Moscow Institute of Physics and Technology

and

P.N.Lebedev Physical Institute

Professor Nikolai Kolachevsky

Precision measurements in quantum optics

Lecture scripts supplementary

The goal of the lecture course is to deliver modern experimental and theoretical methods of precision measurements in quantum optics. The main objectives of the course are the following

- description of stochastic processes in oscillatory systems
- discussion of precision methods in astrophysics and space, introduction to General relativity
- detailed presentation of modern methods of laser cooling, discussion of different trapping methods of atoms and ions
- discussion of modern approaches to laser stabilization and optical frequency measurements, optical clocks

---

Moscow, Russia, 2013

Contents

Lecture 1: Introduction

Frequency and time as most accurately measured quantities in physics. Clocks: from 17th century till today. Mechanical, radiofrequency, microwave and optical oscillators. Accuracy and stability. Phase and amplitude modulation, their mathematical representation and power spectrum.

1.1 Frequency and time as most accurately measured quantities in physics

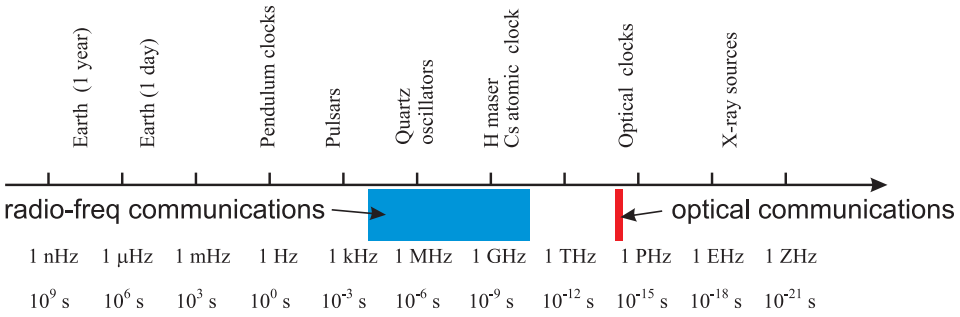


Figure 1.1: Typical frequencies of different oscillator types.

From all known physical quantities, frequency can be measured with the highest accuracy. Today, the accuracy of the frequency measurements reached a few parts in 10^{18}. If someone wants to measure a physical quantity with high accuracy, it is necessary to covert this quantity in frequency.

Examples:
- road radars convert velocity into frequency (the Doppler effect)
- medical tomograph maps the spatial distribution of water containing tissues into frequency spectrum (Nuclear Magnetic Resonance)
- highly accurate measurements of voltages use the Josephson effect: the oscillation frequency from the Josephson junction depends on the potential difference as $U_{DC} = n\frac{\hbar}{2e}\omega$.

For many applications in our life, technology, navigation and fundamental physics stable frequency sources are extremely important. Humans used clocks

from the very beginning of civilization.

First clocks were based on periodicity of day and night and changing seasons. It is directly connected to the rotation of astronomical bodies - Earth, Moon and planets. The rotation period of the Earth around its axis (day), Moon around the Earth (month) and Earth around the Sun (year) were taken as natural units of time. One needs the time scale and the unit of time to discuss events happening our life.

Today the tropical year consists of 365.2422 days and the synodic month – of 29.5306 days. Today's calender bases on Julian (roman) calender accepted in 45 B.C.: the year consists of 365 days, while each 4th year consists of 366 days. The calender was slightly modified in 1582 by pope Gregory. According to this calender the year consists of 365.2425 days which is very close to true number (365.2422 days).

1.2 Clocks: from 17th century till today.

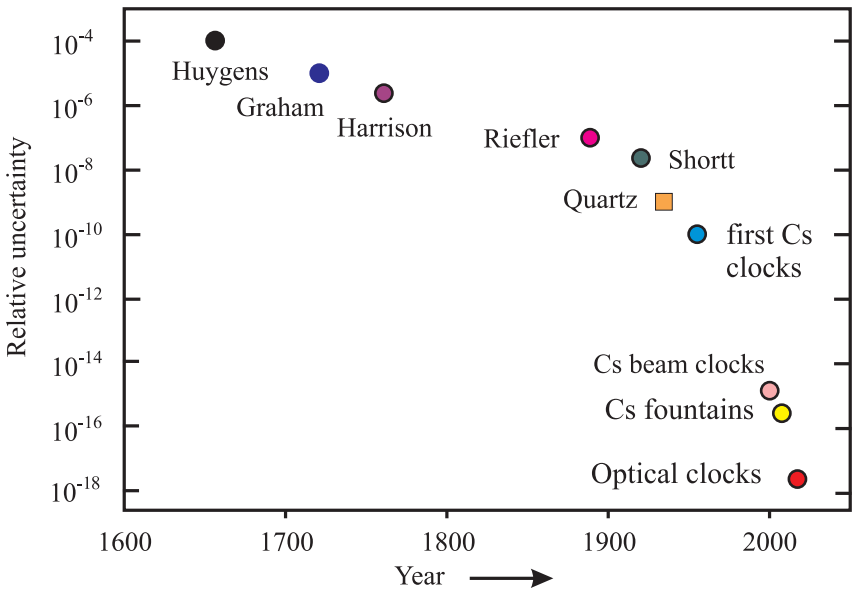


Figure 1.2: Development of clocks over last centuries.

1.2.1 Mechanical clocks

In mechanical clocks the mechanism plays a dual role. It should measure and indicate the frequency of the oscillating system. In addition, it should provide energy to compensate for the losses in the system. It should not influence the frequency of the oscillator! First tower clocks had an accuracy of 15 minutes in a day or $\Delta T/T = \Delta\nu/\nu \approx 10^{-2}$.

Later Galileo Galilei (1564-1642) discovered, that the oscillating period of the pendulum does not depend on amplitude if it is small. He tired to compete in getting prize for "finding the latitude" which was extremely important issue for long-range navigation in the middle of 17th century. Still, first working pendulum clocks were manufactured in 1656 by Chrisitan Huygens. Clocks were accurate to about 10 seconds $\Delta T/T \approx 10^{-4}$. Significant improvement was introduced by George Graham in 1721 who compensated the temperature instability of the pendulum frequency $\Delta T/T \approx 10^{-5}$. Significant breakthrough in the navigation was made by George Harrison (1761) who invented a marine chronometer. The accuracy was 0.2 seconds per day already! $\Delta T/T \approx 10^{-6}$.

Till the early XX century, elaborated mechanical clocks were used in the metrological insinuates. The best mechanical clocks provided instability of $\Delta T/T \approx 2 \times 10^{-8}$ (William Shortt).

1.2.2 Quartz clocks

The beginning of quartz clock era was around 1930. Frequency of quartz oscillator is defined by the piezo-electric oscillations of the elastic quartz crystal. Typical range 100kHz - 10 MHz. Typical frequency drift is 1 ms per day, $\Delta T/T \approx 10^{-8}$. Till 1935 calibration of any clock was done by accurate measurements of astronomical (sun) time.

Later, in 1935 by monitoring the frequency of 3 quartz clock it was shown that the rotation period of the Earth changes (Earth typically decelerates).

Tidal changes of the day duration can be monitored in the past by the *coral growth*. Depending on the season, carbonate concentration in water changes with its temperature and corals structure consists of year rings (like a wood). It was shown that around 135 millon years ago (Jurassic period) the year consisted of 377 days.

1.2.3 Microwave atomic clocks

The main difference between all previous clocks and atomic clocks is that the oscillations there result not from mechanical oscillations of the solid body, but from atomic population oscillations between atomic energy levels. One of the first ideas to use atoms in clocks was given by Isidor Rabi (Nobel Prize, 1944). Cs atomic clocks appeared in the period 1944-1955. There were based on the ideas of Norman Ramsey to excite atoms in spatially separated fields to get narrow unperturbed resonance lines. First commercial Cs clocks – 1958. Development of Cs atomic clocks resulted in re-definition of the second: *the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom.* (CGPM conference, 1967).

The uncertainly of the best beam Cs atomic clock is around $\Delta T/T \approx 10^{-14}$.

Next generation of microwave Cs atomic clocks - "Cs fountain clocks". Atoms are laser cooled and set to ballistic flight for 1 s. It results in spectral line width of the atomic resonance line of 1 Hz. The typical uncertainty is $\Delta T/T \approx 10^{-15}$, the best performance $\Delta T/T = 2 - 3 \times 10^{-16}$.

1.2.4 Optical clocks

We see, that improving the stability of the clocks was connected with increasing the carrier frequency ν_0. The higher the frequency, the higher the stability. Why?

The resonance quality factor is given by

$$Q = \nu_0/\Delta\nu\,, \tag{1.1}$$

where ν_0 is the carrier frequency and $\Delta\nu$ is the resonance spectral width. The higher ν_0, the higher the Q-factor, the higher the stability.

- mechanical: $\nu_0 \sim 1\,\mathrm{Hz}$
- quartz: $\nu_0 \sim 10^7\,\mathrm{Hz}$
- microwave: $\nu_0 \sim 10^{10}\,\mathrm{Hz}$

Further? **Optical!** $\nu_0 \sim 10^{15}\,\mathrm{Hz}$. The resonance line width can still reach small numbers $\Delta\nu \sim 1\,\mathrm{Hz}$, so the Q-factor reaches 10^{15}. What are the advantages?

