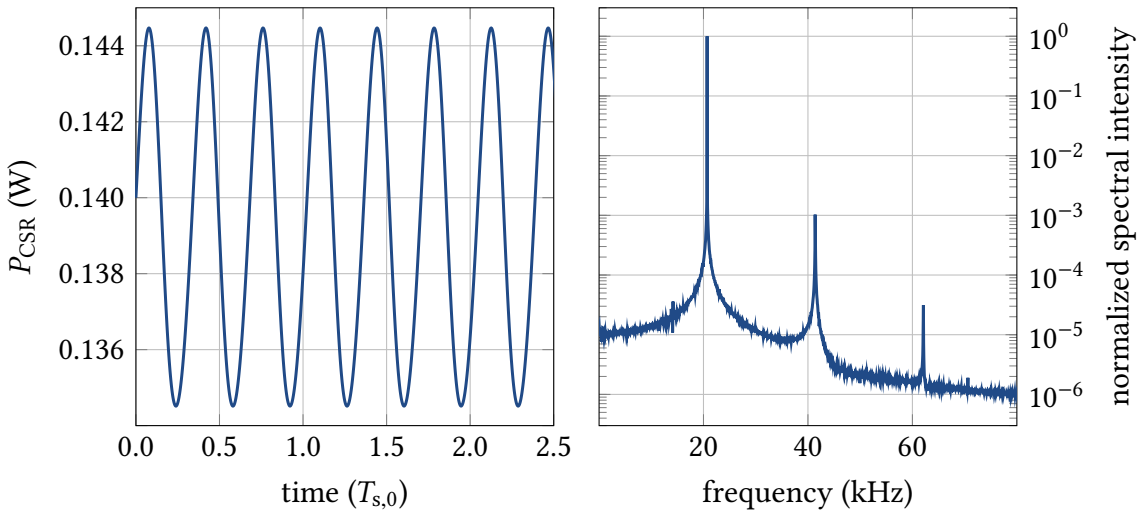



Figure 3.5.: CSR power time signal $P_{\mathrm{CSR}}(t)$ for the exemplary bunch current $I = 115\,\mu\mathrm{A}$ (left) and the magnitude of its Fourier transform $|\widetilde{P}_{\mathrm{CSR}}|$ (right).

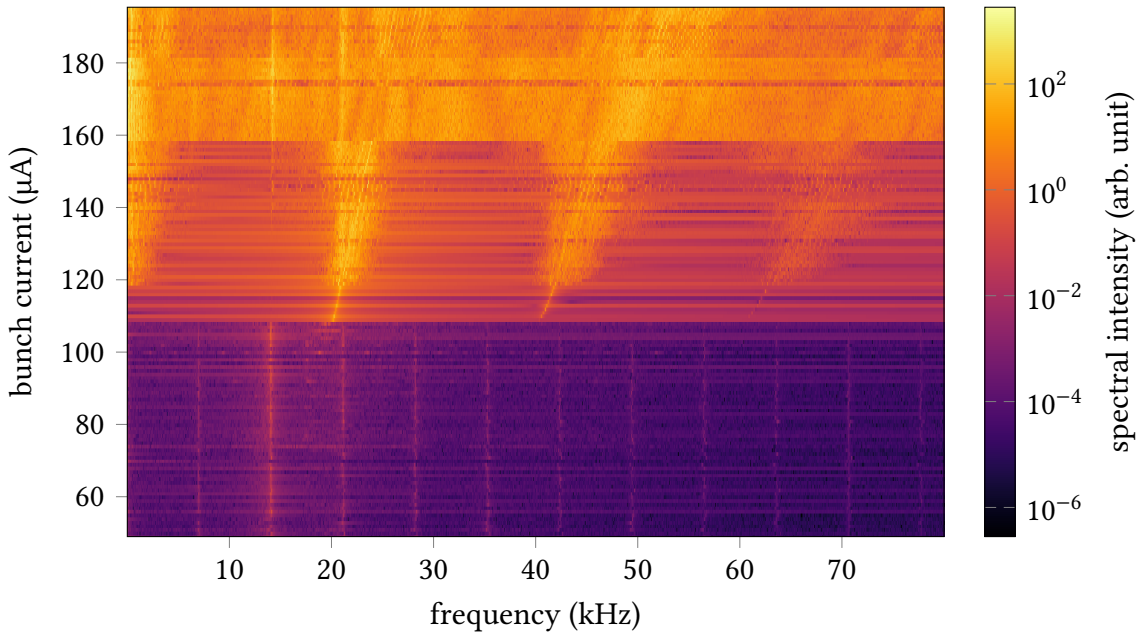



Figure 3.6.: Exemplary CSR power spectrogram, which highlights the dependency of the occurring micro-bunching dynamics on the bunch current. The simulation settings used to generate the shown data correspond to the short-bunch mode at KARA and can be found in appendix A.1 (data set $\mathcal{D}_1$). The instability threshold is found at $I_{\mathrm{th}} = 109\,\mu\mathrm{A}$ with an initial fluctuation frequency of $f_{\mathrm{ms}} = 20.19\,\mathrm{kHz}$. The vertical lines below the threshold current are multiples of the synchrotron frequency ($f_{\mathrm{s},0} = 7\,\mathrm{kHz}$) and indicate minor bunch length oscillations.

characteristics are always observed in both, simulations and measurements, albeit the exact numerical values may differ [32].

### Threshold Current

Regarding the practical operation of electron storage rings and the design of new facilities, the most relevant property of the micro-bunching instability is the threshold current, roughly at $I_{\mathrm{th}} = 110\,\mu\mathrm{A}$ in Fig. 3.6. Although the instability generally doesn't cause sudden beam losses, the longitudinal charge distribution and several derived beam properties start to fluctuate above the instability threshold which may be detrimental to the performance of an accelerator. To reach the required beam stability, it is often desirable to operate at bunch currents below the instability threshold. Then again, the micro-bunching dynamics above the threshold provide intense coherent radiation which may be delivered to dedicated experiments. In any case, precise knowledge of the threshold current is crucial in practical

applications. It has thus been studied theoretically using the VFP formalism, arriving at a predictive model for the instability threshold [8]

$$I_{\mathrm{th}} = \frac{8\pi^2 \varepsilon_0 m_{\mathrm{e}} \gamma c^2 f_{\mathrm{s},0} \sigma_{\delta,0} \sigma_{z,0}^{4/3}}{e R^{1/3}} \left( a_{\mathrm{th}} + b_{\mathrm{th}} R^{1/2} \sigma_{z,0} h^{-3/2} \right) , \tag{3.16}$$

where $h = g/2$ is the half vacuum gap and the parameters $a_{\mathrm{th}} = 0.5$ and $b_{\mathrm{th}} = 0.12$ are determined by a fit to simulated data. Measurements and simulations at BESSY II and MLS [33] as well as KARA [34] have shown good agreement with Eq. (3.16), further validating the model. It is worth noting that for a given beam energy and a fixed geometry (bending radius $R$ and vacuum gap $g$), the predicted threshold current is solely defined by the nominal synchrotron frequency $f_{\mathrm{s},0}$ and the natural bunch length $\sigma_{z,0}$. Alternatively, it is determined by a combination of any two parameters in $\{f_{\mathrm{s},0}, \sigma_{z,0}, \alpha_{\mathrm{c}}, V_0\}$. In practice, the measured values of $V_0$ and $f_{\mathrm{s},0}$ are used to arrive at the desired configuration of the storage ring.

### Regular Bursting Regime

Besides some multiples of the synchrotron frequency, there are no particularly prominent frequencies visible below the instability threshold in Fig. 3.6, which is a consequence of the mostly stationary charge distribution. This changes above the threshold current where a single dominant frequency emerges around 20 kHz. The presence of one dominant frequency in the spectrogram corresponds to a sinusoidal CSR power signal, shown for the exemplary bunch current $I = 115\,\mu\mathrm{A}$ in the upper part of Fig. 3.7. While the main frequency may be accompanied by some of its higher harmonics, this behavior is always observed directly above the threshold and is a characteristic feature of the micro-bunching instability. Simultaneously, the bunch length and the energy spread display a similar behavior, but are opposite in phase. This can be explained by the rotation of the charge distribution in phase space due to the synchrotron oscillation described by the harmonic part of the Hamiltonian in Eq. (3.3). For example, after a rotation of an otherwise stationary distribution by ninety degrees, the distribution displays the same characteristics in the other dimension, respectively.[1] The corresponding temporal evolution of the bunch profile is illustrated in the lower part of Fig. 3.7. While the bunch profile still has a roughly Gaussian shape, there are small deformations visible around the center of the bunch. By subtracting the average bunch profile

$$\overline{\rho}(z) \doteq \frac{1}{n} \sum_{i=1}^{n} \rho(z, t_i) , \tag{3.17}$$

only the non-stationary part remains

$$\Delta\rho(z, t_i) \doteq \rho(z, t_i) - \overline{\rho}(z) , \tag{3.18}$$

which reveals a periodic sequence of minima and maxima. It is worth noting that these extrema are most pronounced at the head of the bunch ($0 < z/\sigma_{z,0} < 2$) and coincide with

[1] Owing to the symmetry of the standard deviation as the square root of the second central moment, that is, $\sigma[\rho(z)] = \sigma[\rho(-z)]$ and $\sigma[\rho(E)] = \sigma[\rho(-E)]$.

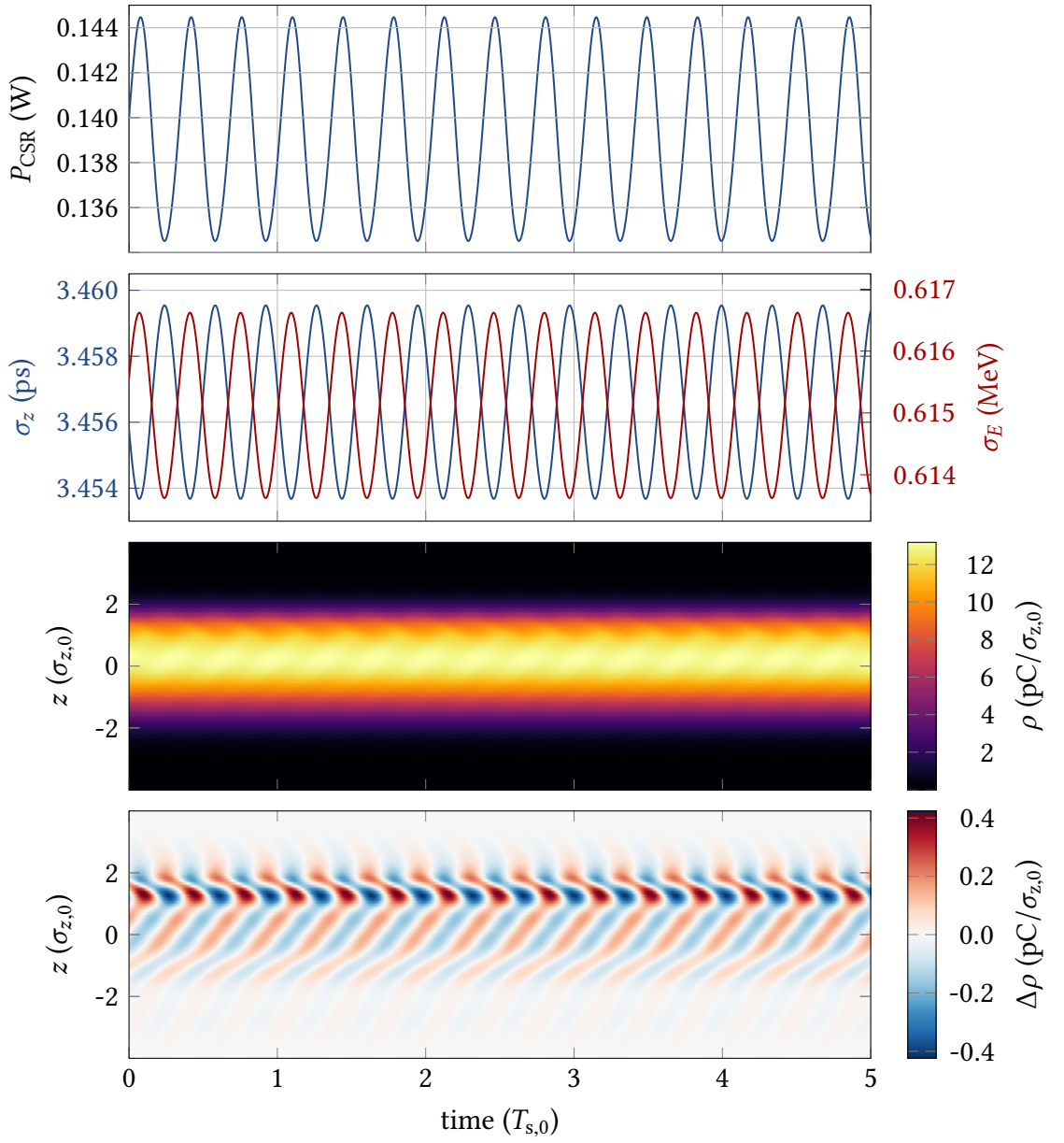



Figure 3.7.: Regular bursting regime ($I = 115\,\mu\mathrm{A}$). Directly above the instability threshold, the emitted CSR power, the bunch length and the energy spread all display a sinusoidal oscillation corresponding to the single dominant frequency in the CSR power spectrogram in Fig. 3.6. These fluctuations are a consequence of the micro-structures arising in the charge distribution, which can be seen in the temporal evolution of the bunch profile shown in the lower part of the figure. The subtraction of the average bunch profile reveals a distinct periodic sequence of minima and maxima.

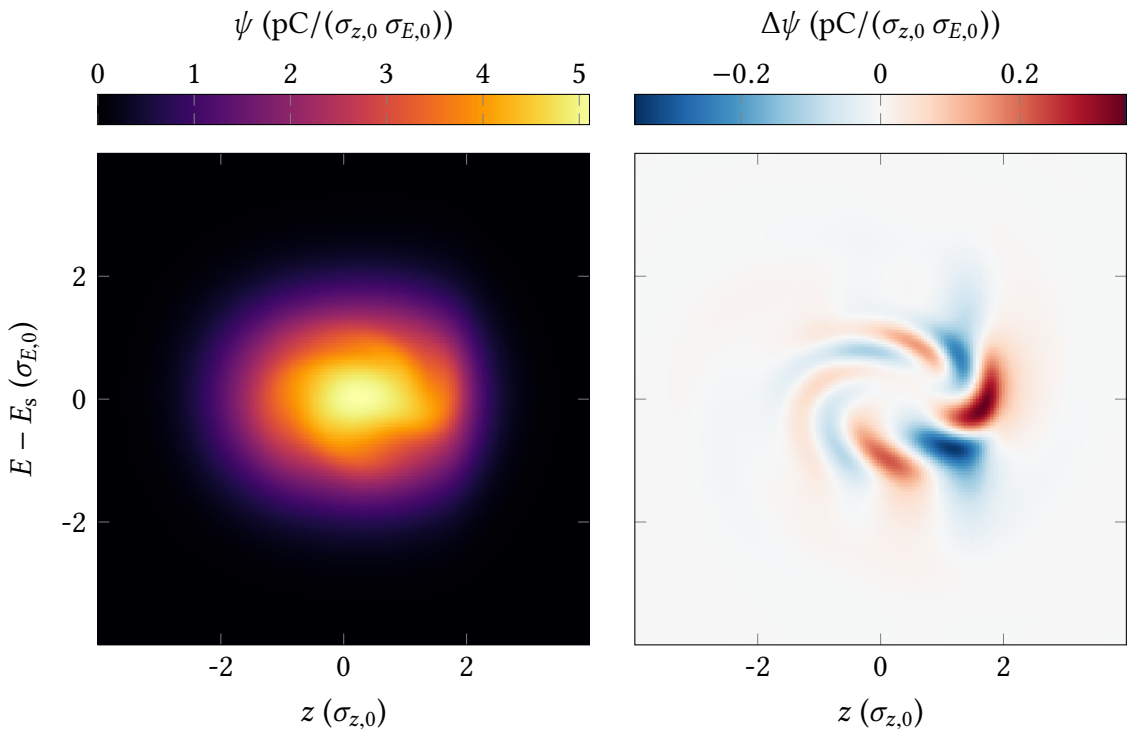



Figure 3.8.: Snapshot of the charge distribution in phase space in the regular bursting regime (left). The deformations of the distribution are emphasized by the subtraction of the temporal average, revealing a characteristic pattern of micro-structures (right).

the minima and maxima of the CSR power signal. All these observations are generated by the dynamic evolution of micro-structures in the longitudinal phase space. Figure 3.8 shows a snapshot of the charge distribution, again with the average distribution

$$\overline{\psi}(z, E) \doteq \frac{1}{n} \sum_{i=1}^{n} \psi(z, E, t_i) \,, \tag{3.19}$$

being subtracted

$$\Delta\psi(z, E, t_i) \doteq \psi(z, E, t_i) - \overline{\psi}(z, E) \,, \tag{3.20}$$

to reveal the occurring micro-structures. Because of synchrotron motion, the structures rotate in phase space and due to the CSR self-interaction continuously evolve over time. Yet, in this regime directly above the threshold current, the micro-structures propagate in a self-sustaining and highly repetitive manner. At each point in time, the micro-structures generate a wake potential which further maintains the present structure. Except for the rotation in phase space and the growth of the micro-structures during their motion from the tail to the head of the bunch, the charge distribution remains mostly constant. After a fraction of the synchrotron period, when the next micro-structure reaches the head of the bunch, the charge distribution is almost identical to before. The time it takes this process to complete corresponds to the oscillation period of the CSR power signal in Fig. 3.7 and the

singular dominant frequency in Fig. 3.8. It will thus be referred to as the micro-structure frequency, initially at about $f_{\mathrm{ms}} = 20\,\mathrm{kHz}$ in Fig. 3.6. The maximum emission of CSR is reached when the state of the micro-structures in phase space is such that it leads to the largest structures on the bunch profile. This typically occurs when two micro-structures are lined up in their projection on the longitudinal axis, leading to the largest contributions to the bunch profile after integrating over the energy. Because of the consistent, periodic nature of the dynamics and the sinusoidal CSR power signal directly above the threshold current, it will be referred to as the regular bursting regime throughout this thesis.

### Sawtooth Bursting Regime

With increasing bunch current the micro-structure frequency in Fig. 3.6 slowly shifts to higher values until it eventually branches out into several contributing frequencies. Simultaneously, there are low frequency contributions emerging at the left edge of the figure (see appendix A.2 for a logarithmically scaled frequency axis). This marks the transition to the sawtooth bursting regime illustrated in Fig. 3.9. Here, the CSR power is emitted in sawtooth-shaped bursts with a relatively slow repetition rate, typically in the range of $f_{\mathrm{burst}} \in [0, 1]\,\mathrm{kHz}$. These bursts are accompanied by an initially decreasing bunch length and energy spread right until the onset of the burst, where both values increase rapidly and eventually reach a maximum shortly after the peak in CSR power emission. At this point in time the charge distribution in phase space is relatively broad and subsequently shrinks again due to radiation damping. Once the bunch length reaches a minimum value, the next burst occurs and the cycle starts anew. The lower part of Fig. 3.9 shows again the corresponding evolution of the bunch profile. While the profiles are relatively smooth in-between bursts, there are strong deformations visible which coincide with the bursts in CSR emission. The subtraction of the average bunch profile reveals again a number of minima and maxima which are most pronounced during the CSR bursts and are washed out immediately afterwards. These dynamics are again generated by the formation and evolution of micro-structures in the longitudinal phase space. The shorter the bunch, the stronger the CSR wake potential, which eventually causes the formation of micro-structures very similar to those occurring in the regular bursting regime, as can be seen in Fig. 3.10. In this case though, the arising micro-structures and the corresponding wake potential lead to a self-amplifying effect, causing the micro-structures to rapidly grow in amplitude. Eventually, the micro-structures reach a maximum amplitude which can no longer be supported by the corresponding wake potential. The micro-structures thus smear out in phase space leading to a smooth but broadened charge distribution.

It is worth pointing out that the main difference to the regular bursting regime is the dynamic growth of the micro-structures over several synchrotron periods. The resulting deformation of the charge distribution is quite comparable, though larger in amplitude, and the underlying driving mechanism in the form of CSR self-interaction is the same.

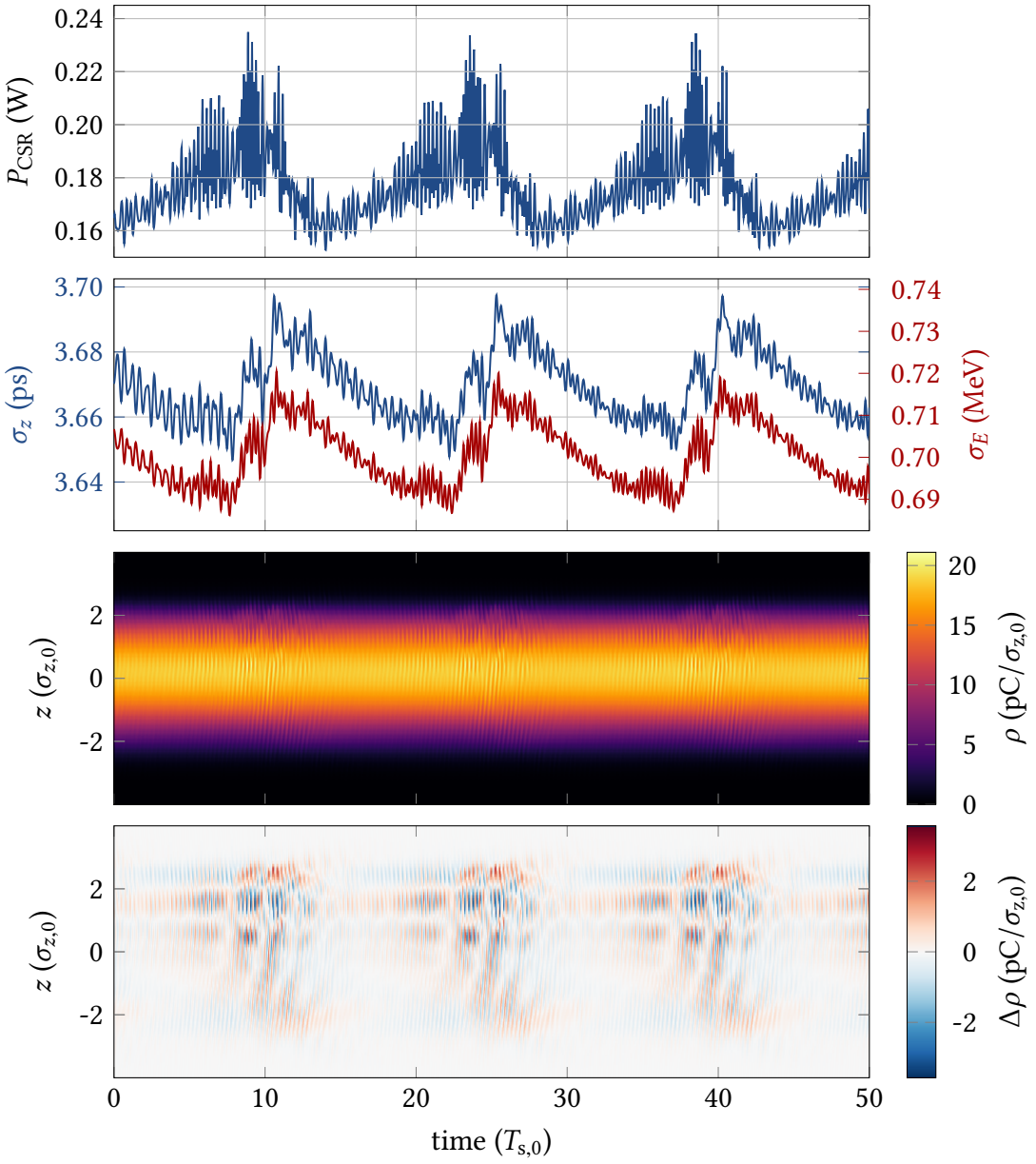



Figure 3.9.: Sawtooth bursting regime ($I = 185\,\mu A$). At large currents above the instability threshold, the CSR power is emitted in sawtooth-shaped bursts, in this case with a repetition rate of $f_{burst} = 0.47\,kHz$. The fast jitter on the signal is caused by the propagation of the occurring micro-structures in phase space at frequencies in the range of $f_{ms} \in [19, 25]\,kHz$. The bunch length and the energy spread increase rapidly during the CSR bursts and are damped afterwards until the onset of the next burst is reached. These dynamics are generated by micro-structures which arise in the charge distribution for short enough bunch lengths and quickly grow in amplitude, causing the increased emission of CSR. After reaching a maximum amplitude, the micro-structures are washed out again leading to a relatively smooth charge distribution in-between the CSR bursts.

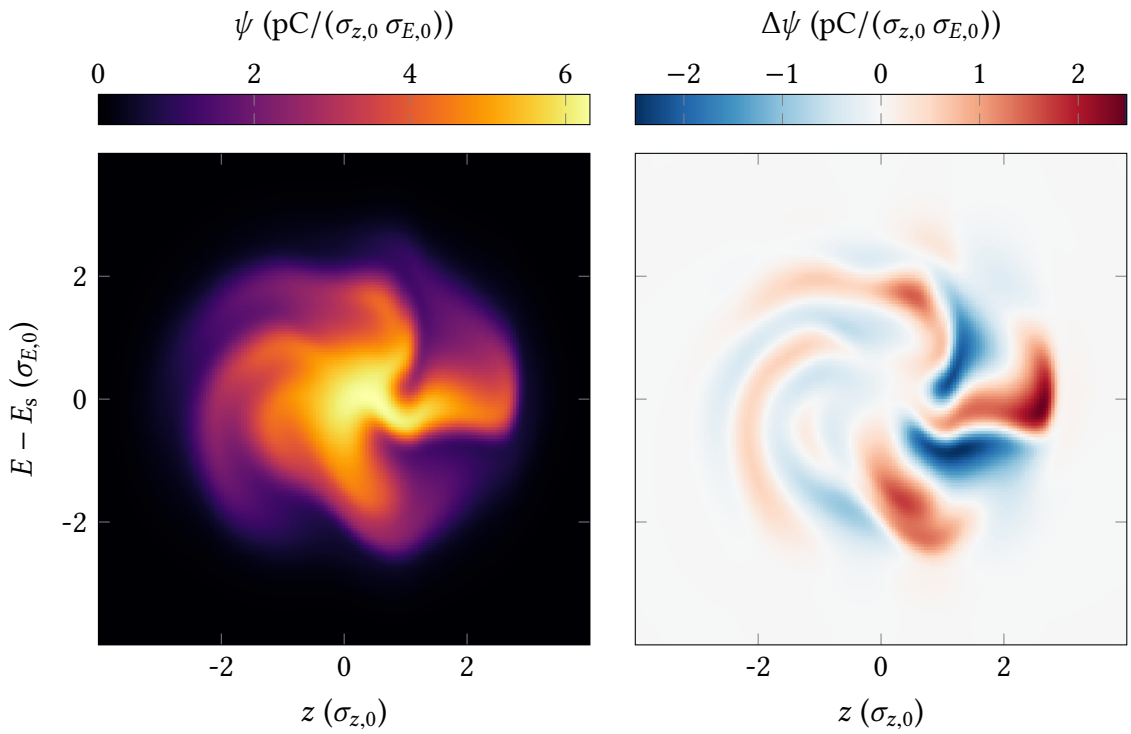



Figure 3.10.: Snapshot of the charge distribution in phase space in the sawtooth bursting regime (left). The subtraction of the temporal average reveals very similar micro-structures compared to Fig. 3.8, though with a significantly larger amplitude (right).

### Short-Bunch-Length Bursting Regime

In some cases, when operating at very small bunch lengths, an additional instability region below the threshold current described by Eq. (3.16) can be observed at KARA. This was already predicted by the theoretical work in [8] and referred to as a weak instability considering its dependence on the longitudinal damping time. It has subsequently been studied in systematic measurements and simulations for KARA, MLS and BESSY II [33, 35] and, though the simulated thresholds were found to be slightly higher than the measured values, the overall behavior is in good agreement. The beam dynamics in this regime are generally comparable to the behavior above the main instability threshold, as one also observes distinct frequencies due to the occurrence of similar micro-structures in the longitudinal charge distribution. However, it is worth noting that the micro-structure frequency in this regime is typically found around the second harmonic of the nominal synchrotron frequency, whereas in the regular bursting regime, it varies. This corresponds to a deformation of the charge distribution which is similar to a quadrupole mode as was found during studies for the SLC damping rings [36] as well as for KARA [37].

### 3.4.3. Mitigation Techniques

Over the course of the past twenty-five years, various efforts towards mitigation of the micro-bunching instability have been undertaken. Nonetheless, to this day, the instability still poses a critical limitation to the operation of electron storage rings with high bunch currents. Although the development of a theoretical description in the VFP formalism and systematic experimental studies have led to a better understanding of the phenomenon, mitigation techniques are still limited to relatively crude and invasive techniques. One of the first attempts to mitigate the instability was the re-design of the entire vacuum chamber of the SLC damping rings and a thorough revision of the design of principal vacuum chamber components for DAΦNE [36]. While a reduction of the impedance budget is a very effective approach, decreasing the strength of the CSR wake potential and therefore increasing the threshold current, this is quite a drastic measure to mitigate the instability at existing machines and poses a major limitation for the design of new synchrotron light sources. Another way to reduce the strength of the CSR self-interaction is to lengthen the electron bunch. This can be achieved in several ways, for example by increasing the momentum compaction factor or by heating the bunch to increase the energy spread [38]. While the former requires a re-design of the magnetic lattice and thereby limits the capabilities to optimize for other beam parameters, the latter simultaneously deteriorates the transverse beam properties. Yet another way to lengthen the bunch is through the RF system, e.g., by using an RF phase modulation or harmonic cavities. In experiments with a strong RF phase modulation at the second harmonic of the synchrotron frequency [39–41], the electron bunch could even be split in two smaller bunchlets leading to a more stable beam. Ultimately, though, the lengthening of the bunch is also not ideal as it limits the capabilities for time-resolved experiments. If there is significant coupling between the longitudinal and the transverse plane it may also affect the transverse beam properties. This is particularly relevant in the development of future diffraction-limited synchrotron light sources of the fourth generation, where the transverse beam size is several orders of magnitude smaller than its longitudinal counterpart. Moreover, there are ongoing studies at KARA to explore the advantages and drawbacks of operating at negative momentum compaction factor, as this may allow a reduction of sextupole magnet strengths in future facilities [42]. Yet, as reported in [43], the current-dependency of the bunch length is found to differ significantly from the operation at positive momentum compaction factor, which may lead to a lower threshold of the micro-bunching instability. In a more recent attempt to mitigate the instability [44], a linear feedback between the measured CSR power and the accelerating voltage was used at SOLEIL to continuously act on the bunch via the RF system. The approach succeeded in mitigating the sawtooth-shaped CSR bursts at lower bunch currents, creating beam dynamics which resemble those in the regular bursting regime. This can also be achieved via different ways of reducing the relative strength of the CSR self-interaction, for example by reducing the accelerating voltage or the momentum compaction factor, which shifts the threshold current and the different instability regimes to higher currents. The approach pursued in this thesis differs from its predecessors as it does not attempt to reduce the strength of the CSR wake potential, but tries to cope with its effects on the beam dynamics by counteracting the most crucial parts of the CSR-induced perturbation. It does so by addressing the phenomenon in a more

subtle way, aiming to smoothen the micro-structures that form in the longitudinal charge distribution without lengthening the bunch. The general concept is derived from a careful analysis of the perturbed synchrotron motion under the influence of CSR self-interaction discussed in chapter 5 and formalized as a reinforcement learning problem in chapter 6.

# 4. Reinforcement Learning

> *I believe there is no philosophical high-road in science, with epistemological signposts. No, we are in a jungle and find our way by trial and error, building our road behind us as we proceed.*
>
> — Max Born, *Experiment and Theory in Physics*

Machine learning (ML) is a subfield of computer science which emerged around the 1950s. In its most general sense, the term refers to the computational task of identifying patterns in data, which usually, but not necessarily, informs some form of decision process when confronted with an unseen problem. Especially over the last decade, machine learning techniques have increasingly found their way into real world applications deeply impacting a large variety of industrial fields. As the capabilities of these algorithms are utilized more effectively across a growing number of domains, machine learning is widely perceived as one of the most disruptive technologies of our time. Ambitious applications in health care, finance, web-based services and autonomous driving are likely to not only transform their respective domains, but society at large. Scientific research in general and physics in particular are no exception to this development. In fact, the general approach of collecting data to design and inform models which predict the behavior of complex systems is a remarkable commonality of the two disciplines. Yet, one apparent distinction is the general aspiration in physics to not only build powerful models, but to simultaneously develop an understanding about the underlying mechanisms. Contrary to this, ML models are commonly quite opaque with respect to the human capability to understand the precise way data patterns are recognized and exploited. Reconciliation of these conflicting aspects is a major challenge for a broader application of ML techniques in modern physics. However, the issue may be addressed in different ways, e.g., by selecting and refining ML approaches which are more suitable for the demands in physics or by improving the means of analysis regarding the trained models. In any case, as scientific research becomes more and more data-driven, ML techniques can be expected to grow into a cornerstone of modern physics. Some of the recent developments and opportunities in the physical sciences are reviewed in [45], and specifically for particle accelerators in [46].

Machine learning approaches are typically divided into three basic categories according to their learning paradigms: supervised learning, unsupervised learning and reinforcement learning. The three subfields aim to tackle different classes of problems and mainly differ in the way directives are given to guide the respective training process. Supervised learning operates under the premise of existing examples with inputs and outputs, where the latter is determined by some form of supervisor, also called the teacher. Based on the provided data, the goal is to learn a general function which maps inputs to outputs and may thus be applied to make predictions on new inputs. The key difference in unsupervised learning is

that there are no given target outputs, that is, no supervisor. Operating solely on some form of input, the goal is to discover intrinsic structure within the data. Successful applications of unsupervised learning may reveal hidden patterns and thereby provide new information about the data or guide a subsequent decision process. While the learning paradigm of supervised learning may be described as learning by example, unsupervised learning could be interpreted as learning by observation [47]. Reinforcement learning (RL) differs from the other two subfields by not requiring a pre-existing data set. Instead, the learning process is based entirely on the interaction with a dynamic environment. Thereby, the goal is defined as the maximization of a scalar reward signal and all learning takes place in an iterative process based on the general concept of trial-and-error search. Although machine learning terms are notoriously difficult to define, a more thorough attempt at describing the domain of reinforcement learning is made in the next section.

As will become clear after the chapters 5 and 6, the general task of micro-bunching control can be approached quite effectively from a reinforcement learning perspective. The following chapter therefore provides a brief introduction to the field, specifies the formal definition of an RL problem and covers a selection of modern RL algorithms. The treatment and notation throughout this thesis is largely based on [48], which is an excellent textbook that covers the subject in much more detail. The chapter concludes with a brief review of ongoing RL efforts in the field of accelerator physics and future opportunities.

## 4.1. Defining Aspects

Reinforcement learning is the computational approach to goal-directed learning from interaction. The learner or decision maker, usually called the agent, iteratively interacts with an environment while seeking to improve its behavior, as illustrated in Fig. 4.1. At each time step $t$, the agent perceives the current situation, the state $S_t$ of the environment and performs an action $A_t$. Based on the chosen action, the agent receives a scalar reward $R_{t+1}$ and finds itself in a new state $S_{t+1}$. The agent's overall goal is defined as to maximize the cumulative reward over time.

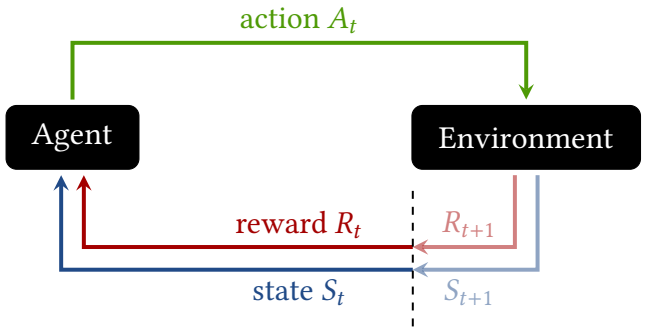



Figure 4.1.: Reinforcement learning is based entirely on the interaction between an agent and an environment. At each iteration, the agent receives a state $S_t$ and reward $R_t$ and executes an action $A_t$. Figure adapted from [48].

One may think about the reward as analogous to the experiences of pleasure and pain in biological systems, where a pleasurable response may reinforce a certain behavior (carrot-and-stick principle). In that sense, reinforcement learning is largely inspired by and probably the closest ML has gotten to the form of learning that is exercised by humans and other animals. It thus has long-standing connections to the fields of neuroscience and psychology, which remain an active area of research.

A key issue that reinforcement learning is concerned with is the trade-off between exploration and exploitation. To discover which strategy yields the highest cumulative reward, the agent has to exploit the knowledge gained by prior experience, but simultaneously needs to explore new actions to improve upon its current strategy. As the brute force method of testing all possible strategies is almost never feasible, the agent has to progressively favor the actions it considers best to narrow down the problem while maintaining sufficient exploration to further improve its decisions.

Moreover, as the overall objective is defined by the maximum cumulative reward, the RL agents have to be able to plan ahead and consider giving up immediate reward at the benefit of a total amount of reward which is higher, but delayed in time. In that sense, they are capable of making sacrifices in the present to improve their chances of reward in the future.

As the training data is generated by the agent's interaction with the environment, the data is sequential and typically not independently and identically distributed. Particularly when neural networks are involved, modern RL algorithms thus incorporate measures to reduce the correlation between samples in order to improve the training process.

Another defining aspect of reinforcement learning is that it explicitly considers the whole problem of finding an optimal strategy in a dynamic and potentially stochastic environment. This distinguishes reinforcement learning from other types of machine learning, which typically operate on more narrow problem definitions. The overall problem reinforcement learning is concerned with is not just a subproblem in the overarching endeavor to reach artificial intelligence (AI). In its most general form, the RL problem is the AI problem [49].

From the very beginning, reinforcement learning has also had a close and mutually beneficial relationship with games. That is partly because of the relative ease with which many games, from the simplest to the most complex, can be framed as a formal RL problem. The reward as some form of score, the state as the current board position and the legal moves as the available actions are frequently applicable, straightforward definitions of the essential RL elements. On the one hand, games offer rich and challenging test scenarios for the development and refinement of RL algorithms. On the other hand, RL-based approaches have led to some of the most powerful computer-based players ever created, which has had a lasting impact on the way these games are played at the highest level. An early example is Samuel's checkers player from 1959 [50], which was pioneering work that already made use of fundamental RL concepts like temporal-difference learning. Perhaps some of the historically most influential work for the application of reinforcement learning in games was the development of TD-Gammon in 1992 [51], which combined temporal-difference learning with neural networks to achieve grandmaster level of play in the game of backgammon. Only recently in 2016, an RL-based program called AlphaGo [52] managed to beat one of the world's best human players in the game of Go. Owing to

the enormous search space and the difficulty to define a position evaluation function, achieving mastery in the game of Go was viewed as a grand challenge for AI and it was expected to take many years more to reach this level of play. AlphaGo managed to defy the odds by combining supervised learning from human expert games with reinforcement learning from games of self-play. It built upon the success of TD-Gammon and recent progress that had been made on playing Atari games with an algorithm called DQN [53], which made use of deep convolutional neural networks. AlphaGo combined these concepts with a novel version of Monte Carlo tree search to improve its selection of moves. A year later, the exceptional performance of AlphaGo was even surpassed by a revised version of the program called AlphaGo Zero [54]. In contrast to its predecessor, AlphaGo Zero used no human data or expertise beyond the basic rules of the game and learned exclusively from self-play reinforcement learning. The universality of the used approach allowed a generalized version, named AlphaZero [55], to simultaneously achieve superhuman performance in the games of chess and shogi, as well as Go. As a consequence of the strict focus on self-play in training, these programs didn't merely differ in proficiency from human play, but also in their style of play. Particularly in the game of Go, this has led to a process of questioning and rethinking of long-standing, well-established Go theory, which intriguingly, is an example of humans learning from a black box type ML system, discussed in more detail in appendix A.3. The remarkable success of AlphaGo has clearly led to a new wave of attention for the field in recent years.

To conclude this section, the following quotation from [48] summarizes the core idea and some of the most defining aspects of reinforcement learning:

> "Reinforcement Learning is learning what to do – how to map situations to actions – so as to maximize a numerical reward signal. The learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them. In the most interesting and challenging cases, actions may affect not only the immediate reward but also the next situation and, through that, all subsequent rewards. These two characteristics – trial-error-search and delayed reward – are the most distinguishing features of reinforcement learning."

## 4.2. Formal Definitions

In all applications of machine learning, and especially in the case of reinforcement learning, it is of the utmost importance to have a precise definition of the problem that one is concerned with. Only in that case can general theoretical statements be derived and different algorithms be compared in their performance. This section thus formalizes the RL problem in a mathematical sense and introduces some of the core terminology that is used to discuss the different aspects of solution methods.

The interaction between the agent and the environment, illustrated in Fig. 4.1, is formally described as a Markov decision process (MDP). An MDP is the mathematical formalization of a sequential decision process in which actions influence not just immediate rewards, but also the subsequently encountered situations or states, and through that also future rewards. The involved decisions thus require a trade-off between immediate reward and

the long-term perspective, that is, delayed reward. Over a sequence of discrete time steps $t = 0, 1, 2, 3, \ldots$, the agent iteratively receives a state $S_t \in \mathcal{S}$ of the environment and is tasked with the decision to chose an action $A_t \in \mathcal{A}$. After the starting state and the first action, and in part as a consequence of its decisions, the agent also receives a numerical reward $R_t \in \mathcal{R} \subset \mathbb{R}$. The experience collected by the agent is thus always a sequence starting with

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots \ . \tag{4.1}$$

The transition from one state to another accompanied by a certain reward can be described by a probabilistic function that determines the dynamics of the MDP. For a particular pair of values, $s' \in \mathcal{S}$ and $r \in \mathcal{R}$, the probability for those values occurring at time step t is

$$p(s', r \mid s, a) \doteq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\} \ , \tag{4.2}$$

with

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r \mid s, a) = 1 \quad \text{for all } s \in \mathcal{S} \text{ and } a \in \mathcal{A} \ . \tag{4.3}$$

This probabilistic function, which is conditioned only on the immediately preceding state and action, $S_{t-1}$ and $A_{t-1}$, fully characterizes the dynamics of the environment. The restriction of the function to depend only on the preceding time step is known as the Markov property, which is best viewed as a condition on the state:

> A state signal that succeeds in retaining all relevant information (regarding the future) is said to be Markov, or to have the Markov property.

This is a pivotal formal condition of all RL problems, which ensures that the agent has sufficient information about the environment to make its decisions. As the transition dynamics of the environment do not, for any situation encountered, depend on its history, knowledge of the current state is sufficient to choose that action which optimizes the future reward. Intuitive examples, where the Markov property is easily fulfilled, are the games of Go and chess. At any point in these games, all that matters for the best course of action and the outcome of the game is the current position on the board, not how it came about.[1] In contrast, finding a state definition which satisfies the Markov property can be more difficult in card games. In those games it is frequently important to track which cards were played already and which remain hidden in the players' hands or in the deck. All of this needs to be accounted for by a valid definition of the state. Yet, in many real world examples, the agent's perception of the environment may be restricted to some form of sensory input, which results in imperfect information. Even in cases where the considered system may be fully described by an MDP with a state definition that satisfies the Markov property, these states may only be partially observable to the agent. In a partially observable Markov decision process (POMDP), which is a generalization of an MDP, the agent receives only observations $O_t \in \mathcal{O}$ instead of the true states of the environment.[2] For any particular state $s \in \mathcal{S}$ of the environment, the corresponding

[1] The minor exceptions being related to conditions for draws by repetition.
[2] This differs from [48] which contains a slightly different definition of POMDPs.

observation $o \in \mathcal{O}$ received by the agent is determined by a probabilistic function, which may only depend on the current state $s$, not on its history

$$p(o \mid s) \doteq \Pr\{O_t = o \mid S_t = s\} \,. \tag{4.4}$$

In a POMDP, the agent's task is to map observations to actions with the unvaried goal to maximize the cumulative reward. In the optimal solution, the agent takes the best action for each possible observation (or belief over the state of the environment).

The agent's way of behaving is formally defined by the so-called policy, a mapping from perceived states (or observations) to actions. While it may simply be a deterministic function $\pi : \mathcal{S} \to \mathcal{A}$, $\pi(s) = a$ in some cases, the more general case is a stochastic policy, where $\pi(a|s)$ denotes the probability of taking action $A_t = a$ when encountering the state $S_t = s$. The policy is the core part of any RL agent, as it alone is sufficient to determine its behavior. The agent's goal may be described as finding the policy which yields the largest amount of reward.

The high degree of generality with which the formal RL problem is defined makes it applicable to a large number of different tasks and domains. This notion is expressed in the reward hypothesis which states [48]:

> "That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward)."

The goal, that is to be pursued by an RL agent, is always defined in the form of a scalar reward function

$$r(s, a, s') \doteq \mathbb{E}\left[R_t \mid S_{t-1} = s, A_{t-1} = a, S_t = s'\right] = \sum_{r \in \mathcal{R}} r \frac{p(s', r \mid s, a)}{p(s' \mid s, a)} \,. \tag{4.5}$$

This is a powerful formalism, however, in some applications, where the definition of the reward is not immediately apparent, it can be difficult to express the goal in just a single scalar function. While one might be able to describe the overall intention, formalizing that idea in a precise mathematical function can be a challenging task, but one that it is essential to the success of any RL-based endeavor.

In the simplest case, the agent will maximize the return, which is simply the sum of the rewards

$$G_t = R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T \,, \tag{4.6}$$

where $T$ denotes a terminal time step. This is a valid approach in applications that are episodic tasks, that is, those which come with a natural notion of a terminal state, such as the end of a game. Although individual episodes may end in different ways, the next episode begins again in some starting state which is independent of the previous episode. In contrast, a continuing task is one which does not naturally break down into individual episodes, but continually goes on without limit. In this case, the return defined in Eq. (4.6) is problematic as the sum of rewards can easily become infinitely large (for example with a reward of +1 at every time step). The discounted return

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,, \tag{4.7}$$

thus incorporates a discount factor $0 \leq \gamma \leq 1$ to ensure convergence. The choice of $\gamma = 1$ restores the undiscounted reward defined in Eq. (4.6), while $\gamma < 1$ guarantees convergence as long as the reward sequence $\{R_t\}$ is bounded. In the extreme case of $\gamma = 0$, the agent focuses exclusively on the immediate reward and neglects any further consequences of its actions on the long-term perspective. The discount factor thus directly affects the agent's trade-off between immediate and delayed rewards. In this way, the discount factor is not a free parameter, but its value changes the overall goal pursued by the agent. A policy which is optimal for the discount factor $\gamma$ may not be optimal for a different value $\gamma'$. The discount factor is thus best viewed as a part of the problem definition, not as a tunable parameter of the solution method.

In order to track the objective pursued by the agent, almost all RL algorithms estimate some form of value function. A value function quantifies how valuable it is for the agent to be in a given state (or to perform a particular action in a given state) in terms of the future reward that can be expected, that is, the expected return. This value always depends on the current policy followed by the agent, which can easily be illustrated by considering again the example of games like chess or Go. Unless the agent knows how to convert a favorable position into a win, the position is clearly not as valuable as it may be otherwise. With suboptimal play, the agent may make a mistake which leads to a draw or even a loss. The expected outcome of the game (or expected return) thus depends not only on the position, but also on the agent's play defined by its policy. Hence, the state-value function for policy $\pi$ is defined as the expected return when starting in state $s$ and following policy $\pi$ thereafter

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right] . \tag{4.8}$$

To help with the agent's task of choosing actions, it can be more convenient to define a function which describes the value of taking the action $a$ when encountering a particular state $s$. The action-value function for policy $\pi$ is thus defined as the expected return when starting in state $s$, taking action $a$, and following policy $\pi$ thereafter

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right] . \tag{4.9}$$

The relation between the two value functions is simply

$$v_\pi(s) = \sum_a \pi(s|a)\, q_\pi(s,a) . \tag{4.10}$$

To estimate these functions one typically makes use of the recursive relationship between the value of a state $S_t = s$ and the value of its successor state $S_{t+1} = s'$. As the return defined in Eq. (4.7) can be decomposed into the immediate reward and the discounted return of the successor state

$$G_t = R_{t+1} + \gamma\,(R_{t+2} + \gamma R_{t+3} + \dots) = R_{t+1} + \gamma G_{t+1} , \tag{4.11}$$

the same can also be done for the value function

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a) \left[ r + \gamma v_\pi(s') \right] . \end{aligned} \tag{4.12}$$

Equation (4.12) is known as the Bellman equation for $v_\pi$. It expresses the value of a given state $s$ in terms of the immediate reward and the values of possible successor states under policy $\pi$ weighted by the transition probabilities of the environment. Analogously, one obtains the Bellman equation for the action-value function

$$q_\pi(s, a) = \sum_{s',r} p(s', r \mid s, a) \left[ r + \gamma \sum_a \pi(s', a')\, q_\pi(s', a') \right] . \tag{4.13}$$

While there are RL algorithms which dispense with the idea of a value function and rely exclusively on direct policy search, most algorithms incorporate some form of value estimation. The Bellman equation forms the basis of a number of ways to approximate value functions and thereby gives rise to different learning concepts, which are introduced in the next section. The optimal state-value function

$$v_*(s) \doteq \max_\pi v_\pi(s) \quad \text{for all } s \in \mathcal{S} , \tag{4.14}$$

describes the maximum return that can be expected for any policy. A policy $\pi$ is optimal if and only if it attains the same amount of expected return, $v_\pi(s) = v_*(s)$ for all $s \in \mathcal{S}$. An optimal policy $\pi_*$ has also an optimal action-value function

$$q_*(s, a) \doteq \max_\pi q_\pi(s, a) \quad \text{for all } s \in \mathcal{S} \text{ and } a \in \mathcal{A} . \tag{4.15}$$

The Bellman equation for optimal value functions can be written in a simpler form, without reference to a particular policy. The Bellman optimality equation for $v_*$ is

$$v_*(s) = \max_a \sum_{s',r} p(s', r \mid s, a) \left[ r + \gamma v_*(s') \right] , \tag{4.16}$$

and for $q_*$

$$q_*(s, a) = \sum_{s',r} p(s', r \mid s, a) \left[ r + \gamma \max_{a'} q_*(s', a') \right] . \tag{4.17}$$

In theory, the Bellman optimality equations for all states $s \in \mathcal{S}$ form a set of $n$ equations with $n$ unknowns and may thus be solved by a variety of methods for solving systems of nonlinear equations. Given either $v_*$ or $q_*$, it is relatively easy to define an optimal policy by acting greedily, that is, by always choosing the action that corresponds to the highest value, e.g.

$$\pi_*(s) = \arg\max_a q_*(s, a) . \tag{4.18}$$

In practice, it is rarely feasible to solve RL problems in this way as it relies on a number of strict assumptions which are frequently violated, like perfect fulfillment of the Markov property or precise knowledge of the environment's dynamics. One might also simply not have sufficient computing resources to solve the problem within a reasonable amount of runtime. While a well-defined notion of optimality is helpful to organize and guide different approaches towards learning, in reinforcement learning applications one commonly has to settle for approximate solutions.