First, the resonance is narrower and the stability is higher.

Second, if one wants to see the discrepancy between two clocks the faster clocks show it faster. Example: two mechanical clocks, one is slower for 1 second in a year (10^{-8}). To see the difference of half a period (π), one needs to wait half a year. For quartz oscillator with $\nu_0 = 10^8\,\mathrm{Hz}$ one needs only half a second!

Best today's clocks:

- *lattice clocks* on ultra-cold laser-cooled atoms (Sr, Yb, Hg): fractional uncertainty 10^{-17}.
- *ion clocks* on laser cooled single ions (Al^+, Sr^+, Yb^+, Hg^+): fractional uncertainty 5×10^{-18}.

Future projects: atomic clocks based on *nuclear optical transitions*. Candidate ^{229}Th with isomeric transition between two nuclear states. Proposed Q-factor is around $Q \sim 10^{20}$. The transition is not yet detected!

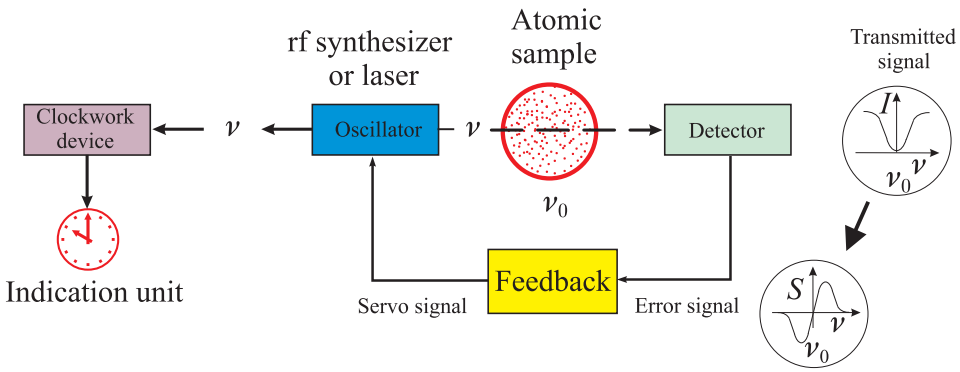


Figure 1.3: Atomic clock schematics.

1.2.5 Accuracy and stability: definition

Frequency is a physical value which fluctuates. To build a good source the frequency should be stable in time. But such a source is not necessarily gives a reproducible value. A high-quality frequency standard should produce (i) highly stable frequency and (ii) this frequency should be known in absolute units (hertz).

Definitions:

- *"Stability"* – frequency of the oscillator is stable in time
- *"Accuracy"* – frequency of the oscillator is reproducible and can be measured in hertz

Good sources: HIGH stability and HIGH accuracy (big numbers) or LOW instability and LOW accuracy (small numbers)

Accuracy of the source cannot be higher than the accuracy of the best primary standards (Cs fountain clock)

If one invents a new oscillator with a superior stability (e.g. optical clocks) better than the primary standard, than one can compare two identical sources to check for reproducibility. It can be later acknowledged as a new *definition* of the second.

1.3 Oscillator. Amplitude and Phase modulation

1.3.1 Harmonic oscillator

Harmonic oscillator is described by the equation

$$U(t) = U_0 \cos(\omega_0 t + \phi)\,. \tag{1.2}$$

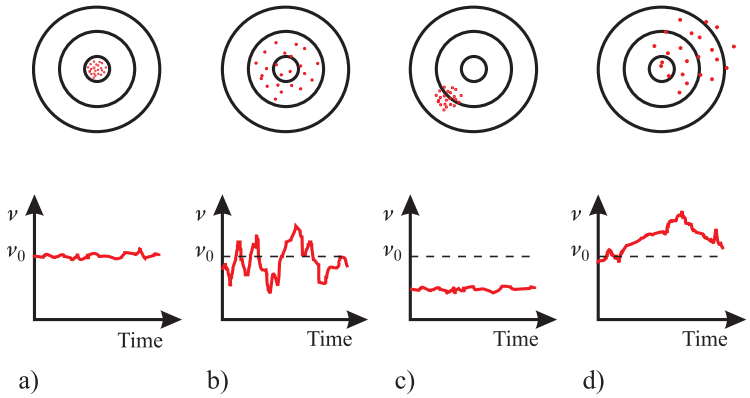


Figure 1.4: Accuracy and stability. a) Accurate and stable signal. b) Accurate and unstable signal. c) Stable, but not accurate signal.

with the amplitude U_0, frequency

$$\nu_0 = \frac{\omega_0}{2\pi} \tag{1.3}$$

and initial phase ϕ.

Let us generalize the equation for harmonic oscillations introducing varying phase and amplitude:

$$U(t) = U_0(t)\,\cos\varphi(t) = [U_0 + \Delta\,U_0(t)]\cos[\omega_0 t + \phi(t)]. \tag{1.4}$$

instant frequency equals

$$\nu(t) \equiv \frac{1}{2\pi}\frac{d\varphi(t)}{dt} = \frac{1}{2\pi}\frac{d}{dt}[2\pi\nu_0 t + \phi(t)] = \nu_0 + \frac{1}{2\pi}\frac{d\phi(t)}{dt} \tag{1.5}$$

which differs from the frequency of the ideal oscillator ν_0 by

$$\Delta\nu(t) \equiv \frac{1}{2\pi}\frac{d\phi(t)}{dt}\,. \tag{1.6}$$

1.3.2 Damped oscillations

Damped oscillations are described by the formula

$$U(t) = U_0\,e^{-\frac{\Gamma}{2}t}\cos\omega_0 t\,, \tag{1.7}$$

where Γ is the damping constant. It is defined by the energy losses of the oscillator per unit time $dW(t) = -\Gamma W(t)\,dt$. The spectrum of damping oscillations is given by Fourier transformation

$$A(\omega) = \int_0^\infty U_0\,e^{-\frac{\Gamma}{2}t}\,\cos(\omega_0 t)\,e^{-i\omega t}\,dt\,, \tag{1.8}$$

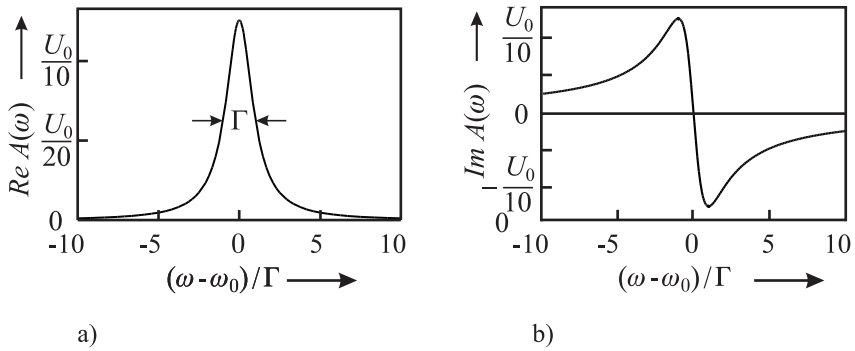


Figure 1.5: Spectrum of damped oscillations.

which will give

$$A(\omega) = \frac{U_0}{2} \frac{-i(\omega-\omega_0)+\frac{\Gamma}{2}}{[i(\omega-\omega_0)+\frac{\Gamma}{2}][-i(\omega-\omega_0)+\frac{\Gamma}{2}]} = \frac{U_0}{2} \frac{-i(\omega-\omega_0)+\frac{\Gamma}{2}}{(\omega-\omega_0)^2+(\frac{\Gamma}{2})^2} \,. \quad (1.9)$$

The spectral function is a complex value with the real and imaginary parts given as the following:

$$\begin{aligned} \Re A(\omega) &= \frac{U_0}{2} \frac{\frac{\Gamma}{2}}{(\omega-\omega_0)^2+(\frac{\Gamma}{2})^2} \\ \Im A(\omega) &= -\frac{U_0}{2} \frac{\omega-\omega_0}{(\omega-\omega_0)^2+(\frac{\Gamma}{2})^2} \,, \end{aligned} \quad (1.10)$$

while the power spectrum is $P(\omega) \propto A(\omega)A^*(\omega) = [\Re A(\omega)]^2 + [\Im A(\omega)]^2$

$$P(\omega) \propto \frac{U_0^2}{4} \frac{(\omega-\omega_0)^2+(\frac{\Gamma}{2})^2}{[(\omega-\omega_0)^2+(\frac{\Gamma}{2})^2]^2} = \frac{U_0^2}{4} \frac{1}{(\omega-\omega_0)^2+(\frac{\Gamma}{2})^2} \,. \quad (1.11)$$