Another concept which may help with the overall objective of finding good policies is the use of a model.[3] In reinforcement learning, a model approximates the dynamics of the environment and thereby allows for predictions of the environment's response to actions taken by the agent. The use of a model provides RL agents with the opportunity to consider future situations before they are actually encountered. This may inform their course of action ahead of time and is called planning. RL algorithms that make use of models and planning are called model-based methods, whereas those that do not are called model-free methods. In model-based methods, the experience collected by the agent in the actual environment can be augmented by simulated experience using the model, which can greatly increase the sample efficiency compared to model-free approaches.

## 4.3. Learning Concepts

Given the above definition of the formal RL problem, this section addresses the question of how to find better policies, that is, how to improve the agent's behavior until the problem can be considered solved. In general, the answer to this question depends heavily on the type of information which is available to the agent and on the nature of the state and action space, $\mathcal{S}$ and $\mathcal{A}$. To introduce some of the most fundamental learning concepts, this section focuses on the special case of a finite MDP, where the state, action and reward sets, $\mathcal{S}$, $\mathcal{A}$ and $\mathcal{R}$, are all finite. Section 4.4 then extends these ideas to the more general case of continuous state and action spaces.

For cases where the transition dynamics of the MDP $p(s', r \mid s, a)$ are known, there is a collection of algorithms, referred to as dynamic programming (DP), which can be used to compute optimal policies. Because of their requirement of a perfect model and their great computational expense, these methods are of limited utility in practical RL applications. However, they provide a good foundation for understanding the methods introduced later on. Dynamic programming and a large part of reinforcement learning methods in general build upon the idea of using a value function to guide the search for good policies. In a first step, one is thus concerned with finding the value function $v_\pi$ for a given policy $\pi$, which is called policy evaluation or the prediction problem. Although applying the Bellman equation in Eq. (4.12) to all states $s \in \mathcal{S}$ leads to a system of linear equations which may be solved for $v_\pi$ by different methods, one is more interested in iterative solution

[3] In contrast to the broad use of the term *model* across the domain of machine learning, it has a quite specific meaning in reinforcement learning which is related to the capability of agents to use planning. A *model-free* RL method may still make use of neural networks or other modeling concepts which are not directly related to planning.

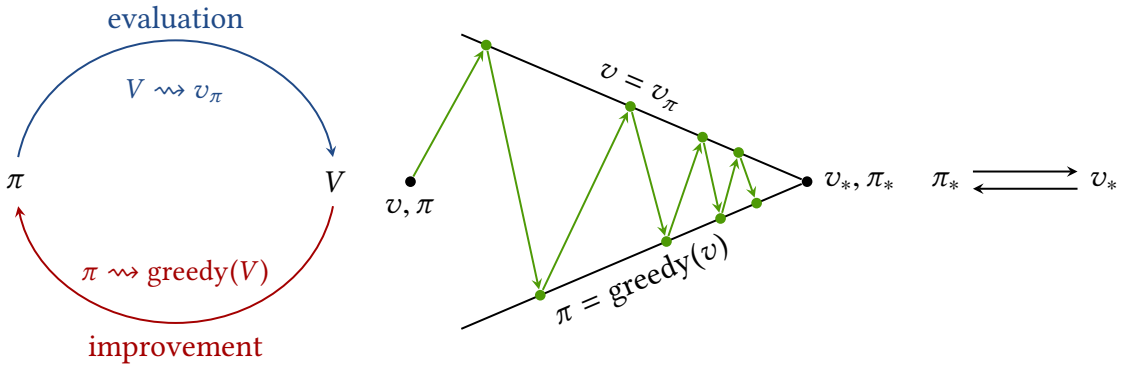



Figure 4.2.: The term generalized policy iteration refers to the concept of alternating between policy evaluation and policy improvement steps (left). Starting with an arbitrary policy $\pi$ and value function $v$, the estimate of the value function is improved towards the true value function $v_\pi$ and then used to improve the policy by making it greedy with respect to $v$ (middle). The interaction between the two processes eventually leads to a stable joint solution, the optimal value function $v_*$ and an optimal policy $\pi_*$ (right). Figure adapted from [48].

methods due to their scalability. Starting with an initial, arbitrarily chosen estimate $v_0$, the Bellman equation can be used as an update rule

$$\begin{aligned} v_{k+1}(s) &\doteq \mathbb{E}[R_{t+1} + \gamma v_k(S_{t+1}) \mid S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r \mid s, a) \left[ r + \gamma v_k(s') \right] \quad \text{for all } s \in \mathcal{S} , \end{aligned} \tag{4.19}$$

to obtain a sequence of approximations $\{v_k\}$, which can be shown to converge to $v_\pi$ as $k \to \infty$. The algorithm defined by Eq. (4.19) is called iterative policy evaluation. Given an estimate of the value function for the currently followed policy, one can simply improve the original policy $\pi$ by making it greedy with respect to the original value function $v_\pi$. This process is referred to as policy improvement. In case of an arbitrary deterministic policy $\pi(s)$, the new greedy policy may be defined as

$$\begin{aligned} \pi'(s) &= \arg\max_a q_\pi(s, a) \\ &= \arg\max_a \sum_{s',r} p(s', r \mid s, a) \left[ r + \gamma v_\pi(s') \right] . \end{aligned} \tag{4.20}$$

An algorithm that alternates between these two processes, policy evaluation and policy improvement, can be expected to converge to the optimal value function and an optimal policy. The general idea of interleaving these two processes is called generalized policy iteration (GPI) and illustrated in Fig. 4.2. The idea of GPI is not limited to just DP methods, but a core concept of many RL solution methods. Another remarkable property of DP methods is that they estimate the values of states based on estimates of the values of successor states, as in Eq. (4.19). This general concept of updating estimates based on other estimates is called bootstrapping and a central idea used by many RL methods.

In cases where the dynamics of the environment are unknown, which is the majority of situations, the RL agent has to learn exclusively from experience, that is, sample sequences of states, actions and rewards. Monte Carlo methods offer exactly that, but may still attain optimal solutions. The term refers to methods which only update value estimates and policies at the end of an episode based on averaged complete returns. They are thus only well-defined for episodic tasks, where all episodes eventually terminate and complete returns can be calculated. In this case, one can adapt the concept of GPI for Monte Carlo methods by estimating the expected return via the average return seen by the agent after visiting a particular state

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] \approx \text{avg}_\pi[G_t \mid S_t = s] \ . \tag{4.21}$$

As the number of visits to the state increases and more returns are observed, the average in Eq. (4.21) should converge to the expected value of the return. In order to perform policy improvement in a similar way as in Eq. (4.20) it is useful to estimate the action-values instead

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] \approx \text{avg}_\pi[G_t \mid S_t = s, A_t = a] \ . \tag{4.22}$$

Using the action-value function, the new greedy policy can simply be constructed as

$$\pi'(s) = \arg\max_a q_\pi(s,a) \ . \tag{4.23}$$

The main problem with this approach is that many state-action pairs may never be visited by the agent. In the case of a deterministic policy, the agent will always choose the same action $\pi(s) = a$ and thus not provide any returns for other actions. The estimates of $q_\pi(s,a')$ will therefore not improve and the agent may be stuck in a suboptimal solution. To prevent this from happening, one has to ensure sufficient exploration, which means that the agent has to eventually try all actions. One way to achieve this, is to guarantee a minimal probability with which each action is taken. An $\varepsilon$-greedy policy

$$\pi'(s,a) = \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}|, & \text{if } a = \arg\max_a q_\pi(s,a) \\ \varepsilon/|\mathcal{A}|, & \text{if } a \neq \arg\max_a q_\pi(s,a) \end{cases} , \tag{4.24}$$

for some $\varepsilon > 0$, thus selects the greedy action most of the time, but retains the minimal probability $\varepsilon/|\mathcal{A}|$ for all other actions. Although policies like the one defined in Eq. (4.24) solve the problem of maintaining exploration, they introduce explicit randomness and thereby give up optimal behavior. This is indicative of a bigger issue, namely, the inevitable trade-off between exploration and exploitation. One approach to deal with this problem more effectively is the idea of using two policies instead of one. While an exploratory behavior policy is used to interact with the environment and to generate experience, the agent also retains and updates a second, greedy target policy, which fully exploits the gained knowledge and approximates optimal behavior. The general concept of learning about one policy while following another to sample data is called off-policy learning. Compared to the on-policy approach of using only one policy, off-policy methods often come at the cost of greater variance and slower convergence. Learning about a policy from data which is distributed according to a different policy also requires special care

and additional measures. On the other hand, off-policy learning is the more powerful and general concept. In fact, off-policy methods include the on-policy case as the special instance where behavior and target policy are the same. In principle, they also allow RL agents to learn from data generated by more conventional controllers or from human behavior, which opens up a variety of additional use cases.

While Monte Carlo methods can learn directly from raw experience, they only perform updates at the end of an episode, which is not very efficient and may thus require a lot of sample data. This issue is addressed by the concept of temporal-difference (TD) learning, which is a very central idea in reinforcement learning. TD methods can perform updates already during episodes because, like DP methods, they base estimates in part on other estimates, that is, they bootstrap. However, unlike DP methods, they can also learn exclusively from experience and don't require any knowledge about the transition dynamics of the environment. In this way, TD methods incorporate ideas from both, DP and Monte Carlo methods. Like the other two, they also use the concept of GPI to find better policies. In order to solve the prediction problem, that is, to estimate the value function, the simplest TD method makes the update[4]

$$V(S_t) \leftarrow V(S_t) + \alpha \left[R_{t+1} + \gamma V(S_{t+1}) - V(S_t)\right] , \tag{4.25}$$

immediately after transitioning from $S_t$ to $S_{t+1}$ and receiving $R_{t+1}$. Here, $\alpha \in (0, 1]$ is a constant parameter, referred to as the learning rate, which determines the step-size of the update. Each iteration moves $V(S_t)$ towards the target $R_{t+1} + \gamma V(S_{t+1})$ by reducing the so-called TD error

$$\delta_t \doteq R_{t+1} + \gamma V(S_{t+1}) - V(S_t) . \tag{4.26}$$

As the TD error approaches zero, the updates become smaller and smaller and the value estimate of $V(S_t)$ converges. The algorithm defined by Eq. (4.25) is called TD(0) or one-step TD as it looks exactly one step ahead to find an estimate which the update is based on. This is a special case of the TD($\lambda$) algorithm which generalizes the idea to an arbitrary number of steps, up to $\lambda = 1$, which corresponds to using complete returns as in the case of Monte Carlo methods. To perform policy improvement, it is again more convenient to estimate the action-value function

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)\right] . \tag{4.27}$$

The algorithm defined by this update rule is called Sarsa, referring to the fact that is makes use of every element in the tuple $(S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1})$. Based on Sarsa prediction of the action-value function one can easily define an on-policy control method by combining it with an $\varepsilon$-greedy policy as defined in Eq. (4.24). One of the early breakthroughs in reinforcement learning was the development of an off-policy version of the Sarsa method. The algorithm, called Q-learning, is defined by the update rule

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right] , \tag{4.28}$$

[4] $V$ and $Q$ denote array estimates of the state-value function $v$ and the action-value function $q$, respectively. In the case of finite MDPs, these are valid representations which may be used for practical implementations.

where the estimate of the action-value function in the next state $Q(S_{t+1}, A_{t+1})$ is replaced by a maximization over all available actions in state $S_{t+1}$. Instead of considering the action that was actually taken by the behavior policy, $A_{t+1}$, the update rule considers the optimal way to continue from this point onwards. Therefore, the action-value function learned via Eq. (4.28) directly approximates $q_*$, independent of the policy being followed. In fact, under the condition of sufficient exploration, Q-learning has been shown to always converge to the optimal action-value function.

The learning concepts introduced in this section, in particular Q-learning and TD learning, are used by many modern RL algorithms. While better policies are generally found through some form of GPI, the discussed DP, Monte Carlo and TD methods mostly differ in the way they approach the prediction problem, that is, how they estimate value functions. By combining the ability to learn directly from experience with the idea of bootstrapping, TD methods are capable of processing experience online while requiring relatively little computation. These benefits make TD learning one of the cornerstones of modern reinforcement learning.

## 4.4. Approximate Solution Methods

The algorithms discussed in the previous section are effective solution methods for finite MDPs with relatively small state and action spaces. This, however, is quite a strong presupposition as many problems naturally come with state spaces which are enormous, as in the example of games like chess or Go. And even worse, in physical problems the state space is typically continuous. In those cases, the solution methods of section 4.3 are no longer applicable as the value functions cannot be implemented as simple lookup tables or arrays. Any attempt to do so would require a lot of resources in terms of sample data, memory and computation time. Yet, for very large state spaces, most of the states will never have been encountered by the agent regardless. The learned value function thus cannot be expected to make usable predictions about the environment. The key issue to solve here is that of generalization. As the agent cannot be expected to encounter every possible state, it has to generalize its experience from those states which were actually visited. While relying on experience from similar states may improve the agent's decision making in new, previously unseen situations, it clearly cannot be expected to act optimally in every instance. Instead, one has to settle for approximate solutions. The kind of generalization required here is usually referred to as function approximation. Using only the experience of a limited subset, the task is to infer an approximate value function for the entire, possibly infinite, set of states. The general task of function approximation is illustrated for a simple one-dimensional function in Fig. 4.3. Fortunately, function approximation is precisely the type of problem which is considered in supervised learning, which makes many of the methods studied in that field applicable to the task at hand. Yet, the reinforcement learning setting involves a number of issues, which usually do not arise in supervised learning, such as bootstrapping, delayed targets or nonstationarity. Reinforcement learning methods thus commonly introduce additional measures to improve the stability of the training process.

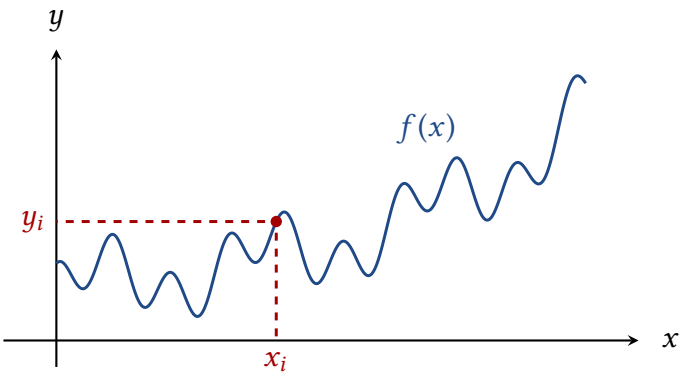



Figure 4.3.: Principle of function approximation. Given only a collection of samples $\{(x_i, y_i)\}$, with $f(x_i) = y_i$, the task is to approximate the function $f(x)$ on the entirety of its domain.

Approximate solution methods which use GPI typically represent the estimated value function as a parametrized function $\hat{v}(s, \boldsymbol{w})$ with weight vector $\boldsymbol{w} \in \mathbb{R}^d$. During policy evaluation, they usually minimize the mean squared value error

$$\overline{\mathrm{VE}}(\boldsymbol{w}) \doteq \sum_{s \in \mathcal{S}} \mu(s) \left[ v_\pi(s) - \hat{v}(s, \boldsymbol{w}) \right]^2 , \tag{4.29}$$

where $\mu(s)$ denotes the on-policy distribution, that is, the distribution of states while following the policy $\pi$. A particularly well suited class of algorithms to optimize the objective function defined by Eq. (4.29) in online reinforcement learning are those based on stochastic gradient descent (SGD). In order to apply gradient descent methods, the approximate value function $\hat{v}(s, \boldsymbol{w})$ has to be a differentiable function of $\boldsymbol{w}$ for all $s \in \mathcal{S}$. If that is the case, SGD methods can be used to iteratively adjust the weight vector $\boldsymbol{w}_t$ over a series of time steps $t$ to reduce the mean squared value error. They do so by moving the weight vector a small amount into the direction of the largest error reduction

$$\begin{aligned} \boldsymbol{w}_{t+1} &\doteq \boldsymbol{w}_t - \frac{1}{2} \alpha \nabla \left[ U_t - \hat{v}(S_t, \boldsymbol{w}_t) \right]^2 \\ &= \boldsymbol{w}_t + \alpha \left[ U_t - \hat{v}(S_t, \boldsymbol{w}_t) \right] \nabla \hat{v}(S_t, \boldsymbol{w}_t) , \end{aligned} \tag{4.30}$$

where $U_t$ is an estimate of the true value function $v_\pi(S_t)$, e.g., the Monte Carlo target $U_t \doteq G_t$ or the TD(0) target $U_t \doteq R_{t+1} + \gamma \hat{v}(S_{t+1}, \boldsymbol{w})$. An important special case of function approximation, where this concept is applicable, is that in which $\hat{v}(s, \boldsymbol{w})$ is a linear function of the weight vector $\boldsymbol{w}$. Here, the states are described by a real-valued feature vector

$$\boldsymbol{x}(s) \doteq (x_1(s), x_2(s), \ldots, x_d(s))^\top , \tag{4.31}$$

where each feature $x_i$ is the value of a function $x_i : \mathcal{S} \to \mathbb{R}$. The linear approximate state-value function is given by the inner product between the weight vector $\boldsymbol{w}$ and the feature vector $\boldsymbol{x}(s)$

$$\hat{v}(s, \boldsymbol{w}) \doteq \boldsymbol{w}^\top \boldsymbol{x}(s) \doteq \sum_{i=1}^{d} w_i x_i(s) . \tag{4.32}$$

Given the simple gradient of this function, $\nabla \hat{v}(s, \boldsymbol{w}) = \boldsymbol{x}(s)$, the general SGD update in Eq. (4.30) reduces to

$$\boldsymbol{w}_{t+1} \doteq \boldsymbol{w}_t + \alpha \left[ U_t - \boldsymbol{w}_t^\top \boldsymbol{x}(S_t) \right] \boldsymbol{x}(S_t) \,, \tag{4.33}$$

which in the special case of using the TD(0) target is

$$\boldsymbol{w}_{t+1} \doteq \boldsymbol{w}_t + \alpha \left[ R_{t+1} + \gamma \boldsymbol{w}_t^\top \boldsymbol{x}(S_{t+1}) - \boldsymbol{w}_t^\top \boldsymbol{x}(S_t) \right] \boldsymbol{x}(S_t) \,. \tag{4.34}$$

Analogously, one can derive similar update rules for an approximate action-value function $\hat{q}(s, a, \boldsymbol{w})$, which can facilitate the construction of a policy, e.g., by using an $\varepsilon$-greedy policy as defined in Eq. (4.24). These linear methods already give rise to powerful approximate solution methods, which are applicable to a large range of problems and can be very efficient in terms of sample data and computation. Thereby, the performance is crucially dependent on the representation of the states in terms of the features $x_i$. The choice of appropriate features for a task depends heavily on the particularities of the considered problem and can be a way of adding prior domain knowledge to RL systems. As constructing features which are ideally suited for these linear methods can be a very challenging task, it is often beneficial to resort to nonlinear function approximation.

The most common way of achieving nonlinear function approximation in reinforcement learning is through the use of artificial neural networks. A neural network (NN) is composed of a number of interconnected units, also called nodes or neurons, which process an incoming real-valued signal and transmit the updated signal to other units. The connections between individual units are called edges, which typically have a weight $w_i$ that scales the transmitted signal. The units are arranged in so-called layers, which may perform different operations on their respective input. An incoming signal travels from the input layer to the output layer while passing an arbitrary number of hidden layers. Overall, a neural network is a parametrized differentiable function $f(\boldsymbol{x}, \boldsymbol{w})$, which maps a real-valued input vector $\boldsymbol{x}$ to a possibly also vector-valued output $\boldsymbol{y}$. One of the simplest forms of a neural network is a feedforward NN as illustrated in Fig. 4.4. Here, the real-valued input vector $\boldsymbol{x}$, which may now have a different dimension as the weight vector $\boldsymbol{w} \in \mathbb{R}^d$, is initially processed by the neurons in the first hidden layer. Each neuron in that layer computes the weighted sum of the input signals, adds a real-valued bias $w_b$ and applies a so-called activation function, e.g., $g(w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_b)$. The such computed outputs of the first hidden layer serve as the inputs of the second hidden layer, where there procedure is repeated. Finally, the outputs of the second hidden layer are the inputs of the single neuron in the output layer, which again computes the weighted sum, adds a bias and applies an activation function. In general, a feedforward neural network may have an arbitrary number of hidden layers, each with an arbitrary number of units. Typically, the number of parameters aggregated in $\boldsymbol{w}$, that is, the number of all weights and biases, is much larger than the number of inputs. Different layers and different nodes may also have different activation functions. Some of the most commonly used functions are the logistic sigmoid function

$$\sigma : \mathbb{R} \to (0, 1) \quad \text{with} \quad \sigma(x) \doteq \frac{1}{1 + \mathrm{e}^{-x}} \,, \tag{4.35}$$

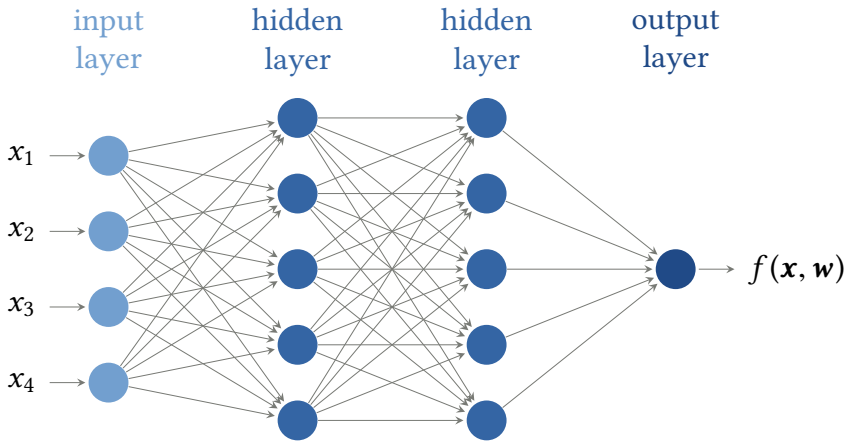



Figure 4.4.: A simple example of a feedforward neural network. The four-dimensional input vector $\boldsymbol{x}$ is processed by two hidden layers with five nodes each, before a single output node computes the output $f(\boldsymbol{x}, \boldsymbol{w})$.

the tangens hyperbolicus, which is simply a rescaled version of the logistic sigmoid function defined above

$$\tanh : \mathbb{R} \to (-1, 1) \quad \text{with} \quad \tanh(x) \doteq 1 - \frac{2}{1 + \mathrm{e}^{2x}} = 2\sigma(2x) - 1 \,, \tag{4.36}$$

the rectified linear unit (ReLU) activation function

$$\mathrm{ReLU} : \mathbb{R} \to [0, \infty) \quad \text{with} \quad \mathrm{ReLU}(x) \doteq \max(0, x) \,, \tag{4.37}$$

or simply the identity function

$$\mathrm{id} : \mathbb{R} \to (-\infty, \infty) \quad \text{with} \quad \mathrm{id}(x) \doteq x \,. \tag{4.38}$$

Perhaps the most widely used activation function nowadays is the ReLU activation function. If only the identity function $\mathrm{id}(x)$ is used for all neurons, then the output is just a linear function of the input vector $\boldsymbol{x}$. The nonlinearity of the overall function $f(\boldsymbol{x}, \boldsymbol{w})$ is achieved by using nonlinear activation functions like those in Eqs. (4.35), (4.36) and (4.37). There are many different architectures beyond these simple feedforward NNs, like recurrent neural networks (RNNs) or convolutional neural networks (CNNs). A more detailed introduction to the general subject, and to these architectures in particular, can be found in [56]. However, in theory, a simple feedforward NN with a single hidden layer, which contains enough units, can already approximate any continuous function. Despite this remarkable property, more sophisticated architectures and additional hidden layers can significantly improve the performance of NNs in practical applications. Particularly the use of NNs with many hidden layers, which are referred to as deep neural networks, has led to impressive results in recent years. One of the most prominent examples is the use of deep convolutional neural networks in the domain of computer vision, which led to significant performance gains in tasks like the classification or segmentation of images. Here, the additional hidden layers progressively extract higher-level features from raw pixel input

and form a hierarchical representation of concepts and patterns. Independent of the used architecture, neural networks are typically trained by some form of back-propagation based on (stochastic) gradient descent. Given an NN-based approximation of the value function $\hat{v}(s, \boldsymbol{w})$ or $\hat{q}(s, a, \boldsymbol{w})$, update rules like the one defined in Eq. (4.30) can be used to train the network and to iteratively improve the value estimation.

A remarkable side effect of using function approximation in reinforcement learning is that it also extends these methods to partially observable problems. If the chosen representation of the approximate value function does not allow the estimated value to depend on certain aspects of the state, then it is just as if those aspects are unobservable.

While the use of a parametrized approximate value function addresses the issue of enormous state spaces, it does not cover the case of very large or continuous action spaces. The greedy policies in Eqs. (4.23) and (4.24) are implicitly defined by the action-value function and rely on a maximization of $q_\pi(s, a)$ over all available actions $a \in \mathcal{A}$ in a given state $s \in \mathcal{S}$. This is no longer feasible for very large action spaces. Instead, the policy is represented explicitly in a parametrized form

$$\pi(a|s, \boldsymbol{\theta}) = \Pr\{A_t = a \mid S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\} , \tag{4.39}$$

where $\boldsymbol{\theta}_t \in \mathbb{R}^d$ denotes the policy's parameter vector $\boldsymbol{\theta}$ at time step $t$. In this case, the RL objective is equivalent to the search for that parameter vector $\boldsymbol{\theta}$ which optimizes the agent's performance. Given a differentiable scalar measure of the agent's performance $J(\boldsymbol{\theta})$, the parameter vector can be improved iteratively based on the gradient with respect to its argument, or more generally

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \alpha \widehat{\nabla J(\boldsymbol{\theta}_t)} , \tag{4.40}$$

where $\widehat{\nabla J(\boldsymbol{\theta}_t)}$ denotes a stochastic estimate of the gradient. RL methods which use a parametrized policy and follow this general policy improvement scheme are called policy gradient methods, independent of whether or not they use an approximate value function. In episodic tasks, the performance measure may simply be defined as

$$J(\boldsymbol{\theta}) \doteq v_{\pi_{\boldsymbol{\theta}}}(s_0) , \tag{4.41}$$

where $v_{\pi_{\boldsymbol{\theta}}}(s)$ is the true value function for the policy $\pi_{\boldsymbol{\theta}}$ determined by the parameter vector $\boldsymbol{\theta}$ and $s_0$ denotes the starting state. In this case, a change in the parameter vector $\boldsymbol{\theta}$ does not merely affect the actions taken by the agent, but also the distribution of the encountered states which depends on the typically unknown transition dynamics of the environment. This makes the policy improvement step more complicated. Fortunately, the so-called policy gradient theorem relates the required gradient of $J(\boldsymbol{\theta})$ to known quantities[5]

$$\begin{aligned} \nabla J(\boldsymbol{\theta}) &\propto \sum_s \mu(s) \sum_a q_\pi(s, a) \nabla \pi(a|s, \boldsymbol{\theta}) \\ &= \mathbb{E}_\pi \left[ \sum_a q_\pi(S_t, a) \nabla \pi(a|S_t, \boldsymbol{\theta}) \right] . \end{aligned} \tag{4.42}$$

[5] Here, and in the remaining section, only the special case of no discounting ($\gamma = 1$) is considered for simplicity of notation.

Based on the expression in Eq. (4.42) and the Monte Carlo estimate of the discounted return, one can derive the update rule

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &\doteq \boldsymbol{\theta}_t + \alpha\, G_t \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)} \\ &= \boldsymbol{\theta}_t + \alpha\, G_t \nabla \ln \pi(A_t|S_t, \boldsymbol{\theta}_t)\,, \end{aligned} \tag{4.43}$$

which defines the so-called REINFORCE algorithm. This is a Monte Carlo policy gradient method which uses complete returns instead of an estimated value function. In contrast to the value-based methods discussed above, REINFORCE is a policy-based algorithm which does not use an explicit value function. Reinforcement learning methods which use both, a parametrized value function $\hat{v}(s, \boldsymbol{w})$ or $\hat{q}(s, a, \boldsymbol{w})$ and a parametrized policy $\pi(a|s, \boldsymbol{\theta})$, are usually referred to as actor-critic methods. The term actor refers to the learned policy $\pi(a|s, \boldsymbol{\theta})$ which determines the chosen actions while the term critic refers to the learned value function, $\hat{v}(s, \boldsymbol{w})$ or $\hat{q}(s, a, \boldsymbol{w})$, which is used to evaluate (or criticize) the action selection. Given an approximate value function, the Monte Carlo target in Eq. (4.43) can be replaced by an expression that allows for updates before the end of an episode. One example is the one-step actor-critic defined by

$$\begin{aligned}\boldsymbol{\theta}_{t+1} &\doteq \boldsymbol{\theta}_t + \alpha \left[R_{t+1} + \gamma \hat{v}(S_{t+1}, \boldsymbol{w}) - \hat{v}(S_t, \boldsymbol{w})\right] \frac{\nabla \pi(A_t|S_t, \boldsymbol{\theta}_t)}{\pi(A_t|S_t, \boldsymbol{\theta}_t)} \\ &= \boldsymbol{\theta}_t + \alpha\, \delta_t \nabla \ln \pi(A_t|S_t, \boldsymbol{\theta}_t)\,, \end{aligned} \tag{4.44}$$

which instead uses the TD error $\delta_t$. Analogously to the TD(0) methods Sarsa and Q-learning, the update rule in Eq. (4.44) uses bootstrapping to allow for updates after each time step. While the TD target reduces variance compared to the Monte Carlo approach, it depends on the approximation of the value function and thus introduces bias. The idea of optimizing this trade-off gives rise to a range of actor-critic variants, some of which resort to estimating the so-called advantage function

$$\mathcal{A}_\pi(s, a) \doteq q_\pi(s, a) - v_\pi(s)\,. \tag{4.45}$$

A selection of modern actor-critic algorithms is introduced in the following section. Overall, policy gradient methods can be used to address problems with very large or continuous action spaces, which extends the collection of RL solution methods to the most general case of a continuous state and action space.

## 4.5. Modern Reinforcement Learning Algorithms

As detailed in the chapters 5 and 6, the problem considered in this thesis is a continuing task with a continuous state and action space. This section thus covers a selection of modern algorithms which are suited for this type of RL problem. Given the availability of a simulation code and the fast data rates expected at the actual storage ring, sample efficiency is not a primary concern in the context of this task, which is why only model-free methods are considered. The fact that these algorithms were all published between 2015 and 2018 is indicative of the rapid ongoing development in the field. It also means that further improvements upon the presented algorithms as well as the publication of new methods can be expected in the coming years.

### 4.5.1. Deep Deterministic Policy Gradient

The combination of deep neural networks with the concept of Q-learning led to the development of an RL algorithm, called Deep Q-Network (DQN), which achieved impressive results in the domain of Atari games [53]. Receiving only raw pixel information and the game score as input, DQN was able to achieve a level of control that was comparable to human play. While DQN is capable of solving problems with high-dimensional observation spaces, it can only handle discrete and low-dimensional action spaces. The extension of the learning concepts underlying DQN to the domain of continuous action spaces led to the development of an algorithm called Deep Deterministic Policy Gradient (DDPG) in 2015 [57], which also builds upon the Deterministic Policy Gradient (DPG) method published the year before [58]. DDPG is a model-free, off-policy actor-critic algorithm which uses deep neural networks for approximation of the action-value function and a deterministic policy. Assuming a deterministic policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$, the action-value function can be expressed as

$$q_\pi(S_t, A_t) = \mathbb{E}[R_{t+1} + \gamma q_\pi(S_{t+1}, \pi(S_{t+1}))] \ . \tag{4.46}$$

This allows for an approximate action-value function $\hat{q}(s, a, \boldsymbol{w})$, which can be learned off-policy, that is, by transitions generated from a different stochastic behavior policy $b(a|s)$. The parametrized action-value function is optimized by minimizing the loss

$$L(\boldsymbol{w}) \doteq \mathbb{E}_b\Big[\big(\hat{q}(S_t, A_t, \boldsymbol{w}) - y_t\big)^2\Big] \ , \tag{4.47}$$

with

$$y_t \doteq R_{t+1} + \gamma \hat{q}(S_{t+1}, \pi(S_{t+1}, \boldsymbol{\theta}), \boldsymbol{w}) \ , \tag{4.48}$$

where the fact that $y_t$ is also dependent on $\boldsymbol{w}$ is ignored. The parametrized policy $\pi(s, \boldsymbol{\theta})$ is updated by using the gradient of the expected return from the start distribution, given by

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) \approx \mathbb{E}_b\Big[\nabla_{\boldsymbol{\theta}} \pi(s, \boldsymbol{\theta}) \, \nabla_a \hat{q}(s, a, \boldsymbol{w})|_{a=\pi(s,\boldsymbol{\theta})}\Big] \ . \tag{4.49}$$

Following the example of DQN, DDPG also makes use of a so-called replay buffer and separate target networks to stabilize the training process of the nonlinear function approximators. The replay buffer is a finite sized cache which stores the transitions $(S_t, A_t, R_{t+1}, S_{t+1})$ sampled by the exploration policy. The updates of the actor and the critic are performed on a minibatch sampled uniformly from the buffer, which decorrelates the transitions. The target networks are copies, $\pi'(s, \boldsymbol{\theta}')$ and $\hat{q}'(s, a, \boldsymbol{w}')$, of the actor and critic networks, respectively. They are used to slowly track the learned networks,

$$\boldsymbol{w}' \leftarrow \tau \boldsymbol{w} + (1 - \tau)\boldsymbol{w}' \quad \text{and} \quad \boldsymbol{\theta}' \leftarrow \tau \boldsymbol{\theta} + (1 - \tau)\boldsymbol{\theta}' \ , \tag{4.50}$$

with $\tau \ll 1$. This constrains the target values to change slowly and improves the stability of learning. As DDPG is an off-policy algorithm, it can be combined with arbitrary exploration policies. In [57], the behavior policy is constructed by adding noise sampled from a noise process $\mathcal{N}$ to the actor policy

$$b(S_t) \doteq \pi(S_t, \boldsymbol{\theta}_t) + \mathcal{N} \ , \tag{4.51}$$

suggesting an Ornstein-Uhlenbeck process [59] for exploration efficiency in physical control problems with inertia. The DDPG algorithm was found to robustly solve a range of simulated physics tasks, including classic problems such as cartpole swing-up and legged locomotion, even when learning directly from raw pixel inputs.

### 4.5.2. Twin Delayed Deep Deterministic Policy Gradient

To address commonly observed issues with DDPG, the so-called Twin Delayed Deep Deterministic Policy Gradient (TD3) method published in 2018 [60] introduces additional measures to improve the stability of learning. Building upon the DDPG algorithm, it addresses potential overestimation of action-values and its deteriorating effects on learning through three appropriate adjustments: clipped double-Q learning, target policy smoothing and delayed policy updates. The first of these refers to the fact that TD3 concurrently learns two approximate action-value functions, $\hat{q}_1$ and $\hat{q}_2$, which are used to adjust the target

$$y_t \doteq R_{t+1} + \gamma \min_{i=1,2} \hat{q}_i(S_{t+1}, \pi(S_{t+1}, \boldsymbol{\theta}), \boldsymbol{w}_i) \ . \tag{4.52}$$

Both action-value functions are trained on this target, that is, they optimize the loss function defined in Eq. (4.47) for the new target. Using only the smaller value of the two functions, $\hat{q}_1$ and $\hat{q}_2$, helps to avoid overestimation of the action-value function. The parametrized policy $\pi(s, \boldsymbol{\theta})$ is trained according to the gradient in Eq. (4.49) using only the first action-value function $\hat{q}_1(s, a, \boldsymbol{w}_1)$. The definition of $y_t$ in Eq. (4.52) is further adjusted by adding additional noise to the Q-learning target computed by the learned target policy $\pi(s, \boldsymbol{\theta})$. While an approximate value function generally leads to similar values for similar actions, this is enforced explicitly by fitting a small area around the target action

$$y_t \doteq R_{t+1} + \gamma \min_{i=1,2} \hat{q}_i(S_{t+1}, \pi(S_{t+1}, \boldsymbol{\theta}) + \epsilon, \boldsymbol{w}_i) \ , \tag{4.53}$$

with the noise parameter

$$\epsilon \sim \text{clip}(\mathcal{N}(0, \sigma^2), -\hat{\epsilon}, \hat{\epsilon}) \ , \tag{4.54}$$

where $\mathcal{N}(0, \sigma^2)$ denotes the normal distribution with zero mean and standard deviation $\sigma$ and $\hat{\epsilon}$ is constraining the magnitude of the noise. This is called target policy smoothing and acts as a form of regularization. Lastly, TD3 also delays the updates of the policy and the target networks. While the approximate action-value functions are updated after each time step, the policy and target networks are only updated every couple iterations, which further stabilizes the training process. In tests on a range of OpenAI gym [61] tasks, TD3 has been found to match or outperform the original version of DDPG.

### 4.5.3. Soft Actor-Critic

The Soft Actor-Critic (SAC) algorithm published in 2018 [62] shares several features with TD3. However, unlike DDPG and TD3, it learns a stochastic policy $\pi(a|s, \boldsymbol{\theta})$ and uses entropy regularization. In maximum entropy reinforcement learning, the agent seeks to

maximize both the expected return and the expected entropy of the policy. The reinforcement learning objective is thus defined as finding the policy

$$\pi_* \doteq \arg\max_\pi \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \Big(R_{t+k+1} + \alpha_\mathrm{T} H(\pi(\cdot|S_{t+k}))\Big)\right] , \tag{4.55}$$

with the entropy

$$H(p) \doteq \mathbb{E}[-\ln p(x)] , \tag{4.56}$$

and the temperature parameter $\alpha_\mathrm{T} > 0$, which determines the trade-off between reward and entropy, and thus controls the stochasticity of the optimal policy. The corresponding state-value function is consequently defined as

$$v_\pi(s) \doteq \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \Big(R_{t+k+1} + \alpha_\mathrm{T} H(\pi(\cdot|S_{t+k}))\Big) \,\Big|\, S_t = s\right] , \tag{4.57}$$

and the action-value function as

$$q_\pi(s,a) \doteq \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k \Big(R_{t+k+1} + \alpha_\mathrm{T} H(\pi(\cdot|S_{t+k}))\Big) \,\Big|\, S_t = s, A_t = a\right] . \tag{4.58}$$

Analogously to TD3, the SAC algorithm trains two parametrized action-value functions, $\hat{q}_1(s,a,\boldsymbol{w}_1)$ and $\hat{q}_2(s,a,\boldsymbol{w}_2)$. They are optimized based on the loss function

$$L(\boldsymbol{w}_i) \doteq \mathbb{E}_\pi\left[\frac{1}{2}\Big(\hat{q}(S_t,A_t,\boldsymbol{w}_i) - y_t\Big)^2\right] , \tag{4.59}$$

where the Q-learning target may be expressed as

$$y_t \doteq R_{t+1} + \gamma(1-\tau)\left(\min_{i=1,2} \hat{q}_i(S_{t+1},A_{t+1},\boldsymbol{w}_i) - \alpha_\mathrm{T} \ln \pi(A_{t+1}|S_{t+1},\boldsymbol{\theta})\right) , \tag{4.60}$$

where $A_{t+1}$ is distributed according to the learned policy, that is, $A_{t+1} \sim \pi(\cdot|S_{t+1},\boldsymbol{\theta})$. The stochastic policy can be implemented as a Gaussian distribution with the mean and covariance given by a neural network. It is optimized based on the gradient of

$$J(\boldsymbol{\theta}) \doteq \mathbb{E}\left[\mathrm{D_{KL}}\left(\pi(\cdot|S_t,\boldsymbol{\theta}) \,\middle\|\, \frac{\exp(\hat{q}_1(S_t,\cdot,\boldsymbol{w}_1))}{Z_\mathrm{KL}(S_t,\boldsymbol{w}_1)}\right)\right] , \tag{4.61}$$

where $\mathrm{D_{KL}}$ denotes the Kullback-Leibler divergence and the generally intractable partition function $Z_\mathrm{KL}(S_t,\boldsymbol{w}_1)$ normalizes the distribution, but does not contribute to the required gradient and can thus be ignored. The inclusion of the entropy in the SAC objective serves the purpose of improving the exploration and robustness of this actor-critic method. The algorithm was tested on a range of OpenAI gym tasks and found to generally outperform DDPG and other state-of-the-art model-free deep RL methods.

### 4.5.4. Proximal Policy Optimization

In contrast to the off-policy methods introduced above, the Proximal Policy Optimization (PPO) algorithm published in 2017 [63] is an on-policy method. While on-policy methods offer the benefit of reduced sample variance, they are not able to use experience which was previously obtained under a different policy, which makes them less sample efficient. PPO generally uses an estimate of the advantage function $\hat{\mathcal{A}}_t$ at time $t$ to express the policy gradient

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}\left[\nabla \pi(A_t|S_t)\hat{\mathcal{A}}_t\right] , \tag{4.62}$$

but introduces constrains to limit the size of the policy updates. Similarly to the Trust Region Policy Optimization (TRPO) method [64], it defines the clipped surrogate objective

$$J^{\mathrm{CLIP}}(\boldsymbol{\theta}) \doteq \mathbb{E}\left[\min\left(r_t(\boldsymbol{\theta})\hat{\mathcal{A}}_t, \mathrm{clip}(r_t, 1-\epsilon_\mathrm{c}, 1+\epsilon_\mathrm{c})\hat{\mathcal{A}}_t\right)\right] , \tag{4.63}$$

with the hyperparameter $\epsilon_\mathrm{c} > 0$ and the probability ratio

$$r_t(\boldsymbol{\theta}) \doteq \frac{\pi(A_t|S_t, \boldsymbol{\theta})}{\pi(A_t|S_t, \boldsymbol{\theta}_\mathrm{old})} , \tag{4.64}$$

where $\boldsymbol{\theta}_\mathrm{old}$ is the policy's parameter vector before the update. While the first term in Eq. (4.63) is equivalent to the TRPO objective, the second term modifies the surrogate objective by clipping the probability ratio. By taking the minimum, the final objective becomes a lower bound on the unclipped objective. An alternative approach uses a penalty on the Kullback-Leibler divergence to achieve some target value $d_\mathrm{KL}^\mathrm{targ}$ each policy update

$$J^{\mathrm{KLPEN}}(\boldsymbol{\theta}) \doteq \mathbb{E}\left[r_t(\boldsymbol{\theta})\hat{\mathcal{A}}_t - \beta_\mathrm{KL}\mathrm{D}_\mathrm{KL}\Big(\pi(\cdot|S_t, \boldsymbol{\theta}_\mathrm{old}) \parallel \pi(\cdot|S_t, \boldsymbol{\theta})\Big)\right] , \tag{4.65}$$

where the hyperparameter $\beta_\mathrm{KL}$ is adjusted based on the value of

$$d_\mathrm{KL} \doteq \mathbb{E}\left[\mathrm{D}_\mathrm{KL}\Big(\pi(\cdot|S_t, \boldsymbol{\theta}_\mathrm{old}) \parallel \pi(\cdot|S_t, \boldsymbol{\theta})\Big)\right] , \tag{4.66}$$

that is,

$$\beta_\mathrm{KL} \leftarrow \begin{cases} \beta_\mathrm{KL}/2, & \text{if } d_\mathrm{KL} < d_\mathrm{KL}^\mathrm{targ}/1.5 \\ 2\,\beta_\mathrm{KL}, & \text{if } d_\mathrm{KL} > 1.5\, d_\mathrm{KL}^\mathrm{targ} \end{cases} . \tag{4.67}$$

In principle, any method of estimating the advantage function $\hat{\mathcal{A}}_t$ can be used in combination with the objective functions in Eqs. (4.63) and (4.65). In [63], a truncated version of generalized advantage estimation, that is, an $n$-step TD method based on a parametrized state-value function is suggested

$$\hat{\mathcal{A}}_t \doteq \hat{\delta}_t + (\gamma\lambda)\hat{\delta}_{t+1} + \cdots + (\gamma\lambda)^{n-t+1}\hat{\delta}_{n-1} , \tag{4.68}$$

with the estimate of the TD error

$$\hat{\delta}_t = R_{t+1} + \gamma\hat{v}(S_{t+1}, \boldsymbol{w}) - \hat{v}(S_t, \boldsymbol{w}) . \tag{4.69}$$

Implemented like this, the PPO algorithm can make use of the experience collected by several parallel actors. After each actor collected $n$ steps of data, the surrogate objective is calculated and optimized by stochastic gradient descent. In tests on a collection of benchmark tasks, including simulated robotic locomotion and Atari game playing, PPO was found to outperform other online policy gradient methods.

## 4.6. Reinforcement Learning and Particle Accelerators

All in all, reinforcement learning offers a variety of methods that are well-suited to tackle issues in and around accelerator physics. This is partly because of the underlying learning paradigm, which allows RL methods to learn purely from interaction with a stochastic system. As large particle accelerators are commonly built as one of a kind machines, and their exact setup and configuration may change over the years, relying on pre-existing, labeled data sets is frequently not an option. In addition, many of the problems encountered at these machines are of a dynamic nature. This may be an intrinsic property of the considered problem, like in tasks which directly address the dynamics of the beam, or because of inevitable drifts in related subsystems, e.g., because of changes in temperature or humidity. On the other hand, particle accelerators represent a particularly demanding and multifaceted domain to develop and test RL methods. As some of the largest, most data-intensive, and most complex machines ever built or conceived, particle accelerators come with a variety of challenging, often continuous control problems. Owing to the large number of interacting subsystems, a priori models are usually insufficient to predict the precise behavior of a particular particle accelerator. Machine commissioning and day-to-day operation thus involve a significant amount of trial-and-error search and online tuning. The application of RL methods simultaneously offers the opportunity to automatize some of these processes as well as to approach problems which could not have been tackled with conventional methods.

Given the recency of some of the advances made in the domain of reinforcement learning, it is remarkable that there is already a range of applications seeking to deploy these methods at particle accelerators [65–71]. These efforts include tasks like control of various RF subsystems, laser alignment or optimization of magnetic field strengths. In these applications and for the future success of RL methods in the field of accelerator physics, it is crucially important that the deployed algorithms are selected with the necessary care and diligence to match the requirements and particularities of the considered task. Questions like whether or not the task can be modeled with a finite state or action space, what an appropriate reward function looks like or how important sample efficiency is to the success of the application all should be answered beforehand. Depending on the specifics of a given problem, different RL methods may have varying levels of effectiveness or may not even be applicable to the task. One important distinction is that between problems that are stationary optimization tasks and those that are not. The learning concepts and methods introduced in this chapter focus on sequential decision problems, that is, situations in which a sequence of actions or decisions is required to reach a predefined goal. This is because the pursued objective of micro-bunching control is inherently of that nature, requiring continuous action to counteract the perturbation caused by CSR self-interaction. Although a stationary solution might be reached in the sense that repetitive action may be sufficient to achieve or maintain the desired level of control, this is a special case and cannot be assumed a priori. Stationary optimization problems, in which one seeks to optimize a finite set of parameters purely through interaction with the system, are an important special case

of the general RL problem.[6] These are generally referred to as (contextual) multi-armed bandit problems and come with their own individually tailored solution methods. One of these methods, which is particularly effective in situations where the interaction with the environment is expensive of some sort, is Bayesian optimization (BO) using Gaussian processes (GPs). BO has been successfully deployed at a number of accelerator facilities for a variety of tasks [72–78], including the optimization of free-electron lasers (FELs) and laser plasma accelerators (LPAs).

In practice, stating an accelerator physics task as a formal RL problem and selecting ideally suited solution methods can be a difficult task as it requires intimate knowledge of both domains. Nonetheless, with the growing expertise in the particle accelerator community and the ongoing rapid development of modern reinforcement learning methods, the number of successful RL applications is likely to increase over the coming years. Their impact on the field at large will, to some extent, depend on the acceptance and willingness to cope with these black box type systems. Ideally used, reinforcement learning methods may support and guide the process of scientific discovery.

[6] Intuitively, one may think about this distinction as the difference between the task of playing the game of chess and the task of solving a mate in one problem (which immediately wins and ends the game).

# 5. Micro-Bunching Instability: An Approach to Control

*So many questions. Humans asked them about everything, but they usually weren't half as good at finding the answers.*

— Guillermo del Toro & Cornelia Funke,
*Pan's Labyrinth: The Labyrinth of the Faun*

In order to develop a better understanding of the longitudinal dynamics underlying the micro-bunching instability, this chapter focuses on the perturbation of the synchrotron motion that is caused by the self-interaction of the electron bunch with its own emitted CSR. Adopting the perspective of a single particle, the unperturbed longitudinal particle motion can be described by a simple one-dimensional harmonic oscillator. The inclusion of the additional CSR wake potential causes a specific deformation of the particle trajectories, which can be described approximately by a perturbation of the strength of the linear restoring force exerted by the RF system. Below the threshold current, this perturbation breaks the homogeneity in the longitudinal phase space and leads to a quadrupole-like deformation of the charge distribution, potentially acting as a seeding mechanism for the micro-bunching instability. Above the threshold current, the charge distribution and thus the perturbation caused by the CSR wake potential are continuously varying in time. Yet again, the corresponding formation process of micro-structures in the longitudinal charge distribution can be shown to be largely driven by a dynamic perturbation of the restoring force. The presented content is largely based on the publications in [79] and [80].

These findings offer a new perspective on different aspects of the observed micro-bunching dynamics and the interpretation of measurements at KARA. Simultaneously, they give rise to a range of further questions, which are promising subjects for future research, as briefly outlined in section 5.4. Eventually, the insights gained by this analysis lead to an approach towards control of the micro-bunching dynamics based on a dedicated RF amplitude modulation scheme. The necessity of dynamic adaptations of the RF signal to counteract the CSR-induced perturbation is discussed in section 5.5 and motivates the RL-based control pursued in the following chapters.