This is the LORENTZIAN function. Full Width On the Half Maximum (FWHM) equals

$$\Delta\omega_{\mathrm{FWHM}} = \Gamma \,. \quad (1.12)$$

Very similar to the *uncertainty relation*

$$\Delta E \Delta t \geq \frac{\hbar}{2} \,. \quad (1.13)$$

Quality factor:

$$Q \equiv \frac{\omega_0 W}{-dW/dt} \,. \quad (1.14)$$

Since $W \propto \overline{U(t)^2} = U_0^2/2 \exp(-\Gamma t)$ and $dW/dt \propto -\Gamma U_0^2/2 \exp(-\Gamma t)$, which gives

$$Q = \frac{\omega_0}{\Gamma} = \frac{\omega_0}{\Delta\omega} \,. \quad (1.15)$$

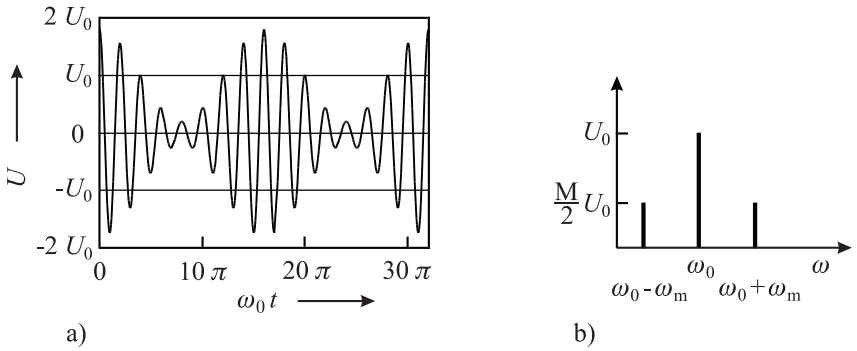


Figure 1.6: Amplitude modulated signal and is spectrum.

1.3.3 Harmonic amplitude modulation

$$\begin{aligned} U_{\mathrm{AM}}(t) &= (U_0 + \Delta U_0 \cos\omega_m t)\cos\omega_0 t \\ &= U_0(1 + M\cos\omega_m t)\cos\omega_0 t\,, \end{aligned} \tag{1.16}$$

where

$$M \equiv \frac{\Delta U_0}{U_0} \tag{1.17}$$

calls the index of the amplitude modulation. Rewriting the formula (1.16) we will get

$$U_{\mathrm{AM}}(t) = U_0\left[\cos\omega_0 t + \frac{M}{2}\cos(\omega_0+\omega_m)t + \frac{M}{2}\cos(\omega_0-\omega_m)t\right]. \tag{1.18}$$

which indicates that the spectrum of amplitude modulated signal consists of three components at the frequencies ω_0 (carrier) and $\omega_0 \pm \omega_m$ (sidebands).

Exercise 1: The laser radiation with amplitude modulated signal at the frequency ω_m and the modulation index M is focused on the photodiode. The spectrum analyzed connected to the photodiode records the signal at the frequency ω_m. What will be the amplitude of this signal?

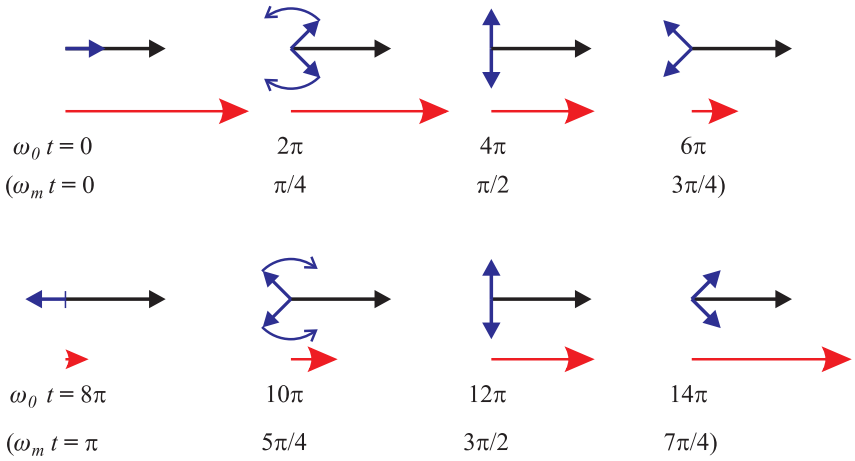


Figure 1.7: Phase plane representation of an amplitude modulated signal.

Solution: The photodiode detects the power of the amplitude modulated signal

$$\begin{aligned} P_{\mathrm{AM}} &\propto U_0 \left[e^{i\omega_0 t} + \frac{M}{2} e^{i(\omega_0+\omega_m)t} + \frac{M}{2} e^{i(\omega_0-\omega_m)t} \right] \\ &\times U_0^* \left[e^{-i\omega_0 t} + \frac{M}{2} e^{-i(\omega_0+\omega_m)t} + \frac{M}{2} e^{-i(\omega_0-\omega_m)t} \right] \\ &= |U_0|^2 \left[1 + 2\frac{M}{2} e^{-i\omega_m t} + 2\frac{M}{2} e^{i\omega_m t} + 2\frac{M^2}{4} + 2\frac{M^2}{4} e^{2i\omega_m t} + 2\frac{M^2}{4} e^{-2i\omega_m t} \right] \\ &= |U_0|^2 \left[1 + \frac{M^2}{2} + 2M\cos\omega_m t + \frac{M^2}{2}\cos(2\omega_m t) \right] . \end{aligned} \tag{1.19}$$

the spectrum analyzer will detect the signal at the frequency ω_m with the amplitude $A_{SA} \propto |U_0|^2 M$.

1.3.4 Harmonic phase modulation

Phase modulated oscillations are described by the expression

$$U_{\mathrm{PM}}(t) = U_0 \cos\varphi = U_0 \cos(\omega_0 t + \delta \cos\omega_m t) . \tag{1.20}$$

Index of phase modulation δ gives the maximal deviation of the phase (hub). The frequency ω_m is the modulation frequency. The instant frequency

$$\omega(t) = \omega_0 - \omega_m \delta \sin\omega_m t \equiv \omega_0 - \Delta\omega \sin\omega_m t . \tag{1.21}$$

Phase and frequency modulation are closely connected and are basically have the same physics. The expression "phase modulation" is used when the coefficient δ does not depend on the modulation frequency ω_m. In this case the

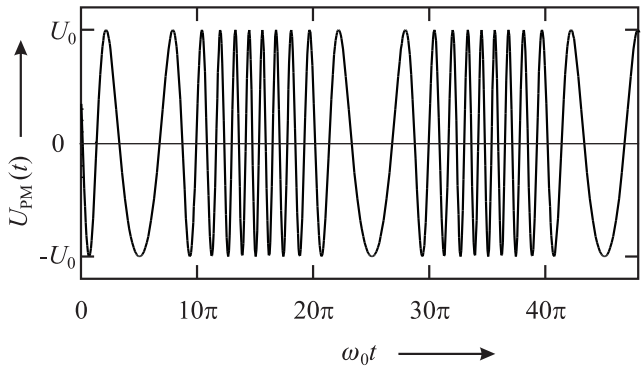


Figure 1.8: Phase modulated signal.

frequency deviation $\Delta\omega$ linearly depends on modulation frequency ω_m. The expression "frequency modulation" is used when the deviation ω_m is constant and the phase deviation δ is reversely proportional to ω_m.

Rewriting the expression into complex form

$$\begin{aligned} U_{\mathrm{PM}}(t) &= U_0 \cos(\omega_0 t + \delta \cos \omega_m t) \\ &= U_0 \Re \left\{ \exp(i\omega_0 t)\, \exp(i\delta \cos \omega_m t) \right\} . \end{aligned} \tag{1.22}$$

We expand the exponent into Teylor series

$$\begin{aligned} &\exp[i\delta \cos(\omega_m t)] \\ &\quad = 1 \;+\; i\delta \cos(\omega_m t) \\ &\qquad +\; i^2 \frac{1}{2!}\, \delta^2 \frac{1}{2} [1 + \cos(2\,\omega_m t)] \\ &\qquad +\; i^3 \frac{1}{3!}\, \delta^3 \frac{1}{4} [3\, \cos(\omega_m t) + \cos(3\,\omega_m t)] \\ &\qquad +\; i^4 \frac{1}{4!}\, \delta^4 \frac{1}{8} [3 + 4\, \cos(2\,\omega_m t) + \cos(4\,\omega_m t)] \\ &\qquad +\; i^5 \frac{1}{5!}\, \delta^5 \frac{1}{16} [10\, \cos(\omega_m t) + 5\, \cos(3\,\omega_m t) + \cos(5\,\omega_m t)] \\ &\qquad +\; i^6 \frac{1}{6!}\, \delta^6 \frac{1}{32} [10 + 15\, \cos(2\,\omega_m t) + 6\, \cos(4\,\omega_m t) + \cos(6\,\omega_m t)] \\ &\qquad +\; i^7 \frac{1}{7!}\, \delta^7 \frac{1}{64} [35\, \cos(\omega_m t) + 21\, \cos(3\,\omega_m t) + 7\, \cos(5\,\omega_m t) + \cos(7\,\omega_m t)] \\ &\qquad +\; \cdots . \end{aligned}$$

and after re-grouping we get

$$\begin{aligned} \exp[i\delta \cos(\omega_m t)] = J_0(\delta) + 2\, i\, J_1(\delta) \cos(\omega_m t) + 2\, i^2\, J_2(\delta) \cos(2\,\omega_m t) \\ + \cdots + 2\, i^n\, J_n(\delta) \cos(n\,\omega_m t) \cdots , \end{aligned} \tag{1.23}$$

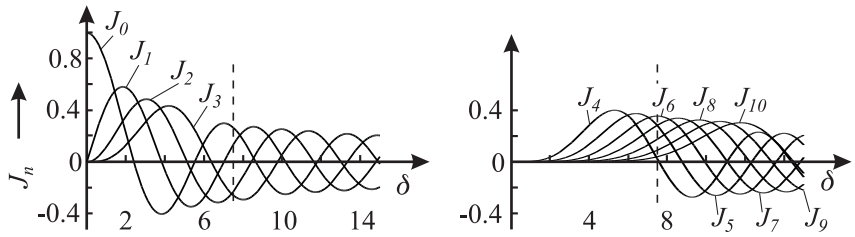


Figure 1.9: Bessel functions.

where J_n are the Bessel functions:

$$\begin{aligned} J_0(\delta) &= 1 - \left(\frac{\delta}{2}\right)^2 + \frac{1}{4}\left(\frac{\delta}{2}\right)^4 - \frac{1}{36}\left(\frac{\delta}{2}\right)^6 + \cdots \\ J_1(\delta) &= \left(\frac{\delta}{2}\right) - \frac{1}{2}\left(\frac{\delta}{2}\right)^3 + \frac{1}{12}\left(\frac{\delta}{2}\right)^5 - \cdots \\ J_2(\delta) &= \frac{1}{2}\left(\frac{\delta}{2}\right)^2 - \frac{1}{6}\left(\frac{\delta}{2}\right)^4 + \frac{1}{48}\left(\frac{\delta}{2}\right)^6 - \cdots \\ J_3(\delta) &= \frac{1}{6}\left(\frac{\delta}{2}\right)^3 + \frac{1}{24}\left(\frac{\delta}{2}\right)^5 + \frac{1}{240}\left(\frac{\delta}{2}\right)^7 - \cdots \end{aligned} \tag{1.24}$$

At the end we will get

$$U_{\mathrm{PM}}(t) = U_0 \sum_{n=-\infty}^{\infty} \Re\{(i)^n J_n(\delta)\, \exp[i(\omega_0 + n\,\omega_m)\, t]\}\,. \tag{1.25}$$

Negative order Bessel functions can be calculated as

$$J_{-n} = (-1)^n J_n\,. \tag{1.26}$$

Explicitly, it will result in

$$\begin{aligned} U_{\mathrm{PM}}(t) = U_0 \Re\{ & J_0(\delta)\, \exp(i\omega_0 t) \\ & + iJ_1(\delta)[\exp\, i(\omega_0 t + \omega_m t) + \exp\, i(\omega_0 t - \omega_m t)] \\ & - J_2(\delta)[\exp\, i(\omega_0 t + 2\omega_m t) + \exp\, i(\omega_0 t - 2\omega_m t)] \\ & - iJ_3(\delta)[\exp\, i(\omega_0 t + 3\omega_m t) + \exp\, i(\omega_0 t - 3\omega_m t)] \\ & + iJ_4(\delta)[\exp\, i(\omega_0 t + 4\omega_m t) + \exp\, i(\omega_0 t - 4\omega_m t)] \\ & + i\cdots\} \\ = U_0\{ & J_0(\delta)\, \cos\omega_0 t \\ & - J_1(\delta)\sin(\omega_0 t + \omega_m t) - J_1(\delta)\sin(\omega_0 t - \omega_m t) \\ & - J_2(\delta)\sin(\omega_0 t + 2\omega_m t) - J_2(\delta)\sin(\omega_0 t - 2\omega_m t) \\ & + J_3(\delta)\sin(\omega_0 t + 3\omega_m t) + J_3(\delta)\sin(\omega_0 t - 3\omega_m t) \\ & + J_4(\delta)\sin(\omega_0 t + 4\omega_m t) + J_4(\delta)\sin(\omega_0 t - 4\omega_m t) \\ & - \cdots\}\,, \end{aligned} \tag{1.27}$$

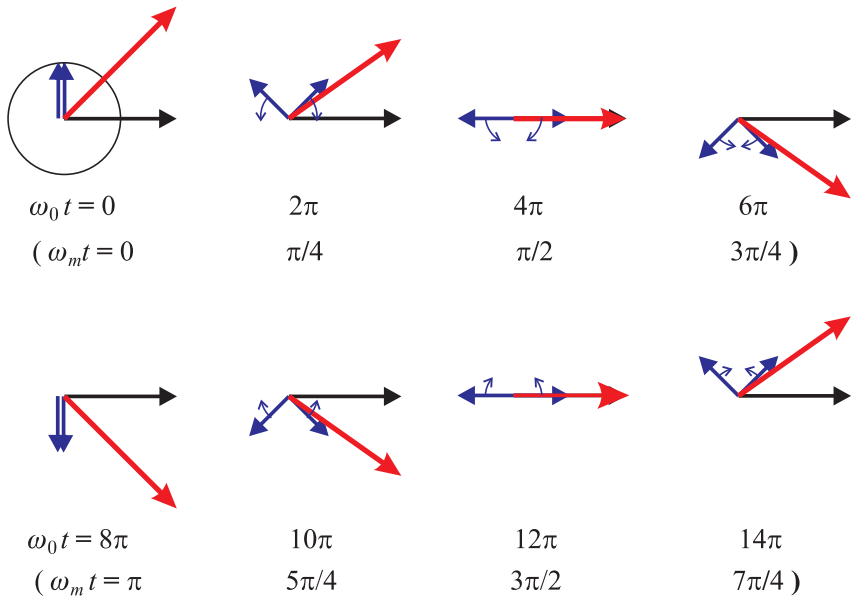


Figure 1.10: Phase plane representation of a phase modulated signal.

One can see that the spectrum of phase modulated signal consists of a central frequency ω_0 and an infinite number of sidebands $\omega_0 \pm n\omega_m$. The spectrum significantly differs from the spectrum of amplitude modulation.

Contribution of higher-order Bessel functions is significant if the modulation index is $\delta > 1$. Roughly, the number of strong sidebands is given by the coefficient δ (e.g. if $\delta = 8$, there will be 8 strong sidebands).

Exercise 2: The laser radiation with pase modulated signal at the frequency ω_m and the modulation index δ is focused on the photodiode. The spectrum analyzed connected to the photodiode records the signal at the frequency ω_m. What will be the amplitude of this signal?

Solution: The photodiode detects the power of the amplitude modulated signal

$$P_{\mathrm{PM}} \propto U_{\mathrm{PM}} \times U^*_{\mathrm{PM}} \approx$$
$$U_0^2\{J_0(\delta)\ \exp(i\omega_0 t) + iJ_1(\delta) \exp\ i(\omega_0 t + \omega_m t) + iJ_1(\delta) \exp\ i(\omega_0 t - \omega_m t)\} \times$$
$$\{J_0(\delta)\ \exp(-i\omega_0 t) - iJ_1(\delta) \exp\ -i(\omega_0 t + \omega_m t) - iJ_1(\delta) \exp\ -i(\omega_0 t - \omega_m t)\} \tag{1.28}$$

Focussing only on the terms at the modulation frequency $+\omega_m$ we will get

$$\begin{aligned} A_{SA}(\omega_m) \propto & -iJ_0J_1 \exp(i\omega_0 t) \exp(-i\omega_0 t) \exp(i\omega_m t) + \\ & iJ_0J_1 \exp(-i\omega_0 t) \exp(i\omega_0 t) \exp(i\omega_m t) = \\ & 0 \end{aligned} \tag{1.29}$$

The spectrum analyzer will detect the signal at the frequency ω_m with ZERO amplitude. It results from the fact that the sidebands have different phases. Compare to result of **the Exercise 1**.

Lecture 2: Amplitude and phase fluctuations

Mathematical description of stochastic processes, distribution function, mean value, dispersion. Allan deviation. Correlated fluctuations. Autocorrelation function. Spectral density. Wiener-Khinchin theorem. Stochastic processes in physical systems. From spectral representation of fluctuations to time representation. Spectral density and Allan deviation of different fluctuation types.

2.1 Mathematical description of stochastic processes, distribution function, mean value, dispersion.

Output frequency even of the best frequency synthesizers is not constant, but fluctuates in time. For example, harmonic amplitude and phase modulation change the output frequency. In real life perturbations have a stochastical nature and should be described in a corresponding framework. The fluctuations can be introduced mathematically as following

$$U(t) = [U_0 + \Delta U(t)] \cos(2\pi\nu_0 t + \phi(t)) \,. \tag{2.1}$$

where $\Delta U(t)$ is the stochastic fluctuations of amplitude and $\phi(t)$ stands for the phase fluctuations. For comparison of different sources oscillating at different frequencies let us introduce normalized phase

$$x(t) \equiv \frac{\phi(t)}{2\pi\nu_0} \,, \tag{2.2}$$

and frequency fluctuations

$$y(t) \equiv \frac{\Delta\nu(t)}{\nu_0} = \frac{dx(t)}{dt} \,. \tag{2.3}$$

Let us then consider some value $y(t)$ fluctuating in time. In physical experiment we always use discretisation, reading the value by some device (the

latter expression is usual for frequency measurement devices)

$$\bar{y_i} = \frac{1}{\tau}\int_{t_i}^{t_i+\tau} y(t)dt\,. \tag{2.4}$$

so we get a set of randomly distributed numbers with some distribution function which will depend on the stochastic process nature.