## 5.1. Perturbation of the Restoring Force

In the absence of collective effects, and assuming a linear momentum compaction and a linear accelerating voltage ($\sin(x) \approx x$ for small $x$), the Hamiltonian given by Eq. (3.6)

takes the form of a simple one-dimensional harmonic oscillator. Here, the RF potential acts as a linear restoring force

$$V_{\mathrm{RF}}(q) = -kq \,, \tag{5.1}$$

with the constant parameter $k$ describing the slope of the RF potential, that is, the strength of the restoring force. Neglecting radiation damping and diffusion, this leads to the simple equations of motion

$$\ddot{q} + \frac{k}{\xi} q = 0 \quad \text{and} \quad \ddot{p} + \frac{k}{\xi} p = 0 \,, \tag{5.2}$$

with the constant scaling parameter $\xi$, and the solution

$$q(t) = a_0 \cos(\omega t + \varphi_0) \,, \tag{5.3}$$

$$\dot{q}(t) = p(t) = -a_0 \, \omega \sin(\omega t + \varphi_0) \,, \tag{5.4}$$

with the angular oscillation frequency $\omega = \sqrt{k/\xi}$, the amplitude $a_0$ and the initial phase $\varphi_0$. Due to the specific choice of the generalized coordinates $q$ and $p$ in Eq. (3.2), and the notation of time in multiples of the synchrotron period, $\Theta = f_{\mathrm{s},0} t$, the expression $k/\xi$ simplifies to 1, yielding the Hamiltonian in Eq. (3.3) and perfectly circular trajectories in phase space as illustrated in Fig. 2.4a. In anticipation of the additional effect of the CSR wake potential, a small perturbation to the strength of the restoring force is introduced

$$k' = k - \varepsilon \quad \text{with} \quad \varepsilon > 0 \,. \tag{5.5}$$

With the reduced restoring force defined by $k'$, the system remains a harmonic oscillator, but now has the altered solution

$$q'(t) = a_0 \cos(\omega' t + \varphi_0) \,, \tag{5.6}$$

$$\dot{q}'(t) = p'(t) = -a_0 \, \omega' \sin(\omega' t + \varphi_0) \,, \tag{5.7}$$

with the angular oscillation frequency

$$\omega' = \sqrt{k'/\xi} = \sqrt{(k - \varepsilon)/\xi} \,. \tag{5.8}$$

While the maximum deviation in $q$ is unaffected

$$\max |q'(t)| = \max |q(t)| = |a_0| \,, \tag{5.9}$$

the maximum deviation in $p$ is decreased by the perturbation

$$\max |p'(t)| = |a_0 \omega'| < |a_0 \omega| = \max |p(t)| \,. \tag{5.10}$$

The particle's trajectory in the phase space spanned by the original definitions of $q$ and $p$ in Eq. (3.2) is thus elliptical, as illustrated in Fig. 5.1, and of altered periodicity

$$|\omega'| < |\omega| \,. \tag{5.11}$$

In the following section, the effect of the additional CSR wake potential on the particle motion below the instability threshold is approximated as a position-dependent perturbation to the strength of the linear restoring force as defined in Eq. (5.5). This serves the purpose of developing a better understanding of how single particle motion relates to the formation of micro-structures and how to counteract it.

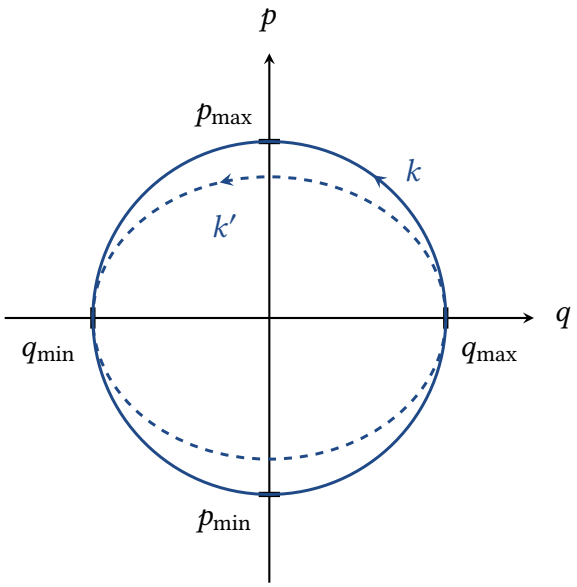



Figure 5.1.: A small perturbation of the strength of the restoring force $k$ leads to an elliptical particle trajectory in the phase space spanned by the generalized coordinates $q$ and $p$.

## 5.2. Particle Motion below Threshold

Given the parallel plates impedance $Z_{\mathrm{CSR}}^{\mathrm{PP}}$ defined in Eq. (3.10) and illustrated in Fig. 3.3, the wake potential of a Gaussian bunch profile takes the form shown in the upper part of Fig. 5.2. While such a perfectly Gaussian electron distribution exists only in the absence of collective effects, that is, in the zero current limit, a higher bunch current leads to an increased perturbation strength and thus distortion of the Gaussian shape. Yet, at a fixed bunch current below the instability threshold, the distribution still remains fairly stationary, $\psi(q,p,t) \approx \psi(q,p)$, which corresponds to a stationary wake potential as depicted in the lower part of Fig. 5.2 for exemplary parameter settings (data set $\mathcal{D}_2$ with $g = 32\,\mathrm{mm}$, defined in appendix A.1) and a range of different bunch currents. It should be noted that the general shape of the wake potential is still similar to that of a Gaussian shaped bunch up until right below the threshold current of $I_{\mathrm{th}} = 260\,\mu\mathrm{A}$, where the wake potential is no longer stationary.

A single particle propagating in the longitudinal phase space is subject to the sum of the linear RF potential and the CSR wake potential, that is, the effective potential $V_{\mathrm{eff}}(q)$ defined in Eq. (3.13). Because of its small deviations from the synchronous phase, the particle is only exposed to $V_{\mathrm{eff}}(q)$ on a part of its domain, $q \in [q_{\mathrm{min}}, q_{\mathrm{max}}]$, where $q_{\mathrm{min}}$ and $q_{\mathrm{max}}$ denote the maximum deviations from the longitudinal position of the synchronous particle as indicated in Fig. 5.1. By approximating $V_{\mathrm{eff}}(q)$ as a linear function on the given interval

$$V_{\mathrm{eff}}(q) \approx -k'\,q \quad \text{with} \quad q \in [q_{\mathrm{min}}, q_{\mathrm{max}}]\,, \tag{5.12}$$

as illustrated in Fig. 5.3, the single particle motion is still harmonic below the threshold current, with the strength of the restoring force $k'$ being dependent on $q_{\mathrm{min}}$ and $q_{\mathrm{max}}$.

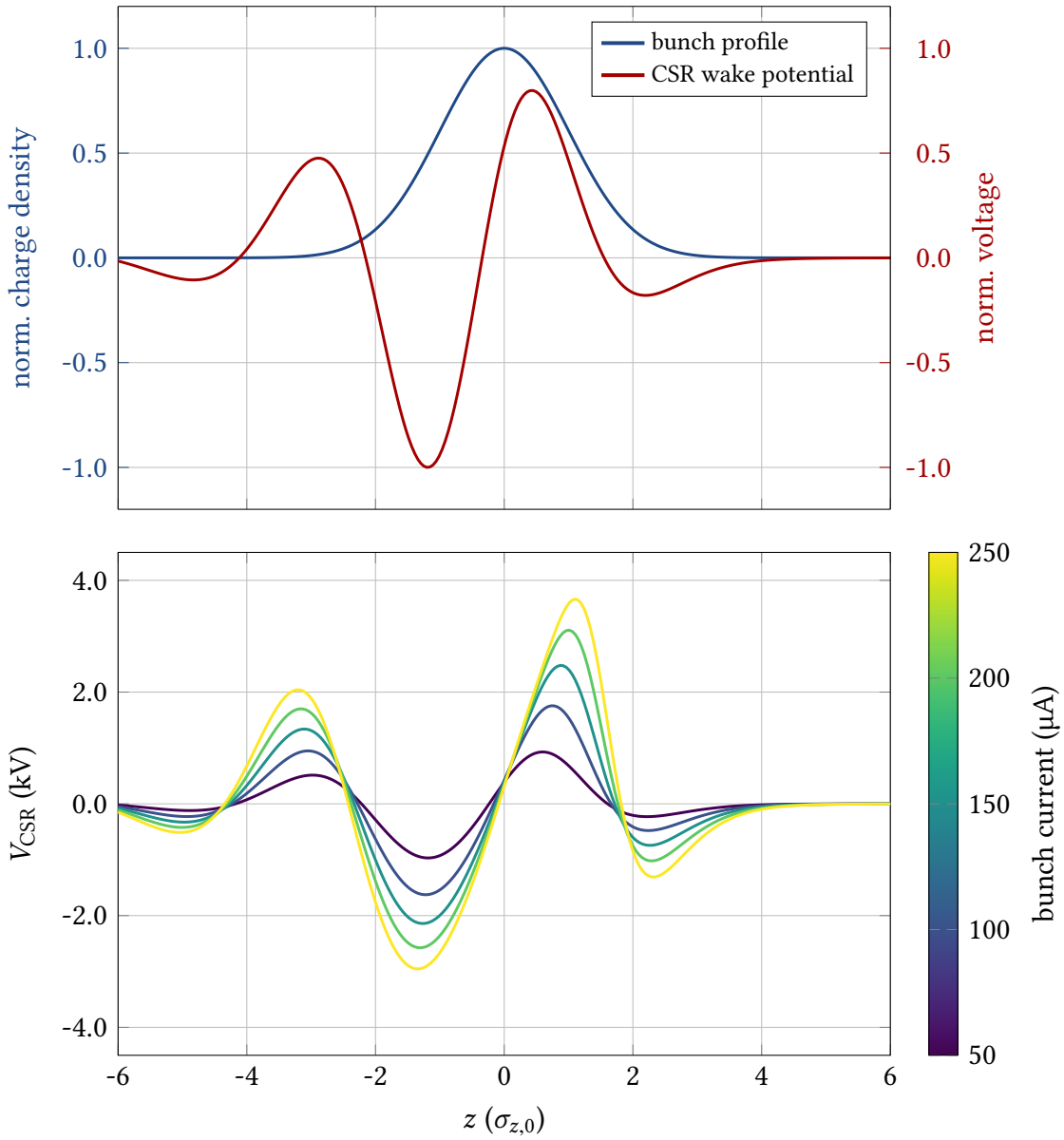



Figure 5.2.: CSR wake potential for a Gaussian bunch profile (red and blue, top) and for the bunch currents $I = (50, 100, 150, 200, 250)$ μA below the instability threshold of $I_{th} = 260$ μA (bottom). Shown are the respective temporal averages for simulations (including damping and diffusion) of the parameter settings defined for $\mathcal{D}_2$ in appendix A.1 (with $g = 32$ mm).

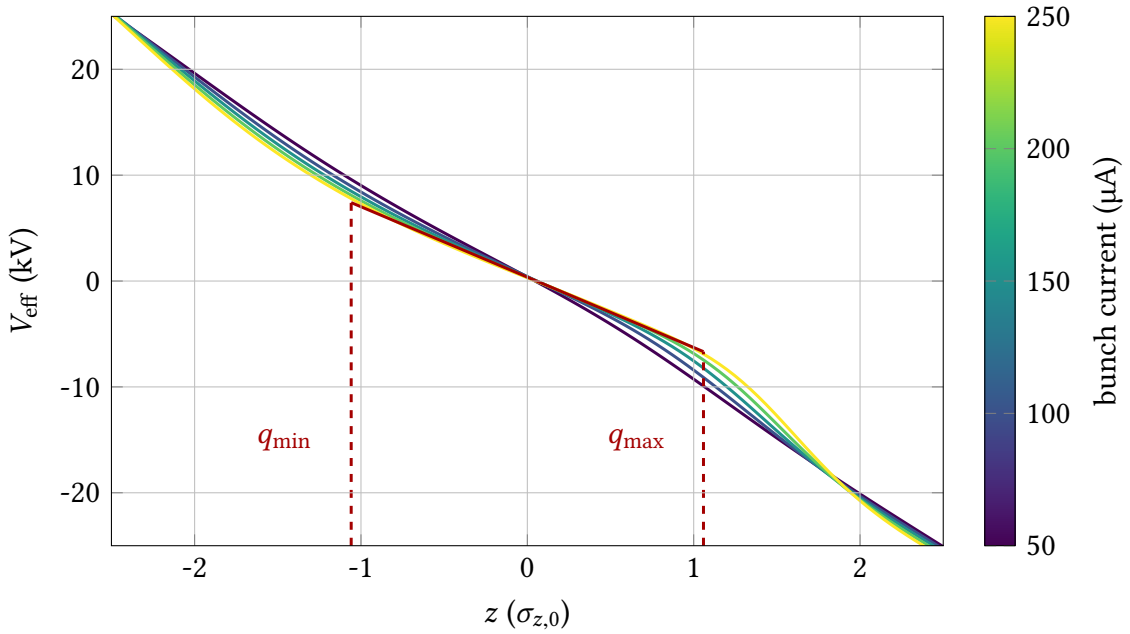



Figure 5.3.: Combining the RF potential $V_{\mathrm{RF}}(q)$ and the CSR wake potential $V_{\mathrm{CSR}}(q)$ yields an effective potential $V_{\mathrm{eff}}(q)$ that the electrons are exposed to during their revolution in the storage ring. Close to the synchronous position $q = 0$, the potential can be approximated by a linear function (shown as solid red line for an exemplary particle with an amplitude of roughly $q_{\mathrm{max}} \approx 1.1$). For larger deviations from $q = 0$ the linear approximation becomes less accurate, but still provides a useful estimate of the perturbed potential.

While the linear approximation seems quite suitable for particles with small deviations from the synchronous particle, the approximation is getting more inaccurate for larger oscillation amplitudes. Yet, as the majority of the charge is located close to $q = 0$, this constitutes the most interesting part of the potential, and the simple model is shown to yield reasonable approximations nonetheless. As is apparent from Fig. 5.3, the CSR wake potential acts as a perturbation to the slope of the RF potential with the strength of the perturbation being dependent on the amplitude of the particle's oscillation. According to Eqs. (5.5)–(5.11), this results in a position-dependent ellipticity of particle trajectories in phase space. Analogously, the oscillation frequency varies as a function of the particle's maximum deviation from the synchronous particle.

These insights can be verified by using the passive particle tracking method[1] that was recently added to Inovesa [81]. To that end, the initial charge distribution $\psi(q, p, t_0)$ is modeled by a particle ensemble of one hundred thousand particles, as illustrated in Fig. 5.4. Subsequently, the temporal evolution under the influence of the CSR wake potential is calculated simultaneously for both, the charge and the particle distribution. The thus

[1] The computation of the CSR wake potential is still based on the charge density function $\psi(q, p, t)$, the particles are tracked accordingly.

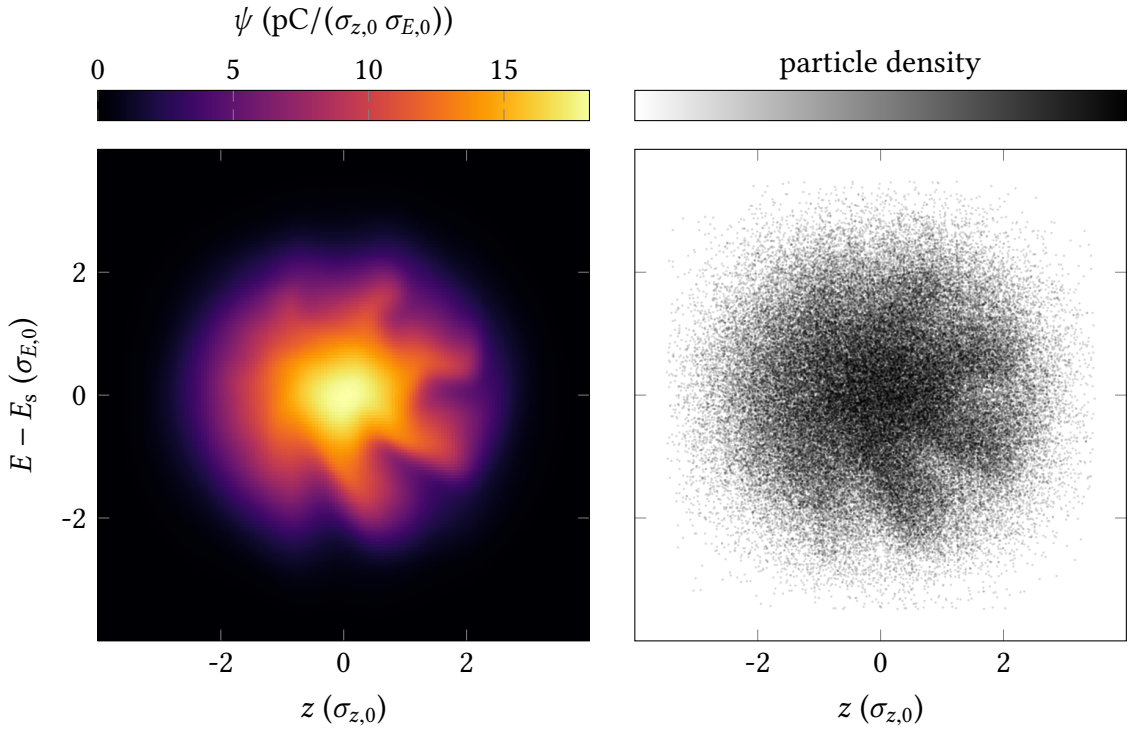



Figure 5.4.: In order to examine single particle trajectories, the initial charge distribution $\psi(q, p, t_0)$ (left) is modeled by a distribution of one hundred thousand particles (right). The single particle motion is then simulated using the passive particle tracking implemented in Inovesa.

obtained simulation results for the particle distribution are displayed in Fig. 5.5. Here, the upper part shows the difference of the maximum amplitude in $q$ and $p$

$$\Delta a_{\max}(q_{\max}) = q_{\max} - p_{\max} \,, \tag{5.13}$$

for the individual particles as a function of their maximum longitudinal deviation $q_{\max}$. The particle trajectories clearly deviate from $\Delta a_{\max} = 0$, which would correspond to a circular trajectory, and thereby show the expected position-dependent ellipticity. While the trajectories are already elliptical close to the origin, the maximum difference in amplitude is reached at values of $q_{\max}$ in the range of 1.0 to 1.5 depending on the bunch current. Both the maximum value of $\Delta a_{\max}$ and the corresponding longitudinal position are increasing with higher bunch currents due to the increased strength of the perturbation. For particles with larger deviation from the synchronous particle the amplitude difference reduces again, indicating a trend to more circular shaped trajectories for larger amplitudes. Additionally, the lower part of Fig. 5.5 displays the corresponding oscillation frequencies. As expected from Eqs. (5.5)–(5.11), the individual oscillation frequencies of particles close to $q = 0$ are significantly lower than the nominal synchrotron frequency $f_{s,0}$ due to the perturbation of the linear restoring force. Yet, this difference diminishes for particle trajectories with larger amplitudes yielding different oscillation frequencies dependent on the position of the particles within the bunch. These results can be directly compared to predictions

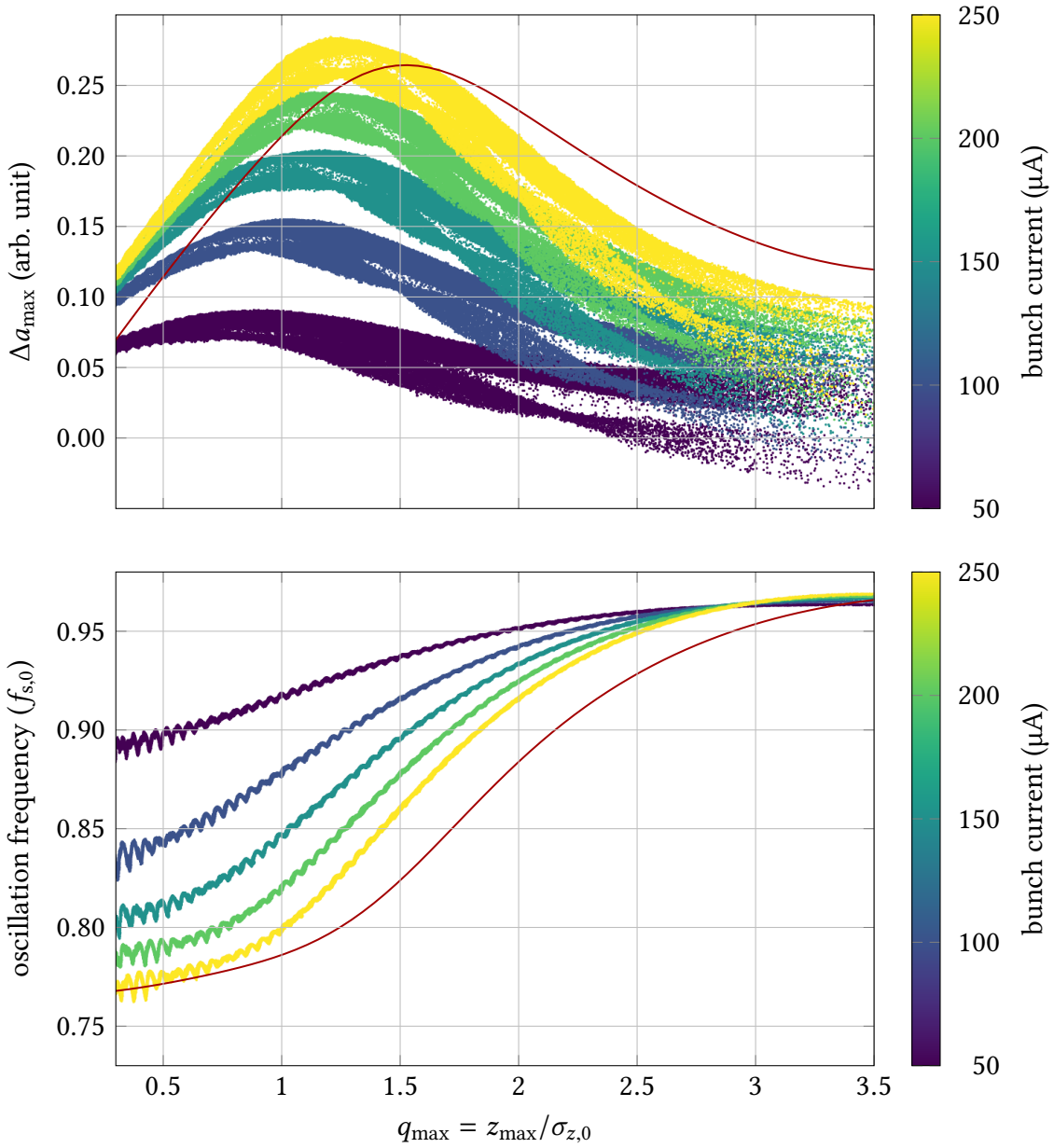



Figure 5.5.: Amplitude differences (top) and oscillation frequencies (bottom) of one hundred thousand particle trajectories simulated with Inovesa for several bunch currents below the instability threshold. The small oscillations of the frequencies, mainly visible in the range of $0 < q_{\max} < 1$, are presumed numerical artifacts of the simulation and data analysis. The solid red lines depict predictions based on the linear approximation of $V_{\text{eff}}(q)$ for the bunch current $I = 250\,\mu\text{A}$.

based on the linear approximation of $V_{\mathrm{eff}}(q)$ in Eq. (5.12). To do so, the estimated value of $k'(q_{\mathrm{max}})$ is used to determine the oscillation frequency at that position

$$f_{\mathrm{s}}(q_{\mathrm{max}}) = \omega_{\mathrm{s}}(q_{\mathrm{max}})/2\pi \approx \sqrt{k'(q_{\mathrm{max}})}/\zeta \ , \tag{5.14}$$

where $\zeta$ is just used for normalization purposes. Given an estimate of the oscillation frequency and using the approximation $p_{\mathrm{max}} \approx q_{\mathrm{max}}\, \omega_{\mathrm{s}}(q_{\mathrm{max}})$, the expected difference in amplitude is

$$\Delta a_{\mathrm{max}}(q_{\mathrm{max}}) \approx q_{\mathrm{max}} \left[1 - \omega_{\mathrm{s}}(q_{\mathrm{max}})\right] \ . \tag{5.15}$$

The calculated estimates are shown as solid red lines in Fig. 5.5 for the case of $I = 250\,\mu\mathrm{A}$. Clearly, the linear approximation of the effective potential in Eq. (5.12) is already sufficient for describing a major part of the perturbation by the CSR wake potential and its effect on the synchrotron motion of single particles. While the estimates of $f_{\mathrm{s}}(q_{\mathrm{max}})$ and $\Delta a_{\mathrm{max}}(q_{\mathrm{max}})$ deviate from the simulated trajectories for larger values of $q_{\mathrm{max}}$, this is expected due to the inaccuracy of the linear approximation of $V_{\mathrm{eff}}(q)$ for values further away from $q = 0$. In case of the amplitude difference $\Delta a_{\mathrm{max}}(q_{\mathrm{max}})$, the deviation already starts at small values

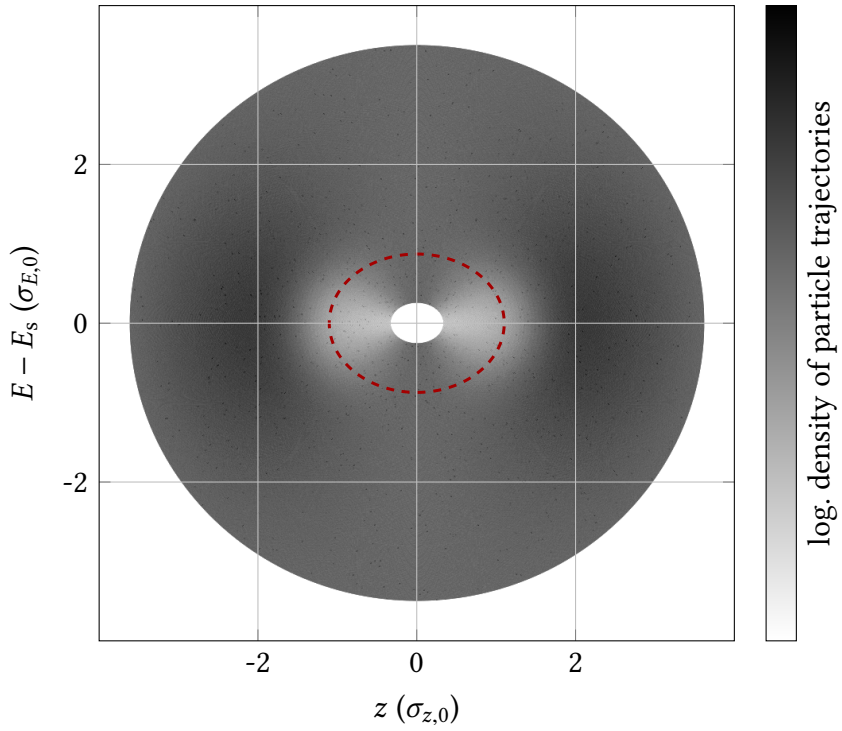



Figure 5.6.: Visualization of the effect of the position-dependent elliptical trajectories in phase space on the resulting charge density. Shown are perfectly elliptical trajectories of one thousand particles with uniformly distributed energies and an amplitude difference given by the estimate shown as solid red line in Fig. 5.5. The large number of trajectories helps to visualize the varying density of trajectories. In addition, the dashed red line depicts a single elliptical particle trajectory with $q_{\mathrm{max}} = z_{\mathrm{max}}/\sigma_{z,0} \approx 1.1$.

of $q_{\max}$, the general shape however, can still be reproduced.

In order to understand the implications of these modified single particle trajectories, an ensemble of particles with uniformly distributed energies and perfectly elliptical trajectories is considered. Thereby, the ellipticity is determined according to the position-dependent amplitude difference $\Delta a_{\max}(q_{\max})$ shown as solid red line in Fig. 5.5. As is apparent from the visualization in Fig. 5.6, this leads to a non-uniform distribution of particle trajectories in phase space. In particular, two distinguished locations of lower particle trajectory concentration are visible close to the origin. Similarly, though harder to identify by eye, there are two locations of higher particle trajectory concentration at larger oscillation amplitudes (roughly at $z \approx \pm 2\,\sigma_{z,0}$). This general pattern is a direct consequence of the basic shape of the CSR wake potential shown in Fig. 5.2. The CSR-induced perturbation of the RF potential thus breaks the homogeneity in phase space and creates local particle densities that form a quadrupole-like modulation of the longitudinal charge distribution. This inhomogeneity introduces a higher frequency component to the longitudinal bunch profile and may thereby initially seed the formation of micro-structures, that is, kick off the micro-bunching instability. It is worth noting that the general notion of dense particle trajectories leading up to a distinct charge modulation within the bunch resembles the caustic expression adopted for micro-bunching phenomena in linear accelerators [82].

This initial excitation of a quadrupole-like mode can be further verified by examining the Fourier transformed longitudinal bunch profiles across different bunch currents below the instability threshold. As is apparent from Fig. 5.7, the current-dependent perturbation by the CSR wake potential introduces a structure at higher frequencies to the longitudinal

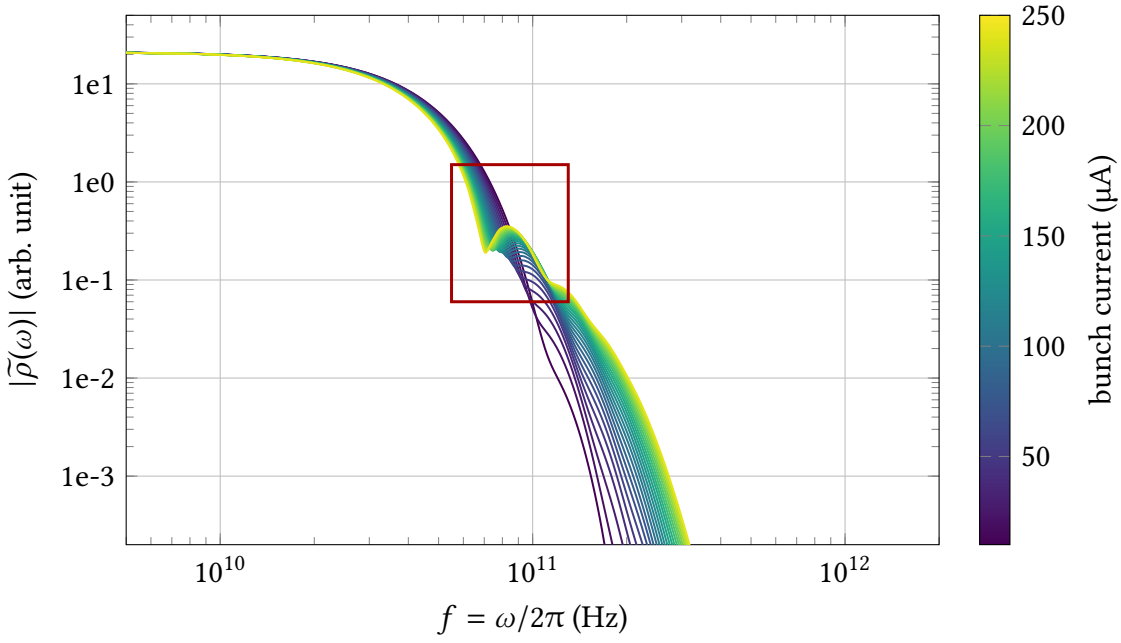



Figure 5.7.: Magnitude of the Fourier transformed bunch profile $|\widetilde{\rho}(\omega)|$ for a range of bunch currents below the instability threshold of $I_{\text{th}} = 260\,\mu\text{A}$. The red rectangle marks the additional frequency component (at roughly 85 GHz) that is excited due to the perturbation by the CSR wake potential.

charge distribution. For the used parameter settings this frequency is found at about 85 GHz corresponding to a modulation of the bunch profile at the wavelength

$$\lambda \approx c/85\,\mathrm{GHz} \approx 2.2\,\sigma_{z,0}\,, \tag{5.16}$$

which is roughly the distance of the two expected distinguished locations of decreased charge density in phase space. As the micro-structures occurring above the threshold current correspond to a significantly higher frequency (roughly at 150 GHz, shown in appendix A.4), the peak in Fig. 5.7 can clearly be attributed to the initial quadrupole mode.

## 5.3. Particle Motion above Threshold

The previous section considered only the perturbation of the single particle synchrotron motion created by the stationary CSR wake potential below the threshold current. Here, the linear approximation of the effective potential in Eq. (5.12) yields a simple model which sufficiently describes major aspects of the resulting particle trajectories and thereby facilitates understanding the implications of this perturbation. Essentially, the underlying longitudinal dynamics can be considered a simple one-dimensional harmonic oscillator with a position-dependent perturbation of the linear restoring force. Above the instability threshold, the longitudinal charge distribution as well as the CSR wake potential are not stationary anymore, which makes this simple model no longer applicable. Nevertheless, the notion of a perturbed restoring force extends to the dynamics just above the instability threshold and eventually motivates an approach towards control of the micro-bunching instability. The provided analysis builds upon extensive, prior studies of the micro-bunching dynamics above the instability threshold at KARA. It extends the gained understanding of the underlying longitudinal beam dynamics and offers a new perspective on the interpretation of previous observations.

In order to examine the single particle trajectories above the instability threshold, the initial charge distribution $\psi(q, p, t_0)$ is again modeled by a distribution of one hundred thousand particles and passively tracked using Inovesa. The upper part of Fig. 5.8 shows the amplitude difference $\Delta a_{\mathrm{max}}(q_{\mathrm{max}})$ of the resulting particle trajectories for a range of bunch currents in comparison to a current below the instability threshold ($I = 250\,\mu\mathrm{A}$, violet line). Due to the more complex dynamics and the variation of the individual particle trajectories in time, the data scatters a lot. In order to enable a comparison between multiple currents nonetheless, a moving average over $q_{\mathrm{max}}$ with a window length of 1000 data points is displayed. The elliptical shape is still apparent and seems quite comparable to the results below the threshold current in Fig. 5.5. The position of the maximum amplitude difference $\Delta a_{\mathrm{max}}(q_{\mathrm{max}})$ however, is slightly shifting to larger values. This implies that the local charge accumulation (see Fig. 5.6) also shifts to a position further away from the origin. Particularly interesting is the change in the distribution of the corresponding oscillation frequencies that is displayed below. While the higher bunch current leads to an abrupt change of the oscillation frequencies in the interval $q_{\mathrm{max}} \in [0.3, 1.5]$, it has a much smaller effect on the oscillation frequencies of particles with larger amplitudes. This region of $q_{\mathrm{max}} \in [0.3, 1.5]$ is precisely where the micro-structures occur in phase space and hints to the additional wake potential caused by the corresponding charge modulation.

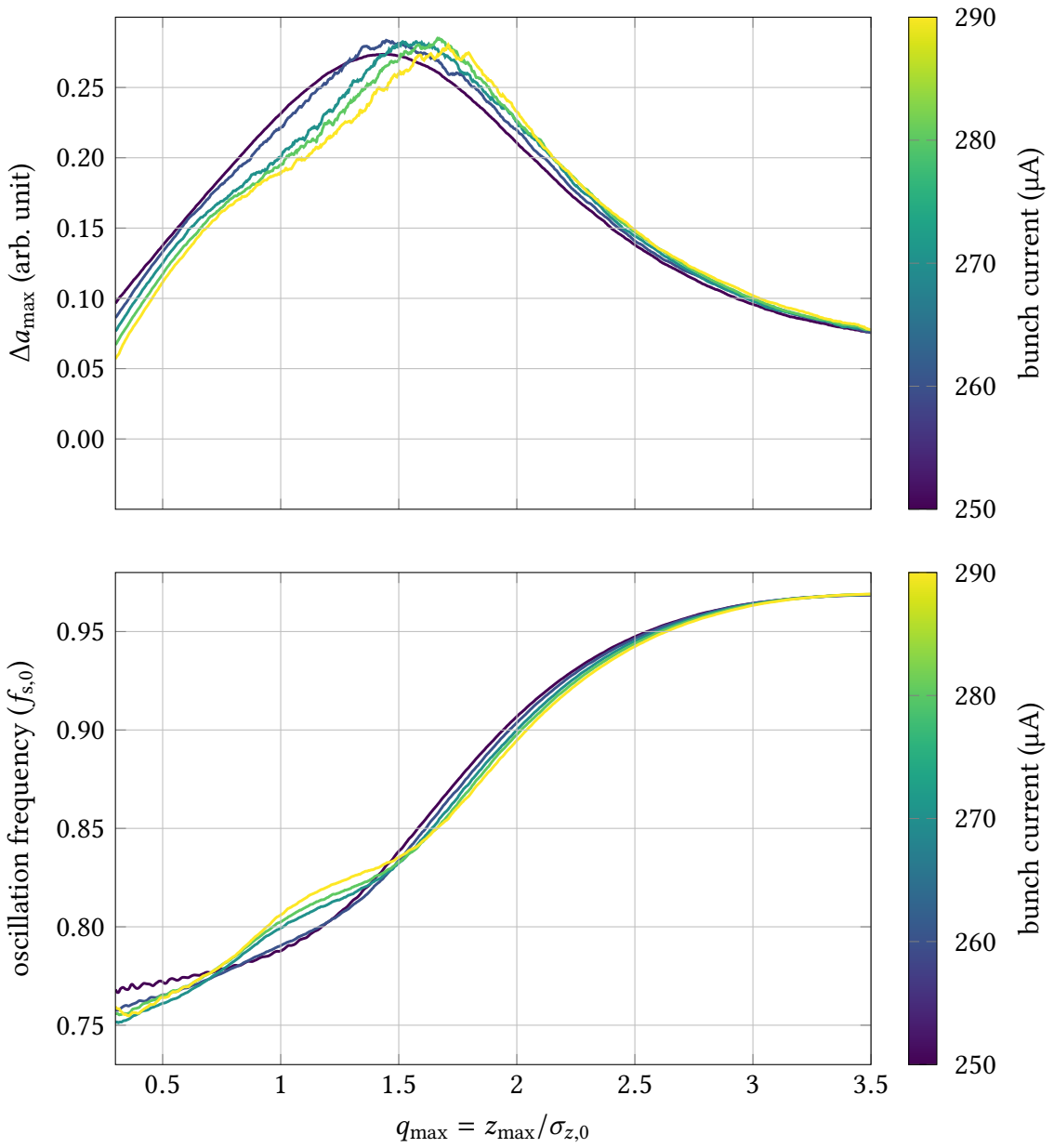



Figure 5.8.: Shown are the amplitude difference (top) and oscillation frequency (bottom) of particles trajectories for the bunch currents $I = (250, 260, 270, 280, 290)$ µA. As the data scatters a lot, only a moving average over $q_{\max}$ is displayed to enable a comparison between different currents.

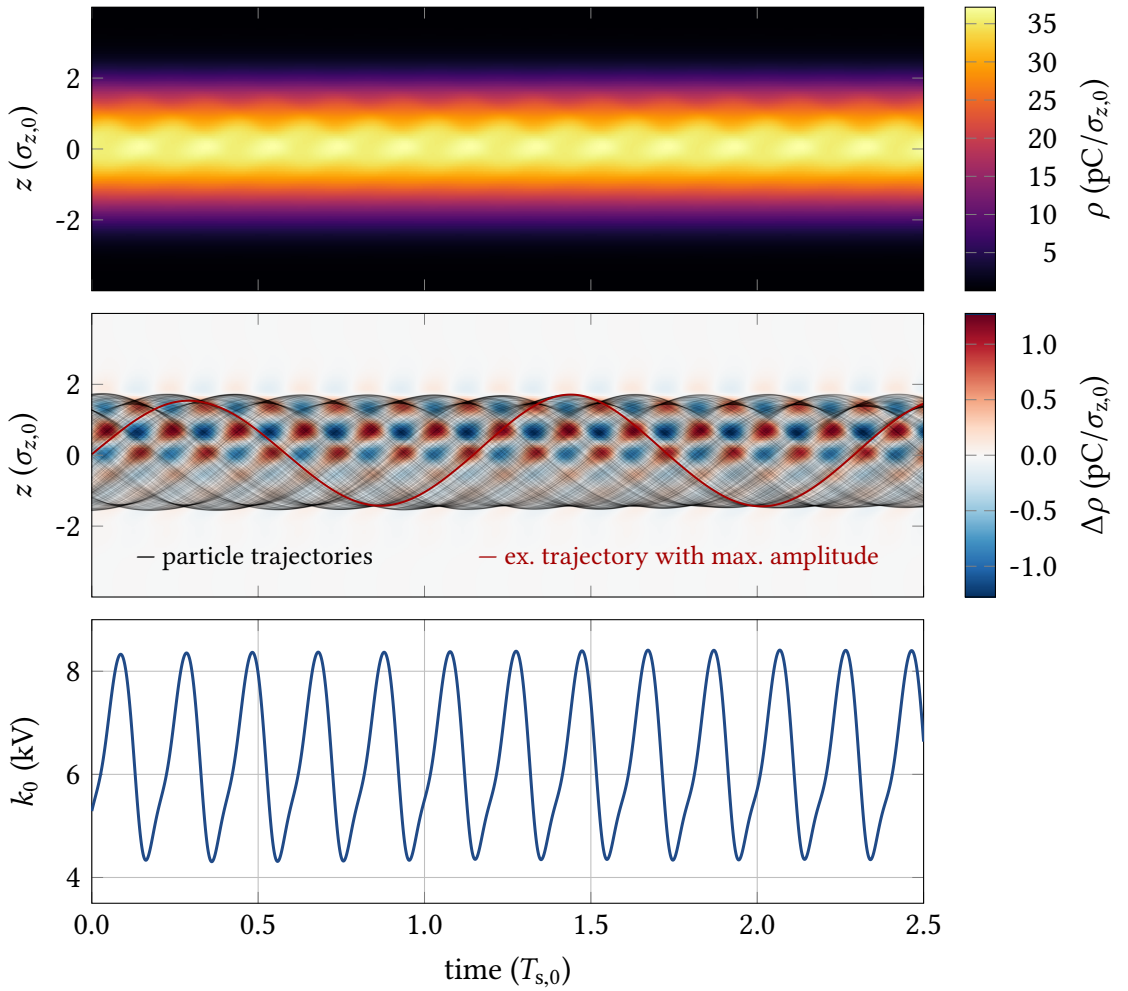



Figure 5.9.: Temporal evolution of the longitudinal bunch profile (top) and its difference to the temporal average (middle) for the bunch current $I = 290\,\mu A$. In order to visualize the corresponding single particle motion one thousand particle trajectories are plotted on top with an opacity of 0.05, displaying a distinct modulation of the maximum deviation in the longitudinal position. The maximum oscillation amplitude (solid red curve) is reached for particles that are exposed to the additional wake potential caused by the local charge modulation. These dynamics are largely driven by the CSR-induced dynamic perturbation of the strength of the restoring force, estimated at the position of the synchronous particle by $k_0$ (bottom, linear fit in the interval $[-0.2, 0.2]\,\sigma_{z,0}$).

To factor in the temporal dynamics above the instability threshold, a different way of visualizing the individual particle trajectories is required. The upper part of Fig. 5.9 thus displays the temporal evolution of the longitudinal bunch profile over a time period of two and a half synchrotron periods for the bunch current $I = 290\,\mu\text{A}$. By close examination, the periodic modulation of the charge density due to the occurring micro-structures can already be identified. Nevertheless, the difference to the temporal average $\Delta\rho(z)$ is again shown below, which displays the micro-bunching dynamics much more explicitly. In order to examine how the motion of single particles relates to these micro-bunching dynamics one thousand particle trajectories are plotted on top with an opacity of 0.05. These trajectories are deliberately chosen to have an average radius in phase space

$$\bar{r} = \frac{1}{n}\sum_{i=1}^{n} r_{t_i} = \frac{1}{n}\sum_{i=1}^{n}\sqrt{q_{t_i}^2 + p_{t_i}^2}\,, \tag{5.17}$$

which is comparable to the estimated distance of the micro-structures from the origin. Finally, the bottom part of Fig. 5.9 shows the estimated slope of the effective potential at the position of the synchronous particle

$$k_0(t) \approx -\left.\frac{\partial V_{\text{eff}}(q,t)}{\partial q}\right|_{q=0}\,. \tag{5.18}$$

The almost sinusoidal oscillation of $k_0(t)$ clearly coincides with the periodic modulation of the charge density. In the following subsections, different aspects of the particle motion above the instability threshold are examined using the insights of previous sections.

### 5.3.1. Head-Tail Asymmetry

Particularly intriguing is the distinct modulation of the maximum amplitude $q_{\text{max}}$ that is visible at the head of the bunch ($q > 0$). This modulation is perfectly synchronized to the micro-structure dynamics in the charge density and reaches its extrema at exactly the same positions in time. However, similarly to the charge modulation, it predominantly occurs at the head and diminishes towards the tail of the bunch. The maximum amplitude in $q$ is reached when a particle travels on a trajectory that leads to its exposure to an additional contribution in the CSR wake potential caused by the local charge modulation. Particles traveling exactly along the position of the maximum local charge density (red areas) while passing through $q \in [0, 2]$ are subsequently driven to the largest deviations $q_{\text{max}}$ (illustrated by the solid red curve in Fig. 5.9). Similarly, particles passing through the minima (blue areas) end up closest to the origin.

This asymmetry can be explained by the different effects a local structure at different positions in the charge density has on the restoring force. For $q > 0$, a positive contribution to the effective potential $V_{\text{eff}}$ results in a further decrease of the restoring force and thus drives particles further outside in phase space. This amplifies the inhomogeneity and can thereby drive and support the local charge modulation. In contrast, a positive contribution at $q < 0$ partially recovers the strength of the restoring force and focuses the particles towards the center of the charge distribution, reducing the inhomogeneity and damping the local structure.

While this explains why the micro-structures are more pronounced at the head of the bunch rather than the tail, it simultaneously illustrates how single particle motion is leading up to these local charge modulations. Particles are driven outside in phase space, cause an excess of charge at that position and thereby form the occurring micro-structures.

### 5.3.2. Formation of Micro-Structures

To investigate how the motion of individual particles relates to the motion of the observed micro-structures, this subsection takes a look at the location of particles one synchrotron period before they form a local structure. To do so, Fig. 5.10 displays the particle distribution at time step $t = 0\,T_{s,0}$ in two different ways. On the left hand side each individual particle is colored according to the relative charge density at this time step

$$\text{color}(n = i) \leftarrow \Delta\psi(q_{n=i}, p_{n=i}, t_{\text{color}} = 0\,T_{s,0}) \,, \tag{5.19}$$

where $n = i$ is the particle index and $(q_{n=i}, p_{n=i})$ denotes its location in phase space at time $t_{\text{color}}$. The color assignment on the right hand side is adjusted to match the relative charge density one synchrotron period afterwards $\Delta\psi(q_{n=i}, p_{n=i}, t_{\text{color}} = 1\,T_{s,0})$ instead. The distribution on the right hand side thus shows where the particles forming the micro-structures at time step $t = 1\,T_{s,0}$ were one synchrotron period before. Thereby, one can analyze where the particles forming the local structures come from and whether or not they stay within these structures. Clearly, the two distributions look quite different. This implies that the particles forming the structures at $t = 0\,T_{s,0}$ do not necessarily participate in forming the same kind of structure one synchrotron period later. They might for example travel from the position of a local maximum (red) to position of a local minimum (blue) or vice versa. This effect was already observed in previous studies [13].

Fig. 5.10 illustrates conceptually that the formation of these micro-structures is not merely caused by the resonant motion of single particles, but rather by the collective effect of many particles traveling on varying trajectories. Moreover, the relation between the motion of individual particles and the occurring micro-structures was found to vary significantly across different bunch currents and parameter settings (an exemplary, different bunch current is shown in appendix A.5).

### 5.3.3. Micro-Structure Frequency

As discussed in section 3.4, the occurrence of micro-structures in the longitudinal phase space results in fluctuations of the emitted CSR power. Right above the instability threshold, this fluctuation is typically dominated by the characteristic frequency $f_{\text{ms}}$. As reported by different facilities, e.g. [29, 32, 36], this frequency is usually observed close to an integer multiple of the nominal synchrotron frequency

$$f_{\text{ms}} \approx m\, f_{s,0} \,, \tag{5.20}$$

but may deviate significantly depending on the parameter settings [29, 32]. This subsection aims to illustrate how this frequency originates from the micro-bunching dynamics in the longitudinal phase space.

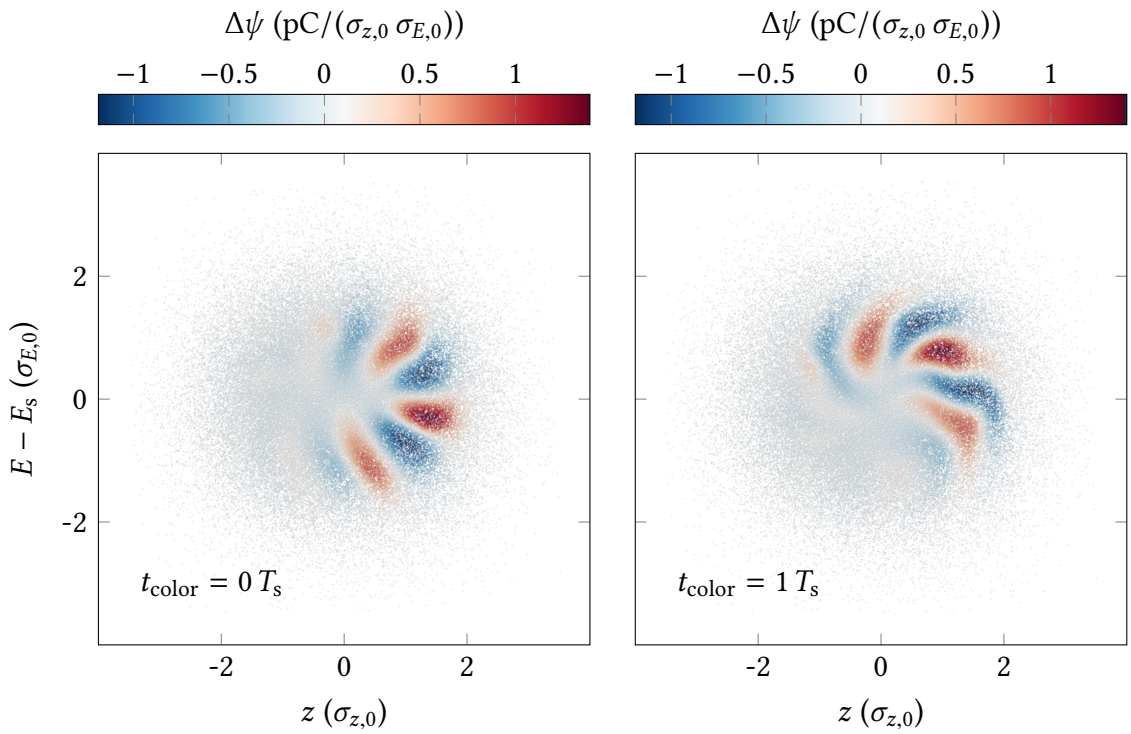



Figure 5.10.: To illustrate the relation between single particle motion to the formation and propagation of micro-structures, the particle distribution at time step $t = 0\,T_{s,0}$ for the bunch current $I = 290\,\mu A$ is depicted in two different ways. On the left hand side, each individual particle is colored according to the corresponding relative charge density $\Delta\psi(q, p, t_{\text{color}} = 0\,T_{s,0})$ at that particle's location (opacity of 0.15). On the right hand side, the color assignment is adjusted to match $\Delta\psi(q, p, t_{\text{color}} = 1\,T_{s,0})$ instead. It thereby illustrates which particles will form the micro-structures one synchrotron period after this time step.