We can calculate the average value

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \tag{2.5}$$

and the dispersion

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\left(\sum_{i=1}^{N} y_i\right)^2\right]\,. \tag{2.6}$$

The width of the distribution will be given by a dispersion

$$s_{\bar{y}} = \frac{s_y}{\sqrt{N}}\,. \tag{2.7}$$

If the process is *stationary* (both $\bar{y}$ and $s_{\bar{y}}$ do not depend on time) than according to the central theorem, the distribution will approach Gaussian distribution if $T \to \infty$.

$$p(y) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(y-\bar{y})^2}{2\sigma^2}\right)\,, \tag{2.8}$$

The stochastic process is characterized by the expectation value (mean value)

$$\langle y\rangle = \int_{-\infty}^{\infty} yp(y)\,dy \tag{2.9}$$

and the dispersion

$$\sigma^2 = \int_{-\infty}^{\infty}(y-\langle y\rangle)^2 p(y)\,dy\,. \tag{2.10}$$

The latter can be rewritten as

$$\sigma^2 = \langle(y-\langle y\rangle)^2\rangle = \langle y^2\rangle - \langle y^2\rangle\,. \tag{2.11}$$

In real experiment we can only *evaluate* expectation value and the dispersion. Expectation and the dispersion can be also evaluated from the measurement results on the *ensemble* of devices. Typically it is impossible. But, for a stationary process the result should not depend on whether one picks up values from an ensemble of the the devices of from a time realization of one of one of

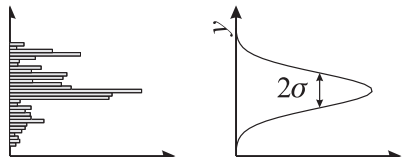


Figure 2.1: Illustration to deriving of the dispersion. a) Recorded signal. b) Digitized signal. c) Averaged signal over intervals τ.d) Histogram. e) its approximation by Gaussian function.

them. These processes are called *ergodic processes*. Very often for modelling some processes one uses the assumptions about *stationarity and ergodicity* implicitly, but one has to be quite accurate doing it. For example, if noise grows continuously in time than the process is not stationary any more.

Another important issue is the possible existence of *correlations*. If fluctuating values are not completely independent, it will result in correlations. If two sets of values, e.g. x_i and y_i are correlated, the plot $x_i(y_i)$ will not be symmetrical. If it is one realization of parameters, the correlation may appear as the fact that the different subsets will have different, statistically inconsistent mean values and dispersion.

2.2 Allan deviation.

To adequately numerically describe a stochastic process in presence of correlations it is possible to use a so-called N*-point dispersion*. For that one has to take N measurements of duration τ each. In principle, there can be a "dead" time interval between the measurements of duration $T - \tau$. The period of the measurement is thus equals to T. The N-point dispersion is thus equals to

$$\sigma^2(N, T, \tau) = \frac{1}{N-1} \sum_{i=1}^{N} \left(\bar{y}_i - \frac{1}{N} \sum_{j=1}^{N} \bar{y}_j \right)^2 . \tag{2.12}$$

It is usually accepted to use dispersion with $N = 2$ and $T = \tau$ according

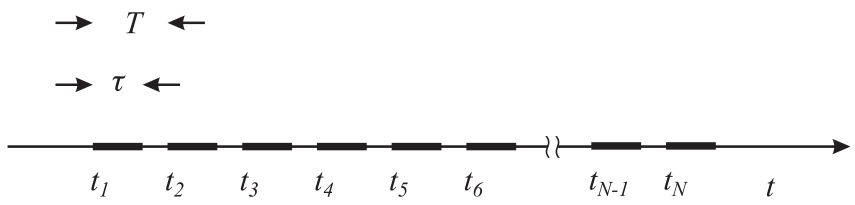


Figure 2.2: Measurement sequence for Allan deviation.

to suggestion of Dave Allan. It is so-called *Allan deviation* which is usually denoted as $\sigma_y^2(2,\tau)$ or $\sigma_y^2(\tau)$

$$\sigma_y^2(\tau) = \left\langle \sum_{i=1}^{2} \left(\bar{y_i} - \frac{1}{2} \sum_{j=1}^{2} \bar{y_j} \right)^2 \right\rangle = \frac{1}{2} \langle (\bar{y_2} - \bar{y_1})^2 \rangle \,. \tag{2.13}$$

It is based on the measurement of the *differences* of the neighboring measurements and not on the deviation from the mean value as in the case of the regular dispersion. The square root of the Allan dispersion is called as *Allan deviation*.

The Allan deviation for the phase is given as

$$\sigma_y^2(\tau) = \frac{1}{2\tau^2} \left\langle (\bar{x}_{i+2} - 2\bar{x}_{i+1} + \bar{x}_i)^2 \right\rangle \,. \tag{2.14}$$

since

$$\bar{y_i} = \frac{\bar{x}_{i+1} - \bar{x}_i}{\tau} \,. \tag{2.15}$$

Practical definition of Allan dispersion

To measure the frequency of some oscillator "1" and, correspondingly, its Allan deviation one has to compare the frequency of this oscillator with some other one ("2"), preferably, much more stable. In general, we can measure only the *frequency ratio*, taking one of the signals as an etalon one.

To measure the Allan dispersion one has to do the following

- measure the frequency of of the oscillator "1" compared to "2"
- the counter should operate in so-called Π-mode without a dead time ($\tau = T$)
- calculate differences of the neighboring readings ν_i and ν_{i+1}, square it and divide by 2

It takes a lot of time to measure the Allan deviation for a set of different time intervals τ. Typically, for saving time the measurement is done by the following procedure:

- one measures the set of ν_i for the minimal time interval τ_{min} which counter can provide (without a dead time!)
- calculation of the Allan deviation for the minimal time $\sigma_y(\tau_{min})$ is given by a standard procedure described above
- one can calculate Allan deviation for $n\tau_{min}$ using the same dataset, n is the integer number. Corresponding frequency is obtained by averaging corresponding n neighboring frequency readings. E.g. for $n = 3$. $\tau = 3\tau_{min}$, $\bar{y}_{1,\tau} = (\bar{y}_{1,\tau_{min}} + \bar{y}_{2,\tau_{min}} + \bar{y}_{3,\tau_{min}})/3$, $\bar{y}_{2,\tau} = (\bar{y}_{2,\tau_{min}} + \bar{y}_{3,\tau_{min}} + \bar{y}_{4,\tau_{min}})/3$, $\bar{y}_{3,\tau} = \ldots$.

If one of the oscillators is much more stable than the other one, than the measurement will give the stability of studied oscillator "1". Another case which is easy to interpret is when one compares two identical oscillators. In that case the Allan dispersion of one of the oscillators will be given as

$$\begin{aligned} \sigma^2_{\mathrm{y,tot}}(\tau) &= \sigma^2_{y,1}(\tau) + \sigma^2_{y,2}(\tau) \qquad \text{and} \\ \sigma_{y,1}(\tau) &= \sigma_{y,2}(\tau) = \frac{1}{\sqrt{2}}\,\sigma_{\mathrm{y,tot}}(\tau). \end{aligned} \tag{2.16}$$

Allan deviation is a very useful measure of an oscillator stability which allows to characterize the stability depending on the observation time. For example, *frequency-stabilized lasers* possess a short time stability of $\sigma_y \leq 5 \times 10^{-16}$ on the time intervals of 1 - 100 s. For longer time intervals the Allan deviation grows due to frequency drifts. For time intervals 1000 - 10000 s *hydrogen maser* offers better stability and lower Allan deviation of $\sigma_y \leq 1 \times 10^{-15}$. One can also distinguish different dependencies for Allan deviation $\sigma_y(\tau)$ which depend on the dominating noise type. We will discuss it laser.

As an example let us first consider some deterministic frequency changes.

Exercise 3: Allan deviation for a linear frequency drift Consider an oscillator which frequency linearly changes in time $y(t) = at$, where a is the drift rate. Calculate the Allan deviation.

Since $\bar{y}_1 = [at_0 + a(t_0 + \tau)]/2$ and $\bar{y}_2 = [a(t_0 + \tau) + a(t_0 + 2\tau)]/2$ we will get

$$\sigma_y(\tau) = \left\langle a\tau/\sqrt{2} \right\rangle = \frac{a}{\sqrt{2}}\,\tau \qquad . \tag{2.17}$$

Hence, linear frequency drift of an oscillator results in the Allan deviation linearly depending on averaging time τ.

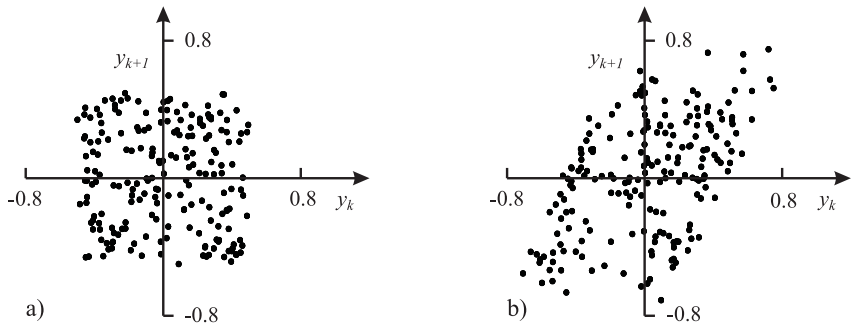


Figure 2.3: a) Uncorrelated data. b) Correlated data.

Exercise 4: Allan deviation for a frequency modulated signal Consider an oscillator with a frequency modulated output

$$y(t) = \frac{\delta\nu_0}{\nu_0}\sin(2\pi f_m t)\,, \tag{2.18}$$

where f_m is the modulation frequency .

After straightforward calculations we will get

$$\sigma_y(\tau) = \frac{\delta\nu_0}{\nu_0}\frac{\sin^2(\pi f_m\tau)}{\pi f_m\tau}\,. \tag{2.19}$$

We see that the Allan deviation results in zero for $\tau = 1/f_m$, i.e time τ is the multiple of the modulation period $1/f_m$ and the influence of modulation becomes zero after averaging over a period. Deviation reaches maximum for $\tau \approx n/(2f_m)$, where n is an integer even number.

2.2.1 Correlated fluctuations

The most simple way to find correlations in experimental data is to plot each measured value as a function of previous one. If the data are correlated, e.g. following the simplest model

$$y_{k+1} = \alpha y_k + \epsilon\,, \tag{2.20}$$

where the fluctuating value y has a pure statistical contribution ϵ. Besides that the value y_{k+1} partly depends on the previous value y_k. The correlation coefficient is $0 \leq \alpha \leq 1$. For $\alpha = 0$ the function $y_{k+1}(y_k)$ is homogeneously distributed over all for quadrants and correlation is absent. If $\alpha > 0$ the correlation will appear as changing the shape of the cloud towards 1s and 3rd quadrants. We will discuss methods which are used for statistical evaluation of the experimental data.

Usually, the fluctuating signal $B(t)$ (e.g. $y(t)$, $U(t)$ or $\Phi(t)$) is represented as a sum of purely fluctuating contribution $b(t)$ and the average value $\overline{B(t)}$:

$$B(t) = b(t) + \overline{B(t)}\,. \tag{2.21}$$

The autocorrelation function is given by

$$R_b(\tau) = \overline{b(t+\tau)b(t)} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} b(t+\tau)b(t)\,dt\,. \tag{2.22}$$

If fluctuations are completely independent, the average value $\overline{b(t+\tau)b(t)}$ is 0 for any $\tau > 0$. For any stationary process $R_b(-\tau) = R_b(\tau)$. It is clear that

$$R_b(\tau = 0) = \sigma_b^2 \tag{2.23}$$

for $\langle B\rangle^2 = 0$.Usually for very large τ correlations are completely lost and $R_b(\tau) \to 0$ for $\tau \to \infty$.

We have shown previously, that the Fourier transformation of some function will give its frequency spectrum. For a fluctuating value the function $U(t)$ is not defined, but the function $R_b(\tau)$ is well defined.

Let us assume $b(t) = \mathcal{F}(a(\omega))$, where the function $a(\omega)$ will be discussed later.

$$\begin{aligned} R_b(\tau) &= \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty} a(\omega) e^{i\omega(t+\tau)}\,d\omega \int_{-\infty}^{\infty} a(\omega') e^{i\omega' t}\,d\omega' dt \\ &= \frac{1}{(2\pi)^2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \left[\lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} e^{it(\omega+\omega')}\,dt\right] a(\omega)a(\omega') e^{i\omega\tau}\,d\omega' d\omega\,, \end{aligned} \tag{2.24}$$

the order of integration was changed in the second raw. In the limit $T \to \infty$ expression in the square brackets is the Dirac delta function, hence

$$\begin{aligned} R_b(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} a(\omega)a(\omega') e^{i\omega\tau} \delta(\omega+\omega')\,d\omega' d\omega \\ &= \int_{-\infty}^{\infty} \frac{|a(\omega)a(\omega)|}{2\pi}\, e^{i\omega\tau}\,d\omega \\ &\equiv \int_{-\infty}^{\infty} S_b(f) e^{i2\pi f\tau}\,df\,. \end{aligned} \tag{2.25}$$

To understand the function $S_b(f)$ assume $\tau = 0$ which will give us

$$R_b(0) = \int_{-\infty}^{\infty} S_b(f)\,df\,. \tag{2.26}$$

The left part of (2.26) equals mean square of the fluctuating function $b(t)$. Hence, S_b is the spectral power density of our fluctuations. E.g. for fluctuating voltage it is measured in V^2/Hz.

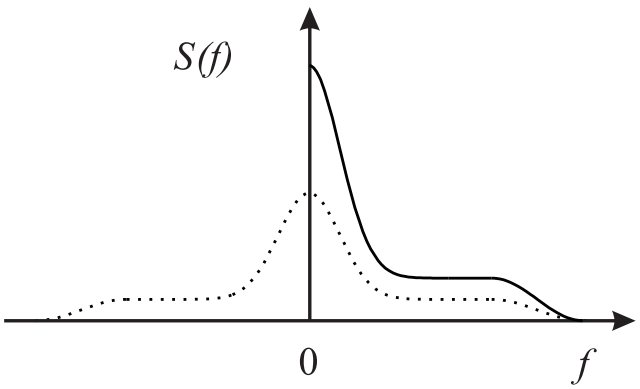


Figure 2.4: Two-sided and one-sided power spectral density.

Autocorrelation function $R_b(t)$ and the spectral density are connected by the Fourier transformation :

$$S_b^{\text{2-sided}}(f) \equiv \mathcal{F}^*\{R_b(\tau)\} = \int_{-\infty}^{\infty} R_b(\tau)\exp(-i2\pi f\tau)\,d\tau\,, \tag{2.27}$$

$$R_b(\tau) \equiv \mathcal{F}\{S_b^{\text{2-sided}}(f)\} = \int_{-\infty}^{\infty} S_b(f)\exp(i2\pi f\tau)\,df\,, \tag{2.28}$$

the index "2-sided" will be discussed later.The expression (2.27) is one of the forms for Wiener-Khinchin theorem. It allows to calculate the spectral power density from autocorrelation function.

If we replace the amplitude $b(t)$ by e.g. phase $\phi(t)$, the spectral density will be given in nits of rad^2/Hz.

The spectral density is given for Fourier frequencies from $-\infty$ to ∞ using both negative and positive frequencies. In this case the spectral density is referred to as "two-sided" $S_b^{\text{2-sided}}(f)$. Since $R_b(\tau) = R_b(-\tau)$, the spectral density is a real positive function and $S_b(-f) = S_b(f)$. In experiment negative frequencies do not exist and sometimes "one-sided" spectral density is used for frequency range $0 \leq f \leq \infty$:

$$S_b^{\text{1-sided}}(f) = 2S_b^{\text{2-sided}}(f)\,. \tag{2.29}$$

2.3 Spectral representation of frequency fluctuations

For an oscillator with enough high stability one can expect that the instant frequency $\nu(t)$ does not significantly deviate from its mean value $\bar{\nu}$ and the following expression is valid:

$$\Delta\nu(t) \equiv \nu(t) - \bar{\nu} \ll \bar{\nu}\,. \tag{2.30}$$

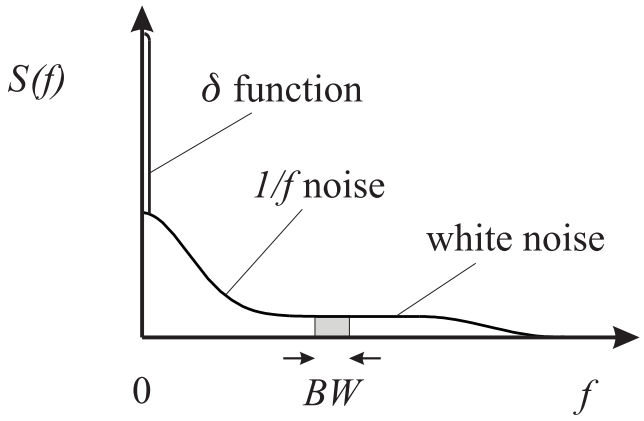


Figure 2.5: Typical shape of a power spectrum.