The integrated power of the emitted CSR, $P_{\text{CSR}}(t)$, is solely determined by the longitudinal bunch profile $\rho(q, t)$, and the CSR impedance. Different charge distributions in phase space at different time steps $t_{i,j}$ thus result in the exact same value of the radiated power if they correspond to identical longitudinal profiles $\rho(q, t_i) \doteq \rho(q, t_j)$. Consequently, the periodic structure of the fluctuating CSR power can be explained by a repetitive sequence of the longitudinal profiles. The most trivial way to produce similar longitudinal profiles at different time steps is by having the charge densities in phase space be similar as well

$$\psi(q, p, t_i) \approx \psi(q, p, t_j)\ . \tag{5.21}$$

In the simulations with Inovesa this was always found to be the case. The charge densities separated by one period of the micro-structure frequency $\Delta t = 1/f_{\text{ms}}$ are nearly indistinguishable by eye and only differ by small numeric values. The observed micro-structure frequency is thus directly determined by the periodic behavior of the micro-bunching dynamics and the corresponding time interval. However, as established above, the propagation of these micro-structures in phase space and in time is a non-trivial subject. With

the varying oscillation frequencies of the single particle trajectories (illustrated in Fig. 5.8) and the collective formation of the occurring micro-structures, very little can be deduced about the oscillation frequency of the micro-structures themselves. Empirically, it was found to deviate up to 20 % from the nominal synchrotron frequency. In that case, the simple estimate in Eq. (5.20) is no longer applicable. In particular, trying to estimate the integer $m$ via

$$m \approx f_{\mathrm{ms}} / f_{\mathrm{s},0} \tag{5.22}$$

will yield inconsistent and unreliable results as the micro-structures may propagate at differing frequencies. Lastly, it is worth noting that the interpretation of $m$ as a simple azimuthal mode number is difficult to align with the head-tail asymmetry of the CSR self-interaction discussed in subsection 5.3.1. This asymmetry explains why the micro-structures always form at the head of the bunch and are more pronounced there. The micro-structure frequency observed in the emitted CSR power is simply determined by the repetition rate of this formation process.

### 5.3.4. Dependence on Shielding

Figures 3.8 and 3.10 as well as Fig. 5.9 illustrate the distinct charge modulation that forms under the influence of CSR self-interaction in the micro-bunching instability. Evidently, the modulation pattern is asymmetric as the micro-structures change shape and notably amplitude depending on their position in phase space. Nonetheless, one might be willing to identify an integer number $n_{\mathrm{str}}$ of maxima constituting the charge modulation. Modifying Eq. (5.20), we can relate this to the micro-structure frequency

$$f_{\mathrm{ms}} = n_{\mathrm{str}}\, f_{\mathrm{str}}\ , \tag{5.23}$$

where $1/f_{\mathrm{str}}$ denotes the time it takes the micro-structures to perform one full revolution in phase space. As discussed above, this is only loosely related to the oscillation frequency of single particles and may deviate from the nominal synchrotron period. As already mentioned in [83], the integer $n_{\mathrm{str}}$ changes if the shielding by the vacuum pipe is altered by varying its height. This subsection considers the corresponding synchrotron motion across these different parameter settings. Figure 5.11 thus displays the oscillation frequencies of the micro-structures (red) and the single particle frequencies at their positions in phase space (blue). All frequencies are estimated directly above the instability threshold $I_{\mathrm{th}}$, which itself is changing due to the varying shielding [8]. The dashed red line connects data points with the same number of maxima $n_{\mathrm{str}}$ in the charge modulation pattern, which was estimated by examining the charge density difference $\Delta\psi(q, p, t)$ by eye. With increased shielding (reduced vacuum gap) the particle frequencies are growing quite smoothly. The oscillation frequencies of the micro-structures, however, show a differing behavior. While they take on very similar values for the vacuum gaps $(19, 22, 24, 27, 32)$ mm, the micro-structure oscillation frequencies are growing much faster with decreasing vacuum gap. This can be observed up until the point where an additional extremum is identified and $n_{\mathrm{str}}$ is incremented. Afterwards, the frequencies of particles and micro-structures are found at very similar values again (for example after the transition from 28 mm to 27 mm). Figure 5.11 thereby illustrates again the partial decoupling of the micro-structure propagation in phase space from the single particle motion.

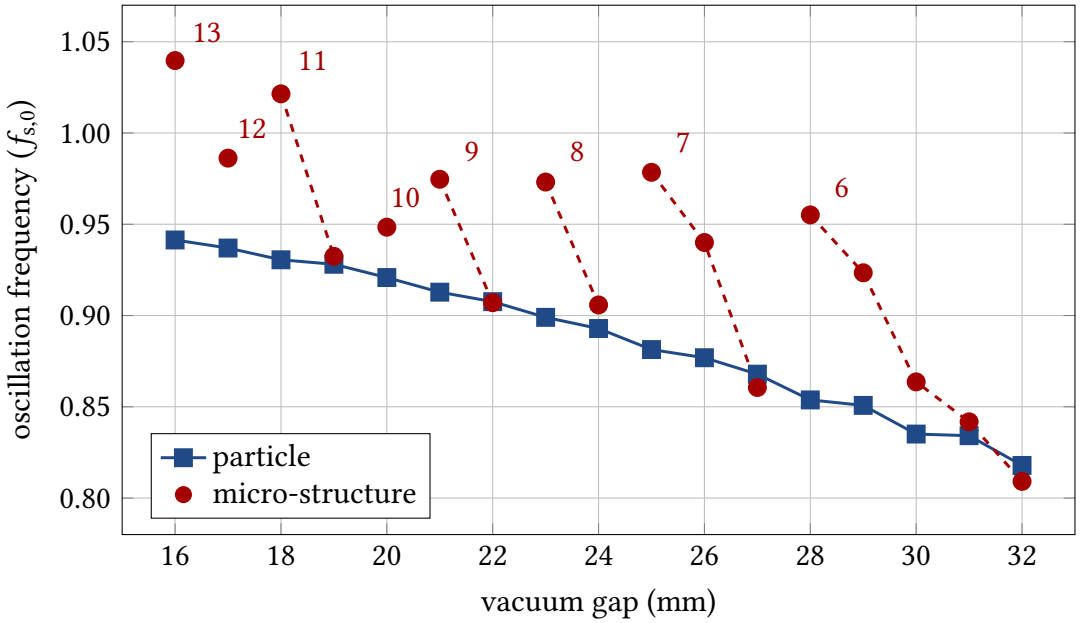



Figure 5.11.: Modifying the CSR shielding by varying the vacuum gap leads to altered micro-bunching dynamics in the longitudinal phase space. While the single particle frequency at the micro-structure position (blue squares) increases smoothly with reduced vacuum gap, the oscillation frequency of the micro-structures themselves (red circles) grows much faster and shows an abrupt change when an additional structure is observed.

### 5.3.5. Amplitude and Position of Micro-Structures

As explained in subsection 5.3.1, the additional wake potential caused by the micro-structures at the head of the bunch can support and drive the micro-bunching dynamics. Larger local charge modulations lead to a larger perturbation by the additional wake potential which then results in the individual particles being driven to larger oscillation amplitudes. Following this chain of thought, naturally we expect a correlation between the maximum amplitude of the occurring micro-structures and their maximum longitudinal deviation from the origin in phase space. In order to verify this, Fig. 5.12 displays the maximum amplitude and maximum longitudinal position of the occurring micro-structures for a range of bunch currents between 260 µA and 1000 µA. With increasing current the strength of the perturbation caused by the CSR self-interaction increases, which leads to larger amplitudes of the micro-structures within the bunch. Figure 5.12 illustrates that this corresponds to a larger deviation from the synchronous particle as the particles are exposed to a stronger CSR wake potential. For the simulation data considered here, a clear linear correlation (as illustrated by the dashed red line) with a correlation coefficient of 0.91 is found.

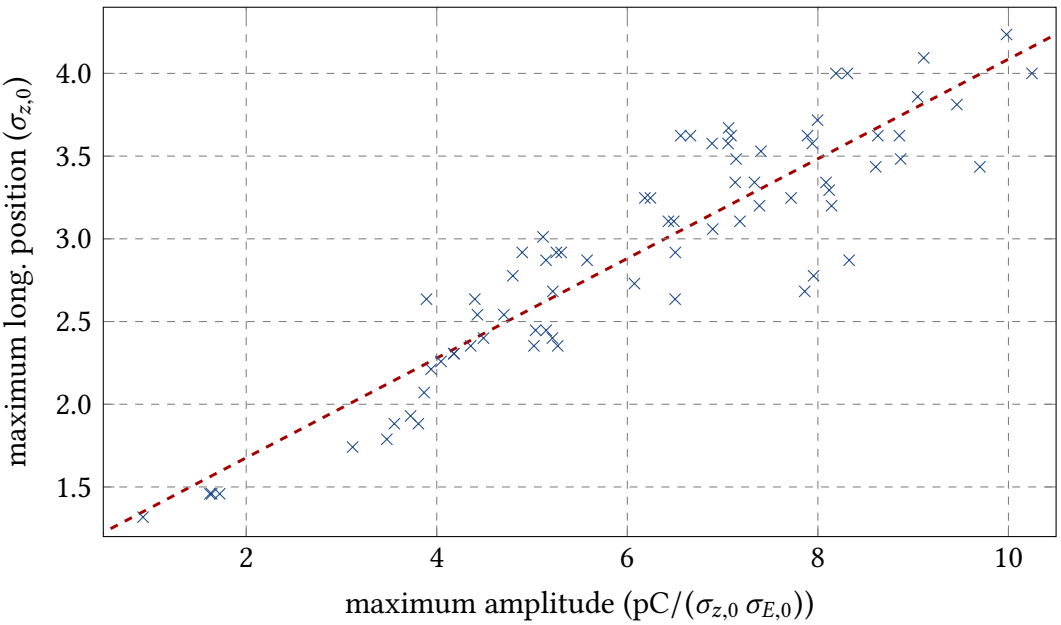



Figure 5.12.: Correlation between the maximum amplitude and maximum longitudinal position of the occurring micro-structures. Shown is simulation data for bunch currents between 260 µA and 1000 µA. The dashed red line illustrates the clear linear correlation with a correlation coefficient of 0.91.

## 5.4. Implications and Further Questions

In summary, the perturbation of the synchrotron motion studied in the preceding sections amounts to an intuitive explanation of how micro-structures form under the influence of CSR self-interaction. The perturbation caused by the stationary CSR wake potential below the threshold current leads to a reduced strength of the restoring force exerted by the RF system. Particles are thus driven to larger deviations from the synchronous particle forming a quadrupole-like deformation of the longitudinal charge distribution. This introduces a higher frequency component to the bunch profile and the corresponding wake potential, which grows with increasing bunch current (as visible in Fig. 5.7). Eventually, the equilibrium between charge distribution and CSR wake potential becomes unstable and their interaction leads to the dynamic formation of micro-structures in the longitudinal phase space. This formation process is again largely driven by a dynamic perturbation of the restoring force as illustrated in Fig. 5.9. Depending on their oscillation phase, individual particles are exposed to additional contributions to the CSR wake potential and are thereby driven to smaller or larger deviations from the synchronous particle. The accumulation of particle trajectories at specific locations in phase space causes an excess of charge at that position and creates local structures. Besides motivating an approach towards control of the micro-bunching dynamics, which is discussed in the final section of this chapter, this offers a new perspective on different aspects of the instability and raises a range of further questions briefly addressed below.

The quadrupole-like deformation of the charge distribution found below the threshold current merits further studies to improve the understanding of the initial formation of dynamic structures from a stationary charge distribution. Moreover, it is worth pointing out that the additional weak instability introduced in subsection 3.4.2, is usually observed with a micro-structure frequency of $f_{\mathrm{ms}} \approx 2\, f_{\mathrm{s},0}$. This corresponds to a non-stationary quadrupole mode in the longitudinal charge distribution, which aligns remarkably well with the stationary quadrupole-like deformation below the threshold current. Moreover, measurements of such a quadrupole-like deformation of the longitudinal charge distribution were already reported in [84], albeit for bunch currents above the instability threshold.

Another promising subject for further studies is the dependence of the micro-bunching dynamics on shielding as illustrated in Fig. 5.11. The distinct transitions in the number of observed micro-structures and the associated changes in their oscillation frequency are particularly interesting as they provide additional insight into the formation process of the micro-structures under different boundary conditions. They may also hint towards an intrinsic discretization of the micro-bunching dynamics as already mentioned in [83].

As primarily discussed in the subsections 5.3.2 and 5.3.4, the propagation of the micro-structures in the longitudinal phase space may deviate from the motion of single particles. Nonetheless, given the general distribution of particle oscillation frequencies in Figs. 5.5 and 5.8, it seems reasonable to assume that also the collectively formed structures propagate faster if they are located further away from the origin in phase space. Based on this hypothesis a basic reasoning can be derived which explains the occurrence of several characteristic features in the simulated and measured CSR power spectrograms (example shown in Fig. 3.6). The increasing bunch current leads to a stronger perturbation by the CSR wake potential and thus to a larger amplitude of the occurring micro-structures. As established in subsection 5.3.5, this corresponds to a drift of the micro-structure position towards larger longitudinal deviations in phase space. Granted that this leads to a faster oscillation of the structures in phase space ($f'_{\mathrm{str}} > f_{\mathrm{str}}$), the micro-structure frequency is, according to Eq. (5.23), increasing as well. While this explains the slight shift of the micro-structure frequency across current directly above the threshold, these simple initial dynamics are only observed in a comparably small current range. In Fig. 3.6, the dominant frequency fans out at roughly 120 µA marking a clear transition in the occurring dynamics. At these higher bunch currents, the micro-structures reach an amplitude that can no longer be supported by the corresponding CSR wake potential. Reaching this amplitude, the bunch blows up in size and the structures smear out in phase space. As the increased bunch length leads to a reduced perturbation by CSR, the bunch is mainly shrinking due to radiation damping afterwards. Once the bunch length is short enough, the micro-structures emerge again and continuously grow in amplitude. When forming initially, the structures are located close to the origin in phase space and propagate rather slowly, which corresponds to a low micro-structure frequency observed in the CSR power signal. With increasing amplitude the micro-structures are then driven further outside in phase space and accelerate in oscillation frequency. This leads to a sawtooth-shaped burst of CSR emission up until the point where the micro-structure amplitude becomes too large and the cycle starts anew. The spread out micro-structure frequency would thus be caused by the varying oscillation frequencies of the micro-structures during their continuous growth in phase space. These considerations motivate a future study that concentrates

on a more detailed analysis of the temporal evolution of the micro-structure frequency in both, simulations and measurements. The rather low frequencies at the edge of the figure, however, correspond to the repetition rate of the described bursting cycle and are thus related to the longitudinal damping time as shown in [85]. The approach towards control pursued in the following chapters instead concentrates on time scales comparable to the formation process of the micro-structures, which is governed by the synchrotron period.

## 5.5. Necessity of Dynamic Control

One of the key insights of the preceding chapter is that the micro-bunching dynamics above the instability threshold correspond to a modulation of the restoring force, that is, the slope of the effective potential $V_{\mathrm{eff}}(q)$. That naturally motivates an approach to control of the micro-bunching dynamics by aiming to increase or counteract this perturbation with an RF amplitude modulation

$$V_{\mathrm{RF}}(t) = \hat{V}(t)\sin(2\pi f_{\mathrm{RF}}\, t)\ , \tag{5.24}$$

$$\hat{V}(t) = V_0 + V_{\mathrm{mod}}\sin(2\pi f_{\mathrm{mod}}\, t + \varphi_{\mathrm{mod}})\ . \tag{5.25}$$

While the perturbation by the CSR wake potential cannot be compensated in its entirety, this aims at stabilizing the strength of the restoring force and thereby mitigating the micro-bunching dynamics, as illustrated in Fig. 5.13. By choosing the natural micro-structure frequency of the occurring instability (can be observed in the emitted CSR signal) as the

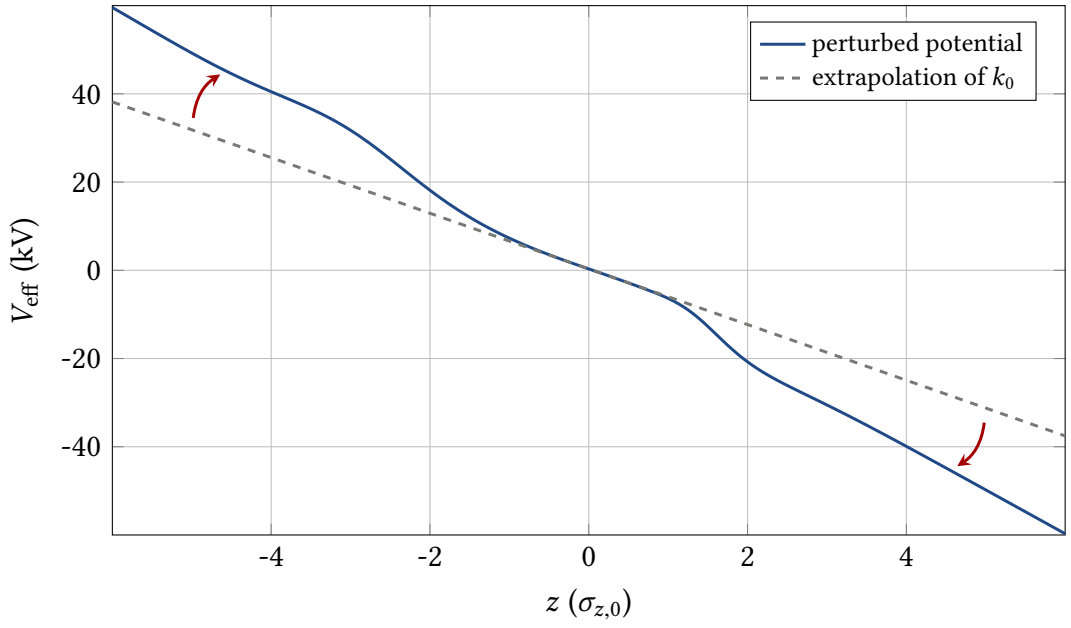



Figure 5.13.: The dynamic perturbation of the slope of the effective potential at the synchronous position $k_0$ is found to be critical for the micro-structure formation process. By modulating the RF amplitude, the CSR-induced perturbation can be partially counteracted.

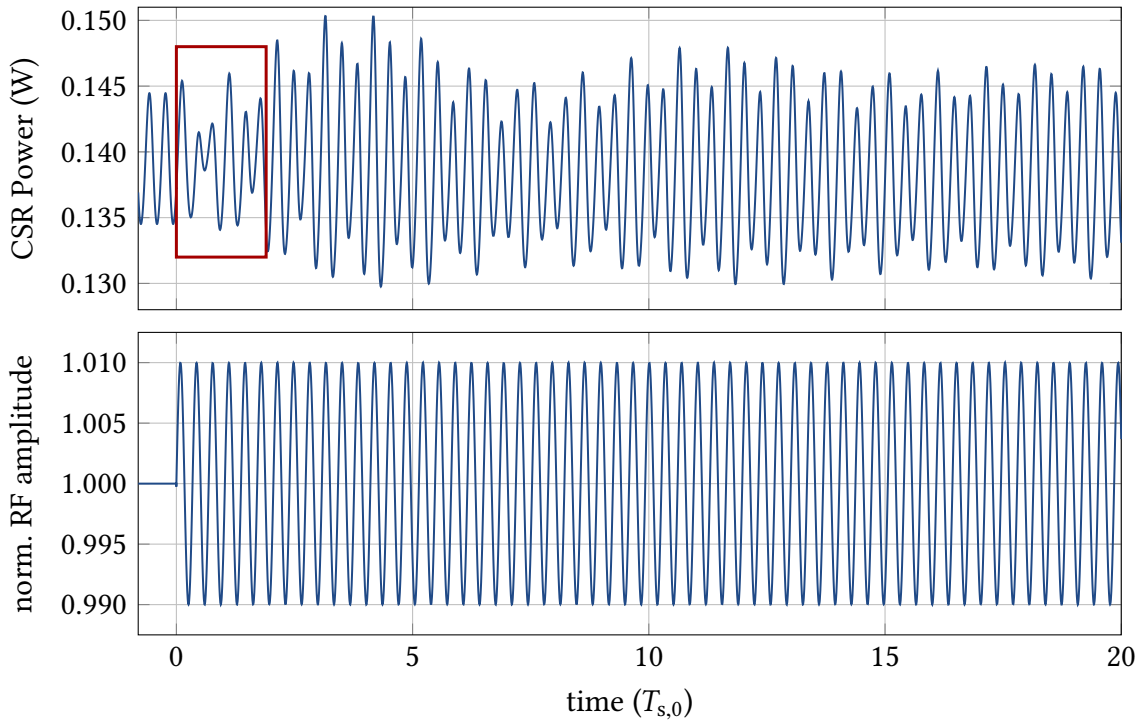



Figure 5.14.: An RF amplitude modulation at the micro-structure frequency with the appropriate phase can partially recover the strength of the restoring force as illustrated in Fig. 5.13. Here, the application of a modulation with $V_{\mathrm{mod}} = 0.01\,V_0$ leads to an initial reduction of the CSR power oscillations, as marked by the red rectangle (parameter settings according to $\mathcal{D}_1$ with $I = 115\,\mu\mathrm{A}$). Yet, after merely a few synchrotron periods, the control signal runs out of sync and amplifies the oscillation instead.

modulation frequency $f_{\mathrm{mod}} \doteq f_{\mathrm{ms}}$, and carefully adjusting the amplitude $V_{\mathrm{mod}}$ and phase $\varphi_{\mathrm{mod}}$, the RF modulation can be expected to partially compensate the perturbation by the CSR wake potential. Figure 5.14 demonstrates how this can lead to a reduction of the oscillations in the CSR power signal, which corresponds to a mitigation of the occurring micro-bunching dynamics. Yet, the initial damping effect only lasts for a few synchrotron periods. In fact, the magnitude of the subsequent oscillation is even slightly higher than that caused by the natural micro-bunching dynamics. This can be explained by the dynamic self-interaction of the charge distribution and the CSR wake potential. Once the applied RF modulation starts interfering with the natural beam dynamics, the evolution of the longitudinal charge distribution and thus the CSR wake potential are changing as well. Particularly the deliberately chosen modification of the restoring force leads to an altered oscillation frequency and thus significantly affects the synchrotron motion of the present micro-structures. Continuously applying this RF amplitude modulation may initially have the desired effect, but will eventually run out of sync with the perturbation it is aiming to counteract. In that case, the RF modulation may no longer stabilize the restoring force,

but actually further drive the instability. As demonstrated in section 7.2, an RF amplitude modulation with a constant frequency and amplitude may thus be used to excite the micro-bunching dynamics to higher amplitudes. Continuous mitigation, however, is a more challenging task as the RF modulation has to be adjusted over time according to the altered micro-bunching dynamics

$$V_{\mathrm{mod}} \rightarrow V_{\mathrm{mod}}(t) \quad \text{and} \quad f_{\mathrm{mod}} \rightarrow f_{\mathrm{mod}}(t)\,, \tag{5.26}$$

which leads to a dynamic RF amplitude modulation scheme

$$\hat{V}(t) = V_0 + V_{\mathrm{mod}}(t)\,\sin(2\pi f_{\mathrm{mod}}(t)\,t + \varphi_{\mathrm{mod}})\,. \tag{5.27}$$

The necessity of continuous adjustment of the modulation amplitude and frequency constitutes a sequential decision problem, which motivates the use of reinforcement learning in the following chapters.

# 6. Feedback Design

*Nothing ever is, everything is becoming. (panta rhei)*

— Heraclitean philosophy

The idea of using a dynamic RF amplitude modulation scheme to counteract the perturbation caused by the CSR self-interaction, which was developed in the previous chapter, inherently leads to a sequential decision problem. At a repetition rate that matches the time scale of the micro-structure formation process, the RF modulation has to be continuously adjusted according to the altered perturbation by the CSR wake potential. In order to apply reinforcement learning solution methods, this chapter aims at defining the task as a formal RL problem. The general idea of addressing the problem via RL methods emanates from an interdisciplinary collaboration at KIT between the Laboratory for Applications of Synchrotron Radiation (LAS), the Institute for Beam Dynamics and Technology (IBPT) and the High Performance Humanoid Technologies (H2T) group in the Institute for Anthropomatics and Robotics (IAR) and was first published in [86]. As the definition of several essential RL elements, like the reward function or the observed feature vector, are not immediately apparent, this is not a straightforward task. In fact, the formulation given below is the result of various iterations over the past years and still subject to change should inconsistencies or a deviation from the intuitive understanding of the physical objective arise. Nonetheless, the final formulation stated below allowed for the proof-of-principle control presented in chapter 7.

From a reinforcement learning perspective, the problem of micro-bunching control constitutes a continuing task with a naturally continuous state and action space. One major advantage in satisfying the conditions of a formal RL problem is that the definition of a Markovian state is straightforward. Yet, this information is only partially available at the actual storage ring. Section 6.2 thus introduces two different formulations of the problem: one that is theoretically sound with a perfect fulfillment of the Markov property, and one that represents a more feasible option tailored towards the implementation at KARA. Furthermore, the complementary objectives of either mitigation or excitation of the micro-bunching dynamics are addressed in section 6.3.

## 6.1. Choice of Action Space

Given the insights of the previous chapter, the action space is simply defined as

$$\mathcal{A} \doteq \{V_{\mathrm{mod}}\} \times \{f_{\mathrm{mod}}\} \,, \tag{6.1}$$

where any action $a \in \mathcal{A}$ defines an RF amplitude modulation for the subsequent time step according to Eq. (5.27). To reduce the size of the action space, the modulation amplitude and frequency are subjected to constraints, respectively

$$V_{\mathrm{mod}} \in [V_{\mathrm{min}}, V_{\mathrm{max}}] \quad \text{and} \quad f_{\mathrm{mod}} \in [f_{\mathrm{min}}, f_{\mathrm{max}}] \,, \tag{6.2}$$

where an amplitude at the percent level is typically sufficient, e.g., $V_{\mathrm{mod}} \in [0, 0.01]\, V_0$, and the modulation frequency is restricted to a small range around the estimated micro-structure frequency, e.g., $f_{\mathrm{mod}} \in [f_{\mathrm{ms}} - 0.5\, f_{\mathrm{s},0}, f_{\mathrm{ms}} + 0.5\, f_{\mathrm{s},0}]$. To facilitate the use of neural networks, the final action space is normalized to

$$\mathcal{A} \doteq [0, 1] \times [0, 1] \quad \text{or} \quad \mathcal{A} \doteq [-1, 1] \times [-1, 1] \,. \tag{6.3}$$

As already discussed in [86], the straightforward choice of giving the RL agent full control over both the RF amplitude and the RF phase

$$\mathcal{A} \doteq \{V_0\} \times \{\varphi_0\} \,, \tag{6.4}$$

leaves the agent with an option for a trivial solution. The micro-bunching dynamics in general and the instability threshold (according to Eq. (3.16)) in particular are dependent on the amplitude of the RF voltage as it directly affects the length of the bunch. The agent may therefore simply lower the RF voltage, which lengthens the bunch, reduces the CSR-induced perturbation and therefore stabilizes the dynamics just naturally. To circumvent that issue for the initial formulation in [86], the action space was restricted to modulations of both quantities, the RF amplitude and the RF phase

$$\mathcal{A} \doteq \{V_{\mathrm{mod}}\} \times \{f_{\mathrm{mod},V}\} \times \{\varphi_{\mathrm{mod}}\} \times \{f_{\mathrm{mod},\varphi}\} \,. \tag{6.5}$$

Although chapter 5 suggests a modulation of particularly the RF amplitude, the modulation of the RF phase was also found to affect the micro-bunching dynamics. It thus remains an interesting option for an extension of the action space defined in Eq. (6.1). Yet, as the effect of an RF amplitude modulation on the micro-bunching dynamics is understood more clearly and extensive control could already be achieved without involvement of the RF phase, the empirical studies in chapter 7 focus exclusively on the action space defined in Eq. (6.1).

## 6.2. State Definition and Markov Property

In case of the micro-bunching instability and its simulation via VFP solvers, the definition of a Markovian state is straightforward. At any given point in time, the charge distribution $\psi(q, p, t)$ and its evolution solely depend on a set of constant parameters and the preceding charge distribution. Defining the temporal sequence of charge distributions as the state signal[1]

$$S_t \doteq \psi(q, p, t_i) \tag{6.6}$$

[1] To distinguish between the RL time step parameter $t$ and the time step indicating the discrete time series of simulated or measured physical quantities, the latter is denoted with $t_i$. Equation (6.6) should thus be read as: The state at time step $t$ is given by the $i$-th element of the discrete time series of simulated or measured charge distributions.

thus yields a Markov process, fully satisfying Eq. (4.2). As constant parameters can be neglected, and in order to tailor the definition of the state space to the use of neural networks, the state definition is slightly modified to

$$S_t \doteq \Delta\hat{\psi}(q, p, t_i) \,, \tag{6.7}$$

with the difference of the normalized charge distribution to the normalized, initial temporal average $\overline{\psi}_{\text{init}}(q, p)/Q$ as observed for the natural dynamics of the instability

$$\Delta\hat{\psi}(q, p, t_i) \doteq \hat{\psi}(q, p, t_i) - \overline{\psi}_{\text{init}}(q, p)/Q \,. \tag{6.8}$$

It is worth noting that the straightforward fulfillment of the Markov property is not a special property of the micro-bunching instability nor of VFP solvers, but rather a general characteristic of a physical phase space. Classically speaking, in the absence of external forces, the knowledge of a particle's position and its velocity allows for predictions about its trajectory. Analogously, it provides an RL agent with the necessary information to make predictions about the environment. As many accelerator problems are described using a charge density or particle distribution in an up to six-dimensional phase space, this motivates the use of RL algorithms for a larger range of problems. Typically though, the exact charge distribution in phase space is difficult to measure at an actual storage ring. Although first efforts towards phase space tomography at KARA have shown promising results [87], this type of information is still not fully accessible.

There are, however, several diagnostic systems in place which can measure derived beam properties and thereby provide information about the longitudinal charge distribution. An electro-optical near field setup is capable of measuring the longitudinal bunch profile on a turn-by-turn basis [88]. Complementary information about the charge distribution in energy can be obtained by measuring the horizontal bunch profile in a dispersive section of the storage ring using a fast-gated camera [89]. Yet, the simplest and most reliable way to acquire information about the micro-bunching dynamics is to measure the emitted CSR power $P_{\text{CSR}}(t)$ [32, 40], which is typically done using Schottky diodes and the in-house developed data acquisition system KAPTURE (Karlsruhe Pulse Taking Ultra-fast Readout Electronics) [90]. Compared to the use of the full longitudinal charge distribution in phase space, this condenses the information to a single scalar number, the integrated CSR power emitted by that distribution. Whether or not an observation based on solely the CSR power signal

$$O_t \doteq O_t(P_{\text{CSR}}(t)) \,, \tag{6.9}$$

yields sufficient information to successfully apply reinforcement learning solution methods is unknown a priori. Ideally, the condensed information yields a fast learning rate and convergence to a satisfying amount of control over the longitudinal beam dynamics. In anticipation of that issue and based on the accumulated experience about the dynamics of the instability at KARA, a hand-crafted eight-dimensional feature vector is introduced

$$\boldsymbol{x}(t) \doteq (x_1(t), x_2(t), \ldots, x_8(t))^\top \,, \tag{6.10}$$

which aims to capture the most relevant information about the state of the micro-bunching dynamics. As the CSR signal can be measured on a turn-by-turn basis, the sample rate is

equal to the revolution frequency. As $f_{\mathrm{rev}}$ is about a factor of 400 larger than the nominal synchrotron frequency $f_{\mathrm{s},0}$, several samples can be acquired between the agent's actions

$$\{\overbrace{P_{\mathrm{CSR}}(t_0)}^{S_{t-1},A_{t-1}}, P_{\mathrm{CSR}}(t_1), \ldots, \overbrace{P_{\mathrm{CSR}}(t_n)}^{S_t,A_t}\} \,, \tag{6.11}$$

where $t_0$ coincides with the preceding RL time step $t-1$ and $t_n$ corresponds to the current RL time step $t$. Based on this, the first feature is defined as the mean of the CSR power signal observed since the last RL time step

$$\mu_{\mathrm{CSR}}(t) \doteq \frac{1}{n}\sum_{i=1}^{n} P_{\mathrm{CSR}}(t_i) \,, \tag{6.12}$$

normalized by the initial mean $\mu_{\mathrm{CSR}}^{\mathrm{init}}$ as observed during the natural behavior of the instability

$$x_1(t) \doteq (\mu_{\mathrm{CSR}}(t) - \mu_{\mathrm{CSR}}^{\mathrm{init}})/\mu_{\mathrm{CSR}}^{\mathrm{init}} \,. \tag{6.13}$$

Analogously, the second feature is defined as the normalized standard deviation of the preceding CSR signal

$$x_2(t) \doteq (\sigma_{\mathrm{CSR}}(t) - \sigma_{\mathrm{CSR}}^{\mathrm{init}})/\sigma_{\mathrm{CSR}}^{\mathrm{init}} \,. \tag{6.14}$$

The third feature is designed to indicate a slow trend of the emitted CSR power and is defined as

$$x_3(t) \doteq \frac{2}{\pi}\arctan\left(\frac{\mu_{\mathrm{CSR,end}}(t) - \mu_{\mathrm{CSR,start}}(t)}{n}\right) \,, \tag{6.15}$$

where $\mu_{\mathrm{CSR,start}}(t)$ and $\mu_{\mathrm{CSR,end}}(t)$ denote the mean of the CSR power signal around the start and end of the sequence in Eq. (6.11), respectively[2]. Given the clear signature of the instability in the frequency domain, features four to six aim to encode the dominant frequency contribution in the Fourier transformed CSR power signal. To attain a minimum resolution in the frequency domain, the preceding 1024 values are considered to calculate the Fourier transform $\widetilde{P}_{\mathrm{CSR}}(\omega)$, even if the sequence in Eq. (6.11) has less entries. The fourth feature indicates the relative strength of the main peak in the frequency distribution

$$x_4(t) \doteq |\widetilde{P}_{0,\mathrm{CSR}}(\omega_{\mathrm{main}})| \,/\, \sum_i |\widetilde{P}_{0,\mathrm{CSR}}(\omega_i)| \,, \tag{6.16}$$

where $|\widetilde{P}_{0,\mathrm{CSR}}(\omega_i)|$ denotes the magnitude of the Fourier transform of the normalized time signal defined by the preceding 1024 values of $P_{\mathrm{CSR}}(t_n)$

$$\widetilde{P}_{0,\mathrm{CSR}}(\omega_i) \doteq \mathcal{F}(P_{\mathrm{CSR}}(t_j) - \mu_{\mathrm{CSR}}(t_j)) \quad \text{with} \quad j \in [n-1024, n] \,, \tag{6.17}$$

and $\omega_{\mathrm{main}}$ the main contributing angular frequency

$$|\widetilde{P}_{0,\mathrm{CSR}}(\omega_{\mathrm{main}})| \doteq \max_{\omega_i} |\widetilde{P}_{0,\mathrm{CSR}}(\omega_i)| \,. \tag{6.18}$$

[2] Typically, the first and last ten values in the sequence are considered.

The fifth feature is defined as the normalized frequency

$$x_5(t) \doteq |\widetilde{P}_{0,\mathrm{CSR}}(\omega_{\mathrm{main}})| \,/\, \omega_{\mathrm{max}} \,, \tag{6.19}$$

where $\omega_{\mathrm{max}}$ denotes the maximum valid frequency in the spectrum. The sixth feature also adds the corresponding complex phase

$$x_6(t) \doteq \frac{1}{2\pi}\left[2\pi + \arctan\left(\frac{\mathrm{Im}(\widetilde{P}_{0,\mathrm{CSR}}(\omega_{\mathrm{main}}))}{\mathrm{Re}(\widetilde{P}_{0,\mathrm{CSR}}(\omega_{\mathrm{main}}))}\right)\right] \bmod 2\pi \,. \tag{6.20}$$

The seventh feature represents an estimate of the phase difference $\Delta\vartheta$ between the applied sinusoidal RF amplitude modulation and the CSR power signal in the preceding time step, determined via cross-correlation

$$x_7(t) \doteq \begin{cases} \Delta\vartheta_0/\pi, & \text{if } \Delta\vartheta_0 \leq \pi \\ (\Delta\vartheta_0 - 2\pi)/\pi, & \text{if } \Delta\vartheta_0 > \pi \end{cases} \,. \tag{6.21}$$

with

$$\Delta\vartheta_0 \doteq \Delta\vartheta \bmod 2\pi \,. \tag{6.22}$$

The definition of this feature in particular is based on the analysis in chapter 5 and is expected to provide critical information about the synchronization between the applied control signal and the ongoing micro-bunching dynamics. Finally, the eighth and last feature is reserved for the termination condition introduced in section 6.4. Altogether, the features of vector $\boldsymbol{x}_t$ define an observation at time $t$

$$O_t \doteq \boldsymbol{x}(t) \,, \tag{6.23}$$

which can be provided by the diagnostic systems already in place at KARA.

As particularly the feature $x_7(t)$ is expected to carry crucial information, the more theoretical definition in Eq. (6.7) is also augmented by an additional five-dimensional feature vector

$$\boldsymbol{x}^{\mathrm{a}}(t) \doteq (x_1^{\mathrm{a}}(t), x_2^{\mathrm{a}}(t), \ldots, x_5^{\mathrm{a}}(t))^{\top} \,. \tag{6.24}$$

Here, the first two features directly encode the phase of the applied RF amplitude modulation

$$x_1^{\mathrm{a}}(t) \doteq \cos(\vartheta) \quad \text{and} \quad x_2^{\mathrm{a}}(t) \doteq \sin(\vartheta) \,, \tag{6.25}$$

and the third and fourth represent the phase difference to the CSR power signal

$$x_3^{\mathrm{a}}(t) \doteq \cos(\Delta\vartheta) \quad \text{and} \quad x_4^{\mathrm{a}}(t) \doteq \sin(\Delta\vartheta) \,. \tag{6.26}$$

Analogously to the feature vector in Eq. (6.10), the fifth and last feature is reserved for the termination condition introduced in section 6.4. The full information given to the agent in this more theoretical formulation of the problem is finally given by the tuple

$$S_t \doteq (\Delta\hat{\psi}(q, p, t_i), \boldsymbol{x}^{\mathrm{a}}(t)) \,. \tag{6.27}$$

## 6.3. Choice of Reward Function

The definition of a proper reward function is a crucial component of any RL based endeavor as it solely defines the objective pursued by the agent. It is thus of critical importance that the optimization of this function corresponds to the desired behavior of the agent. Only in that case, can the agent be expected to make progress towards solving the real world task. For the overarching objective of achieving extensive control over the micro-bunching instability, both mitigation and excitation of the micro-bunching dynamics are considered throughout this thesis. As it turns out, the former is fundamentally more challenging than the latter. An excitation of the dynamics can already be achieved via a constant RF amplitude modulation as demonstrated in section 7.2, whereas mitigation requires a dynamic adjustment of the modulation amplitude and frequency. The definition of a reward function which describes the objective of mitigating the instability is thus considered first and more extensively. The alternative objective of exciting the micro-bunching dynamics is briefly addressed afterwards.

As much of the interest around the pursued micro-bunching control is centered around the corresponding emission of CSR, it seems natural to construct a reward function based on the CSR power signal

$$R_t \doteq R_t(P_{\mathrm{CSR}}(t)) \ . \tag{6.28}$$

An additional benefit of doing so is that this definition and the observation defined in Eq. (6.23) are based on the same measurements, which facilitates the practical implementation. As the occurring micro-bunching dynamics lead to distinct oscillations in the CSR power signal, the reduction of these fluctuations corresponds to a mitigation of the underlying dynamics. As outlined in section 3.4.3, lengthening the bunch is a way of achieving this mitigation, but one that comes at the cost of limiting the operation in other regards, e.g., it leads to a decreased emission of CSR and reduces the capability to support time-resolved experiments. To maintain the average intensity of the emitted CSR power and to prevent a simple lengthening of the bunch, the average power of the signal should thus be considered simultaneously. An early version of the reward was therefore defined as

$$R_t \doteq \frac{(\mu_{\mathrm{CSR}}(t) - \mu_{\mathrm{CSR}}^{\mathrm{init}})/\mu_{\mathrm{CSR}}^{\mathrm{init}}}{(\sigma_{\mathrm{CSR}}(t) - \sigma_{\mathrm{CSR}}^{\mathrm{init}})/\sigma_{\mathrm{CSR}}^{\mathrm{init}}} \ . \tag{6.29}$$

Yet, it was found that this does not constitute a good trade-off between mean and standard deviation of the signal as it essentially neglects the mean for small values of the standard deviation. The definition was therefore changed to

$$R_t \doteq \frac{\mu_{\mathrm{CSR}}(t) - \mu_{\mathrm{CSR}}^{\mathrm{init}}}{\mu_{\mathrm{CSR}}^{\mathrm{init}}} - w_\sigma \frac{\sigma_{\mathrm{CSR}}(t) - \sigma_{\mathrm{CSR}}^{\mathrm{init}}}{\sigma_{\mathrm{CSR}}^{\mathrm{init}}} \ , \tag{6.30}$$

with the trade-off parameter $w_\sigma$. In empirical studies this definition was found to closely match the intuitive understanding of the objective. Typically, with an equal weighting for the average power and the amount of fluctuation, that is, $w_\sigma = 1$.

Given the apparent oscillation of the CSR power signal, the estimates of its mean and standard deviation depend on the time window considered and the corresponding phase

of the oscillation. To reduce the susceptibility to this issue, a smoother version of the definition in Eq. (6.30) can be obtained by considering a longer sequence of prior values of $P_{\mathrm{CSR}}(t_i)$. While this reduces the fluctuation of the reward signal, it also dilutes the information about the more recent signal. Another way of reducing the fluctuations of the reward signal is by replacing the calculation of the mean and standard deviation in Eq. (6.30) by more robust proxies. For example, the mean may be approximated by

$$\mu_{\mathrm{CSR}}(t) \approx \mu_{\mathrm{CSR}}^{\mathrm{proxy}}(t) \doteq \frac{1}{2}\left(P_{\max}(t) + P_{\min}(t)\right) \,, \tag{6.31}$$

with the maximum and minimum values of the CSR power signal in the preceding time frame

$$P_{\max}(t) \doteq \max_{t_i} P_{\mathrm{CSR}}(t_i) \quad \text{and} \quad P_{\min}(t) \doteq \min_{t_i} P_{\mathrm{CSR}}(t_i) \,. \tag{6.32}$$

Analogously, the standard deviation can be approximated by

$$\sigma_{\mathrm{CSR}}(t) \approx \sigma_{\mathrm{CSR}}^{\mathrm{proxy}}(t) \doteq \sqrt{\frac{1}{2}\left[\left(P_{\max}(t) - \mu_{\mathrm{CSR}}^{\mathrm{proxy}}(t)\right)^2 + \left(P_{\min}(t) - \mu_{\mathrm{CSR}}^{\mathrm{proxy}}(t)\right)^2\right]} \,. \tag{6.33}$$

The use of these proxies leads to a reduced oscillation of the reward function, but comes at the cost of blurring the objective and a higher susceptibility to noisy extrema, as can be expected in measurements. Both modifications of Eq. (6.30), smoothing via a larger time frame and the use of proxies, were found to yield good results in empirical testing. They are thus both maintained until studies at the actual accelerator may lead to a preference for one or the other.

The alternative objective of exciting the micro-bunching dynamics is typically pursued in order to obtain higher intensities of the CSR power emitted in the corresponding frequency range. The reward at time $t$ may thus simply be defined as

$$R_t \doteq \int_{\omega_1}^{\omega_2} \mathcal{P}_{\mathrm{CSR}}(\omega, t_i)\, \mathrm{d}\omega \,, \tag{6.34}$$

where $[\omega_1, \omega_2]$ defines the frequency range of interest. Section 7.2 considers the optimization of this objective under the influence of constant RF amplitude modulations, which corresponds to an agent that always selects the same action from the action space defined in Eq. (6.1). Although this is sufficient for reaching a significantly higher emitted CSR power, dynamic adjustments of the RF signal may be even more effective in driving the micro-bunching dynamics and are thus a promising subject for further research.

All in all, the actions defined in Eq. (6.1), the Markovian states in Eq. (6.27) and the rewards in Eq. (6.30) or (6.34) yield a fully functional MDP. The replacement of the Markovian states with the observation of a feature vector in Eq. (6.23) leads to a partially observable MDP, which can be implemented at KARA. Figure 6.1 illustrates the resultant general feedback scheme. The agent is informed by measurements of the CSR signal, from which both the observation and the reward are constructed, and interacts with the longitudinal beam dynamics via the RF system. The repetition rate of the entire feedback loop has to match the time scale of the micro-structure formation process, that is, it has to be in the order of the synchrotron frequency. In practice, this yields challenging time constraint for the involved data transmission and data processing, as well as the decision making process of the agent. The topic is thus revisited in more detail in chapter 8.

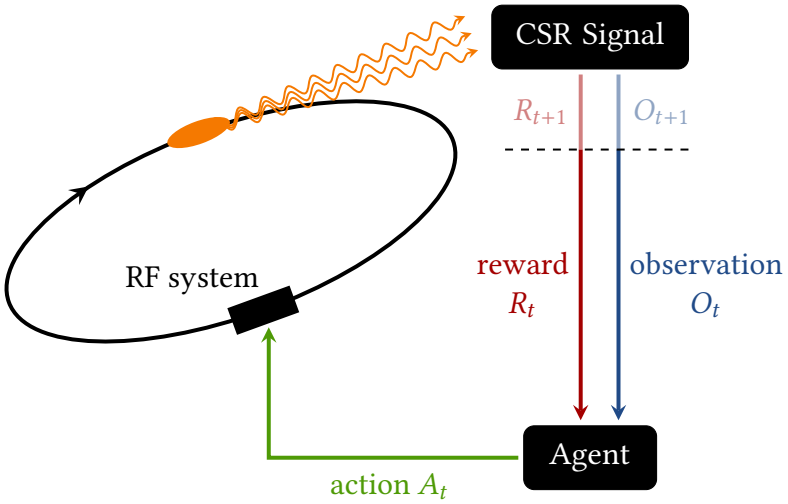



Figure 6.1.: General feedback scheme for the implementation of RL-based micro-bunching control at KARA. The agent receives an observation $O_t$ and reward $R_t$ calculated based on the measured CSR power signal $P_{\mathrm{CSR}}(t_i)$ and choses an action $A_t$ that defines the RF amplitude modulation for the subsequent time step.

## 6.4. Termination Condition

The fact that the pursued micro-bunching control constitutes a continuing task leads to an additional complication for training an RL agent in practice. Although the concept of temporal difference learning allows updates during the episode and the agent may thus be trained in a very long episode, it is more feasible in practice to break down the learning process into episodes of reasonable length. This facilitates the data handling and the analysis of the agent's performance. Yet, terminating the agent's experience after an arbitrary number of steps leads to an ill-posed problem, as best illustrated in an example: Say the agent encounters a particular state $s$ as the terminal state of the episode $S_T = s$. Per definition, the state- or action-value of that state would be equal to zero. In the given MDP, however, the same state could also occur at the beginning of the episode or at any step in between. Depending on the number of steps remaining until termination of the episode, the cumulative reward and thus the expected value can take on very different values. Were the reward simply defined as $R_t \doteq +1$ at each step, the expected value would be equal to the expected number of steps after the state is encountered. As the agent only receives the state, but not the index of the time step at which it is encountered, this can lead to large fluctuations in the estimates of the value function independent of the quality of the currently followed policy. Even worse, as function approximation is going to be applied to deal with the continuous state and action spaces, not only the estimate of $v_\pi(s)$ is going to be affected by this ambiguity, but also states which are somewhat similar to $s$. With value estimates this arbitrary the agent cannot be expected to learn efficiently.

To address this issue, an explicit termination condition is introduced. Instead of a fixed number of steps, the episode is terminated based on an additional performance measure. A textbook example of this approach is the well-known cart-pole balancing task. Here,

the episode is typically terminated after the pole exceeds a certain angle from its ideal vertical position. For the RL problem considered in this thesis, the termination condition is based on the reward function. After the first ten initial steps, the episode is terminated if the agent does not meet a minimum performance requirement, that is, if the termination condition

$$c_t^{\text{term}} \geq 0 \quad \text{with} \quad c_t^{\text{term}} \doteq \begin{cases} 0 & \text{if } t < 10 \\ R_t - R_{t-10} & \text{if } t \geq 10 \end{cases}, \tag{6.35}$$

is violated. Thereby, performance is measured by the latest reward $R_t$ compared to the one ten steps prior $R_{t-10}$. This requirement of improvement is relaxed in cases where a high enough level of control is already achieved, which in this application corresponds to a high enough immediate reward

$$R_t \geq r_{\text{relax}}\,, \tag{6.36}$$

where the relaxation parameter is usually set to $r_{\text{relax}} = 0.5$. The episode is thus terminated at the first time step the boolean statement

$$(c_t^{\text{term}} < 0) \;\wedge\; (R_t < r_{\text{relax}})\,, \tag{6.37}$$

is found to be true. This fixes the ambiguity of the value estimates, but introduces a dependency on a reward in the agent's past, which violates the Markov property. The parameter $c_t^{\text{term}}$ is thus added as the final feature of the eight-dimensional feature vector in Eq. (6.10) and the five-dimensional feature vector in Eq. (6.24).

One drawback of the definition of $c_t^{\text{term}}$ in Eq. (6.35) is that it focuses solely on a single reward in the past $R_{t-10}$. This may lead to a termination of the episode just because a single pair $(R_{t-10}, R_t)$ did not satisfy the termination condition although good average progress was made by the agent. In an alternative definition, a quantile $Q_p$ of the reward gradients is used instead to address the potential stochasticity of the learning process

$$c_t^{\text{term}} \doteq \begin{cases} 0 & \text{if } t < 10 \\ Q_p(\{R_{t-9} - R_{t-10}, \ldots, R_t - R_{t-1}\}) & \text{if } t \geq 10 \end{cases}. \tag{6.38}$$

With this definition the agent may continue the episode if the ratio $p$ of reward gradients is positive. In empirical testing the best results were achieved for $p = 0.67$, but this version of the termination condition generally allowed for long episodes where the agent was visibly not improving towards the general objective. Overall the definition of the termination condition in Eq. (6.35) was found to be more reliable.

# 7. Micro-Bunching Control in Simulations

*If a machine is expected to be infallible,*
*it cannot also be intelligent.*

— Alan Turing

The approach towards micro-bunching control developed in chapter 5 and formalized as a reinforcement learning problem in chapter 6 is based on the idea of using a modulation of the RF amplitude to interact with the occurring micro-bunching dynamics. In order to test the validity of this approach and the effectiveness of applying reinforcement learning methods, the general feedback scheme is implemented in a virtual environment using the VFP solver Inovesa for the simulation of the underlying longitudinal beam dynamics. The modular implementation described in section 7.1 allows for tests of various combinations of the different definitions introduced in chapter 6. Thereby, the general objective of micro-bunching control is split into two complementary formulations of the task: In section 7.2, a deliberate and controlled excitation of the micro-bunching dynamics is pursued, which serves the purpose of providing intense coherent radiation to dedicated applications. As this can already be achieved with an RF modulation of constant amplitude and frequency, the limits of this approach are explored without the employment of RL solution methods. The more challenging task of mitigating the micro-bunching dynamics is pursued in section 7.3. A practical way of mitigating the instability is desirable as it extends the regime of stable operation at electron storage rings to shorter bunch lengths and higher bunch currents. Besides expanding the range of sustainable beam parameters at existing machines, and thus their capability to support further research, it also enables a more effective optimization of related beam properties and thereby facilitates the design of new faclities. To achieve this mitigation, the modulation of the RF amplitude has to be continuously adjusted to the varying perturbation caused by the CSR wake potential. As a general proof of feasibility and in order to provide a benchmark scenario, subsection 7.3.1 demonstrates the effectiveness of the dynamic RF amplitude modulation scheme via manual control. The reinforcement learning control presented in the subsections 7.3.2 and 7.3.3 is evaluated along this benchmark scenario. In subsection 7.3.2, the agent is given full access to the Markov states, that is, the charge distribution in phase space, which constitutes a theoretically sound formulation of the problem. In subsection 7.3.3, the information provided to the agent is restricted to a feature vector derived from the observed CSR power signal, which represents a more feasible formulation tailored towards the implementation at KARA.

Overall, the presented results serve as a proof-of-principle for the approach to micro-bunching control developed in this thesis. Both excitation and mitigation of the micro-

bunching dynamics are successfully demonstrated, indicating the extensive control which can be achieved by careful, dynamic adjustments of the RF amplitude. However, there are also recurring instabilities observed in the RL training process, which are not yet fully understood. Furthermore, the demonstrated RL-based control has to be generalized across continuous time, different bunch currents and varying machine parameters. These subjects are addressed in the final section of this chapter and further expanded on in chapter 8.

## 7.1. General Implementation Scheme

In order to facilitate the application of different RL solution methods, the reinforcement learning environment is implemented in Python as an OpenAI gym [61] environment. This offers the benefit of standardized interfaces the agent can interact with and thus supports a variety of reinforcement learning libraries. The general implementation scheme is illustrated in Fig. 7.1. The longitudinal beam dynamics underlying the micro-bunching instability are simulated using the VFP solver Inovesa written in C++. The Python package InovesaIPC (Inovesa Inter-Process Communication) written by Patrick Schreiber allows for communication with Inovesa during runtime. As indicated in Fig. 7.1, the environment is implemented in a modular approach to support different definitions of all essential RL elements. The reward handler implements the different reward functions described in section 6.3. Based on the choice of the reward function, the termination condition is derived according to the definitions in section 6.4. In order to fulfill the Markov property as closely as possible, the termination condition is added to the observation provided by the observation handler, which primarily supports the definitions in Eqs. (6.27) and (6.23). Finally, the action handler implements different ways to manipulate the RF signal, most importantly modulations of the RF amplitude. As the phase difference between the sinusoidal RF modulation and the oscillation of the CSR power signal is expected to carry crucial information about the state of the micro-bunching dynamics, the observation is also augmented by this particular feature. In order to ensure the validity of the physics simulation during the interaction with an RL agent, several checks are incorporated into the environment. One such example is a regular inspection of the charge loss. As Inovesa describes the charge density in phase space on a fixed grid, large deviations from the synchronous position may lead to a cut-off at the tails of the distribution. If this results in a charge loss of more than 0.1 %, a large negative number is added as a penalty to the reward at that time step $R_t$. This serves the purpose of discouraging the agent of pursuing solutions that involve non-physical distributions. For the action space defined in Eq. (6.1) this boundary condition was violated very rarely. Yet, it gains in relevance when the action space is extended to also include RF phase modulations. The entire interaction process between an agent and the environment during an episode can be stored in a history-file of HDF5-format [91]. Simultaneously, Inovesa supports storage of the data related to the physics simulation in an additional HDF5-file. The RL agents used in conjunction with this environment are largely based on three different RL libraries: keras-rl [92], Stable Baselines [93] and TF-Agents [94]. The back end computation of the involved neural networks for all three libraries is done in TensorFlow [95]. While Stable Baselines and

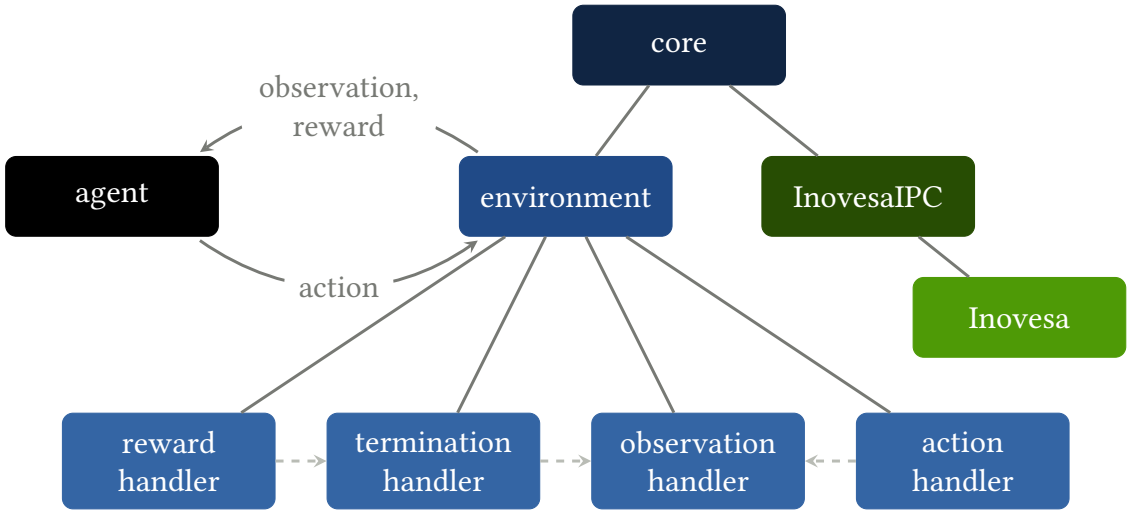



Figure 7.1.: Implementation of the RL feedback scheme in simulations (InovesaRL package). At the center is an OpenAI gym environment which communicates via the InovesaIPC package with the VFP solver Inovesa, used for the simulation of the longitudinal beam dynamics. Because of its modular design, the environment supports various definitions and combinations of the different RL elements. The standardized interface facilitates the application of RL algorithms from different RL libraries, including keras-rl, Stable Baselines and TF-Agents.

TF-Agents support all four RL algorithms introduced in section 4.5, DDPG, TD3, SAC and PPO, the older library keras-rl only offers an implementation of the DDPG algorithm.

All these modules are combined into an installable Python package called InovesaRL and virtualized in a Docker [96] container to improve the reproducibility of the obtained results.

## 7.2. Excitation of Micro-Bunching Dynamics

In chapter 5, the perturbation of the restoring force exerted by the RF system was found to be critical for the formation process of the occurring micro-structures. This immediately suggests the use of an RF amplitude modulation to amplify the perturbation by the CSR wake potential and thereby excite the micro-bunching dynamics, as published in [97]. Given that the strength of the restoring force at the synchronous position, $k_0$, is naturally modulated at the micro-structure frequency, a straightforward approach leads to an RF amplitude modulation at that frequency.

Neglecting the relative phase between the natural perturbation and the applied RF signal, this idea is tested on an exemplary data set ($\mathcal{D}_2$ with $g = 32\,\mathrm{mm}$) at a bunch current directly above the instability threshold, that is, $I = 260\,\mu\mathrm{A}$. Figure 7.2 illustrates the results of applying an RF amplitude modulation at the micro-structure frequency $f_\mathrm{mod} = f_\mathrm{ms} = 4.74\,f_{\mathrm{s},0}$ with an amplitude of $V_\mathrm{mod} = 0.05\,V_0$. The application of the RF amplitude modulation immediately increases the oscillation amplitude of the CSR power signal. Subsequent to an initial transition phase, the oscillation settles for a new quasi-

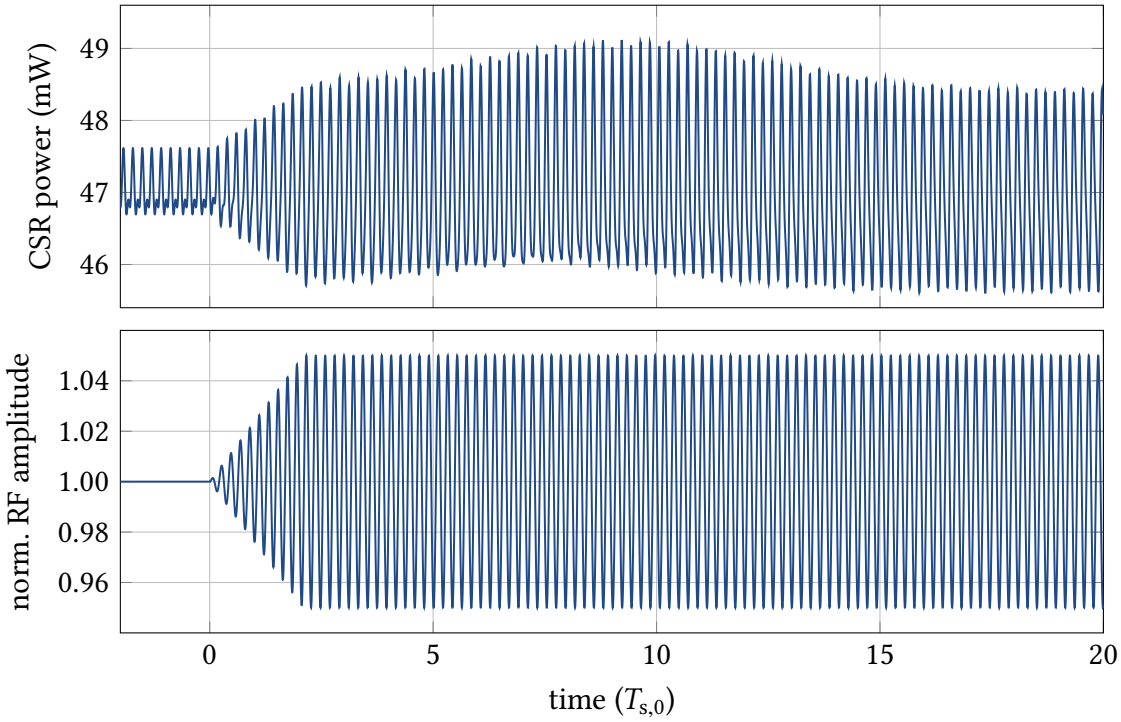
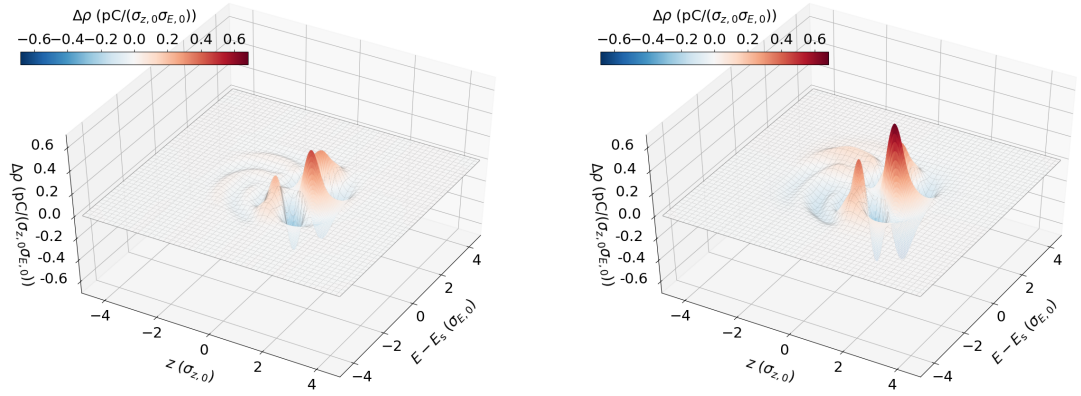



Figure 7.2.: The natural micro-bunching dynamics can be amplified with an RF amplitude modulation at the micro-structure frequency $f_{\mathrm{mod}} = f_{\mathrm{ms}}$. The oscillation of the CSR power signal (top) is immediately amplified by the RF modulation applied after $t = 0\,T_{s,0}$ (bottom). After an initial transition phase, roughly from $t = 0\,T_{s,0}$ to $t = 15\,T_{s,0}$, the micro-bunching dynamics settle for a new quasi-equilibrium with higher oscillation amplitude.

equilibrium (around $t = 15\,T_{s,0}$). Compared to the natural oscillation, the signal takes on a more sinusoidal shape with a clearly amplified oscillation amplitude. With a constant application of the RF amplitude modulation, the observed dynamics also continue in the same, highly repetitive manner after the time frame displayed in Fig. 7.2. At this point, the oscillation of the CSR power signal is largely driven by the external excitation. Independent of the initial phase difference between the natural perturbation of the restoring force and the applied RF modulation, this behavior is always observed in the simulations conducted throughout this thesis. After an initial adjustment period, the micro-bunching dynamics follow the external excitation which amplifies the natural perturbation by the CSR wake potential and eventually drives the micro-bunching dynamics. This increased oscillation of the CSR power signal is the result of an amplification of the micro-structures in the longitudinal phase space as illustrated in Fig. 7.3. While the overall shape of the micro-structures is almost identical to those occurring naturally, the applied RF amplitude modulation clearly leads to a growth of the micro-structures in amplitude. Compared to the natural micro-structures, the maximum amplitude in Fig. 7.3b is increased by nearly 50 percent. This is expected given the analysis of the perturbed synchrotron motion in chapter 5. After the initial transition phase, the perturbation by the CSR wake potential is synchronized to the external excitation. As a consequence of the additional effect of the




(a) Natural micro-structures

(b) Excitation of micro-structures

Figure 7.3.: (a) The naturally occurring micro-bunching dynamics lead to the formation of distinct micro-structures in the longitudinal phase space. (b) An RF amplitude modulation at the micro-structure frequency, $f_{\mathrm{mod}} = f_{\mathrm{ms}}$, and with $V_{\mathrm{mod}} = 0.05\, V_0$, leads to an amplification of the already naturally occurring structures by nearly 50 percent.

RF amplitude modulation, the individual particles forming the micro-structures are driven further outside in phase space, which amplifies the inhomogeneity and leads to a growth of the local structures. As the RF amplitude modulation affects the charge distribution in phase space and the corresponding longitudinal bunch profile, it also leads to an altered emission of CSR. In order to assess the effects on the radiated CSR power spectrum, the temporal average of the spectrum

$$\overline{\mathcal{P}}_{\mathrm{CSR}}(\omega) \doteq \frac{1}{n} \sum_{i=1}^{n} \mathcal{P}_{\mathrm{CSR}}(\omega, t_i) \,, \tag{7.1}$$

is shown in Fig. 7.4 for both, the natural behavior of the instability and under the influence of the RF amplitude modulation. In large parts the two spectra are very similar. Yet, in the frequency range corresponding to the spatial extent of the structure, that is, around 150 GHz the emitted power is notably increased (up to 30 percent) by the RF modulation. This is expected due to the similar shape but increased amplitude of the micro-structures in Fig. 7.3b. With a reward function defined according to Eq. (6.34), $R_t = \int_{\omega_1}^{\omega_2} \mathcal{P}_{\mathrm{CSR}}(\omega, t_i)\, \mathrm{d}\omega$, on the frequency range $[140, 200]$ GHz, the RF modulation increases the average reward by about 25 percent. Applications which rely on intense coherent radiation in that specific frequency range may thus be supported through the additional RF modulation. It is worth noting that the total integrated CSR power, e.g. over the frequency range $[1, 1000]$ GHz, is about the same compared to the natural micro-bunching dynamics. The main difference achieved via the RF amplitude modulation is that the part of the spectrum which corresponds to the spatial extent of the micro-structures is notably amplified. These results confirm the effectiveness of interacting with the micro-bunching dynamics via an RF amplitude modulation and thereby verify the insights of chapter 5.

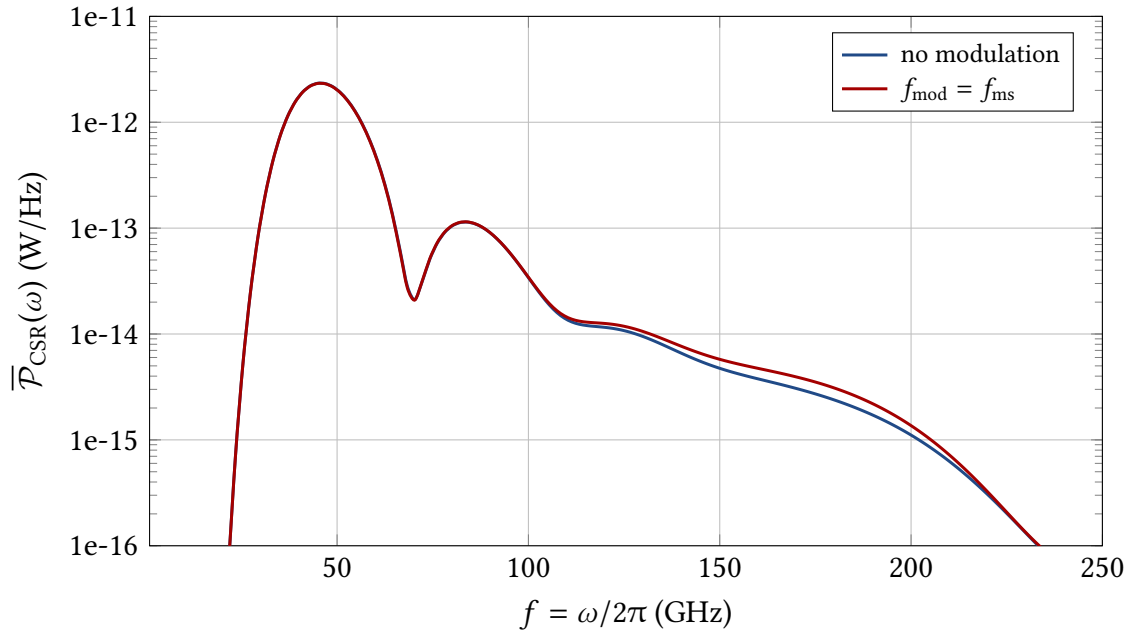



Figure 7.4.: The applied RF amplitude modulation amplifies the average CSR power spectrum in the frequency range corresponding to the spatial extent of the occurring micro-structures. Here, the emitted power is increased by up to 30 percent in the frequency range around 150 GHz.

Besides the excitation of the naturally occurring micro-bunching dynamics, the application of an RF amplitude modulation was also found to offer the possibility of imprinting a new set of micro-structures on the charge distribution. In a scan across different modulation frequencies, another strong response to the applied modulation was found close to the third harmonic of the nominal synchrotron frequency. In fact, the effect of an RF amplitude modulation at $f_{\mathrm{mod}} = 3.06\, f_{\mathrm{s},0}$ with the same modulation amplitude $V_{\mathrm{mod}} = 0.05\, V_0$ on the oscillation of the CSR power signal, shown in Fig. 7.5, is even stronger compared to the RF modulation at the natural micro-structure frequency. Immediately after the RF amplitude modulation is applied, the oscillation amplitude of the CSR power signal starts growing and quickly reaches significantly larger values. Even at the end of the displayed time frame, the oscillation amplitude is still increasing. After about 50 synchrotron periods it eventually settles between the values of 38.2 mW and 66.4 mW. Thereby, the CSR power signal does not take on the almost perfectly sinusoidal shape displayed in Fig. 7.2, instead it features a stronger second harmonic which slightly deforms the signal. Yet, the oscillation is clearly driven by the external excitation directly after the RF modulation is applied. Similarly to the excitation at the micro-structure frequency, these increased oscillations of the CSR power signal are the result of an amplification of the micro-bunching dynamics in the longitudinal phase space. In this case, the micro-structures shown in Fig. 7.6 are not only of increased amplitude, but also of altered shape. Compared to the about five micro-structures occurring due to the natural behavior of the instability, the applied RF amplitude modulation imprints an entirely new set of merely three micro-structures on the charge distribution. The maximum amplitude of these new micro-structures is more than

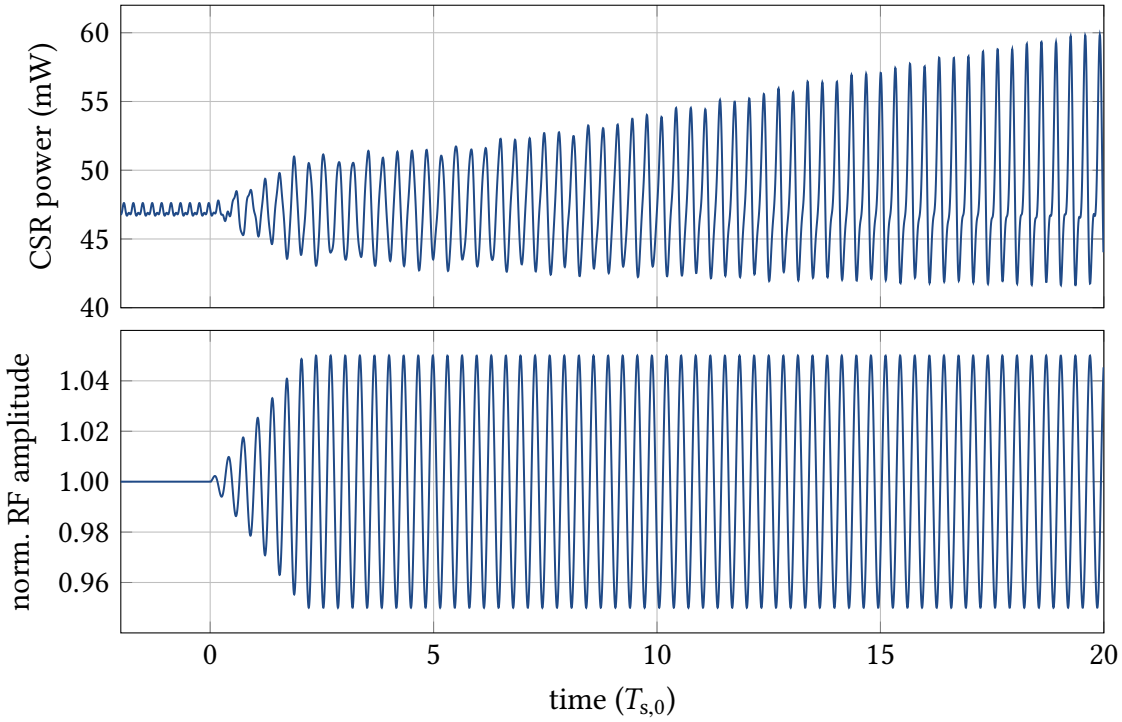
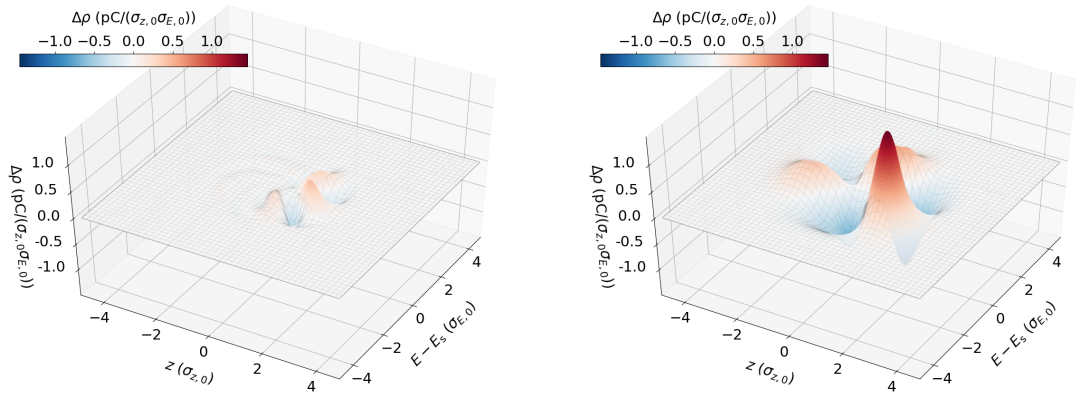



Figure 7.5.: Besides the excitation of the natural micro-bunching dynamics, an RF amplitude modulation at the third harmonic of the nominal synchrotron frequency, $f_{\mathrm{mod}} = 3.06\, f_{s,0}$, was found to have a strong effect on the longitudinal beam dynamics. The oscillation of the CSR power signal (top) immediately responds to the external excitation by the RF amplitude modulation (bottom) and quickly reaches much larger amplitudes. Eventually, the oscillation slightly deviates from a purely sinusoidal signal, as visible towards the end of the displayed time frame. At this point, the oscillation of the signal continues to grow in amplitude for 30 more synchrotron periods before it settles for a regular oscillation between 38.2 mW and 66.4 mW.

three times larger than that of the naturally occurring structures, which leads to the significantly higher peak intensity of the emitted CSR power. Here, the natural micro-bunching dynamics are replaced by a new periodic perturbation of the restoring force induced by the applied RF amplitude modulation. The original micro-structures are no longer visible in the longitudinal phase space and the corresponding frequency $f_{\mathrm{ms}}$ is strongly suppressed in the oscillation of the CSR power signal. These altered micro-bunching dynamics lead to substantial changes in the corresponding emission of CSR. As the naturally occurring micro-structures are replaced by a new set of structures imprinted on the charge distribution, the emission of CSR at frequencies corresponding to the spatial extent of the original structure, that is around 150 GHz, is visibly reduced in the average CSR power spectrum shown in Fig. 7.7. Yet, the emission at frequencies corresponding to the spatial extent of the new micro-structures, between 60 GHz and 140 GHz, is increased up to a full order of magnitude. With a reward function defined according to Eq. (6.34) on the frequency range [60, 140] GHz, the RF modulation leads to an increase of the average reward by more than




(a) Natural micro-structures

(b) Excitation of micro-structures

Figure 7.6.: (a) The naturally occurring set of micro-structures consists of about five structures with relatively small amplitudes. (b) An RF amplitude modulation at the third harmonic of the nominal synchrotron frequency, $f_{\mathrm{mod}} = 3.06\, f_{\mathrm{s},0}$, and with $V_{\mathrm{mod}} = 0.05\, V_0$, leads to a new set micro-structures in the longitudinal phase space, which consists of only three structures, but of much larger amplitude.

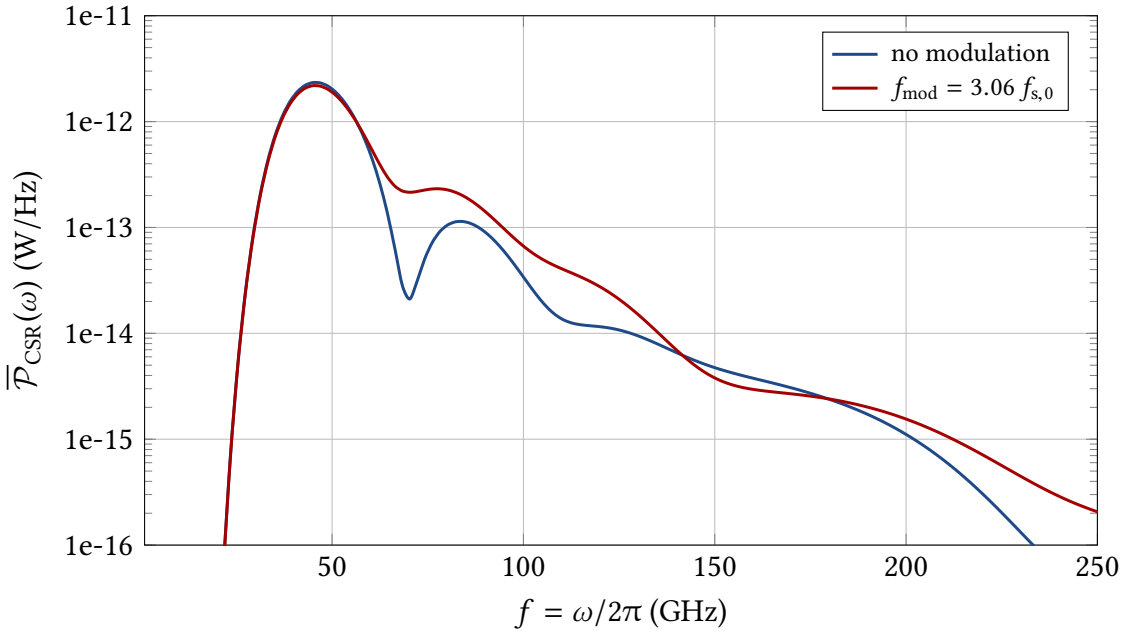



Figure 7.7.: The altered micro-bunching dynamics due to the applied RF amplitude modulation at $f_{\mathrm{mod}} = 3.06\, f_{\mathrm{s},0}$ lead to substantial changes in the radiated CSR power spectrum. While the intensity is decreased in the frequency range corresponding to the spatial extent of the original micro-structure, the new set of large micro-structures shown in Fig. 7.6 increases the emission between 60 GHz and 140 GHz by a large margin.

200 percent. For the parameter settings in this example, the RF amplitude modulation at $f_{\mathrm{mod}} = 3.06\, f_{\mathrm{s},0}$ would thus be of particular interest to experiments with a requirement of intense CSR in that frequency range.

These examples demonstrate how an RF amplitude modulation can be used to excite the micro-bunching dynamics, either by amplifying the naturally occurring micro-structures or by imprinting a new set of structures. For the considered set of simulation parameters, the ladder could only be achieved by an RF modulation at the third harmonic of the synchrotron frequency, alternative modulation frequencies did not have the same, strong effect on the micro-bunching dynamics. In order to determine the conditions under which a new set of micro-structures can be imprinted on the charge distribution and to assess the full potential of this approach, a more systematic study is required. Nonetheless, the amplification of the natural micro-bunching dynamics via an RF amplitude modulation at the micro-structure frequency verifies the insights developed in chapter 5 and provides a simple test scenario for first experiments at KARA as discussed in chapter 8.

## 7.3. Mitigation of Micro-Bunching Dynamics

Compared to an excitation of the occurring micro-structures, their mitigation is a more challenging objective. As illustrated in section 5.5, the application of an RF amplitude modulation with constant amplitude and frequency is not sufficient to counteract the dynamic perturbation caused by the CSR wake potential. Instead, the modulation parameters have to be adjusted according to the altered micro-bunching dynamics caused by previous interactions with the beam. This leads to the sequential decision problem formalized in chapter 6. In a first step, the objective of mitigation is pursued through manual control, that is, the actions are chosen manually in a trial-and-error process, not by an RL agent. This serves as proof for the general feasibility and the soundness of the problem formulation. The achieved level of control is also used as a performance benchmark for the subsequent RL efforts. Initially, the agent is given full access to the Markov states of the MDP, which provides the theoretical comfort of a problem formulation that can expected to be solvable by RL algorithms. Although the final amount of immediate reward is larger under manual control, the benchmark level of total reward is reached, and even slightly exceeded, in this formulation. The best training runs thus yield an RL agent which is clearly capable of mitigating the micro-bunching dynamics. The restriction of the agent's information to an observable feature vector in subsection 7.3.3 comes at the cost of losing some theoretical comfort, but facilitates the practical implementation at KARA. Unexpectedly, the performance achieved with this approach even surpasses the prior results.

### 7.3.1. Proof of Feasibility: Manual Control

To verify the feasibility of mitigating the micro-bunching dynamics via a dynamically adjusted modulation of the RF amplitude, this idea was tested on an exemplary set of simulation parameters (data set $\mathcal{D}_1$) at a bunch current directly above the threshold current, $I = 115\,\mu\mathrm{A}$. As the formation of micro-structures happens at the time scale of the synchrotron period, the time between two consecutive actions $\Delta t$ should be chosen

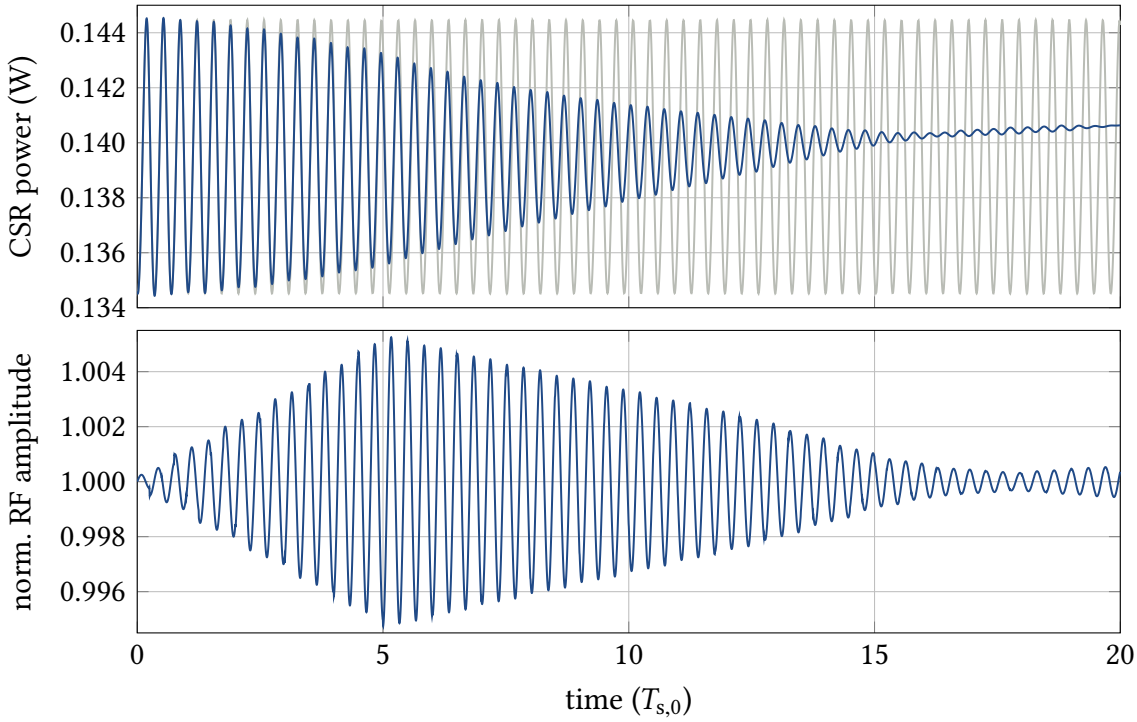
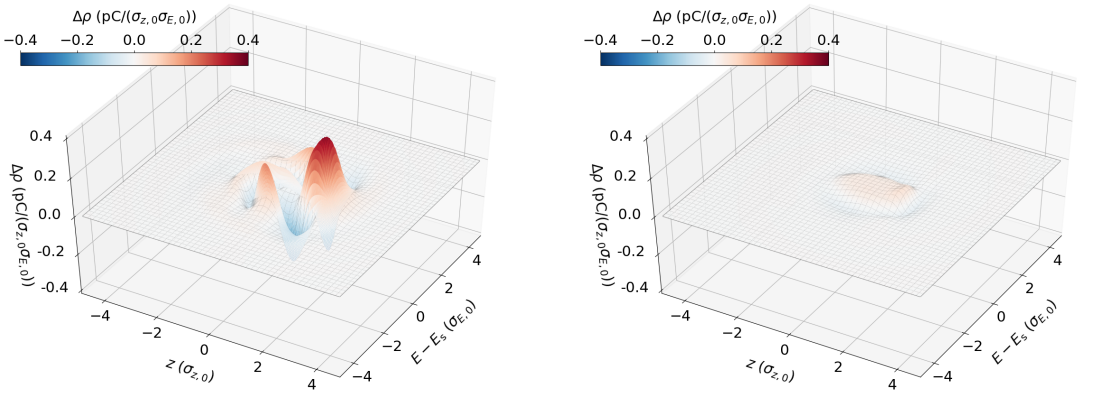



Figure 7.8.: Result of a carefully selected sequence of RF amplitude modulations, found in a trial-and-error process. The initial oscillation of the CSR power signal (top) is continuously damped by the applied RF modulation (bottom). At the end of the considered time frame, the oscillation of the CSR power signal is reduced to a minimum and clearly damped compared to the natural behavior of the instability indicated in gray.

in the same order of magnitude to efficiently counteract the micro-bunching dynamics. For the empirical studies presented in this chapter, $\Delta t$ was therefore set to a quarter of the nominal synchrotron period. Figure 7.8 illustrates the results achieved by carefully adjusting the modulation parameters during a total time frame of 20 synchrotron periods. The oscillation of the CSR power signal shown in the upper part of the figure is continuously reduced down to the point where it is barely visible for the last five synchrotron periods. Compared to the natural behavior of the instability, indicated by the gray line, this clearly demonstrates a mitigation of the underlying micro-bunching dynamics. The corresponding micro-structures at $t = 0\,T_{s,0}$ and $t = 20\,T_{s,0}$ are shown in Fig. 7.9, illustrating the effect on the charge distribution in the longitudinal phase space. To achieve these results, the amplitude of the RF modulation is initially increased until it starts damping the oscillation in the CSR power signal. As the micro-bunching dynamics are gradually mitigated, the strength of the perturbation by the CSR wake potential is reduced and the modulation amplitude can thus be lowered for the subsequent synchrotron periods. Eventually, the amplitude is slightly increased again at the very end of the considered time frame. Simultaneously, though barely visible to the unaided eye, the modulation frequency is also slightly adjusted over the entire sequence. The corresponding action values chosen from




(a) Initial micro-structures at $t = 0\,T_{s,0}$ (b) Mitigated micro-structures at $t = 20\,T_{s,0}$

Figure 7.9.: (a) The initial micro-structures for the sequence displayed in Fig. 7.8 are quickly damped by the applied modulation of the RF amplitude. (b) After 20 synchrotron periods, the micro-structures are mitigated significantly, which yields a smooth charge distribution in the longitudinal phase space.

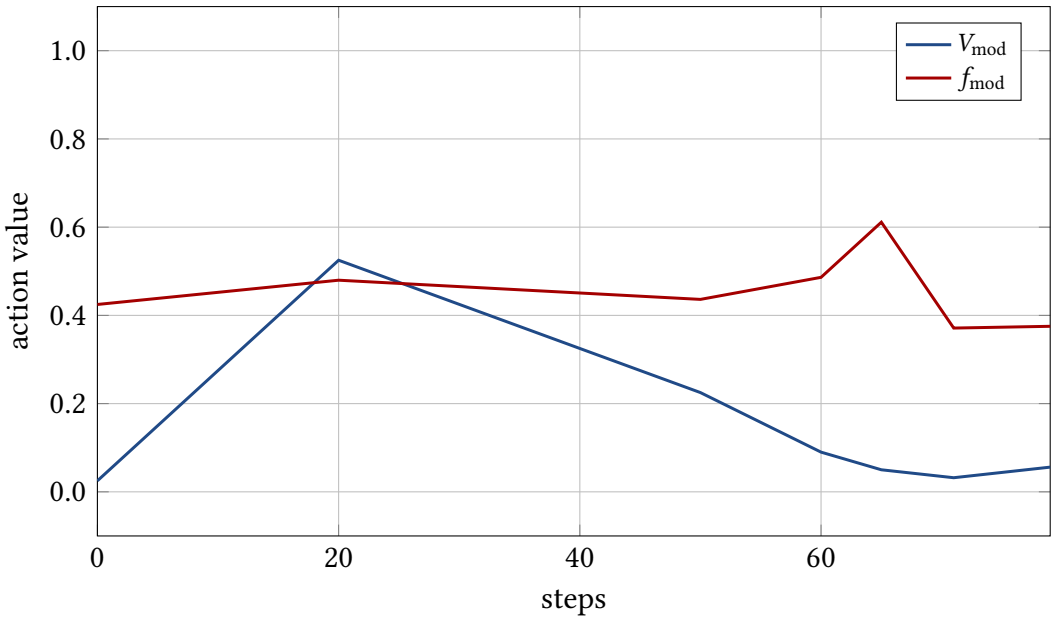



Figure 7.10.: Action values chosen during the sequence displayed in Fig. 7.8. The modulation amplitude $V_{\mathrm{mod}}$ is initially increased until the actions show the desired effect and the perturbation by the CSR wake potential is reduced. After the first 20 steps, it can therefore be lowered again. The modulation frequency is kept close to the natural micro-structure frequency at $f_{\mathrm{ms}} = 2.99\,f_{s,0}$.

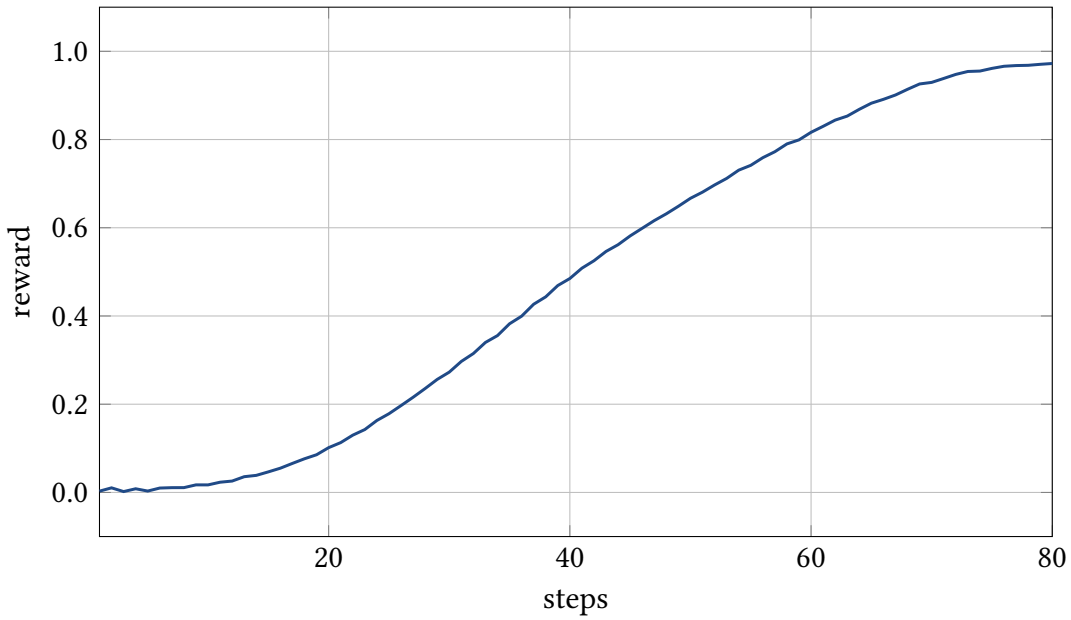



Figure 7.11.: Rewards corresponding to the sequence displayed in Fig. 7.8, calculated according to the definition in Eq. (6.30). To reduce oscillations in the reward signal, the CSR signal of four prior steps is used to calculate the mean and standard deviation.

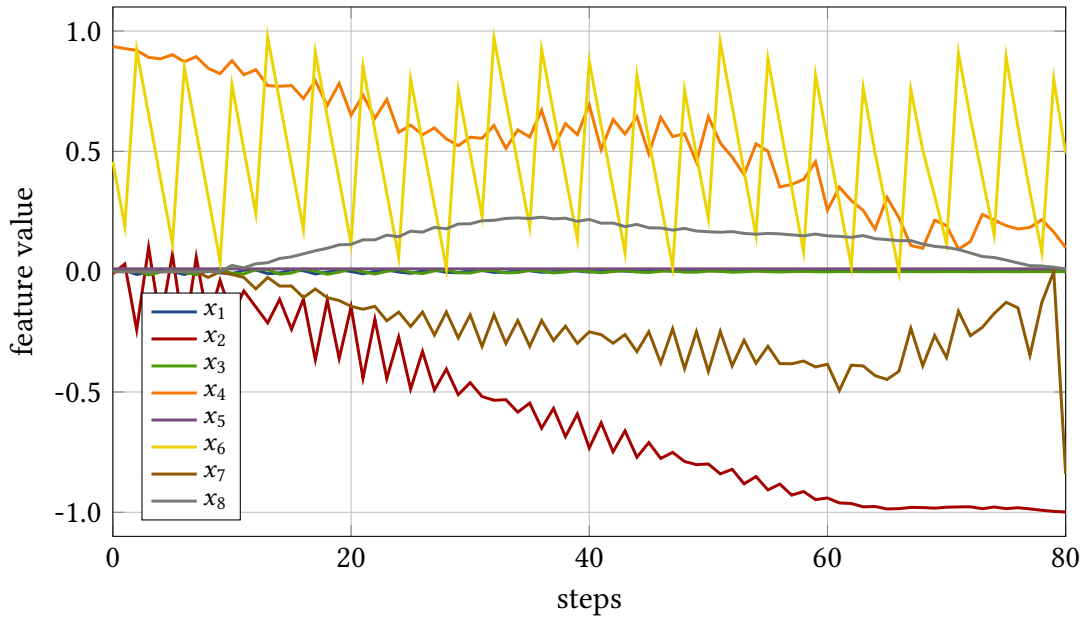



Figure 7.12.: Observed features corresponding to the sequence displayed in Fig. 7.8, calculated according to the definitions in Eqs. (6.13)–(6.21) and Eq. (6.35). The reduced standard deviation of the CSR power signal, encoded in feature $x_2$ (red line), is the main reason for the high rewards achieved towards the end of the sequence.

the normalized action space defined in Eq. (6.3) are displayed in Fig. 7.10. Thereby, the modulation amplitude and frequency are constrained to the intervals $V_{\mathrm{mod}} \in [0, 0.1]\, V_0$ and $f_{\mathrm{mod}} \in [2.5, 3.5]\, f_{\mathrm{s},0}$, with the natural micro-structure frequency in this case being at $f_{\mathrm{ms}} = 2.99\, f_{\mathrm{s},0}$. In an attempt to directly counteract the perturbation by the CSR wake potential, the modulation frequency was kept around the value of 0.5 for most of the chosen actions, that is, it does not deviate much from the natural micro-structure frequency. Yet, it has to be slightly adjusted as the micro-bunching dynamics are altered by previous actions.

The corresponding sequence of rewards is calculated according to the definition in Eq. (6.30) and shown in Fig. 7.11. To obtain a smoother reward signal, the CSR signal of four prior steps is considered for the calculation of the mean and standard deviation. Under these conditions the reward function seems to closely match the physical objective. As the CSR signal in Fig. 7.8 is improving, the reward continuously grows from close to zero, $R_1 = 0.01$, to almost one, $R_{80} = 0.98$, at the end of the sequence. The total undiscounted return obtained over the full sequence is $G_0 = 41.65$, and as a reference for the subsequent RL efforts, the discounted return with $\gamma = 0.99$ is $G_0(\gamma = 0.99) = 24.26$. Although the mean of the CSR power signal is slightly increasing over the sequence, the large majority of the reward is obtained due to a reduction of the standard deviation. As the normalized mean and standard deviation are features in the observation vector, this is visible in Fig. 7.12 where all eight features, calculated according to the definitions in Eqs. (6.13)–(6.21) and Eq. (6.35), are displayed for the considered sequence. While the mean encoded in feature $x_1$ is almost zero, that is, unchanged compared to the natural behavior of the instability, the standard deviation encoded in feature $x_2$ is gradually reduced from zero to minus one. The mitigation of the micro-bunching dynamics is also indicated by feature $x_4$, which encodes the relative strength of the main oscillation frequency in the Fourier transformed CSR power signal. As the micro-structures and the corresponding fluctuations of the CSR power signal are damped, the feature reduces from close to one at the start of the sequence to almost zero towards the end. Although the phase of this CSR power oscillation encoded in feature $x_6$ varies periodically throughout the entire sequence, the phase difference to the applied RF modulation indicated by feature $x_7$ shows a different behavior. In fact, the mitigation of the micro-bunching dynamics seems most effective for negative values down to about $x_7 = -0.5$. Finally, the feature $x_8$ displays the termination condition according to the definition in Eq. (6.35). As it is greater or equal to zero for the entire sequence, the reward signal is continuously improving and the episode would not be terminated. Given that the obtained reward exceeds the value of the relaxation parameter $r_{\mathrm{relax}} = 0.5$ after about 40 steps, the episode would not be terminated regardless.

Overall, these results verify the feasibility of mitigating the micro-bunching dynamics via a modulation of the RF amplitude. It is also important to point out that this not achieved at the cost of lengthening the bunch. In fact, the bunch length is even slightly decreased by the applied RF amplitude modulation. This is indicated by the slowly increasing mean of the CSR power in Fig. 7.8 and shown explicitly in appendix A.6. In the following two sections, RL algorithms are applied to the same task, for which the manual control discussed above serves as a performance benchmark. Particularly with the agent having full access to the Markov states of the system, these results also guarantee that the problem is solvable, at least to the level of control achieved here.