We assume that fluctuation process $\Delta\nu(t)$ is stationary, i.e., its probability density does not depend on time. According to (2.22), let us define the auto-correlation function for frequency fluctuations:

$$R_\nu(\tau) = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{T} \Delta\nu(t+\tau)\Delta\nu(t)\, dt \tag{2.31}$$

and now use Wiener-Khinchin theorem:

$$S_\nu^{\text{2-sided}}(f) = \int_{\infty}^{\infty} R_\nu(\tau) \exp(-i2\pi f\tau)\, d\tau\,. \tag{2.32}$$

Besides spectral density of frequency fluctuations one can use its relative value ((2.31) and (2.32)),

$$S_y(f) = \frac{1}{\nu_0^2}\, S_\nu(f)\,. \tag{2.33}$$

Similar we can define the spectral density of *phase* fluctuations $S_\phi(f)$. From (2.31), (2.32) and the fact, that the phase is the time derivative of frequency ($2\pi\Delta\nu = d/dt\Delta\phi(t)$) we get

$$S_\nu(f) = f^2 S_\phi(f)\,. \tag{2.34}$$

Combing equations we see, that

$$S_y(f) = \left(\frac{f}{\nu_0}\right)^2 S_\phi(f)\,. \tag{2.35}$$

All three spectral densities carry similar information.

A typical shape of the spectral density function is given in fig. 2.5. There are a few characteristic parts: δ-function around $f = 0$ shows up is $B(t)$ possesses a non-zer0 average value $\bar{B}(t)$. Low-frequency part, falling w at

higher frequencies is referred to as $1/f$ noise. A flat, frequency-independent part corresponds to white noise. Full power in the fluctuations is given by:

$$\int_0^\infty S_\nu^{\text{1-sided}}(f)\,df = \int_{-\infty}^\infty S_\nu^{\text{2-sided}}(f)\,df = \langle[\Delta\nu(t)]^2\rangle = \sigma_\nu^2\,. \tag{2.36}$$

Here we used expressions (2.23) and (2.26). Since power should be finite, at higher frequencies the spectral density should vanish (fig 2.5).

Measurements of different stable frequency oscillators (from quartz oscillators to atomic clocks) show, that the fluctuations of noise spectral density may be well approximated by combination of 5 independent noise processes with spectral density functions represented by power series of f (see table 2.1):

$$S_y(f) = \sum_{\alpha=-2}^{2} h_\alpha f^\alpha\,. \tag{2.37}$$

These noise components correspondingly have typical shape in time representation as shown in fig. 2.6.

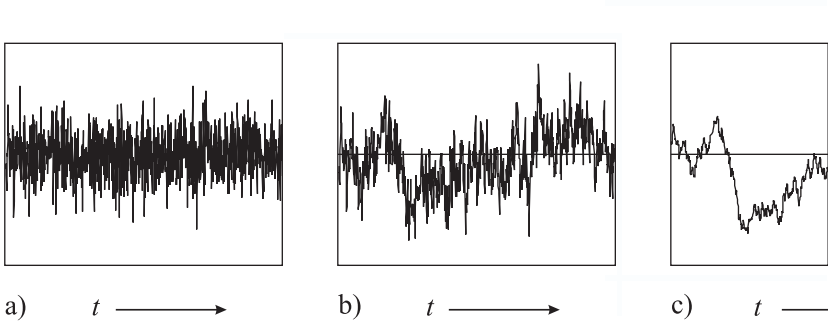


Figure 2.6: Typical time dependencies for noise signals. a) – white noise, b) – noise $1/f$, c) – noise $1/f^2$.

Plot Fig. 2.7 shows in a double-logarithmic scale the fluctuation processes. They are easily distinguished in this plot by different tiles corresponding to (2.37) which allows to identify it. Frequency random walk ($\alpha = -2$) is is often caused by environment (e.g. temperature fluctuations, vibrations).Frequency flicker noise ($\alpha = -1$) is usually observed in active devices like quartz oscillators, hydrogen masers and semiconductor lasers, sometimes also in Cs atomic clock (the latter is a passive device). White frequency noise ($\alpha = 0$) can be caused by thermal noise in the feedback loop in active standards. It is also observed in passive standards due to Poissonian noise from photons or atoms. in this case it corresponds to a quantum noise limit. Phase flicker noise ($\alpha = 1$) comes from noises in electronic circuits, it can be reduced by improving noise characteristics of components. White phase noise ($\alpha = 2$) is important on high frequency and can be reduced by low-pass filtering.

Please note, that mentioned dependencies in (2.37) are only the theoretical model and the real shape can differ from theoretical one.

Table 2.1: Contributions to frequency spectral density with the power dependency of $S_y(f) = h_\alpha f^\alpha$ and corresponding spectral density of phase fluctuations $S_\phi(f)$. Allan dispersion is calculated further under assumption of additional low-pass filter with cut frequency of f_h, where $2\pi f_h \tau \gg 1$.

$S_y(f)$	$S_\phi(f)$	noise type	$\sigma_y^2(\tau)$
$h_{-2}f^{-2}$	$\nu_0^2 h_{-2} f^{-4}$	Frequency random walk	$(2\pi^2 h_{-2}/3)\tau^{+1}$
$h_{-1}f^{-1}$	$\nu_0^2 h_{-2} f^{-3}$	Frequency flicker noise	$2h_{-1}\ln 2\tau^0$
$h_0 f^0$	$\nu_0^2 h_0 f^{-2}$	Frequency white noise (phase random walk)	$(h_0/2)\tau^{-1}$
$h_1 f$	$\nu_0^2 h_1 f^{-1}$	Phase flicker noise	$h_1[1.038 + 3\ln(2\pi f_h \tau)]\cdot \tau^{-2}/4\pi^2$
$h_2 f^2$	$\nu_0^2 h_2 f^0$	Phase white noise	$[3h_2 f_h/(4\pi^2)]\tau^{-2}$

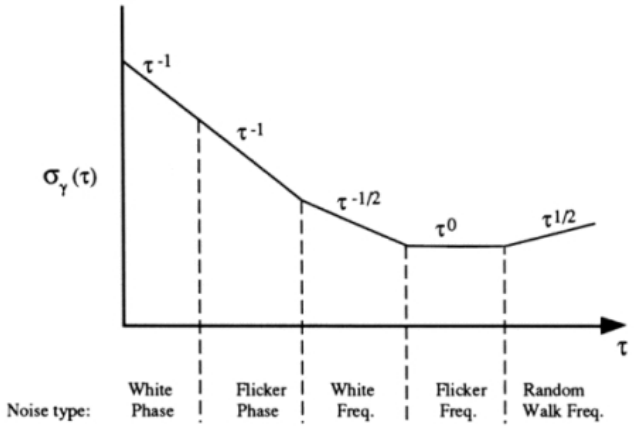


Figure 2.7: Allan deviation for different noise types.

2.4 From spectral representation of fluctuations to time representation

Up to now we described the instability of oscillators either as Fourier transformation (spectral density) or as Allan deviation (time representation). We show here how to calculate Allan deviation from a known spectral density.

Allan dispersion given by (2.13) can be written as

$$\sigma_y^2(\tau) = \frac{1}{2}\left\langle(\bar{y}_2 - \bar{y}_1)^2\right\rangle = \frac{1}{2}\left\langle\left(\frac{1}{\tau}\int_{t_{k+1}}^{t_{k+2}} y(t')\,dt' - \frac{1}{\tau}\int_{t_k}^{t_{k+1}} y(t')\,dt'\right)^2\right\rangle, \tag{2.38}$$

where $t_{k+i} - t_k = i\tau$ for integer i. In the expression (2.38) each counted value is equal to half of difference of two squared values of $y(t)$ for two next intervals of length τ and Allan dispersion is obtained as a mean expected value. To get more information one can substitute a discreet values of $y(t')$ by an integral representation :

$$\sigma_y^2(\tau) = \left\langle\frac{1}{2}\left(\frac{1}{\tau}\int_{t}^{t+\tau} y(t')\,dt' - \frac{1}{\tau}\int_{t-\tau}^{t} y(t')\,dt'\right)^2\right\rangle. \tag{2.39}$$

Expression (2.39) can be rewritten as following:

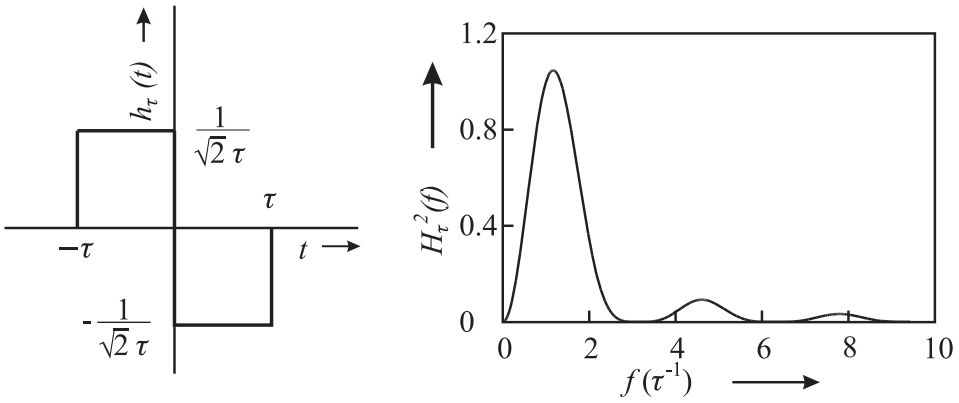
$$\sigma_y^2(\tau) = \left\langle\left(\int_{-\infty}^{\infty} y(t')h_\tau(t-t')\,dt'\right)^2\right\rangle, \tag{2.40}$$

where we introduced a function $h_\tau(t)$ shown in fig. 2.8 a:

$$h_\tau(t) = \begin{cases} -\dfrac{1}{\sqrt{2}\tau} & \text{for} \quad -\tau < t < 0\,, \\ +\dfrac{1}{\sqrt{2}\tau} & \text{for} \quad 0 < t < \tau\,, \\ 0 & \text{for all oter cases} \end{cases} \tag{2.41}$$

The integral in (2.40) is the convolution of $y(t)$ with a function $h_\tau(t)$. One can intuitively understand the impact of $h_\tau(t)$ by substituting the narrow pulse (approximated by the Dirac δ-function) instead of $y(t)$ which will give at output the function $h_\tau(t)$. The convolution integral (2.40) can be interpreted as a time response of a hypothetic linear filter with the pulse characteristic of $h_\tau(t)$. Hence, Allan dispersion is the mean square of fluctuations transmitted by such a filter.