### 7.3.2. RL Results with Phase Space Information

Given a fully functional implementation of the OpenAI gym environment described in section 7.1 and a basic understanding of the involved physics interactions, the application of reinforcement learning methods was pursued in an interdisciplinary collaboration between the Laboratory for Applications of Synchrotron Radiation (LAS), the Institute for Beam Dynamics and Technology (IBPT) and the High Performance Humanoid Technologies (H2T) group in the Institute for Anthropomatics and Robotics (IAR). In particular, the task of training an RL agent with full access to the Markov states of the system was pursued in a Master's thesis conducted by Melvin Klein, at the KIT Department of Informatics under shared supervision by both institutes [98]. His work resulted in the first RL agent reaching the benchmark level of control illustrated in the previous section and provided vital improvements to the formal problem definition given in chapter 6. This subsection reviews the most important results of those joint efforts and presents the peak performance of an RL agent trained using the PPO algorithm. While similar results were achieved using the DDPG or the SAC algorithm, all these algorithms were found to heavily depend on the used random seed. Random number generation is required for the calculation of the exploration noise, the sampling from a replay buffer and other random elements of the used RL algorithms. Moreover, the learning process was frequently unstable, that is, the agent's performance seemed to degrade with continued training time. Also the TD3 algorithm, which is build on DDPG and meant to further stabilize the algorithm, was not found to improve these issues. Of the four algorithms, PPO has the unique characteristic of being an on-policy algorithm which, in theory, should result in reduced variance at the cost of being less sample efficient. Yet, the instabilities in the training process seemed to occur as frequently as for the other algorithms. As these issues also persist under the use of a feature vector, which is discussed in the next subsection, they are further addressed in the final section of this chapter.

The state defined in Eq. (6.27) is a tuple consisting of a matrix $\Delta\hat{\psi}(q, p, t_i)$ describing the charge density in the longitudinal phase space and a five-dimensional feature vector $\boldsymbol{x}^{\mathrm{a}}(t)$. In order to efficiently process this information in a neural network, a convolutional architecture is chosen. In empirical testing, the layout illustrated in Fig. 7.13 was found to yield the best performance and was thus generally used for the actor network, and if required, slightly adapted for the critic network. Initially, the charge density matrix is processed by five convolutional layers with $3 \times 3$-filters, each followed by a max pooling layer. The output of the final layer is concatenated with the additional feature vector $\boldsymbol{x}^{\mathrm{a}}(t)$ and further processed in four fully connected layers, eventually yielding a two dimensional action vector as output. The activation function used for all hidden layers, convolutional and fully connected, is the leaky ReLU function defined by

$$\mathrm{ReLU}_{\mathrm{leaky}} : \mathbb{R} \to (-\infty, \infty) \quad \text{with} \quad \mathrm{ReLU}_{\mathrm{leaky}}(x) \doteq \begin{cases} x & \text{if } x > 0 \\ 0.01\,x & \text{if } x \leq 0 \end{cases}, \tag{7.2}$$

which allows for a small positive gradient when the unit would otherwise not be active. This was found to yield slightly better results than the standard ReLU function defined in Eq. (4.37). The activation function of the output layer is the logistic sigmoid function defined in Eq. (4.35), which restricts the resulting action to $a \in [0, 1] \times [0, 1]$. Given the

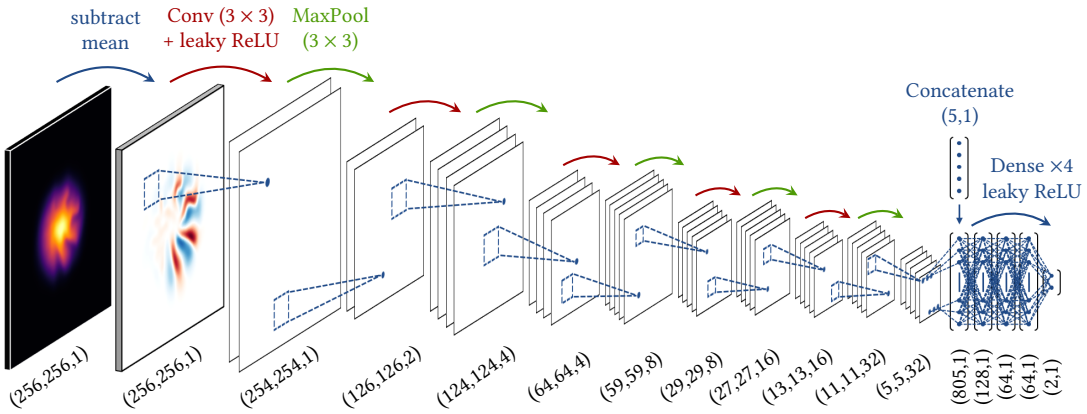



Figure 7.13.: Layout of the actor network used to process the states defined by Eq. (6.27). The charge density matrix is processed by alternating convolution and max pooling layers and subsequently concatenated with the additional feature vector, which serves as the input for the final fully connected layers.

exemplary actions chosen under manual control, the actor and critic network may also be pre-trained on a small data set. This was partially used to optimize the architecture of the network, but did not result in a warm start or an improvement of the final performance for the pre-trained agents. The initially well-performing network is quickly altered by the updates of the agent and degrades in performance, essentially starting from scratch again after only a few episodes.

In order to evaluate the outcome of individual training sessions, the agents are put into evaluation mode, that is, all randomness involved in the exploration process is turned off and the agent chooses greedily at every time step. As the underlying physics simulation is fully deterministic, this yields perfectly reproducible results once the actions are fixed. Even for the largest used networks, the computation time of the training process is primarily determined by the physics simulation. This is still the case when running the parallelized version of Inovesa and ultimately limited by the latency in-between individual simulation steps. Nonetheless, a typical training session consisting of 50 000 to 500 000 steps of the RL environment can be conducted within a single day on standard desktop computers[1], or even faster if more powerful hardware is available. Extending the training session to a larger number of steps was not found to be particularly beneficial due to the aforementioned instabilities leading to degrading performances. Although the maximum episode length is typically restricted to 80 steps during training, the pursued micro-bunching control fundamentally constitutes a continuing task. As the objective is not to obtain high immediate rewards at the cost of future return, a discount factor close to one, $\gamma = 0.99$, is used to keep the return bounded independent of the episode length.

As an important milestone, Fig. 7.14 displays one of the first results where the benchmark performance was exceeded by an RL agent. Under peak performance, the PPO agent managed to perform 80 steps without termination of the episode. In fact, it even manages

[1] As an example, and for better comparability, one of the used systems was an ESPRIMO P920 manufactured by FUJITSU featuring an Intel(R) Core(TM) i7-4770 CPU @ 3.40 GHz.

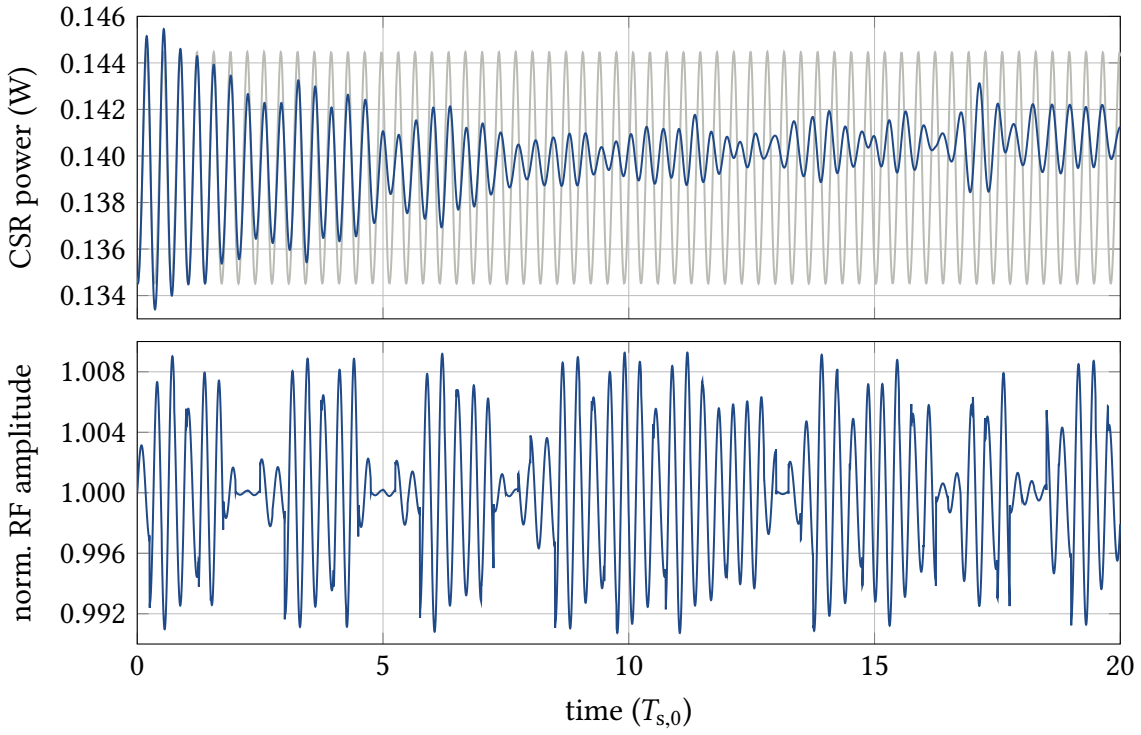
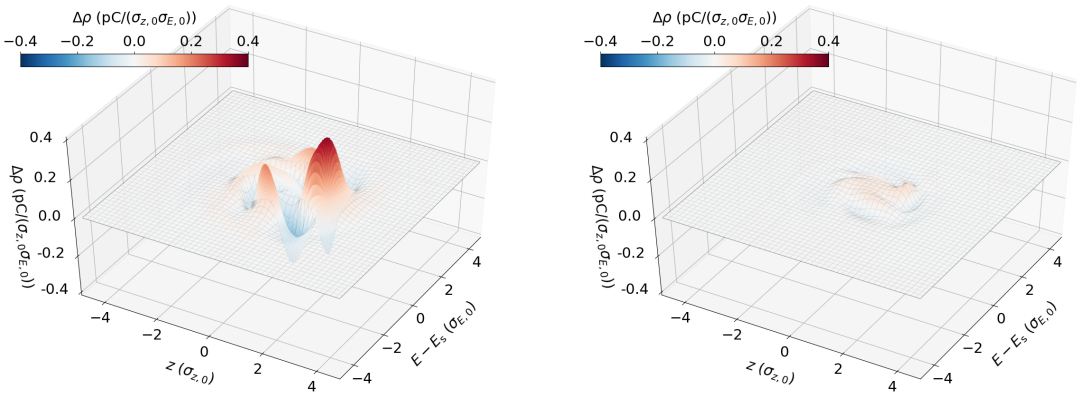



Figure 7.14.: Mitigation of the micro-bunching dynamics and the corresponding oscillations in the CSR power signal (top) by an RL agent trained with the PPO algorithm. The chosen actions and the resulting modulation of the RF amplitude (bottom) differ quite significantly from the manual control shown in Fig. 7.8. Yet, the agent clearly achieves a mitigation compared to the natural behavior of the instability (gray line) with an overall performance that is comparable.

to continue for several steps, reaching a total of 116 steps before violating the termination condition. While the CSR power signal is deteriorated in the first synchrotron period, the agent subsequently recovers and succeeds in damping the oscillation quite effectively with the best performance reached after about 15 synchrotron periods. Afterwards, the oscillation reaches slightly larger amplitudes again. The effect on the corresponding charge distribution in the longitudinal phase space is shown in Fig. 7.15. The naturally occurring micro-structures are again clearly reduced in amplitude, resulting in a much smoother distribution. Intriguingly, these results are achieved by a very different manipulation of the RF amplitude compared to the manual control shown in Fig. 7.8. The agent alternates between small and large modulation amplitudes and simultaneously varies the modulation frequency, as can be seen Fig. 7.16. The agent makes use of a large part of the available action space while dynamically adjusting its actions to the changing micro-bunching dynamics. This suggests that different strategies to mitigate the micro-bunching dynamics are feasible within the defined actions space. Given that the agent aims to maximize the discounted return, the exact definition of the reward function assigns priority to one




(a) Initial micro-structures at $t = 0\,T_{s,0}$ (b) Mitigated micro-structures at $t = 20\,T_{s,0}$

Figure 7.15.: (a) The initial micro-structures for the sequence displayed in Fig. 7.14 are quickly damped by the PPO agent. (b) Similarly to Fig. 7.9b, after 20 synchrotron periods, the micro-structures are clearly reduced in amplitude.

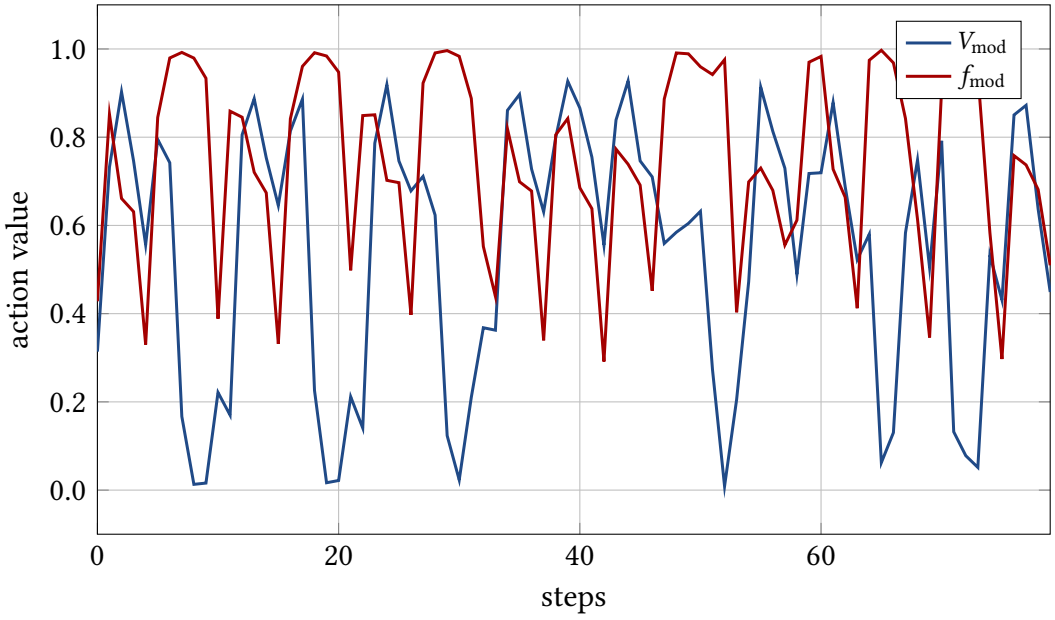



Figure 7.16.: Actions chosen by the PPO agent for the sequence displayed in Fig. 7.14. Both, the amplitude and the frequency of the modulation are dynamically adjusted by the RL agent to mitigate the micro-bunching dynamics.

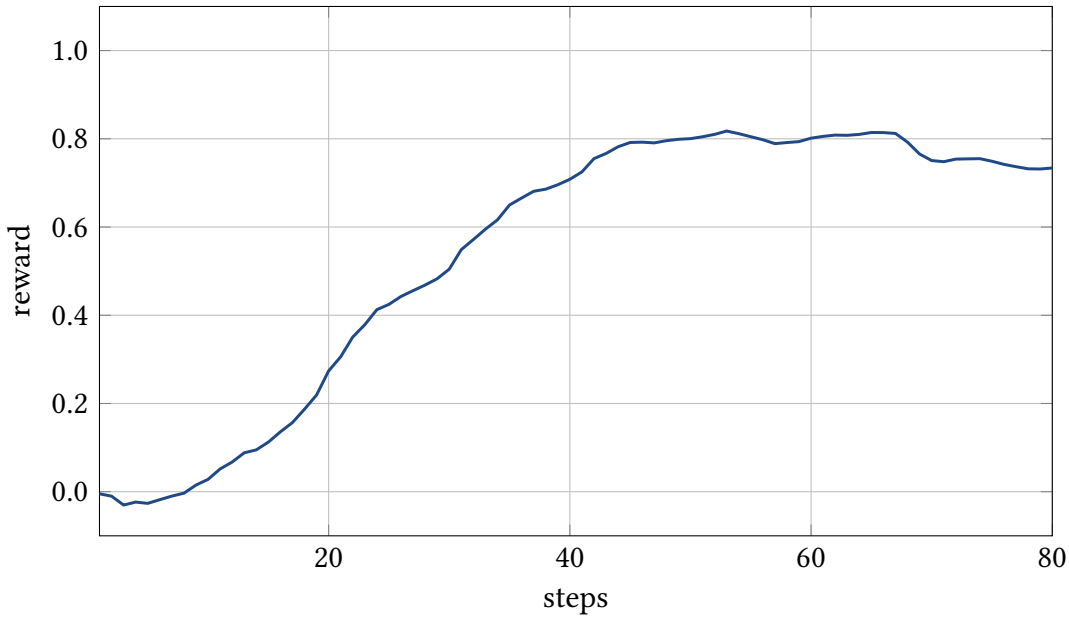



Figure 7.17.: Rewards obtained by the PPO agent during the sequence displayed in Fig. 7.14. While the maximum value of immediate reward at $R_{67} = 0.82$ is lower than under manual control, the total return is slightly higher, $G_0 = 45.66$, due to the fast improvement at the beginning of the sequence.

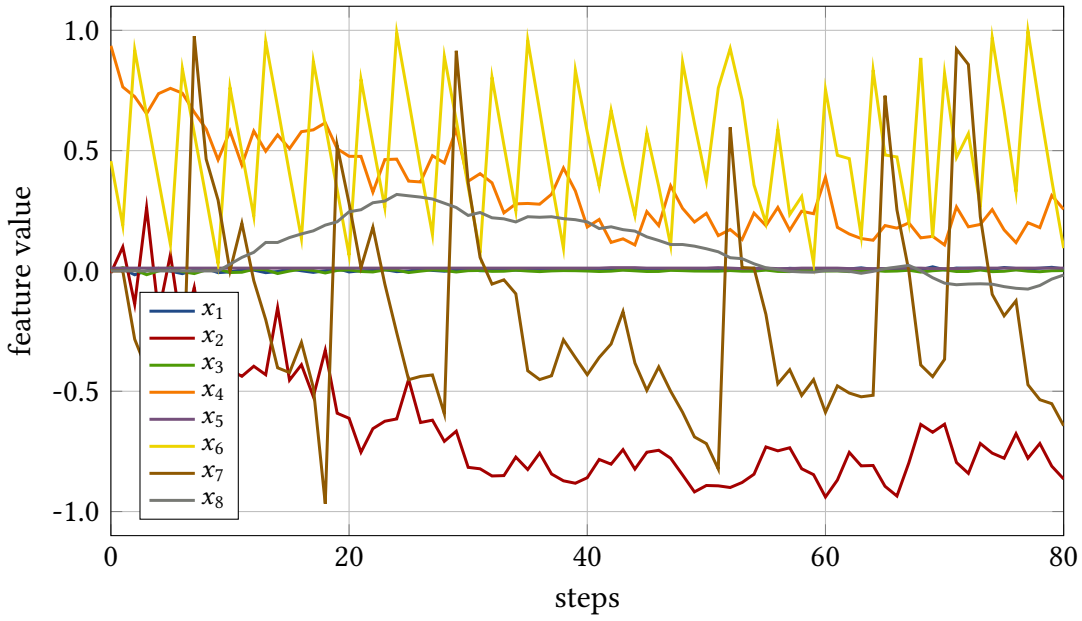



Figure 7.18.: Evolution of the feature vector corresponding to the sequence displayed in Fig. 7.14. Although the RL agent is provided with states defined according to Eq. (6.27), these features are still used for the purpose of analysis and comparison.

strategy or the other. While the choice in Eq. (6.30) seems to match the general physics objective, an additional refinement of the definition may be beneficial to create a higher incentive for the desired solution. The sequence of rewards generated by these actions is shown in Fig. 7.17. Although the maximum achieved immediate reward, $R_{67} = 0.82$, is lower than for manual control, the total undiscounted and discounted return, $G_0 = 45.66$ and $G_0(\gamma = 0.99) = 28.24$, are slightly higher. This is because the PPO agent is faster in achieving a high level of reward, reaching $R_t = 0.5$ after merely 30 thirty steps. At about 45 steps it reaches a plateau at $R_t = 0.8$ before the obtained reward decreases again towards the end of the sequence. Yet, it stays above the value of the relaxation parameter $r_{\text{relax}} = 0.5$, which prevents the termination of the episode.

Although the agent is provided with states defined according to Eq. (6.27), for the purpose of analysis and better comparability, the feature vector defined in Eq. (6.10) is calculated for the given data set and displayed in Fig. 7.18. Here, the fast reduction of the fluctuation of the CSR power signal is clearly indicated by feature $x_2$ encoding its standard deviation. After about 30 steps it reaches a value around $x_2 \approx -0.8$ and stays at that level for the remaining sequence. Simultaneously, the relative amplitude of the main oscillation frequency encoded by feature $x_4$ is reduced again. What differs quite significantly from the features displayed in Fig. 7.12, is the relative phase between the applied modulation and the CSR power signal. Instead of staying at negative values around $x_7 \approx -0.5$, it periodically jumps to large positive values again. This seems to be correlated to the periodic adjustment of the modulation amplitude shown in Fig. 7.16. Whenever the phase is mismatched, with feature $x_7$ jumping to large positive values, the modulation amplitude is lowered to almost zero in order to not degrade the signal. Yet, the agent is capable of establishing the correct phase relation again by adjusting the modulation frequency and subsequently ramps up the modulation amplitude, damping the micro-bunching dynamics in the process.

### 7.3.3. RL Results with Solely CSR Information

The results illustrated in the previous subsection demonstrate the effectiveness of RL methods to solve the sequential decision process arising from the required dynamic adjustment of the RF amplitude modulation. While this verifies the general approach and provides some theoretical comfort, the final objective is a solution which can be implemented at KARA. Unfortunately, providing the RL agent with precise and reliable measurements of the charge distribution in the longitudinal phase space is not yet an available option. The RL efforts in this subsection are thus focused on solving the same task while replacing the state definition in Eq. (6.27) with the feature vector in Eq. (6.10), which relies solely on the observed CSR power signal. As this is quite a substantial change which undermines the fulfillment of the Markov property, the feasibility of this approach is unknown a priori. If the information encoded in the eight-dimensional feature vector were found to be insufficient to decide on the necessary adjustments of the RF modulation, the problem would be ill-posed and not solvable for any RL agent. Yet, that did not turn out to be the case in practice. In fact, the peak performance presented in this subsection even slightly exceeds the results of the previous subsection. The reason for these findings may be related to the considerable, additional computational effort involved in deriving meaningful features from the charge density matrix. The architecture of the neural networks used in

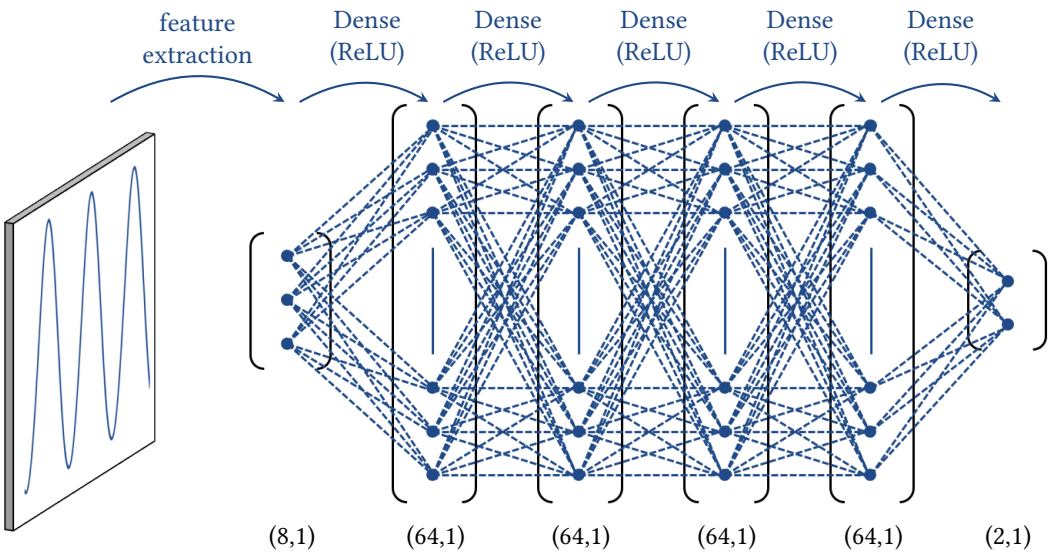



Figure 7.19.: Layout of the actor network used to process the observations defined by the eight-dimensional feature vector derived from the CSR power signal.

conjunction with the state definition in Eq. (6.27) is far more complex than the networks used to process the eight-dimensional feature vector. Typically, a simple fully connected network with three to four hidden layers, as illustrated in Fig. 7.19, was used and found to be sufficient. While the best results were achieved with four layers consisting of 64 units each, going down to three layers with as few as 16 units still led to reasonable results. Given the challenging repetition rate of the RL feedback loop in practice, there is a clear incentive to reduce the complexity of the network as much as possible. The activation function used for all hidden layers is the standard ReLU function defined in Eq. (4.37). In order to restrict the two-dimensional output of the network to the normalized action space, $\mathcal{A} \doteq [0, 1] \times [0, 1]$, the activation function of the final layer is the logistic sigmoid function defined in Eq. (4.35).

Figure 7.20 displays one of the earlier results obtained while training an RL agent on the eight-dimensional CSR feature vector. The termination condition in this particular training session was based on the reward gradients of previous steps as defined in Eq. (6.38). While this was later found to facilitate long episodes in which the agent was not improving and therefore neglected in favor of the definition in Eq. (6.35), it led to some early successes in training. Under peak performance, the DDPG agent manages to reach 73 steps, or 18.25 synchrotron periods, before eventually violating the termination condition. During that time the oscillation of the CSR power signal is visibly reduced, albeit not to the same level as in Figs. 7.8 or 7.14. After the first five synchrotron periods, in which the oscillation amplitude is temporarily even larger than for the natural behavior of the instability, the agent manages to damp the oscillations for a time window of roughly ten synchrotron periods before the oscillation amplitude grows again at the end of the episode. At $t = 18.25\, T_{s,0}$, the episode is eventually terminated due to the immediate reward dropping below the value of the relaxation parameter, $r_{\text{relax}} = 0.5$. The effect on the corresponding charge distribution in the longitudinal phase space is illustrated in Fig. 7.21.

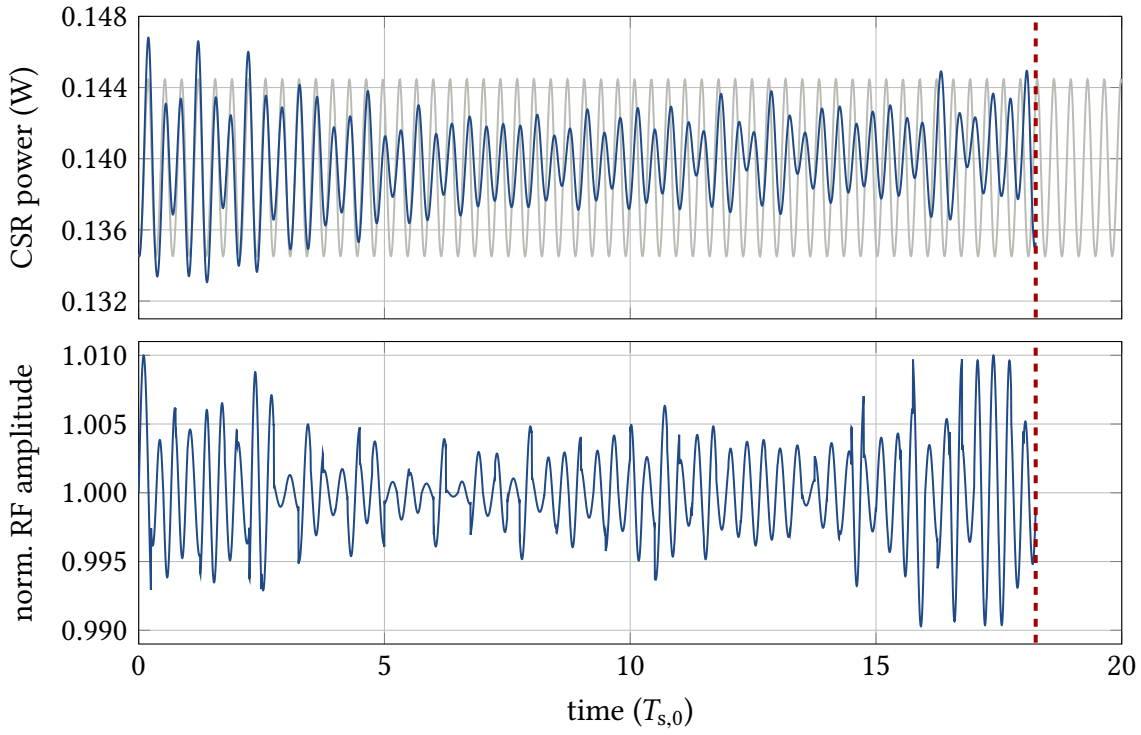
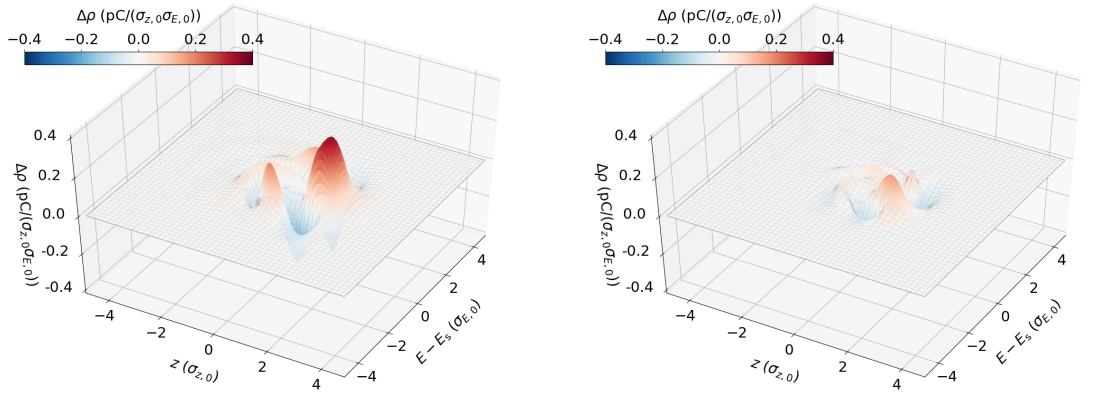



Figure 7.20.: Mitigation of the micro-bunching dynamics and the corresponding fluctuations in the CSR power signal (top) by the DDPG agent. The applied modulation of the RF amplitude (bottom) differs again from the manual control in Fig. 7.9b and also from that of the PPO agent in Fig. 7.14. The dashed red lines indicate the termination of the episode due to insufficient performance by the agent.

Shortly before the oscillations in the CSR power signal grow again, at $t = 15\,T_{s,0}$, the micro-structures are significantly damped but can still be identified in the charge distribution. Over the remaining three synchrotron periods the amplitude of the micro-structures reaches about the same level again as for the initial distribution shown in Fig. 7.21a. These results are achieved through a sequence of actions which frequently includes large changes between consecutive steps, as shown in Fig. 7.22, and therefore seems quite chaotic. Although there is a trend to smaller modulation amplitudes and frequencies halfway through the episode, both action values still fluctuate during the entire sequence. This behavior is quite different from the sequence of actions under manual control shown in Fig. 7.10 and also from those selected by the PPO agent displayed in Fig. 7.16. The sequence of rewards achieved by these actions is shown in Fig. 7.23. Both, the maximum immediate reward at $R_{64} = 0.57$ and the total return, $G_0 = 26.15$ and $G_0(\gamma = 0.99) = 16.52$, are substantially lower compared to previous results. The general distribution of rewards resembles that in Fig. 7.28. After dropping below the baseline level of $R_t = 0$ initially, the agent quickly improves during the subsequent steps and eventually reaches a plateau, here merely around $R_t = 0.5$. Eventually, the performance is reduced again, which in this case




(a) Initial micro-structures at $t = 0\,T_{s,0}$ (b) Mitigated micro-structures at $t = 15\,T_{s,0}$

Figure 7.21.: (a) The initial micro-structures are damped by the DDPD agent during the sequence displayed in Fig. 7.20. (b) Shortly before the episode is terminated due to insufficient performance by the agent, the micro-structures are visibly reduced in amplitude.

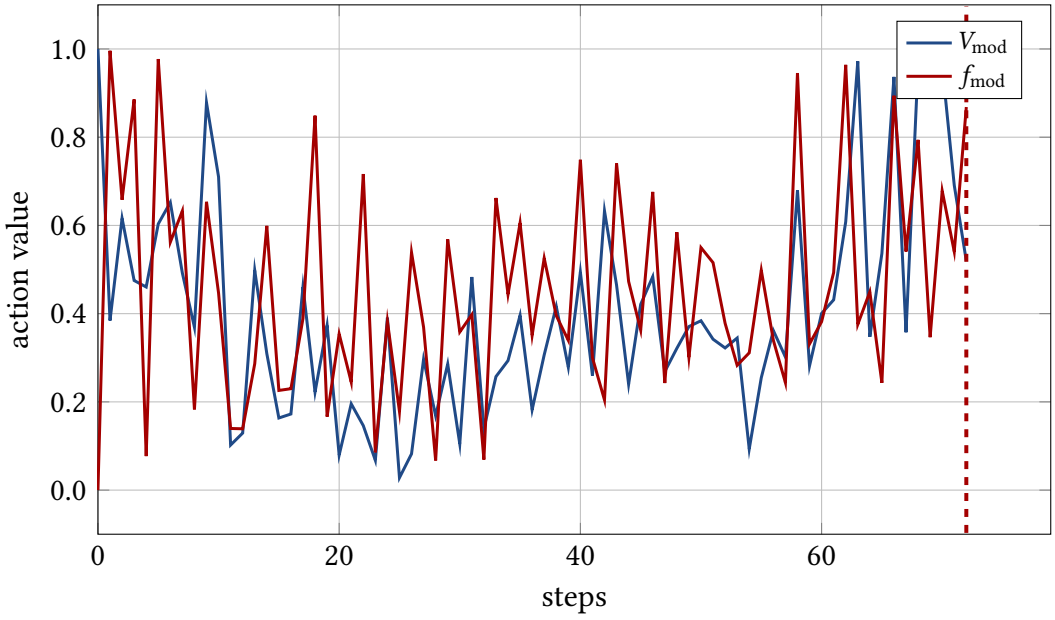



Figure 7.22.: Actions chosen by the DDPG agent during the sequence displayed in Fig. 7.20. Besides a slight trend to smaller action values halfway through the episode, the modulation amplitude and frequency mainly display rapid, irregular oscillations.

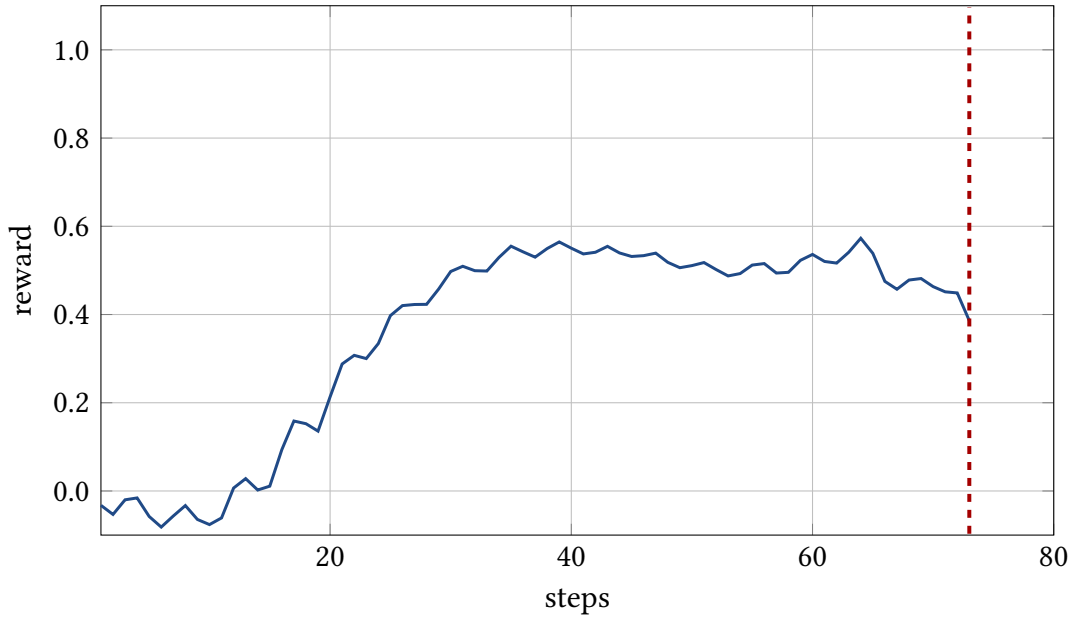



Figure 7.23.: Rewards corresponding to the sequence displayed in Fig. 7.20. The maximum immediate reward at $R_{64} = 0.57$ as well as the total return $G_0 = 26.15$ are substantially lower than in the previous subsections, yet the distribution of rewards resembles that in Fig. 7.17.

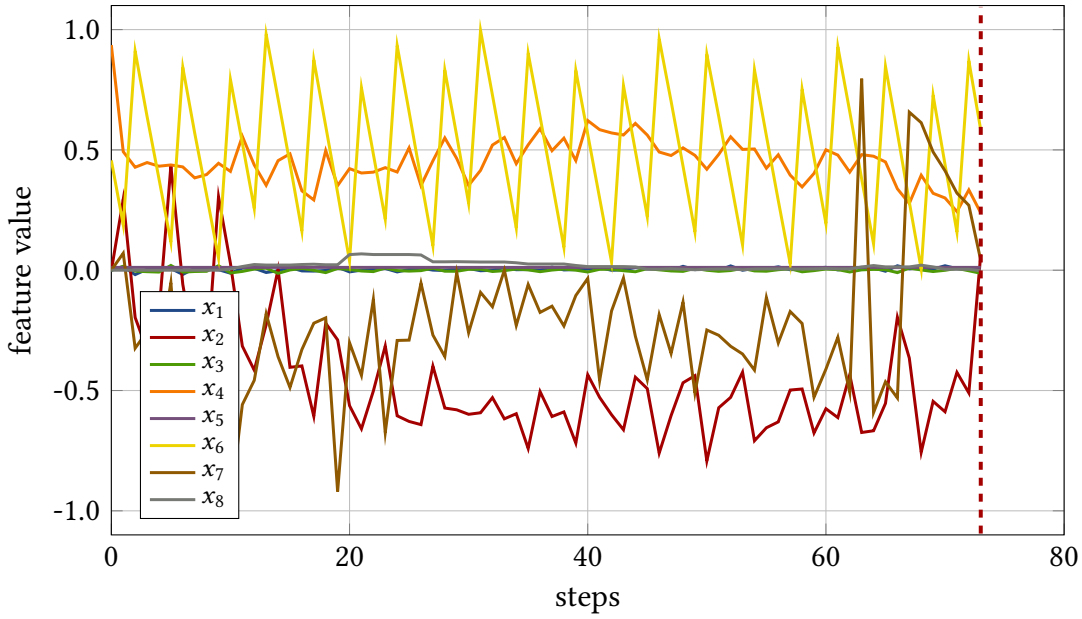



Figure 7.24.: Features as observed by the DDPG agent during the sequence displayed in Fig. 7.20. The termination condition encoded in feature $x_8$ is based on the alternative definition in Eq. (6.38).

terminates the episode as the reward drops below the value of the relaxation parameter. The corresponding sequence of the CSR feature vector, as observed by the agent, is shown in Fig. 7.24. While the relative amplitude of the main oscillation frequency, encoded in feature $x_4$, is immediately decreased to about $x_4 = 0.5$ at the start of the episode, this does not result in a reduced standard deviation, encoded in feature $x_2$, or a high reward. This indicates that the oscillation of the CSR power signal initially continues at roughly the same level, but with a modified frequency distribution. After 20 steps, the standard deviation is eventually reduced corresponding to the increased reward in Fig. 7.23. For the period between step 20 and step 60, where the mitigation of the micro-bunching dynamics is most successful and the reward is at its highest value, the phase difference between the applied RF modulation and the CSR power signal is again at similar values compared to Figs. 7.12 and 7.18. The corresponding feature mainly stays at small negative values, $-0.5 < x_7 < 0$, until jumping to $x_7 = 0.8$ at step 63, where control over the micro-bunching dynamics is gradually lost and the corresponding reward declines. Although these results do not quite reach the level of control demonstrated in subsections 7.3.1 and 7.3.2, given the significantly reduced information available to the agent, they were considered an important intermediate step. Yet, they were exceeded considerably shortly afterwards.

The results of the up to now most successful training session, using again the PPO algorithm, are displayed in Fig. 7.25.[2] Here, the oscillations of the reward function are not mitigated by averaging the mean and standard deviation of the CSR signal over prior steps, but by using the proxies defined in Eq. (6.31) and Eq. (6.33). Although these two complementary definitions generally lead to very similar values of the reward function, to ensure comparability, the performance is again evaluated using the same reward function as above. Similarly to the results in the previous subsection, the agent also quickly manages to damp the oscillation of the CSR power signal and the corresponding micro-structures. After merely five synchrotron periods, the agent has already reached a high level of control and manages to maintain it until the end of the displayed time frame. In fact, the agent even managed to reach a total number of 160 steps and was only terminated because of the a priori defined maximum episode length. The level of control demonstrated after $t = 5\,T_{s,0}$ in Fig. 7.25 was maintained for about 35 synchrotron periods. To achieve this, the agent almost exclusively uses the modulation frequency for dynamic adjustments while generally setting the modulation amplitude to the maximum value. The effect on the corresponding charge distribution in the longitudinal phase space is illustrated in Fig. 7.26. The mitigation of the oscillations in the CSR power signal in Fig. 7.25 is again accompanied by a significant reduction of the micro-structures in amplitude. After five initial synchrotron periods, the charge distribution is already quite smooth and kept that way for the remaining episode. The actions chosen by the agent during that sequence are shown in Fig. 7.27. Again, the choice of actions is quite different from previous results. With the exception of actions $A_{12}$ and $A_{16}$, the modulation amplitude is set to values close to the maximum throughout the entire episode. The necessary dynamic adjustments to the varying perturbation by the CSR wake potential are made via the modulation frequency. Starting at smaller values initially, the chosen modulation frequency eventually oscillates around the value of 0.5, which corresponds to the natural micro-structure frequency. The

[2] Courtesy of Melvin Klein, who briefly worked on the subject as a research assistant.

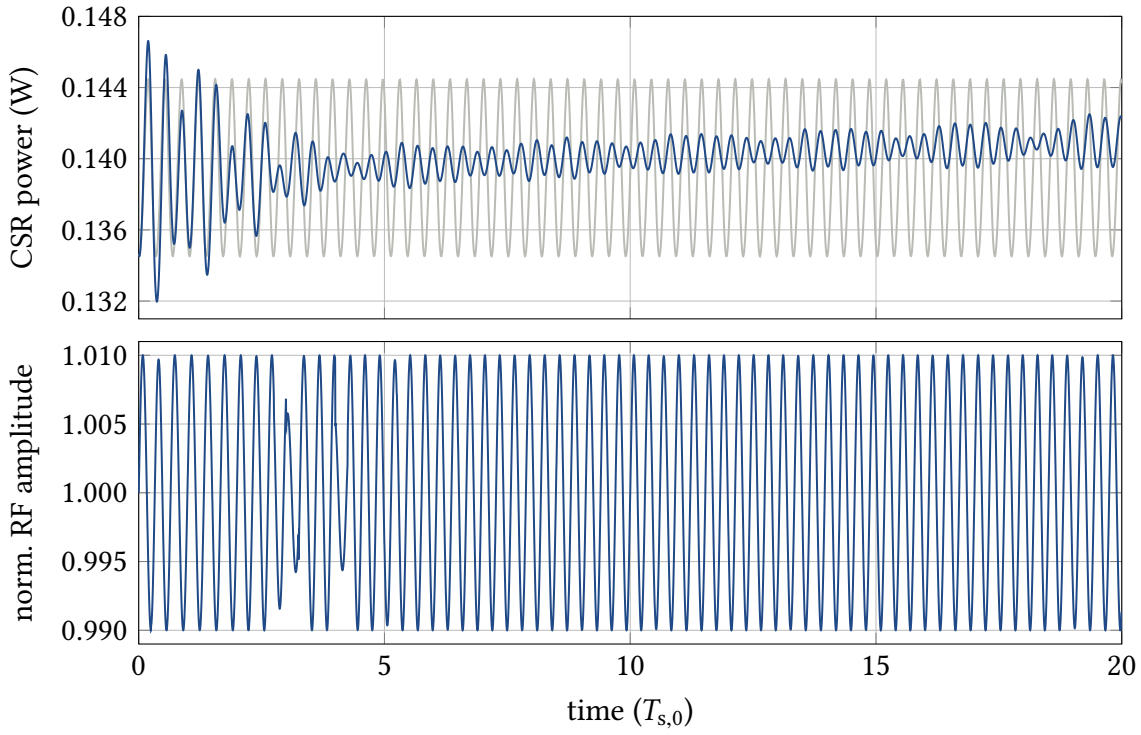
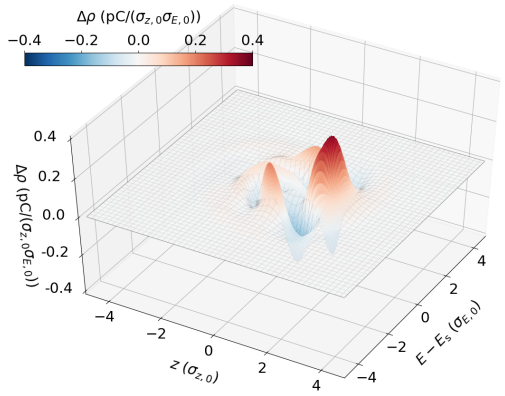
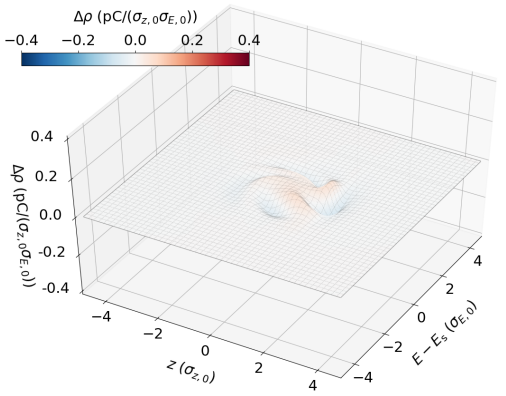



Figure 7.25.: Mitigation of the oscillations in the emitted CSR power (top) and the underlying micro-bunching dynamics by a PPO agent trained on the feature vector derived from the CSR power signal. The modulation of the RF amplitude chosen by the agent (bottom), is almost always at the maximum amplitude. Dynamic adjustments to the changing micro-bunching dynamics are made almost exclusively via the modulation frequency as shown in Fig. 7.27.

sequence of rewards, calculated according to the same definition as for previous agents (smoothing via prior time steps), is shown in Fig. 7.28. While the maximum immediate reward obtained by the agent, $R_{43} = 0.83$ is only slightly higher than in Fig. 7.17 and still below the level of manual control, the total return, $G_0 = 52.18$ and $G_0(\gamma = 0.99) = 32.92$, is increased by a considerable margin. This additional gain in cumulative reward is again achieved by reaching high levels of immediate reward more quickly. The agent reaches values around $R_t = 0.5$ after merely 15 steps, and $R_t = 0.8$ after about 25 steps. This high level of reward is subsequently maintained in a very stable manner until the end of the episode. Analogously to the previously presented agents, the high return is primarily achieved by a reduction of the fluctuations in the CSR power signal and the resulting standard deviation. This is apparent from Fig. 7.29, which displays the feature vector observed by the agent (termination condition $x_8$ derived from proxies). The features $x_4$ and $x_7$ clearly indicate again the reduced oscillation of the CSR power signal. Moreover, the phase difference between the applied RF modulation and the CSR power signal quickly reaches and stabilizes at values around $x = -0.5$, which has been found to be quite effective throughout this thesis. This result is particularly intriguing as it confirms an expectation




(a) Initial micro-structures at $t = 0\,T_{s,0}$




(b) Mitigated micro-structures at $t = 20\,T_{s,0}$

Figure 7.26.: (a) The natural micro-structures occurring at the beginning of the sequence displayed in Fig. 7.25 are quickly damped by the RL agent. (b) After 20 synchrotron periods, the micro-structures are significantly reduced in amplitude.

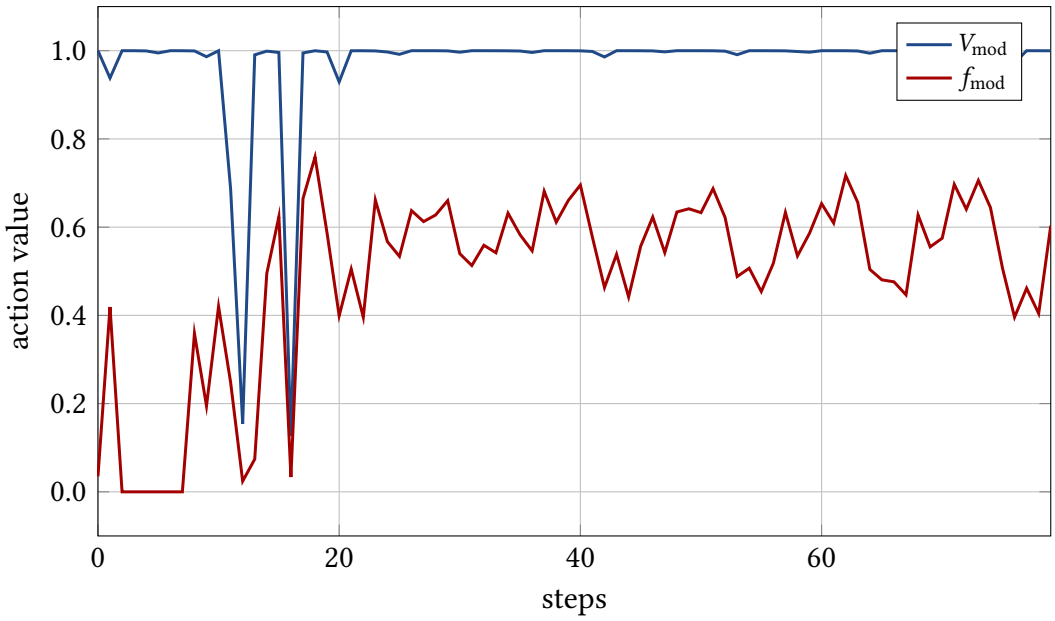



Figure 7.27.: Actions chosen by the PPO agent for the sequence displayed in Fig. 7.25. While the modulation frequency is dynamically adjusted according to the changing micro-bunching dynamics, the modulation amplitude is set close to the maximum value throughout the episode.

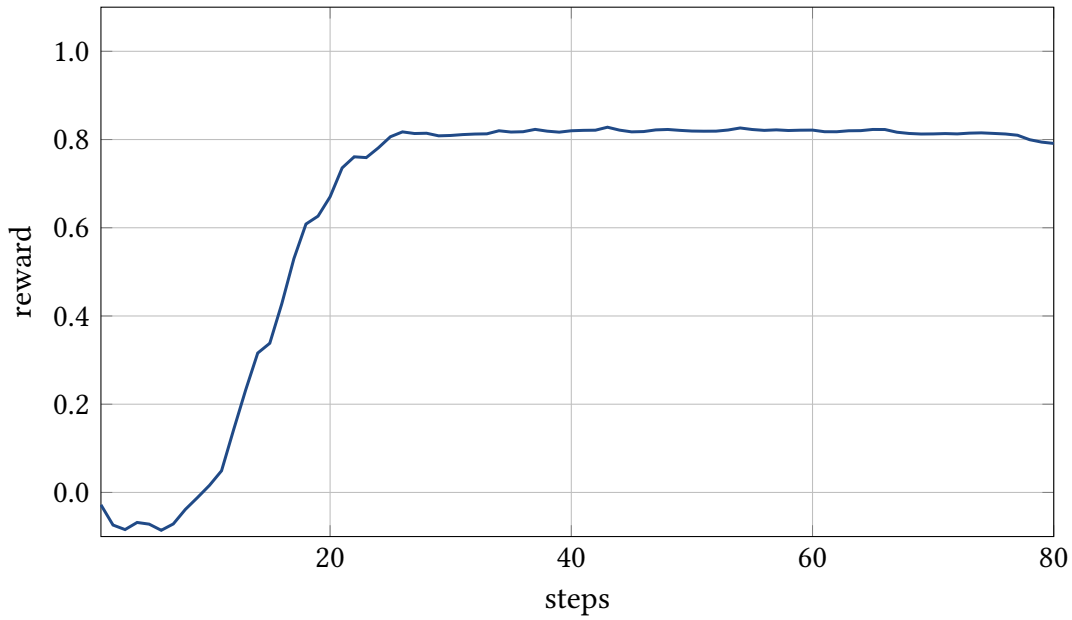



Figure 7.28.: Sequence of rewards corresponding to the mitigation of the micro-bunching dynamics shown Fig. 7.25. The obtained reward quickly reaches and stabilizes at values around $R_t = 0.8$, which allows the agent to achieve the highest return yet, $G_0 = 52.18$.

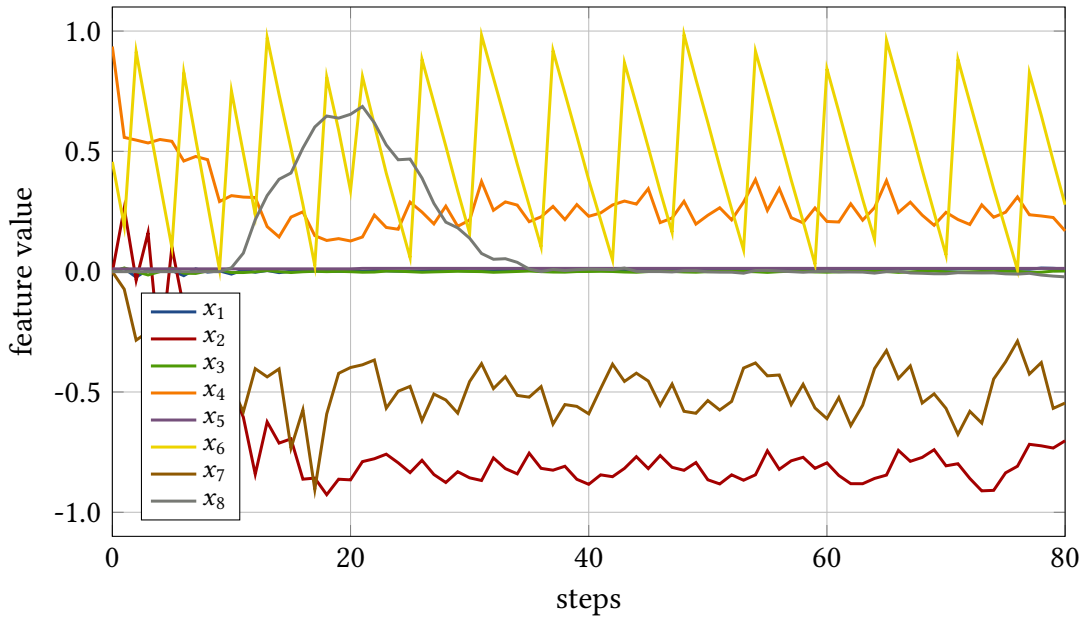



Figure 7.29.: Features as observed by the PPO agent. The termination condition is calculated using the proxies defined in Eq. (6.31) and Eq. (6.33). After an initial adjustment period, the phase difference between the applied RF amplitude modulation and the observed CSR power signal is kept almost constant by the agent, that is, $x_7 \approx -0.5$.

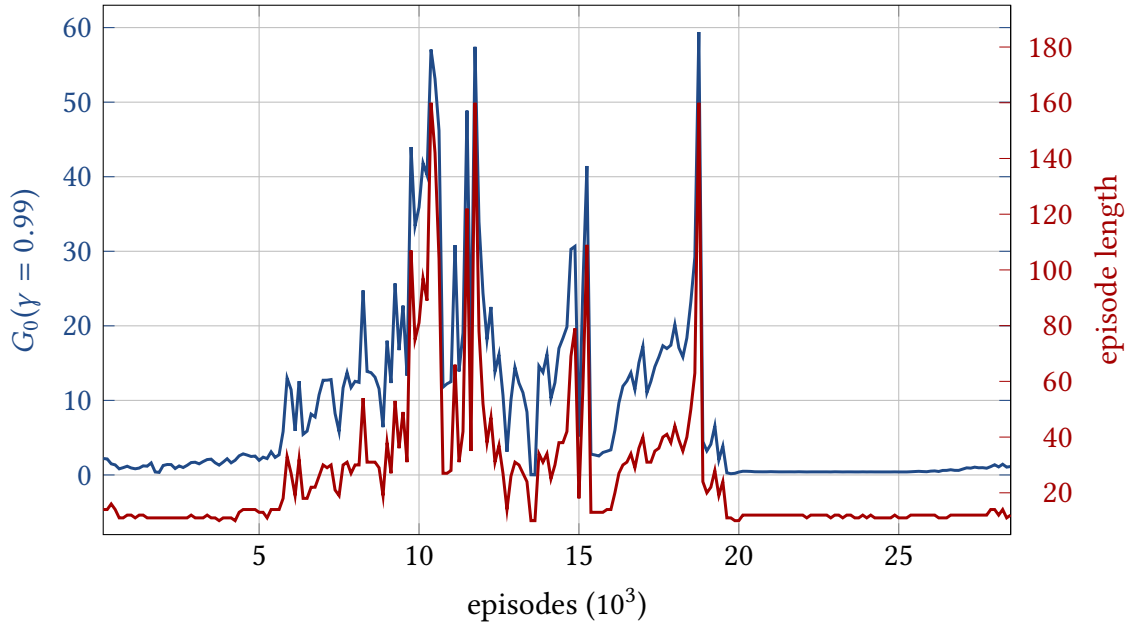



Figure 7.30.: Learning process of the PPO agent trained on the CSR feature vector (in evaluation mode). After about 10 000 episodes, the performance starts to fluctuate and eventually drops to the baseline level again after episode 18 750.

based on the analysis of the underlying longitudinal beam dynamics in chapter 5. With a constant phase relation to the perturbation caused by the CSR wake potential, the applied RF amplitude modulation succeeds in partially recovering the strength of the restoring force and thereby mitigates the formation of micro-structures. As the RL agents are not given any prior knowledge about the environment, these results confirm the general hypothesis and simultaneously demonstrate the agent's ability of identifying crucial information about the underlying dynamics.

While the RL agents presented in this and the previous subsection clearly demonstrate the capability of mitigating the micro-bunching dynamics, this high level of control is only achieved under peak performance. The exemplary learning process of the PPO agent trained on the CSR feature vector is illustrated in Fig. 7.30. Shown are the discounted return (using proxies) and the episode length achieved by the agent after being set into evaluation mode. After about 5000 episodes, the agent starts to continuously improve and eventually reaches the maximum a priori defined episode length of 160 steps and a corresponding return of $G_0(\gamma = 0.99) = 57.06$ at episode 10 375. Afterwards, the performance starts to fluctuate and slightly degrades until reaching episode 18 750, which constitutes the peak performance shown in Figs. 7.25–7.29. Subsequently, the performance suddenly drops again to the baseline level of $G_0(\gamma = 0.99) = 0$ and an episode length of ten steps. Although a small improvement may be identified towards the end of the training session, the agent essential stays at the baseline level for the remaining episodes. A detailed investigation of the causes for these sudden losses in performance was beyond the scope of this thesis. Yet, the final section of this chapter discusses are range of measures to potentially stabilize the agent's learning process, some of which were already rudimentarily tested during

this thesis while others are meant as an outlook towards future work. Overall though, the achieved peak performances clearly demonstrate the feasibility of the developed approach in simulations and provide a proof-of-principle study for the implementation at KARA.

## 7.4. Remarks on Stability and Generalization

The instabilities encountered in the agent's learning process, as illustrated in Fig. 7.30, are undesirable but not strictly prohibitive for the implementation of the developed RL feedback scheme. After the agent is trained for a sufficient number of steps, one may simply select that version of the agent which performs best during evaluation. Nonetheless, the ideal case would be a stable learning process in which the performance of the agent gradually increases until a satisfying amount of control is reached and training can be stopped. While various adjustments of the available hyperparameters, in particular the agent's learning rate, were tested, they generally were not found to significantly improve the stability of the learning process. An additional, promising option is to tailor the agent's exploration noise to the given environment in order to reduce undesirable randomness during training. In the case of the DDPG agent, a slight performance improvement was achieved by changing the Ornstein-Uhlenbeck process suggested in [57] to a Gaussian white noise process. Yet, the learning process was still found to be unstable. One additional complication encountered during these efforts is the difficulty to fix the random seed for the agents implemented in Keras-RL, Stable Baselines and TF-Agents. A fixed random seed offers the benefit of a reproducible test scenario, where the learning process is observed to be unstable. The effect of different measures may then be tested and compared to the initial performance. Yet, due to a variety of reasons, this option is typically not foreseen in the three different RL libraries. While perfect reproducibly was achieved for some algorithms, such as the DDPG algorithm implemented in Keras-RL, others required substantial changes to the implementation of the algorithm and were beyond the scope of this thesis. An alternative approach would be to perform a sufficient number of training sessions with the same parameter settings in order to compare the frequency and severity of the occurring instabilities. While the available time and hardware did not permit such an extensive study, it may be considered more feasible for an implementation at the actual storage ring. The data rate expected at KARA exceeds that of Inovesa by several orders of magnitude. In fact, sample efficiency is expected to be of very little concern in practice. While the best overall performance found in [98] was attributed to the SAC algorithm, the four tested algorithms all achieve very comparable performances and the PPO algorithm is thus generally favored for an implementation at KARA. Due to its nature as an on-policy algorithm, its training process is expected to be more stable. The comparably poor sample efficiency may hinder preparatory simulation efforts, but is expected to be negligible for the performance at KARA. Additional stability may also be gained by a modification of the used reward function or by a restriction of the action space. The different action sequences shown in sections 7.3.1 to 7.3.3 suggest that different strategies to mitigate the micro-bunching dynamics are feasible within the available action space. While this is not an issue in itself, it may lead to contradictory updates during training and thereby reduce the stability of the agent's learning process. An additional term in the reward function,

which favors a continuous action sequence where changes between consecutive actions are small, may be beneficial in assigning priority to more favorable solutions and clarifying the objective. A similar effect may be achievable through a restriction of the available action space. As demonstrated by the agent in Fig. 7.27, assigning a constant value to the modulation amplitude may still provide sufficient flexibility to dynamically adjust the RF modulation to the varying perturbation by the CSR wake potential. A smaller action space is generally desirable, as it reduces the required amount of exploration and the overall complexity of the task. Another opportunity to improve the overall performance is a careful analysis of the different features constituting the eight-dimensional CSR feature vector and their respective relevance for the agent's learning process. While feature $x_7$ is assumed to carry crucial information about the micro-bunching dynamics, other features may not be particularly relevant for the agent's decision making. As the features $x_3$ and $x_5$ are typically found at values close to zero, a rescaling may also be beneficial.

Besides these open questions regarding the stability of the learning process, there are further challenges in exploiting the full potential of the presented approach to micro-bunching control. In order to achieve extensive control over the micro-bunching dynamics at KARA, the presented results results have to be generalized to a larger range of parameters. One such example is the set of machine parameters used for the proof-of-principle studies in this chapter (defined in appendix A.1, data sets $\mathcal{D}_1$ and $\mathcal{D}_2$). As the instability threshold and the micro-bunching dynamics in general depend on several machine parameters, including the accelerating voltage $V_0$ and the momentum compaction factor $\alpha_c$, the demonstrated control has to be verified for alternative configurations of the accelerator. Moreover, as the dynamics also change with the bunch current, leading to the different instability regimes described in subsection 3.4.2, the same holds also true for alternative values of the bunch current. Yet, as the micro-structure formation process is expected to follow a similar mechanism, that is, being largely driven by a CSR-induced perturbation of the restoring force, the developed approach should also be applicable to the dynamics at larger bunch currents. The observed CSR bursts in the sawtooth-bursting regime are generated by rapidly growing micro-structures. These eventually reach considerably larger amplitudes, but their overall shape, and that of the corresponding CSR wake potential, is found to be very similar to those at lower currents. While the perturbation generated by the CSR wake potential may be larger at higher bunch currents, it should still be feasible to counteract these dynamics by an adaptive RF amplitude modulation scheme, although this may require higher modulation amplitudes. As the final objective is continuous control of the occurring micro-bunching dynamics, the achieved simulation results also have to be extended to longer time frames. While the benchmark challenge with a length of 20 synchrotron periods (equal to 2.86 ms) was met and even surpassed by the most successful RL agents, this still constitutes a relatively short time frame compared to the minutes or even hours required at the real storage ring. The 160 steps reached by the PPO agent presented in the previous section are a promising first step in this direction. Finally, the general RL feedback scheme has to be tested for its robustness against noise and stochasticity. The two most relevant sources of noise are expected to result from inaccuracies in realizing the intended RF potential and measurement uncertainties on the observed CSR power signal. As an option for adding RF noise has already been implemented in Inovesa [99], and the uncertainty on the CSR measurements may easily be modeled by the InovesaRL

environment, these additional influences can already be tested in preparatory simulations. Given the generally stochastic treatment of RL problems, most RL algorithms, including those presented in this chapter, should be capable of dealing with a low to medium level of noise.

# 8. Towards Micro-Bunching Control at KARA

*What does his lucid explanation amount to but this, that in theory there is no difference between theory and practice, while in practice there is?*

— Benjamin Brewster,
*The Yale Literary Magazine Vol. 47, Portfolio: Theory and Praxis*

The simulation results presented in the previous chapter demonstrate the feasibility of controlling the micro-bunching dynamics by a carefully adjusted modulation of the RF amplitude. While an RF modulation with constant amplitude and frequency was found to be sufficient for an excitation of the micro-bunching dynamics, their mitigation required dynamic adjustments according to the varying perturbation by the CSR wake potential. The implementation of this general feedback scheme at KARA involves a range of further challenges. In particular, the high frequency at which the RF amplitude modulation has to be adjusted to the evolving beam dynamics leads to a major challenge in practice. The required interaction rate with the beam is directly determined by the time scale of the micro-structure formation process, which is governed by the synchrotron frequency. Although mitigation of the micro-structures was first demonstrated by manual control in subsection 7.3.1, the time difference between consecutive actions in the considered example is $\Delta t = T_{s,0}/4 = 36\,\mu s$. Even tough $\Delta t$ may be set to slightly larger values, it has to be in the same order of magnitude as the synchrotron period, which essentially precludes any form of human involvement in the decision making process. Tasking an RL agent with these decisions still leads to a requirement of very performant hardware to enable the entire feedback loop to run at the required repetition rate. The planned implementation of the RL-based feedback loop, as detailed in section 8.1, thus involves specially designed electronics developed by the Institute of Data Processing and Electronics (IPE) at KIT. While fast read-out electronics for the THz detectors were already developed over the past years, and the second generation thereof, called KAPTURE-2 (Karlsruhe Pulse Taking Ultra-fast Readout Electronics) [100], is available for use, the implementation of the RL agent requires further efforts. In order to meet the requirements resulting from the short computation time between consecutive actions, the RL algorithm and the corresponding neural networks are implemented on an FPGA (Field-Programmable Gate Array). In close cooperation between the IPE and LAS, this general task was taken on by Weijia Wang in his PhD thesis at the KIT Department of Electrical Engineering and Information Technology [101]. Section 8.2 of this chapter provides a brief review of these efforts and summarizes the final results. Following the simulation results in section 7.2, first experiments to verify the possibility of influencing the micro-bunching dynamics at KARA by a modulation of the RF amplitude

are presented in section 8.3. The chapter finally concludes with a brief outlook towards future steps and key challenges in implementing the proposed feedback scheme at KARA.

## 8.1. Implementation of the RL Feedback Scheme

As a test facility for new beam and acceleration technologies and due to its extensive sensor network, the Karlsruhe Research Accelerator (KARA) is ideally suited for the implementation and first tests of the proposed feedback scheme. The information about the state of the micro-bunching dynamics, which is required for the agent's decision making process, is provided by extensively tested and well-established THz diagnostics to measure the emitted CSR power. Using a broadband Schottky diode and the KAPTURE-2 sampling system, the CSR power signal can be measured on a turn-by-turn or even on a bunch-by-bunch basis. To achieve the required high data throughput and fast data processing, the KAPTURE-2 front-end is read out by an FPGA DAQ board on which different reinforcement learning algorithms can be implemented. First results with an implementation of the DDPG algorithm are briefly discussed in the subsequent section. Eventually, the actions provided by the RL agent have to be passed to the RF system of the storage ring in order to realize the corresponding modulation of the RF amplitude. As the more feasible option compared to the main RF system at KARA, a small kicker cavity, which is regularly used in the bunch-by-bunch (BBB) feedback system of the storage ring, is chosen for the interaction with the electron beam. Using the BBB control system, the action vector chosen by the RL agent finally has to be mapped to the corresponding cavity signal. The planned implementation of the entire feedback loop is illustrated in Fig. 8.1, and was recently published in [102].

One crucial question regarding the feasibility of the proposed hardware implementation is whether or not the BBB cavity can reach the necessary modulation amplitudes. Given the simulation results in section 7.3, the required modulation amplitude is roughly $V_{\mathrm{max}} \in [1, 10]\,\mathrm{kV}$. Although there is no calibration data available for the BBB cavity installed at KARA, the maximum reachable voltage can still be estimated. In simulation studies for the almost identical kicker cavity installed at BESSY II [103], the authors arrive at a shunt impedance of $R_{\mathrm{shunt}} \approx 1100\,\Omega$. Given the relation

$$R_{\mathrm{shunt}} = \frac{|V_0|^2}{2P}\,, \tag{8.1}$$

and the input power of $P = 200\,\mathrm{W}$ used at KARA, the maximum voltage is estimated at $V_0 \approx 663\,\mathrm{kV}$, which is slightly below the value used for the simulations in section 7.3, but may still be sufficient. Furthermore, the maximum reachable cavity voltage was also tested in a basic experiment at KARA. As RF modulations can be applied by either the main RF system or the BBB system, one may simply compare their effect on the beam. Using the known amplitude of the modulation set via the main RF system as a reference, a rough guess can also be derived from these measurements. To do so, the electron beam was subjected to an RF phase modulation at the second harmonic of the nominal synchrotron frequency as this typically caused as strong response of the beam. The effect of the modulation was observed via the BBB spectrum, which relies on the measurements of several beam

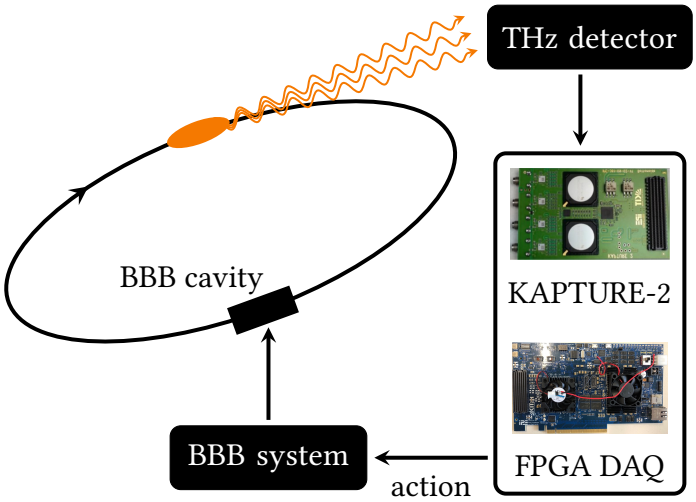



Figure 8.1.: Planned hardware implementation of the proposed RL feedback scheme. The CSR power signal is measured by a THz detector, sampled by the KAPTURE-2 board and read out by the FPGA DAQ board, on which the RL algorithm is implemented. The chosen action is passed to the BBB control system, which generates the corresponding modulation of the RF amplitude. Images: Courtesy of Weijia Wang.

position monitors (BPMs) installed in the storage ring. To apply an RF phase modulation via the main RF system, the KARA control system offers two different options. While it can directly be set in the CSS (Control System Studio) panel, the resulting modulation is expected to also partially involve a modulation of the RF amplitude. An additional Matlab script decouples the RF phase modulation from the RF amplitude and is thus expected to be more reliable. For both options, the peak intensity in the BBB spectrum was measured for a range of modulation amplitudes. These are compared to phase modulations via the BBB kicker cavity at modulation amplitudes of 0.5 and 1.0, as illustrated in Fig. 8.2. Assuming the effect of the RF phase modulation can be described as a simple displacement from the synchronous phase, the required voltage is given by

$$\Delta V = V_0 \left[\sin(\varphi_s + \varphi_{\mathrm{mod}}) - \sin(\varphi_s)\right] \ . \tag{8.2}$$

Using the measurements of the RF modulation set via the CSS panel, one arrives at an estimate of $V_0 \in [5.45, 6.13]$ kV. Yet, the reached peak intensity is clearly lower compared to the RF phase modulation applied via the Matlab script, which is attributed to the partially induced amplitude modulation. Using the measurements of the latter as a reference, one arrives at the slightly lower value of $V_0 \in [4.54, 5.23]$ kV. Regardless which reference is chosen, the estimated maximum cavity voltage is considerably higher than that derived from the shunt impedance. Yet, both estimates roughly put the maximum BBB cavity voltage at the kV-level, which is expected to be sufficient for mitigating the micro-bunching dynamics at low bunch currents.

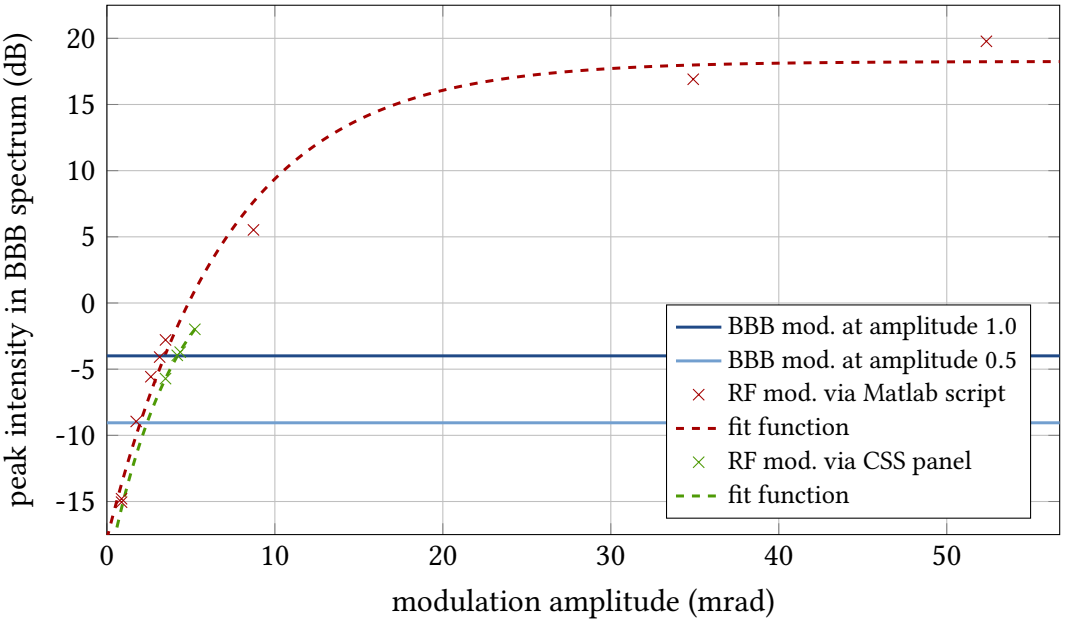



Figure 8.2.: Estimate of the BBB cavity voltage. By comparing the effect of RF phase modulations via the main RF system and the BBB system, the maximum BBB cavity voltage can roughly be estimated using Eq. (8.2). The measurements are interpolated by fitting a limited growth function, $f(x) \doteq a - (a - b)\,\mathrm{e}^{-cx}$. Data: Courtesy of Edmund Blomley.

## 8.2. Meeting the Necessary Time Constraints

One of the major challenges in implementing the feedback scheme illustrated in Fig. 8.1 is meeting the necessary time constraints. In principle, every single iteration requires the extraction of the CSR feature vector from the data sampled by KAPTURE-2, the computation of the next action chosen by the RL agent and an update of the involved neural networks based on the previous experience acquired in the environment. While the latter may possibly be relaxed to updates at every $n$ steps, the computation of the CSR feature vector and the chosen action are strictly bound to time constraints determined by the underlying beam dynamics. The computation time required for the action selection by the agent, that is, inference of the actor network, is thus of crucial importance for the feasibility of the proposed feedback loop. In order to meet these requirements, two different solutions for the implementation of the RL agent are considered in [102]. The first option consists of a heterogeneous architecture, in which the FPGA card is connected to an external GPU. For faster processing, the data is directly transfered from the FPGA into the GPU memory, bypassing the CPU memory system. One benefit of this implementation is that standard machine learning frameworks like TensorFlow are inherently supported by this architecture. The second option directly employs the ARM processor embedded in the FPGA. As this does not require any data transfer to external processing units, it generally yielded better performances and was thus considered the more favorable option for this application. After initial tests of the implemented DDPG agent on the textbook

CartPole problem, the FPGA was connected to the simulation environment introduced in section 7.1 in order to determine the agent's performance. The calculation of the underlying beam dynamics using Inovesa and the computation of the CSR feature vector, as well as the reward function, were conducted on a PC connected by an Ethernet link. With an actor network consisting of four fully-connected hidden layers with 64 units each, the time for a single step of inference was found at $\Delta t = 16.93\,\mu s$. This is significantly below the time difference between consecutive actions used for the simulations in chapter 7 and thus constitutes an important milestone regarding the feasibility of the proposed implementation. Yet, the average time required for a full training step, which also involves backpropagation of the neural networks, was $\Delta t = 1648\,\mu s$. This means the computation of a single training step takes about the time of 50 steps of the environment. Given the high data rate expected at KARA, this may still yield sufficient updates to train an RL agent in practice, but it does slow down the learning process. Ideally, updates should be calculated after every step of the environment. As an additional refinement of the implementation may further reduce the computation time and the performance of the available hardware can be expected to further improve over the next years, this may eventually become feasible. For the time being, the frequency of the training steps has to be restricted to match the required computation time. The actions calculated by the RL agent are finally passed on to the BBB system to realize the corresponding modulation of the RF amplitude. As achieving fast enough data processing requires a modification of the BBB source code, this final step in implementing the RL feedback loop could not be completed within the scope of this thesis. Yet, the required adjustments are considered technically feasible and discussions with the manufacturer of the BBB control system, Dimtel, Inc., are ongoing.

## 8.3. First Experimental Results

Although the implementation of the complete feedback loop is delayed by the required adjustments of the BBB source code, and the RL-based mitigation could thus not be tested in practice, several preliminary experiments could still be conducted at KARA. While the mitigation of the micro-bunching dynamics relies on fast adaptations of the RF amplitude modulation, this is not necessarily required for an excitation of the occurring micro-structures, as shown in section 7.2. The expected effect of an RF amplitude modulation on the micro-bunching dynamics can thus be confirmed without completion of the full feedback loop described in section 8.1. Using the main RF system, the electron beam can simply be exposed to a constant RF amplitude modulation at the micro-structure frequency. Based on the simulations in section 7.2, this should result in an excitation of the micro-bunching dynamics and an amplification of the corresponding fluctuation of the CSR power signal. Given the low-frequency noise on the measured time signals, the results are analyzed in the frequency domain. Figure 8.3 thus displays the spectrum of fluctuations in the measured CSR power signal. Without external excitation, the naturally occurring micro-bunching dynamics already result in a distinct peak at the micro-structure frequency, here roughly at $f_{\mathrm{ms}} = 18.45\,\mathrm{kHz}$. Yet, the RF amplitude modulation at that frequency, $f_{\mathrm{mod}} = f_{\mathrm{ms}}$ and $V_{\mathrm{mod}} = 0.05\,V_0$, amplifies that peak significantly. Compared to the natural micro-bunching dynamics, the spectral intensity of $f_{\mathrm{ms}}$ is increased by more

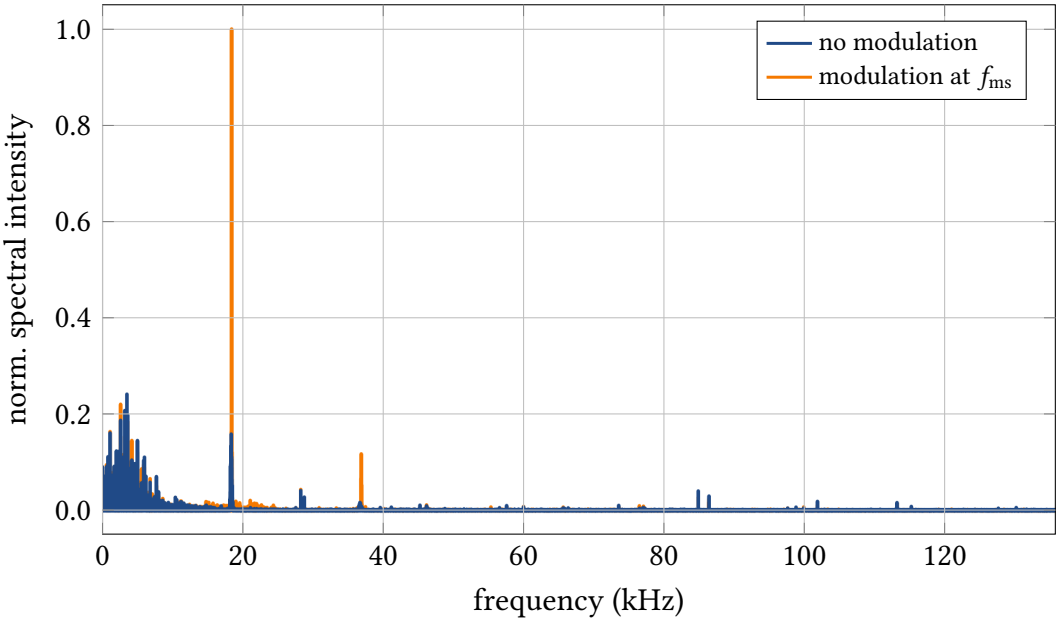



Figure 8.3.: Frequency distribution of the fluctuations in the measured CSR power signal. Besides low-frequency noise up to 10 kHz, the natural spectrum mainly features a distinct peak at the micro-structure frequency $f_{ms} = 18.45$ kHz. The applied RF amplitude modulation amplifies the spectral intensity of that frequency by more than a factor of five. The data may partially show oscillations of the arrival time at the THz detector. Data: Courtesy of Miriam Brosi.

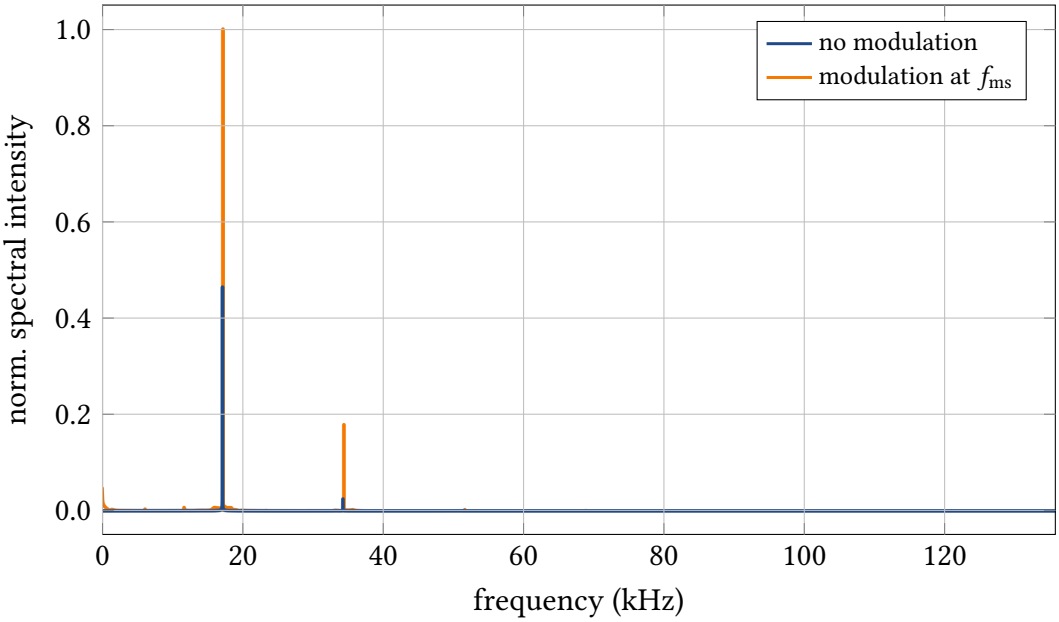



Figure 8.4.: Simulated fluctuation spectrum corresponding to the measurements in Fig. 8.3. The amplified peak intensity at $f_{ms}$ corresponds to an increased amplitude of the occurring micro-structures as illustrated in section 7.2.

than a factor of five. Also visible is an excitation of the second harmonic at 36.9 kHz. In the corresponding Inovesa simulations, shown in Fig. 8.4, the micro-structure frequency is found at a slightly different value, $f_{\mathrm{ms}} = 17.21\,\mathrm{kHz}$. The effect of the RF amplitude modulation on the spectrum, however, is similar to that in the measurements. While the peak intensity at $f_{\mathrm{ms}}$ is only increased by a factor of two, it is also accompanied by an excitation of the second harmonic at 34.42 kHz. Given the availability of the corresponding charge distribution in the longitudinal phase space in simulations, these oscillations of the CSR power signal can directly be attributed to an increased amplitude of the occurring micro-structures, as illustrated in section 7.2. The measurements in Fig. 8.3 are thus also interpreted as the result of an excitation of the underlying micro-bunching dynamics. Thereby, they qualitatively confirm the expected interaction with the electron beam, albeit the exact numerical values differ between simulation and measurement.

In order to prepare for first tests of the RL-based mitigation of the micro-bunching dynamics, the exact machine configuration used for the simulations in section 7.3 and detailed in appendix A.1, data set $\mathcal{D}_1$, is reproduced at KARA. By setting the amplitude of the RF voltage to the simulated value of $V_0 = 1.0\,\mathrm{MV}$ and adjusting the momentum compaction factor $\alpha_{\mathrm{c}}$ via the quadrupole magnets, the synchrotron frequency is adjusted to match the simulations as closely as possible. Thereby, the synchrotron frequency is

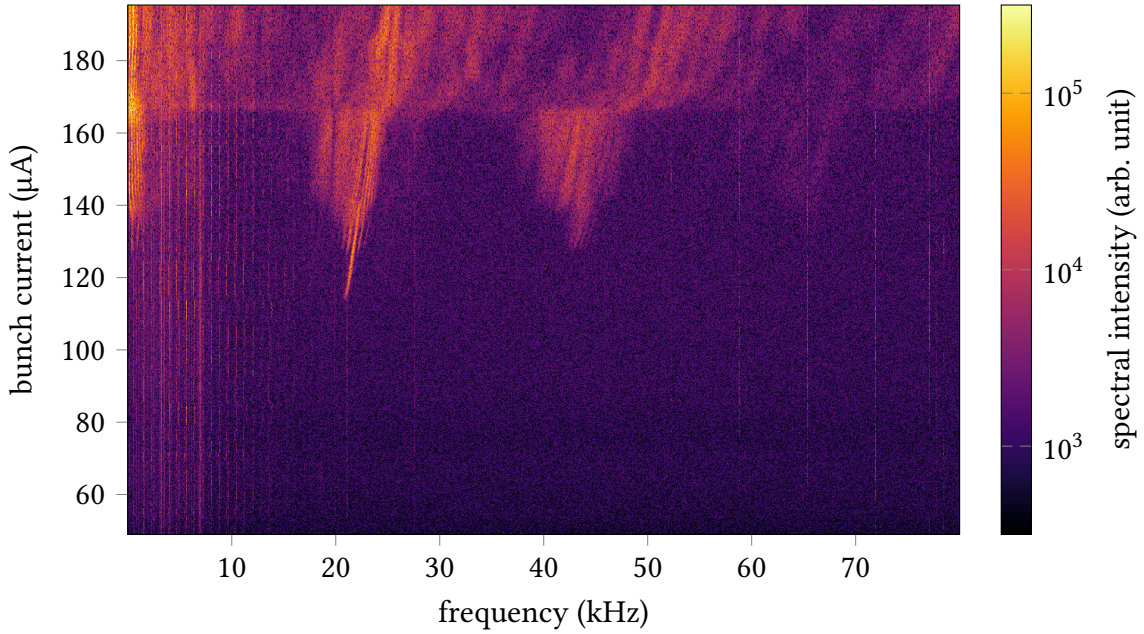



Figure 8.5.: CSR power spectrogram measured at KARA for the same configuration of machine parameters that is used for the simulations in section 7.3 (detailed in appendix A.1, data set $\mathcal{D}_1$). Compared to the CSR power spectrogram generated from simulation data, shown in Fig. 3.6, the observed micro-bunching dynamics are in high qualitative agreement. Several characteristic features, such as the threshold current at $I_{\mathrm{th}} = 114\,\mu\mathrm{A}$ and the micro-structure frequency at $f_{\mathrm{ms}} = 20.95\,\mathrm{kHz}$, also match quantitatively. Data: Courtesy of Miriam Brosi.

measured via the BBB spectrum and compared to the nominal synchrotron frequency, $f_{s,0} = 7.0\,\mathrm{kHz}$. To compare the natural beam dynamics in this particular configuration to the simulations conducted with Inovesa, the emitted CSR power signal is measured in a single bunch operation across different bunch currents. Aggregating these measurements, Fig. 8.5 displays the corresponding CSR power spectrogram. In comparison, the simulated CSR power spectrogram shown in Fig. 3.6 matches the measurements quite closely. The threshold current, $I_{th} = 114\,\mu\mathrm{A}$, and the micro-structure frequency, $f_{ms} = 20.95\,\mathrm{kHz}$, are found at very similar values in both, simulation and measurement. To achieve optimal comparability, first tests of the RL-based mitigation of the micro-bunching dynamics should be conducted in the current range directly above the instability threshold.

## 8.4. Future Steps

The implementation of the general feedback scheme at KARA is accompanied by a range of technical difficulties to reach the required repetition rates. Yet, given the specially designed electronics and the implementation of the RL agent on an FPGA, these challenges can be met in practice. Wherever feasible, the involved components were tested and found to meet the requirements of the RL-based feedback loop. Although, in theory, the use of the PPO algorithm is preferred due its advantages regarding stability, the performances of the four different RL algorithms tested on simulation data are quite comparable and the existing FPGA implementation of the DDPG algorithm is thus expected to allow for first proof-of-principle experiments. Beyond the technical feasibility, the transition from the simulation environment to the actual storage ring will also involve a range of challenges regarding generalization and robustness of the used algorithms. In principle, the performance of the RL agents trained in simulations may directly be tested at KARA. Yet, given the still existing differences between the virtual and the practical environment, the agents presumably have to be trained directly by interacting with the accelerator. Using the pre-trained neural networks may still be beneficial though, as it may result in a warm start and a faster learning process. An additional concern is the robustness of the RL algorithms against different sources of noise that can be expected at KARA. As mentioned in section 7.4, the most prominent sources of noise can be modeled in the virtual environment and thus tested in preparatory simulations. Ultimately though, the practical feasibility of the feedback scheme can only be fully demonstrated at the actual storage ring. The high level of agreement between the Inovesa simulations and the measurements at KARA constitutes a promising starting point to reproduce the results of section 7.3 in practice.

While an excitation of the micro-bunching dynamics and the underlying interaction of an RF amplitude modulation with the electron bunch were demonstrated in first experiments, further studies are required to explore the limitations of this approach. As the possibility to apply an RF amplitude modulation with constant modulation amplitude and frequency is already provided by the KARA control system, these studies are not reliant on the completion of the feedback loop described in section 8.1. In principle, the available diagnostics offer the possibility to directly study the effect of the RF modulations on the emitted CSR spectrum. Such an analysis would provide the opportunity of tailoring the

operation of the storage ring to applications with a requirement of intense CSR in the frequency ranges corresponding to the spatial extent of the occurring micro-structures. Moreover, a systematic study across different machine parameters, such as the RF voltage or the momentum compaction factor, may yield additional insights into the formation process of the occurring micro-structures and its dependencies. While an RF amplitude modulation at the micro-structure frequency should generally result in an excitation of the natural micro-bunching dynamics, the possibility to imprint new structures on the charge distribution, as described in section 7.2, offers a range of further opportunities to manipulate the longitudinal beam dynamics and the corresponding emission of CSR. It is thus considered a promising subject for further research.

# 9. Summary and Outlook

The operation of storage rings with short electron bunches leads to the emission of intense coherent synchrotron radiation which can be provided to a large range of experiments. Yet, the self-interaction of these bunches with their own emitted CSR in the bending magnets of the storage ring leads to complex longitudinal dynamics and, at high enough bunch currents, to the formation of dynamically varying micro-structures in the longitudinal charge distribution. Referred to as the micro-bunching instability throughout this thesis, the overall phenomenon poses a critical limitation to the operation of electron storage rings with high bunch currents. The overarching objective of this thesis was thus to identify an avenue towards control of the occurring micro-bunching dynamics. As the presence of these micro-structures leads to an increased emission of radiation at frequencies corresponding to the spatial extent of the structure, the benefits of extensive control over these dynamics are twofold. A deliberate and controlled excitation of the occurring micro-structures can enhance the emitted radiation in a narrow frequency range and thereby improve the conditions for applications with a demand for intense CSR at these frequencies. On the other hand, practical mitigation of the CSR-induced micro-bunching dynamics extends the regime of stable operation to shorter bunches and higher bunch currents. Besides improving the operating conditions at existing machines, it also allows for a more effective optimization of related beam properties and thereby faciliates the design of new facilities.

In order to better understand the formation process of the occurring micro-structures, chapter 5 focused on an analysis of the synchrotron motion under the influence of CSR self-interaction. Taking the perspective of a single particle, the CSR-induced wake potential was found to cause a position-dependent perturbation of the restoring force exerted by the RF system. Below the instability threshold, this leads to a quadrupole-like modulation of the longitudinal charge distribution. By introducing a higher frequency component to the bunch profile, this amplifies the CSR self-interaction and may thereby act as a seeding mechanism for the micro-bunching instability. Above the instability threshold, the varying perturbation of the restoring force repetitively drives particles to larger deviations from the synchronous position. As the particles cause an excess of charge at these positions, they create local charge modulations and thereby form the occurring micro-structures. Given that the entire process is largely driven by the perturbation of the restoring force, this naturally motivates the application of an RF amplitude modulation. By amplifying the CSR-induced perturbation, the applied modulation can be used to excite the micro-bunching dynamics, that is, to create larger micro-structures in the longitudinal charge distribution. By counteracting the perturbation and thereby recovering the strength of the restoring force, the micro-bunching dynamics can instead be mitigated. Yet, as the interaction with the beam also alters the CSR-induced perturbation, the RF amplitude modulation has to be continuously adjusted according to the evolving charge distribution and the

corresponding wake potential. As the task of identifying these adjustments constitutes a sequential decision problem, it motivates the application of reinforcement learning methods pursued in the subsequent chapters. Regardless of whether or not RL methods are employed, the modulation of the RF amplitude represents a very effective tool to influence the micro-structure formation process. The identification of this approach towards control of the micro-bunching dynamics occurring in electron storage rings is considered a major result of this thesis.

In collaboration with the H2T group at KIT and based on the experience with the micro-bunching instability acquired at KARA, the obtained sequential decision problem has been formulated as a formal RL problem in chapter 6. Thereby, the general problem is split into two complementary formulations. In the more theoretical approach, the agent is given access to the full Markovian states of system, including the charge distribution in the longitudinal phase space, which can be obtained in the simulations using Inovesa. While the information is very difficult to obtain at an actual storage ring, this served the purpose of verifying the general feasibility of the pursued approach. In a second formulation, tailored towards a practical implementation at KARA, the information provided to the agent is restricted to the observed CSR power signal, from which an eight-dimensional feature vector is derived.

Both excitation and mitigation of the CSR-driven micro-bunching dynamics were successfully demonstrated on simulation data in chapter 7. Besides the amplification of the naturally occurring micro-structures through an RF amplitude modulation at the micro-structure frequency, an additional option to imprint a new set of micro-structures on the beam was discovered. Using a modulation frequency close to the third harmonic of the nominal synchrotron frequency, the generated micro-structures reached an amplitude which was more than three times larger than that of the naturally occurring micro-structures. As this leads to substantial changes in the radiated CSR power spectrum, most importantly an increase of the radiated power in the frequency range corresponding to the spatial extent of the micro-structures, it is considered a promising option to tailor the emission of CSR to individual experiments. The practically more challenging task of mitigating the occurring micro-bunching dynamics was initially tested by manually selecting the required actions. While this verified the solvability of the task set for the RL agents, it also served as a performance benchmark for subsequent studies. A first milestone was reached when the first RL agent exceeded the total return obtained under manual control. Given access to the full Markovian states of the system, the PPO agent managed to gradually reduce the micro-structure amplitude and the corresponding oscillation of the CSR power signal over a time frame of 20 synchrotron periods. Unexpectedly, these results were even exceeded by an RL agent trained on the eight-dimensional CSR feature vector. Beyond an overall improved performance and a higher total return, the PPO agent used in this case managed to extend its control to twice the number of steps as defined by the benchmark scenario. After 40 synchrotron periods, the episode was only terminated because the a priori defined maximum episode length was reached. Overall, these results not only verify the feasibility of mitigating the micro-bunching dynamics via a dynamically adjusted RF amplitude modulation, but also the effectiveness of RL methods to solve the corresponding sequential decision problem. Yet, one major difficulty encountered throughout these studies were instabilities in the agent's learning process. Instead of gradually improving with continued

training time, the agents frequently drop to the baseline level of performance, in some cases not recovering for the remainder of the training session. A detailed analysis of the reasons for these instabilities was beyond the scope of this thesis, but is considered an important task in preparing for future implementations of the proposed RL-based feedback scheme.

The main challenge for an implementation of the feedback loop at KARA is reaching the required repetition rates. As the micro-structure formation process happens at at a fraction of the synchrotron period, effective counteraction is bound to the same time scale. In order to reach repetition rates below the ms-level, specially designed electronics developed by the IPE at KIT are incorporated into the hardware implementation of the feedback scheme. With an implementation of the DDPG agent on an FPGA, the mandatory time constraints for inference of the actor network could be met. While a full training step still takes longer than the time window available between consecutive actions, this is not a strict limitation for the feasibility of the pursued approach. By restricting the training process to updates at every $n$ steps the time constraints can still be met, albeit at the cost of slowing down the learning process. Ultimately though, the implementation of the complete feedback loop was beyond the scope of this thesis and the RL-based mitigation of the micro-bunching dynamics could thus not be tested in practice. However, the expected interaction of an RF amplitude modulation with the micro-structure formation process could still be verified for an excitation of the micro-bunching dynamics. Furthermore, in preparation for first tests of the RL-based feedback, the exact configuration of machine parameters used for the simulation studies was reproduced at KARA. The micro-bunching dynamics observed in this configuration were generally found to be in high qualitative and quantitative agreement with the corresponding simulations, providing a solid basis for future efforts to reproduce the achieved results in practice.

While the work summarized in this thesis expands the understanding of the longitudinal dynamics underlying the micro-bunching instability and offers an effective approach to control the formation of micro-structures, it simultaneously raises a range of further questions. The interplay between the oscillation of individual particles and the collectively generated micro-structures, as well as the briefly mentioned dependence on shielding thereof is one such example that warrants further studies. As the provided analysis mostly focuses on the dynamics directly above the threshold, an extension to higher beam currents and the additional weak instability occurring under specific operating conditions may also yield further insights. Regarding the overarching objective of gaining control over the occurring micro-bunching dynamics, there may also be alternative approaches to identifying the required dynamic adjustments of the applied RF signal. While reinforcement learning methods have proven very effective in solving the corresponding sequential decision problem, a deeper understanding of the underlying dynamics may allow for different formulations of the task. Additional insights may also be gained by a careful analysis of the different features constituting the eight-dimensional CSR feature vector and their respective relevance for the agent's decisions. Focusing in particular on the relative phase between the oscillation of the CSR power signal and the applied RF modulation, it may eventually be possible to derive simpler heuristics to solve the underlying control problem. Given the complex dynamics created by the CSR self-interaction across different bunch currents, this is certainly not a straightforward task, and reinforcement learning

methods may thus still be beneficial in supporting the overall process. With the rapid progress being made within the field of reinforcement learning, there may also be new, more suitable algorithms available for this particular task over the course of the upcoming years. The stability of the learning process and the robustness against noise are important criteria for selecting such algorithms. Finally, given the generality of the approach pursued in this thesis, it may also be applicable or transferable to control problems at particle accelerators in and outside the domain of longitudinal beam dynamics.

# A. Appendix

## A.1. Simulation Settings

The Inovesa simulations conducted in the context of this thesis yield to major sets of data, $\mathcal{D}_1$ and $\mathcal{D}_2$, that result from the different parameter settings listed in Table A.1. With the exception of a scan across different vacuum gaps in $\mathcal{D}_2$, these parameters correspond to attainable machine configurations of the storage ring KARA in its short-bunch operation mode. While there is no general preference for one or the other, the respective values of the accelerating voltage yield a slightly different behavior of the micro-bunching instability and some of its characteristic features.

Table A.1.: Simulation parameters used to generate the data discussed in this thesis. The data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ mainly differ in the values used for the accelerating voltage and the vacuum gap.

| Parameter | Unit | $\mathcal{D}_1$ | $\mathcal{D}_2$ |
|---|---|---|---|
| amplitude of RF voltage $V_0$ | MV | 1.0 | 0.6 |
| beam energy $E$ | GeV | 1.3 | 1.3 |
| beam energy spread $\sigma_{\delta,0}$ | | $4.7 \times 10^{-4}$ | $4.7 \times 10^{-4}$ |
| bending radius $R$ | m | 5.559 | 5.559 |
| bunch current $I$ | µA | $1 - 200$ | $1 - 1000$ |
| longitudinal damping time $\tau_\mathrm{d}$ | ms | 10.4 | 10.4 |
| harmonic number $h$ | | 184 | 184 |
| initial charge distribution $\rho(z, E, t_0)$ | | Gaussian | Gaussian |
| revolution frequency $f_\mathrm{rev}$ | MHz | 2.7 | 2.7 |
| nominal synchrotron frequency $f_{\mathrm{s},0}$ | kHz | 7.0 | 7.0 |
| vacuum gap $g$ | mm | 32 (KARA) | $16 - 32$ |

## A.2. CSR Power Spectrogram with Logarithmic Frequency Axis

As the low frequency contributions corresponding to the slow repetition rate of the CSR bursts in the sawtooth bursting regime are difficult to identify in Fig. 3.6, the same data is shown with a logarithmically scaled frequency axis in Fig. A.1. Here, the contributions in the order of 1 kHz are clearly visible. For the exemplary bunch current of $I = 185\,\mu\mathrm{A}$

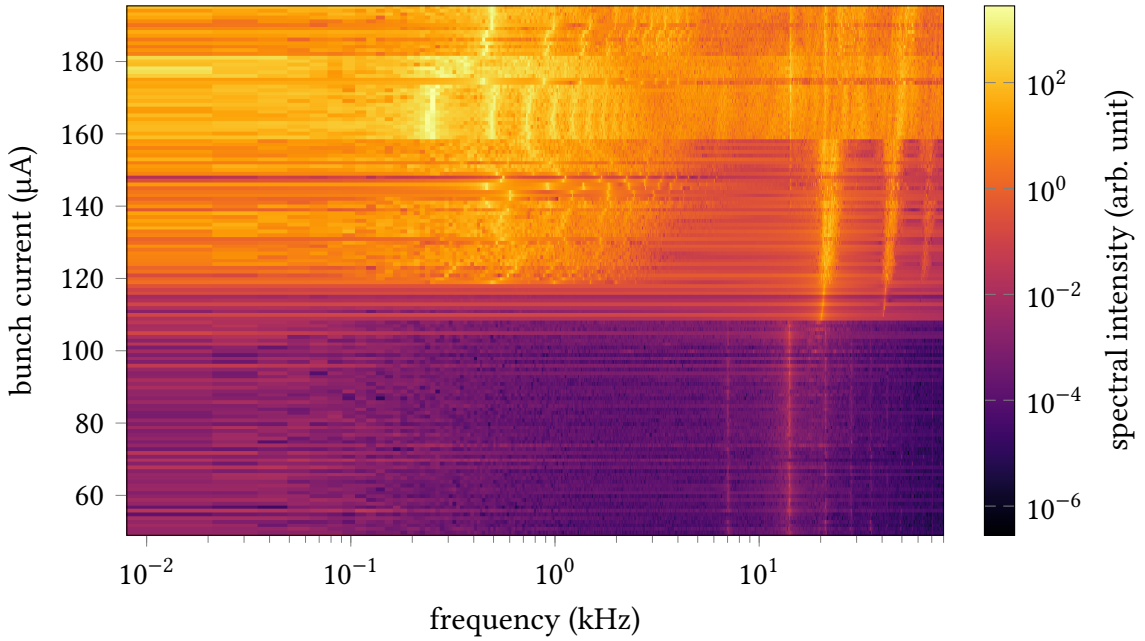



Figure A.1.: Shown is the same CSR Power Spectrogram as in Fig. 3.6 but with a logarithmic frequency axis, highlighting the low frequency contributions emerging around $I = 120\,\mu A$.

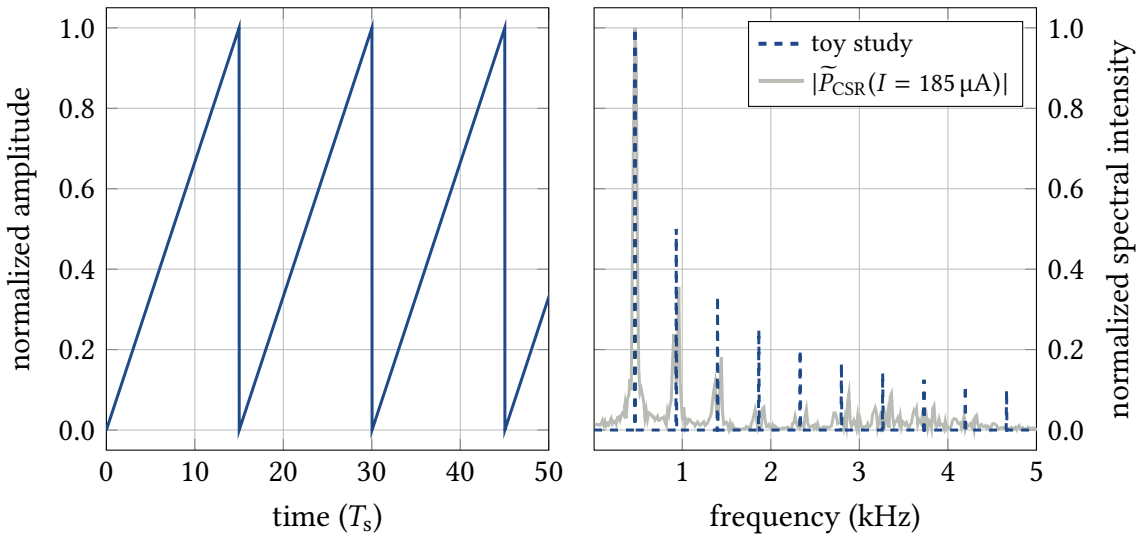
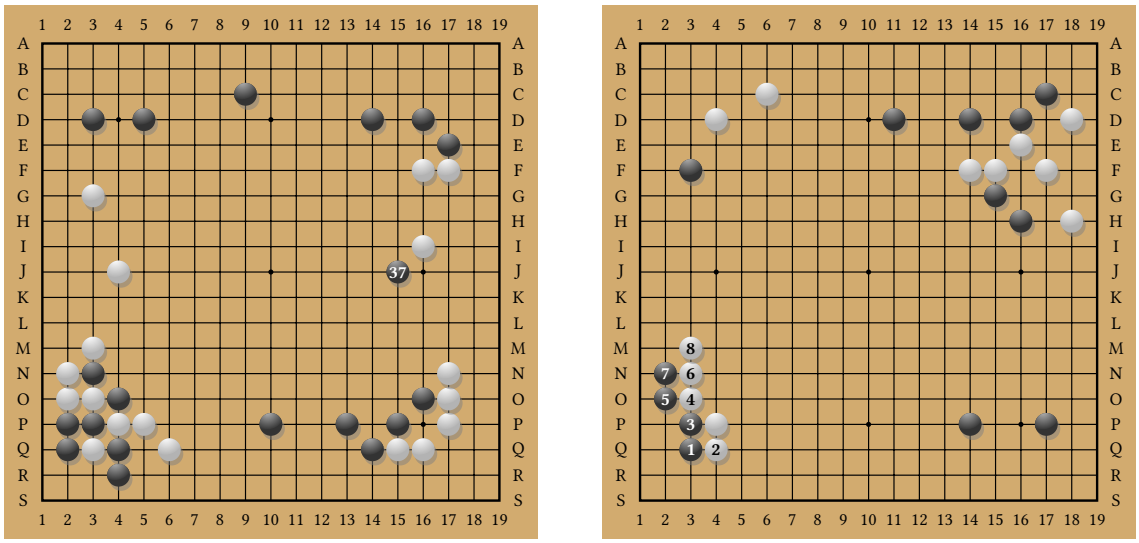



Figure A.2.: Toy study to illustrate the Fourier transform of a sawtooth wave, explaining the higher harmonics of $f_{burst}$ in Fig. A.1. The pure sawtooth wave on the left approximates the CSR power signal in Fig. 3.9. Its Fourier transform shown on the right is composed of the sawtooth repetition rate and its higher harmonics. The gray line depicts the magnitude of the Fourier transform of the CSR power signal at the bunch current $I = 185\,\mu A$, displaying similar characteristics.

discussed in section 3.4.2, the repetition rate of the bursts corresponds to the dominant frequency around $f_{\text{burst}} = 0.47\,\text{kHz}$ while the remaining lines are higher harmonics thereof. These additional harmonics are expected in a Fourier transformed sawtooth wave as illustrated in Fig. A.2. At higher currents, the CSR bursts develop a sharper peak structure, which leads to a multitude of contributing frequencies between $I = 160\,\mu\text{A}$ and $I = 180\,\mu\text{A}$.

## A.3. AlphaGo and the Black Box Issue

In the five-game Go match between AlphaGo and Lee Sedol, who was at the time considered to be one of the strongest human Go players, AlphaGo won all but the fourth game (AlphaGo 4 – 1 Lee Sedol). While the RL-based program won game one convincingly, its capability to deviate from traditional Go theory was more clearly demonstrated by a move in the second game. AlphaGo's move 37 of that game, shown in Fig. A.3a, was described by professional commentator Michael Redmond as "unique" and "creative". Surprised by this, Lee Sedol took an unusually long time to respond to the move. Although it was at the time suspected to be a mistake by the program, AlphaGo subsequently managed to win the game and, in retrospect, move 37 became one of the early signs for AlphaGo's innovative play. This unique, innovative perspective on the game was also found in some of the opening sequences played by AlphaGo, one example being the so-called early 3-3 invasion shown in Fig. A.3b. Defying conventional Go theory, AlphaGo regularly found success with this opening, which led to professional players studying and trying the sequence in their own games. In an attempt to describe the underlying reasoning, a book on the subject was published merely two years later [104].




(a) AlphaGo's move 37

(b) AlphaGo's early 3x3 invasion

Figure A.3.: (a) During the second match against Lee Sedol, AlphaGo made a surprising move 37, which may be seen as an early indication of AlphaGo's innovative play style. (b) The early 3-3 invasion played by AlphaGo defied conventional Go theory and let to a reconsideration of the sequence in professional play.

Overall, AlphaGo's innovative play and its impact on professional Go is an example of humans learning from a black box type system. Although AlphaGo does not provide any reasoning for its decisions, studying the various lines of play advances modern Go theory and helps professional players to reach a higher level of play. Instead of looking at problems like opening sequences with all its possible variations, they may study the choices made by AlphaGo and their implications. This notion of studying a solution to improve the understanding of a problem may also be transferable to RL applications in physics. For the task pursued in this thesis, that is, achieving control over the CSR-induced micro-bunching dynamics, this may imply studying the agent's actions to derive simpler heuristics or even an analytical description for appropriate control signals. It may also simply contribute to the understanding of the underlying longitudinal beam dynamics.

## A.4. Frequency Component of Micro-Structures

The frequency component corresponding to the formation of micro-structures differs from that of the quadrupole-like mode identified in section 5.2. As marked by the red rectangle in Fig. A.4, the micro-structures forming above the threshold current $I_{\mathrm{th}} = 260\,\mu\mathrm{A}$ mainly correspond to frequency contributions around 150 GHz.

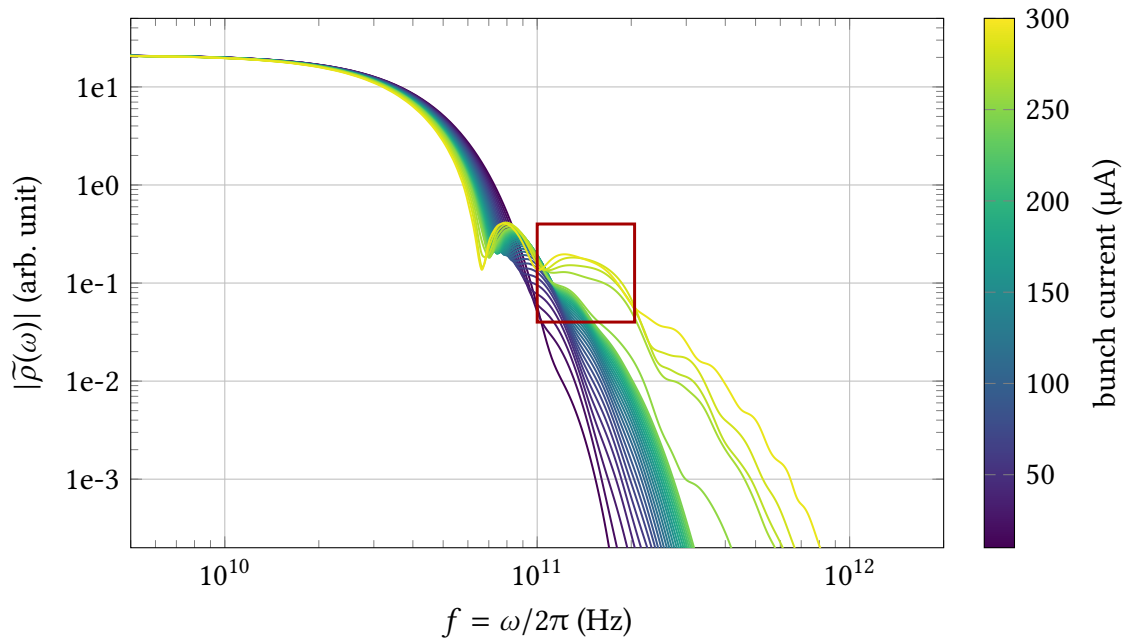



Figure A.4.: Magnitude of the averaged Fourier transformed bunch profile $|\widetilde{\rho}(\omega)|$ for a range of bunch currents. The peak at roughly 85 GHz corresponds to the quadrupole-like modulation already forming below the instability threshold of $I_{\mathrm{th}} = 260\,\mu\mathrm{A}$. The red rectangle marks the additional main frequency component that corresponds to the micro-structures forming in the longitudinal charge distribution.

## A.5. Origin of Particles forming the Micro-Structures

In contrast to the formation process indicated by Fig. 5.10 in subsection 5.3.2, individual particles may instead stay in their respective structure for different parameter settings or bunch currents. Figure A.5 shows one such example. While the formation process is changing over time at this higher current located in the sawtooth bursting regime, for the time frame displayed in Fig. A.5, the distributions clearly show a partial overlap. This means a major part of the particles contributing to the formation of the micro-structures stays within their respective structure for at least one synchrotron period. Although one might expect this behavior to correspond to the growth of the micro-structures in amplitude in the sawtooth bursting regime, this could not be confirmed in further studies across different bunch currents.

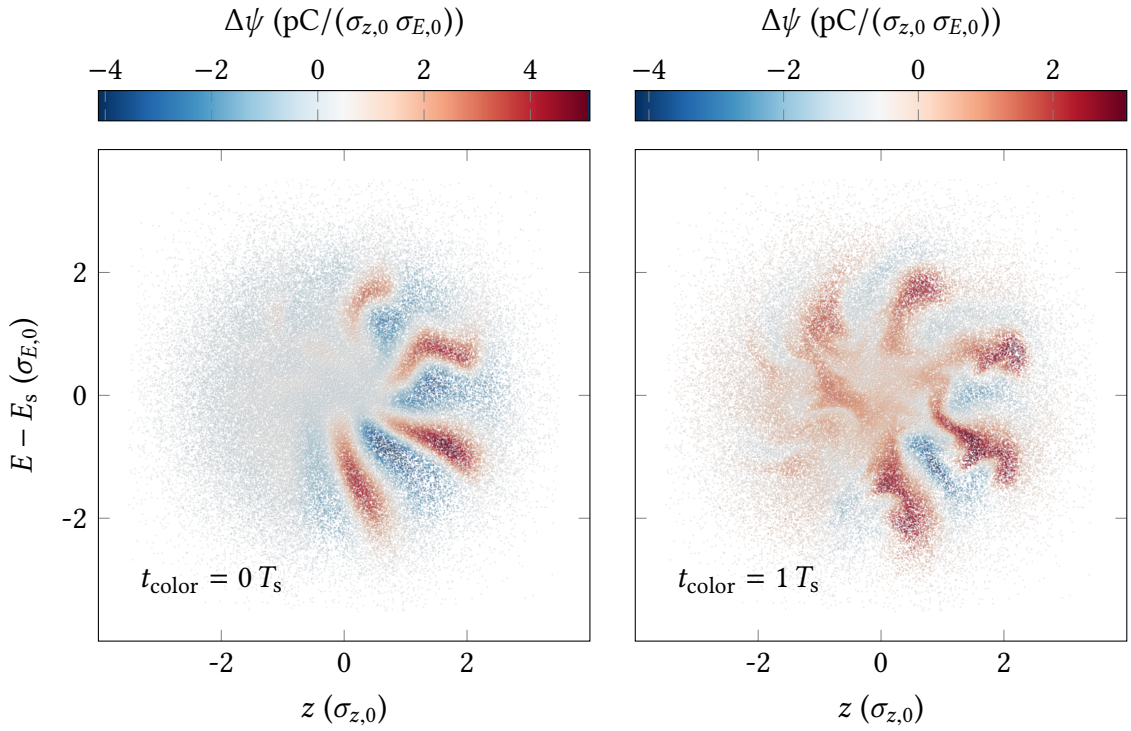



Figure A.5.: Analogously to Fig. 5.10, the particle distribution at time step $t = 0\,T_s$ is depicted with the color assignment $\Delta\psi(q, p, t_{\mathrm{color}} = 0\,T_s)$ (left) and $\Delta\psi(q, p, t_{\mathrm{color}} = 1\,T_s)$ (right) for the bunch current $I = 500\,\mu\mathrm{A}$. In contrast to the lower bunch current, the distributions show a partial overlap, indicating that particles stay within their respective structure over at least one synchrotron period.

## A.6. Bunch Length during Mitigation of Micro-Bunching Dynamics

The mitigation of the micro-bunching dynamics demonstrated in the section 7.3 changes the distribution of charge in the longitudinal phase space. It is thus important to verify that the mitigation is not achieved at the cost of or due to an increased bunch length. Figure A.6 therefore displays the bunch length corresponding to the manual control illustrated in Fig. 7.8 over the same time period. Owing to the reduction of the micro-structures in amplitude, the initial oscillation of the bunch length is continuously damped as the charge density in phase space reaches a smoother distribution. Simultaneously, the average bunch length is even slightly decreased by the applied RF modulation. This is a consequence of the improved focusing towards the position of the synchronous particle, which is achieved by partially recovering the restoring force. As a shorter bunch generally leads to an increased emission of CSR, this effect is already indicated in Fig. 7.8 by the slightly increasing average CSR power at the end of the displayed time frame.

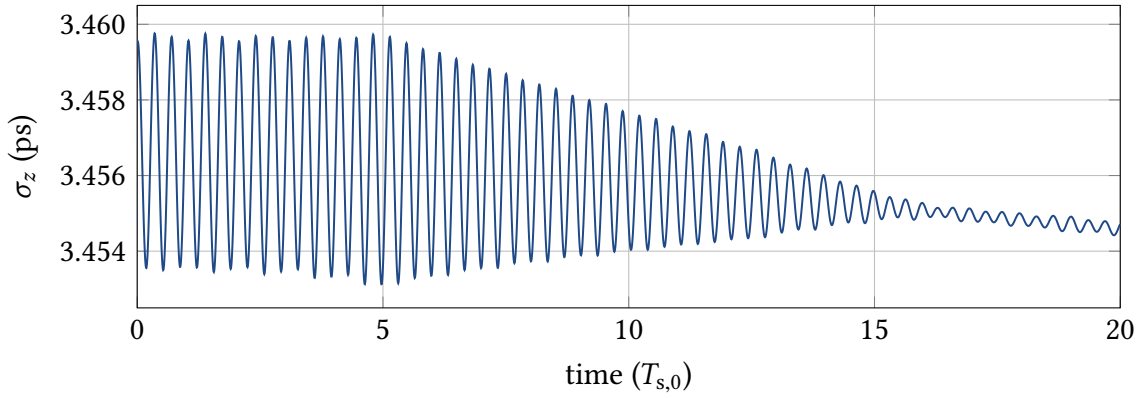



Figure A.6.: Evolution of the bunch length for the sequence displayed in Fig. 7.8. Both, the oscillation and the average value of the bunch length are decreased by the applied modulation of the RF amplitude.

# List of Figures

# Bibliography


[1] Lightsources.org Collaboration. *Lightsource research on SARS-CoV-2*. 2021. URL: http://www.lightsources.org/ (visited on 09/10/2021).

[2] K. Wille. *The Physics of Particle Accelerators: An Introduction*. 1st. Clarendon Press, 2001. ISBN: 978-0-19-850549-5.

[3] H. Wiedemann. *Particle Accelerator Physics*. 4th. Springer, 2015. ISBN: 978-3-319-18316-9. DOI: 10.1007/978-3-319-18317-6.

[4] F. R. Elder et al. "Radiation from Electrons in a Synchrotron". In: *Phys. Rev.* 71 (11 1947), pp. 829–830. DOI: 10.1103/PhysRev.71.829.5.

[5] E. M. Rowe and F. E. Mills. "Tantalus I: A Dedicated Storage Ring Synchrotron Radiation Source". In: *Particle Accelerators* 4 (1973), pp. 211–227. URL: https://cds.cern.ch/record/1107919/files/p211.pdf.

[6] M. Sands. "The Physics of Electron Storage Rings: An Introduction". In: *Conf. Proc. C* 6906161 (1969), pp. 257–411. URL: http://slac.stanford.edu/pubs/slacreports/reports02/slac-r-121.pdf.

[7] A. W. Chao. *Physics of collective beam instabilities in high-energy accelerators*. Wiley, 1993. ISBN: 978-0-471-55184-3. URL: https://www.slac.stanford.edu/~achao/wileybook.html.

[8] K. L. F. Bane, Y. Cai, and G. Stupakov. "Threshold studies of the microwave instability in electron storage rings". In: *Phys. Rev. ST Accel. Beams* 13 (10 Oct. 2010), p. 104402. DOI: 10.1103/PhysRevSTAB.13.104402.

[9] J. B. Murphy, S. Krinsky, and R. L. Gluckstern. "Longitudinal Wake Field for An Electron Moving on A Circular orbit". In: *Part. Accel.* 57 (Apr. 1996), pp. 9–64. URL: https://cds.cern.ch/record/1120287.

[10] Y. Cai. "Theory of Microwave Instability and Coherent Synchrotron Radiation in Electron Storage Rings". In: *Proc. 2nd Int. Particle Accelerator Conf. (IPAC'11)*. (San Sebastian, Spain). JACoW Publishing, pp. 3774–3778. URL: https://accelconf.web.cern.ch/IPAC2011/papers/frxaa01.pdf.

[11] P. Schönfeldt et al. "Parallelized Vlasov-Fokker-Planck solver for desktop personal computers". In: *Phys. Rev. Accel. Beams* 20 (3 Mar. 2017), p. 030704. DOI: 10.1103/PhysRevAccelBeams.20.030704. URL: https://github.com/Inovesa/Inovesa.

[12] R. L. Warnock and J. A. Ellison. *A General Method for Propagation of the Phase Space Distribution, with Application to the Sawtooth Instability*. SLAC Technical Report No. SLAC-PUB-8404. 2000.

[13] P. Schönfeldt. "Simulation and measurement of the dynamics of ultra-short electron bunch profiles for the generation of coherent THz radiation". 54.01.01; LK 01. PhD thesis. Karlsruhe Institute of Technology (KIT), 2018. 142 pp. DOI: 10.5445/IR/1000084466.

[14] J. L. Steinmann et al. "Continuous bunch-by-bunch spectroscopic investigation of the microbunching instability". In: *Phys. Rev. Accel. Beams* 21 (11 Nov. 2018), p. 110705. DOI: 10.1103/PhysRevAccelBeams.21.110705.

[15] P. Krejcik et al. "High Intensity Bunch Length Instabilities in the SLC Damping Rings". In: *Proc. 15th Particle Accelerator Conf. (PAC'93)*. (Washington D.C., USA, May 17–20, 1993). Vol. 5. JACoW Publishing, 1993, pp. 3240–3242. DOI: 10.1109/PAC.1993.309612.

[16] M. Abo-Bakr et al. "Steady-State Far-Infrared Coherent Synchrotron Radiation detected at BESSY II". In: *Phys. Rev. Lett.* 88 (25 June 2002), p. 254801. DOI: 10.1103/PhysRevLett.88.254801.

[17] A. R. Hight Walker et al. "New infrared beamline at the NIST SURF II storage ring". In: *Accelerator-Based Infrared Sources and Applications*. Ed. by Gwyn P. Williams and Paul Dumas. Vol. 3153. International Society for Optics and Photonics. SPIE, 1997, pp. 42–50. DOI: 10.1117/12.290261.

[18] A. Andersson, M. S. Johnson, and B. Nelander. "Coherent synchrotron radiation in the far-infrared from a 1 mm electron bunch". In: *Optical Engineering* 39 (Dec. 2000), pp. 3099–3105. DOI: 10.1117/1.1327498.

[19] G. L. Carr et al. "Observation of coherent synchrotron radiation from the NSLS VUV ring". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 463.1 (2001), pp. 387–392. ISSN: 0168-9002. DOI: 10.1016/S0168-9002(01)00521-6.

[20] U. Arp et al. "Spontaneous coherent microwave emission and the sawtooth instability in a compact storage ring". In: *Phys. Rev. ST Accel. Beams* 4 (5 May 2001), p. 054401. DOI: 10.1103/PhysRevSTAB.4.054401.

[21] J. M. Byrd et al. "Observation of Broadband Self-Amplified Spontaneous Coherent Terahertz Synchrotron Radiation in a Storage Ring". In: *Phys. Rev. Lett.* 89 (22 Nov. 2002), p. 224801. DOI: 10.1103/PhysRevLett.89.224801.

[22] F. Wang et al. "Coherent THz Synchrotron Radiation from a Storage Ring with High-Frequency RF System". In: *Phys. Rev. Lett.* 96 (6 Feb. 2006), p. 064801. DOI: 10.1103/PhysRevLett.96.064801.

[23] A. Mochihashi et al. "Observation of THz Synchrotron Radiation Burst in UVSOR-II Electron Storage Ring". In: *Proc. 10th European Particle Accelerator Conf. (EPAC'06), Edinburgh, UK, Jun. 2006*. (Edinburgh, UK). European Particle Accelerator Conference 10. Geneva, Switzerland: JACoW Publishing, June 2006, pp. 3380–3382. URL: http://accelconf.web.cern.ch/AccelConf/e06/PAPERS/THPLS042.PDF.

[24] E. Karantzoulis et al. "Characterization of coherent THz radiation bursting regime at Elettra". In: *Infrared Physics & Technology* 53.4 (2010), pp. 300–303. ISSN: 1350-4495. DOI: 10.1016/j.infrared.2010.04.006.

[25] J. Feikes et al. "Metrology Light Source: The first electron storage ring optimized for generating coherent THz radiation". In: *Phys. Rev. ST Accel. Beams* 14 (3 Mar. 2011), p. 030705. DOI: 10.1103/PhysRevSTAB.14.030705.

[26] A.-S. Müller. "Accelerator-Based Sources of Infrared and Terahertz Radiation". In: *Reviews of Accelerator Science and Technology*. 2011, pp. 165–183. DOI: 10.1142/9789814340397_0009.

[27] C. Evain et al. "Spatio-temporal dynamics of relativistic electron bunches during the micro-bunching instability in storage rings". In: *EPL (Europhysics Letters)* 98.4 (May 2012), p. 40006. DOI: 10.1209/0295-5075/98/40006.

[28] W. Shields et al. "Microbunch Instability Observations from a THz Detector at Diamond Light Source". In: *Journal of Physics: Conference Series* 357 (May 2012), p. 012037. DOI: 10.1088/1742-6596/357/1/012037.

[29] B. E. Billinghurst et al. "Longitudinal bunch dynamics study with coherent synchrotron radiation". In: *Phys. Rev. Accel. Beams* 19 (2 Feb. 2016), p. 020704. DOI: 10.1103/PhysRevAccelBeams.19.020704.

[30] G. Stupakov and S. Heifets. "Beam instability and microbunching due to coherent synchrotron radiation". In: *Phys. Rev. ST Accel. Beams* 5 (5 May 2002), p. 054402. DOI: 10.1103/PhysRevSTAB.5.054402.

[31] G. Stupakov and R. Warnock. *Microbunch instability theory and simulations*. Tech. rep. Stanford Linear Accelerator Center (SLAC), Menlo Park, CA, May 2005. URL: https://slac.stanford.edu/pubs/slacpubs/10750/slac-pub-10997.pdf.

[32] M. Brosi. "In-Depth Analysis of the Micro-Bunching Characteristics in Single and Multi-Bunch Operation at KARA". PhD thesis. Karlsruhe Institute of Technology (KIT), 2020. 198 pp. DOI: 10.5445/IR/1000120018.

[33] P. Kuske. "CSR-driven Longitudinal Single Bunch Instability Thresholds". In: *Proc. 4th Int. Particle Accelerator Conf. (IPAC'13)*. (Shanghai, China). JACoW Publishing, pp. 2041–2043. URL: http://accelconf.web.cern.ch/IPAC2013/papers/weoab102.pdf.

[34] M. Brosi et al. "Fast mapping of terahertz bursting thresholds and characteristics at synchrotron light sources". In: *Phys. Rev. Accel. Beams* 19 (11 Nov. 2016), p. 110701. DOI: 10.1103/PhysRevAccelBeams.19.110701.

[35] M. Brosi et al. "Systematic studies of the microbunching instability at very low bunch charges". In: *Phys. Rev. Accel. Beams* 22 (2 Feb. 2019), p. 020701. DOI: 10.1103/PhysRevAccelBeams.22.020701.

[36] K. L. F. Bane, K. Oide, and M. Zobov. "Impedance Calculation and Verification in Storage Rings". In: *Proc. 1st CARE-HHH-APD Workshop on Beam Dynamics in Future Hadron Colliders and Rapidly Cycling High-Intensity Synchrotrons (HHH 2004)*. (Geneva, Switzerland). CERN, 2005, pp. 143–157. URL: https://cds.cern.ch/record/925928/files/p143.pdf.

[37] T. Boltz. "Comprehensive Analysis of Micro-Structure Dynamics in Longitudinal Electron Bunch Profiles". MA thesis. Karlsruhe Institute of Technology (KIT), 2017. 77 pp. DOI: 10.5445/IR/1000068253.

[38] M. Migliorati, E. Métral, and M. Zobov. "Review of impedance-induced instabilities and their possible mitigation techniques". In: *Proceedings of the ICFA mini-Workshop on Mitigation of Coherent Beam Instabilities in Particle Accelerators (MCBI 2019)*. (Zermatt, Switzerland, Sept. 23–27, 2019). Vol. 9. CERN Yellow Reports. CERN, Dec. 2020, pp. 1–8. DOI: 10.23732/CYRCP-2020-009.1.

[39] Y. Shoji and T. Takahashi. "Coherent Synchrotron Radiation Burst from Electron Storage Ring under External RF Modulation". In: *Proc. 11th European Particle Accelerator Conf. (EPAC'08)*. (Genoa, Italy). JACoW Publishing, June 2008, pp. 178–180. URL: http://accelconf.web.cern.ch/e08/papers/mopc048.pdf.

[40] J. L. Steinmann. "Diagnostics of Short Electron Bunches with THz Detectors in Particle Accelerators". PhD thesis. Karlsruhe Institute of Technology (KIT), 2019. 226 pp. ISBN: 978-3-7315-0889-2. DOI: 10.5445/KSP/1000090017.

[41] J. L. Steinmann et al. "Increasing the Single-Bunch Instability Threshold by Bunch Splitting Due to RF Phase Modulation". In: *Proc. 12th Int. Particle Accelerator Conf. (IPAC'21)*. (Campinas, Brazil). JACoW Publishing, Aug. 2021, pp. 3193–3196. DOI: 10.18429/JACoW-IPAC2021-WEPAB240.

[42] P. Schreiber et al. "Status of Operation With Negative Momentum Compaction at KARA". In: *Proc. 10th Int. Particle Accelerator Conf. (IPAC'19)*. (Melbourne, Australia). JACoW Publishing, May 2019, pp. 878–881. DOI: 10.18429/JACoW-IPAC2019-MOPTS017.

[43] P. Schreiber et al. "Effect of Negative Momentum Compaction Operation on the Current-Dependent Bunch Length". In: *Proc. 12th Int. Particle Accelerator Conf. (IPAC'21)*. (Campinas, Brazil). JACoW Publishing, Aug. 2021, pp. 2786–2789. DOI: 10.18429/JACoW-IPAC2021-WEPAB083.

[44] C. Evain et al. "Stable coherent terahertz synchrotron radiation from controlled relativistic electron bunches". In: *Nature Physics* 15 (Apr. 2019), pp. 635–639. DOI: 10.1038/s41567-019-0488-6.

[45] G. Carleo et al. "Machine learning and the physical sciences". In: *Rev. Mod. Phys.* 91 (4 Dec. 2019), p. 045002. DOI: 10.1103/RevModPhys.91.045002.

[46] A. Edelen et al. *Opportunities in Machine Learning for Particle Accelerators*. 2018. arXiv: 1811.03172 [physics.acc-ph].

[47] J. Han, J. Pei, and M. Kamber. *Data Mining: Concepts and Techniques*. 3rd. Morgan Kaufmann, 2011. ISBN: 978-0-12-381479-1.

[48] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. 2nd. Cambridge, MA, USA: MIT Press, 2018. URL: https://incompleteideas.net/book/the-book.html.

[49] S. Levine. *Deep Reinforcement Learning*. Lecture Series at UC Berkeley (CS 285). Accessed: 2021–06-09. 2020. URL: http://rail.eecs.berkeley.edu/deeprlcourse/static/slides/lec-1.pdf.

[50] A. L. Samuel. "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM Journal of Research and Development* 3.3 (1959), pp. 210–229. DOI: 10.1147/rd.33.0210.

[51] G. Tesauro. "Practical Issues in Temporal Difference Learning". In: *Machine Learning* 8 (3 May 1992), pp. 257–277. DOI: 10.1023/A:1022624705476.

[52] D. Silver et al. "Mastering the game of Go with deep neural networks and tree search". In: *Nature* 529 (7587 Jan. 2016), pp. 484–489. DOI: 10.1038/nature16961.

[53] V. Mnih et al. "Human-level control through deep reinforcement learning". In: *Nature* 518 (7540 Feb. 2015), pp. 529–533. DOI: 10.1038/nature14236.

[54] D. Silver et al. "Mastering the game of Go without human knowledge". In: *Nature* 550 (7676 Jan. 2017), pp. 354–359. DOI: 10.1038/nature24270.

[55] D. Silver et al. "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play". In: *Science* 362.6419 (2018), pp. 1140–1144. ISSN: 0036-8075. DOI: 10.1126/science.aar6404.

[56] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. URL: http://www.deeplearningbook.org.

[57] T. P. Lillicrap et al. *Continuous control with deep reinforcement learning*. Sept. 2015. arXiv: 1509.02971 [cs.LG].

[58] D. Silver et al. "Deterministic Policy Gradient Algorithms". In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32. Proceedings of Machine Learning Research 1. Bejing, China: PMLR, June 2014, pp. 387–395. URL: http://proceedings.mlr.press/v32/silver14.pdf.

[59] G. E. Uhlenbeck and L. S. Ornstein. "On the Theory of the Brownian Motion". In: *Phys. Rev.* 36 (5 Sept. 1930), pp. 823–841. DOI: 10.1103/PhysRev.36.823.

[60] S. Fujimoto, H. van Hoof, and D. Meger. *Addressing Function Approximation Error in Actor-Critic Methods*. Feb. 2018. arXiv: 1802.09477 [cs.AI].

[61] G. Brockman et al. *OpenAI Gym*. June 2016. arXiv: 1606.01540 [cs.LG].

[62] T. Haarnoja et al. *Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor*. Jan. 2018. arXiv: 1801.01290 [cs.LG].

[63] John Schulman et al. *Proximal Policy Optimization Algorithms*. July 2017. arXiv: 1707.06347 [cs.LG].

[64] J. Schulman et al. *Trust Region Policy Optimization*. Feb. 2015. arXiv: 1502.05477 [cs.LG].

[65] A. L. Edelen et al. "Neural Networks for Modeling and Control of Particle Accelerators". In: *IEEE Transactions on Nuclear Science* 63.2 (2016), pp. 878–897. DOI: 10.1109/TNS.2016.2543203.

[66] J. H. Hanten et al. "Enhancement of the S-DALINAC Control System with Machine Learning Methods". In: *Proc. 8th Int. Beam Instrumentation Conf. (IBIC'19)*. (Malmö, Sweden). JACoW Publishing, Sept. 2019, pp. 475–478. DOI: 10.18429/JACoW-IBIC2019-WEB004.

[67] L. Vera Ramirez et al. "Adding Machine Learning to the Analysis and Optimization Toolsets at the Light Source BESSY II". In: *Proc. 17th Int. Conf. on Accelerator and Large Experimental Physics Control Systems (ICALEPCS'19)*. (New York, NY, USA). JACoW Publishing, Oct. 2019, pp. 754–760. URL: https://accelconf.web.cern.ch/icalepcs2019/papers/tucpl01.pdf.

[68] X. Pang, S. Thulasidasan, and L. Rybarcyk. *Autonomous Control of a Particle Accelerator using Deep Reinforcement Learning*. Oct. 2020. arXiv: 2010.08141 [cs.AI].

[69] J. St. John et al. *Real-time Artificial Intelligence for Accelerator Control: A Study at the Fermilab Booster*. Nov. 2020. arXiv: 2011.07371 [physics.acc-ph].

[70] V. Kain et al. "Sample-efficient reinforcement learning for CERN accelerator control". In: *Phys. Rev. Accel. Beams* 23 (12 Dec. 2020), p. 124801. DOI: 10.1103/PhysRevAccelBeams.23.124801.

[71] N. Bruchon et al. "Basic Reinforcement Learning Techniques to Control the Intensity of a Seeded Free-Electron Laser". In: *Electronics* 9.5 (2020). ISSN: 2079-9292. DOI: 10.3390/electronics9050781.

[72] M. W. McIntire et al. "Bayesian Optimization of FEL Performance at LCLS". In: *Proc. 7th Int. Particle Accelerator Conf. (IPAC'16)*. (Busan, Korea). JACoW Publishing, May 2016, pp. 2972–2975. DOI: doi:10.18429/JACoW-IPAC2016-WEPOW055.

[73] J. Kirschner et al. "Adaptive and Safe Bayesian Optimization in High Dimensions via One-Dimensional Subspaces". In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97. Proceedings of Machine Learning Research. PMLR, June 2019, pp. 3429–3438. URL: http://proceedings.mlr.press/v97/kirschner19a/kirschner19a.pdf.

[74] J. Duris et al. "Bayesian Optimization of a Free-Electron Laser". In: *Phys. Rev. Lett.* 124 (12 Mar. 2020), p. 124801. DOI: 10.1103/PhysRevLett.124.124801.

[75] R. J. Shalloo et al. "Automation and control of laser wakefield accelerators using Bayesian optimization". In: *Nature Communications* 11 (1 Dec. 2020), p. 6355. DOI: 10.1038/s41467-020-20245-6.

[76] S. Jalas et al. "Bayesian Optimization of a Laser-Plasma Accelerator". In: *Phys. Rev. Lett.* 126 (10 Mar. 2021), p. 104801. DOI: 10.1103/PhysRevLett.126.104801.

[77] R. Roussel, A. Hanuka, and A. Edelen. "Multiobjective Bayesian optimization for online accelerator tuning". In: *Phys. Rev. Accel. Beams* 24 (6 June 2021), p. 062801. DOI: 10.1103/PhysRevAccelBeams.24.062801.

[78] A. Eichler et al. “First Steps Toward an Autonomous Accelerator, a Common Project Between DESY and KIT”. In: *Proc. 12th Int. Particle Accelerator Conf. (IPAC’21)*. (Campinas, Brazil). JACoW Publishing, Aug. 2021, pp. 2182–2185. DOI: 10.18429/JACoW-IPAC2021-TUPAB298.

[79] T. Boltz et al. “Perturbation of Synchrotron Motion in the Micro-Bunching Instability”. In: *Proc. 10th Int. Particle Accelerator Conf. (IPAC’19)*. (Melbourne, Australia). JACoW Publishing, May 2019, pp. 108–111. DOI: 10.18429/JACoW-IPAC2019-MOPGW018.

[80] T. Boltz et al. “Perturbation of Synchrotron Motion in the Micro-Bunching Instability”. In: *Phys. Rev. Accel. Beams – manuscript submitted for publication* (2021).

[81] P. Schönfeldt et al. “Elaborated Modeling of Synchrotron Motion in Vlasov-Fokker-Planck Solvers”. In: *Proc. 9th Int. Particle Accelerator Conf. (IPAC’18)*. (Vancouver, Canada). JACoW Publishing, Apr. 2018, pp. 3283–3286. DOI: 10.18429/JACoW-IPAC2018-THPAK032.

[82] T. K. Charles, D. M. Paganin, and R. T. Dowd. “Caustic-based approach to understanding bunching dynamics and current spike formation in particle bunches”. In: *Phys. Rev. Accel. Beams* 19 (10 Oct. 2016), p. 104402. DOI: 10.1103/PhysRevAccelBeams.19.104402.

[83] T. Boltz et al. “Studies of Longitudinal Dynamics in the Micro-Bunching Instability Using Machine Learning”. In: *Proc. 9th Int. Particle Accelerator Conf. (IPAC’18)*. (Vancouver, Canada). JACoW Publishing, Apr. 2018, pp. 3277–3279. DOI: 10.18429/JACoW-IPAC2018-THPAK030.

[84] B. V. Podobedov and R. H. Siemann. “Saw-Tooth Instability Studies in the Stanford Linear Collider Damping Rings”. In: *Proc. 17th Particle Accelerator Conf. (PAC’97)*. (Vancouver, Canada). JACoW Publishing, May 1997. URL: https://accelconf.web.cern.ch/pac97/papers/pdf/2V019.PDF.

[85] M. Brosi et al. “Studies of the Micro-Bunching Instability in the Presence of a Damping Wiggler”. In: *Proc. 9th Int. Particle Accelerator Conf. (IPAC’18)*. (Vancouver, Canada). JACoW Publishing, Apr. 2018, pp. 3273–3276. DOI: 10.18429/JACoW-IPAC2018-THPAK029.

[86] T. Boltz et al. “Feedback Design for Control of the Micro-Bunching Instability based on Reinforcement Learning”. In: *Proc. 10th Int. Particle Accelerator Conf. (IPAC’19)*. (Melbourne, Australia). JACoW Publishing, May 2019, pp. 104–107. DOI: doi:10.18429/JACoW-IPAC2019-MOPGW017.

[87] S. Funkner et al. *Revealing the dynamics of ultrarelativistic non-equilibrium many-electron systems with phase space tomography*. 2019. arXiv: 1912.01323 [physics.acc-ph].

[88] S. Funkner et al. “High throughput data streaming of individual longitudinal electron bunch profiles”. In: *Phys. Rev. Accel. Beams* 22 (2 Feb. 2019), p. 022801. DOI: 10.1103/PhysRevAccelBeams.22.022801.

[89] B. Kehrer. "Time-resolved studies of the micro-bunching instability at KARA". PhD thesis. Karlsruhe Institute of Technology (KIT), 2019. 143 pp. DOI: 10.5445/IR/1000098584.

[90] M. Caselle et al. "A Picosecond Sampling Electronic "KAPTURE" for Terahertz Synchrotron Radiation". In: *Proc. 3rd Int. Beam Instrumentation Conf. (IBIC'14)*. (Monterey, CA, USA). JACoW Publishing, Sept. 2014, pp. 24–28. URL: http://accelconf.web.cern.ch/IBIC2014/papers/moczb1.pdf.

[91] The HDF Group. *Hierarchical data format 5: HDF5*. 1998. URL: http://www.hdfgroup.org/HDF5.

[92] M. Plappert. *keras-rl*. https://github.com/keras-rl/keras-rl. 2016.

[93] A. Hill et al. *Stable Baselines*. https://github.com/hill-a/stable-baselines. 2018.

[94] S. Guadarrama et al. *TF-Agents: A library for Reinforcement Learning in TensorFlow*. https://github.com/tensorflow/agents. 2018.

[95] M. Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2015. URL: https://www.tensorflow.org/.

[96] D. Merkel. "Docker: Lightweight Linux Containers for Consistent Development and Deployment". In: *Linux Journal* 2014.239 (Mar. 2014), p. 5.

[97] T. Boltz et al. "Excitation of Micro-Bunching in Short Electron Bunches Using RF Amplitude Modulation". In: *Proc. 12th Int. Particle Accelerator Conf. (IPAC'21)*. (Campinas, Brazil). JACoW Publishing, Aug. 2021, pp. 3173–3176. DOI: 10.18429/JACoW-IPAC2021-WEPAB233.

[98] M. Klein. "Reinforcement Learning for Micro-Bunching Control in Particle Accelerators". MA thesis. Karlsruhe Institute of Technology (KIT), 2020.

[99] J. Schestag. "Study of Noise in the Simulation of Collective Effects of Relativistic Electron Bunches". Karlsruhe Institute of Technology (KIT), 2018. 38 pp. DOI: 10.5445/IR/1000083507.

[100] M. Caselle et al. "KAPTURE-2. A picosecond sampling system for individual THz pulses with high repetition rate". In: *Journal of Instrumentation* 12.01 (Jan. 2017), pp. C01040–C01040. DOI: 10.1088/1748-0221/12/01/c01040.

[101] W. Wang. "Towards Intelligent Data Acquisition Systems with Embedded Deep Learning on MPSoC". PhD thesis. Karlsruhe Institute of Technology (KIT), 2021. 166 pp. DOI: 10.5445/IR/1000133898.

[102] W. Wang et al. "Accelerated Deep Reinforcement Learning for Fast Feedback of Beam Dynamics at KARA". In: *IEEE Transactions on Nuclear Science* 68.8 (2021), pp. 1794–1800. DOI: 10.1109/TNS.2021.3084515.

[103] T. Knuth et al. "Longitudinal and Transverse Feedback Kickers for the BESSY II Storage Ring". In: *Proc. 18th Particle Accelerator Conf. (PAC'99)*. (New York, NY, USA). JACoW Publishing, Mar. 1999. URL: http://accelconf.web.cern.ch/p99/PAPERS/TUA33.PDF.

[104] Y. Zhou. *Playing AlphaGo's Early 3-3 Invasion*. Independently published, Nov. 2018. ISBN: 978-1731135292.

# Acknowledgements

During the time this thesis was conducted, I was very fortunate to be supported by many kind and capable people. They made working on this project an easy and truly enjoyable task for which I am eternally grateful.

First and foremost, I would like to thank Prof. Dr. Anke-Susanne Müller for providing me with the opportunity to conduct this thesis. Her support of novel ideas and interdisciplinary collaborations is what made this project possible in the first place. The most intriguing and enjoyable phase of my work was when I was piecing together various concepts of accelerator physics and reinforcement learning.

On the same note, I would like to thank Prof. Dr. Tamim Asfour for engaging in this cross-disciplinary subject and for kindly taking over the task of being the second reviewer for this thesis.

Over the time of working on this subject, I was very lucky to be supported not by one, but by two wonderful advisors who were always on hand with their help and advice. A heartfelt thank you goes to Dr. Minjie Yan for convincing me to engage in this thesis to begin with. Even a distance of more than nine thousand kilometers and a time difference of up to nine hours could not keep her from supporting my work with her insight and invaluable advice. In the same way, I want to thank Dr. Bastian Härer whose gift of separating the essential from the negligible always made sure that I did not get lost in the details of my work, opportunities for which there were many. Without him, this thesis may have never been completed.

I would also like to thank Dr. Erik Bründermann whose advice on my research, but also on scientific institutions, conferences and procedures has proven invaluable at numerous occasions.

Furthermore, I feel grateful for the support of the Helmholtz International Research School for Teratronics (HIRST) and the Karlsruhe School of Elementary Particle and Astroparticle Physics: Science and Technology (KSETA) that I received over the years.

A big thank you goes to all my colleagues in the THz working group at IBPT, which provided a friendly, creative, humorous and very vivid working environment. I would like to thank Dr. Marcel Schuh who provided vital help and tips on dealing with those problems of everyday life that can easily drive you mad otherwise. As the creator of the simulation code Inovesa, I want to thank Dr. Patrik Schönfeldt, who made so much of my work possible in the first place. Moreover, I would like to thank Dr. Miriam Brosi, who thoroughly shares my curiosity about the micro-bunching instability and its underlying

dynamics which led to numerous intriguing discussions over the years. I want to thank Patrick Schreiber for being my personal Python guru and for providing support whenever I ran into issues related to computing. Dr. Benjamin Kehrer, Meghana Patil and my office mate Thiemo Schmelzer I want to thank for their humorous and reassuring comments and advice over the years. They made my everyday work a true pleasure. A special thank you also goes to all the colleagues who supported me in compiling this document by proof-reading early versions and offering their feedback on various parts of this thesis. Furthermore, I would like to thank Dr. Robert Ruprecht for his words of encouragement during my extended writing period in times of the COVID-19 pandemic.

I also want to thank all the people at H2T and IPE who contributed to a very fruitful collaboration. Firstly, I would like to thank Dr. Christian Mandery and Dr. Peter Kaiser for suggesting the use of reinforcement learning methods and for initiating our joint work on micro-bunching control. I want to thank Christoph Pohl for his frequent input of ideas and the shared organization and supervision of a master's thesis on the subject. I would like to thank Melvin Klein for his enthusiasm and persistence in getting the various RL algorithms to work. Furthermore, I want to thank Dr. Weijia Wang and Dr. Michele Caselle who worked tirelessly on making the idea of an RL-based feedback realizable in practice. Without their efforts to develop specially tailored electronics, the implementation of the feedback scheme at KARA would still seem like an unattainable goal.

At last, I want thank my family for their consistent support over all this time. My parents always encouraged me to follow my interests and were there for me when things did not go as planned. Yet, if anyone is to be blamed for my curiosity, my enjoyment of lengthy discussions and my motivation for getting to the bottom of things, it is my twin brother. Thank you for being the way you are!

To all of you and everyone else that contributed to this thesis: Thank you very much!