To take into account the filter function we use the convolution theorem, which shows that the convolution of functions $y(t)$ and $h_\tau(t)$ in time representation corresponds to the product of their Fourier transformations $\mathcal{F}(y(t))$



a) b)

Figure 2.8: a) Filter function $h_\tau(t)$ according to (2.41). b) Transmission function $|H_\tau(f)|^2$, corresponding fig. 2.8 a).

and $\mathcal{F}(h_\tau(t))$ in frequency domain. Hence, the spectral density of the signal at filter output is the product of the input spectral density and the filter spectral function (modulus squared):

$$\sigma_y^2(\tau) = \int_0^\infty |H_\tau(f)|^2 S_y^{\text{1-sided}}(f)\, df\,, \tag{2.42}$$

where the function

$$H_\tau(f) = \mathcal{F}\{h_\tau(t)\} \tag{2.43}$$

is the Fourier-transform of the filter function $h(t)$.

Let us calculate the filter transfer function (2.41):

$$\begin{aligned} H(f) &= -\int_\tau^0 \frac{1}{\sqrt{2}\tau}\,\exp(i2\pi ft)\,dt + \int_0^\tau \frac{1}{\sqrt{2}\tau}\,\exp(i2\pi ft)\,dt \\ &= \frac{1}{\sqrt{2}\tau}\left\{-\frac{1}{i2\pi f}[\exp(i2\pi ft)]_{-\tau}^0 + \frac{1}{i2\pi f}[\exp(i2\pi ft)]_0^\tau\right\} \\ &= \frac{1}{\sqrt{2}i2\pi f\tau}[-1+\exp(-i2\pi f\tau)+\exp(i2\pi f\tau)-1] \\ &= \frac{1}{\sqrt{2}i2\pi f\tau}\,2[\cos(2\pi f\tau)-1] = \frac{1}{\sqrt{2}i\pi f\tau}\,2\sin^2(\pi f\tau)\,. \end{aligned} \tag{2.44}$$

We see that

$$|H_\tau(f)|^2 = 2\frac{\sin^4(\pi f\tau)}{(\pi f\tau)^2} \tag{2.45}$$

and

$$\sigma_y^2(\tau) = 2\int_0^\infty S_y(f)\frac{\sin^4(\pi f\tau)}{(\pi f\tau)^2}\,df\,. \tag{2.46}$$

This expression allows to calculate Allan dispersion directly from (one-sided) spectral density $S_y(f)$.

For example let us calculate Allan dispersion for phase white noise ($S_y = h_2 f^2$). Expression (2.46) gives:

$$\sigma_y^2(\tau) = 2\int_0^\infty h_2 f^2 \frac{\sin^4(\pi f\tau)}{(\pi f\tau)^2}\, df = \frac{2h_2}{\pi^2\tau^2}\int_0^\infty \sin^4(\pi f\tau)\, df\,. \tag{2.47}$$

Integral in (2.47) does not converge at $f \to \infty$. In the experiment it does not pose a problem since for any device the frequency bandwidth is restricted at higher frequencies. If we model this restriction by the low pass filter with the cut frequency of f_h, the integral (2.47) can be calculated with the help of the expression $\int \sin^4 ax\, dx = 3/8x - 1/(4a)\sin 2ax + 1/(32a)\sin 4ax$. We get:

$$\sigma_y^2(\tau) = \frac{2h_2}{\pi^2\tau^2}\int_0^{f_h} \sin^4(\pi f\tau)\, df = \frac{3h_2 f_h}{4\pi^2\tau^2} + \mathcal{O}(\tau^{-3})\,. \tag{2.48}$$

Since we can neglect the contribution $\mathcal{O}(\tau^{-3})$ for $f_h \gg 1/(2\pi\tau)$ the Allan deviation for the phase white noise is the power function $\propto \tau^{-2}$. Similarly one can calculate $\sigma_y(t)$ for other spectral shapes. It is summarized in the table 2.1.

Integral (2.46) diverges also for the phase flicker noise ($S_y(f) = h_1 f$). The low pass filtering helps to solve this problem.

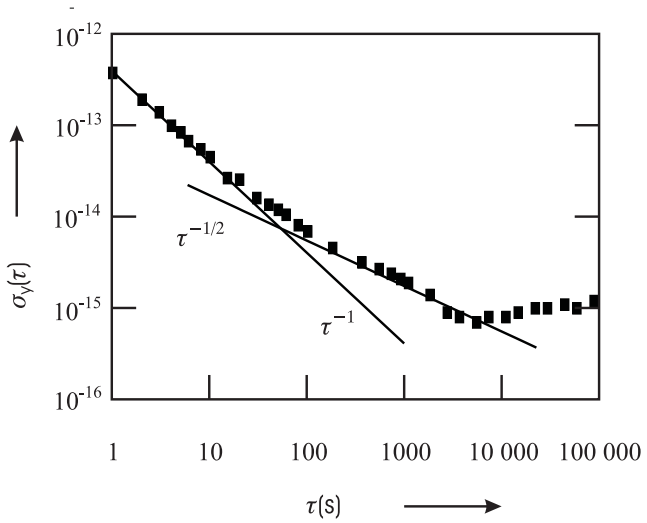


Figure 2.9: Typical Allan deviation of two hydrogen masers compared to each other.

In general, the integral (2.46) diverges at $f \to \infty$ for all functions in (2.37) with $\alpha \geq -1$. Contributions with $\alpha = -1$ and $\alpha = -2$ diverge also at $f \to 0$. In real experiment this divergency is not observed - neither infinite observation time, nor infinite bandwidth can be implemented.

Allan dispersion can be unambiguously calculated from the spectral density, but it is not reversible.

Representation with the help of Allan dispersion is widely used since it is easily measured and calculated. At the other hand, the spectral representation contains *all* information about noises. E.g. consider the Allan dispersion of for the hydrogen maser (fig 2.9). For short integration times the phase white noise dominates ($\propto \tau^{-1}$) and also flicker phase noise (approx. prop to τ^{-1}), for longer integration times – white frequency noise ($\propto \tau^{-1/2}$). Then the Allan dispersion reaches its minimum called the flicker noise floor and then starts growing because of frequency drifts.

Lecture 3: From frequency fluctuations to spectral line shape

Power spectral density of a quasimonochromatic signal with a fluctuating phase. Autocorrelation function representation. Spectral line shape. Line shape in the cases of (i) shallow high-frequency fluctuations and (ii) strong low-frequency phase fluctuations. Line width. Transformation of the line shape in non-linear processes like second harmonic generation.

3.1 Power spectral density of a quasimonochromatic signal with a fluctuating phase.

Regularly by studying spectral properties of a laser or rf oscillator one is interested in a narrow spectral region around the carrier at frequency ν_0. For perfect oscillator it should be a Dirac-δ function, but for a real oscillator phase fluctuations result in spreading of the power around some frequency band around ν_0.

The spectrum can be measured by different tools. For example, it can be a narrow-band filter which central frequency can be tuned around the carrier (in optical region one can use a Fabri-Perot cavity). Another approach is to use a number of narrow band filters in parallel. Also it is possible to implement Fast Fourier Transformation for the signal of interest.

It is still necessary to note that the concept of power spectrum with fixed shape and envelope is not applicable to all fluctuation processes. For example, the $1/f$ noise will not result in the well-defined line shape because of its drifting nature. The line shape will depend on the observation time in this case.

We will show here how one can calculate the line shape knowing a phase noise spectral density $S_\phi(\nu)$. According to (2.27) and (2.28) this (two-sided) spectral density is given via Fourier transformation

$$S_E(\nu) = \int_{-\infty}^{\infty} \exp(-i2\pi\nu\tau) R_E(\tau)\, d\tau \tag{3.1}$$

from the correlation function

$$R_E(\tau) = \langle E(t+\tau)E^*(t)\rangle \tag{3.2}$$

for the electric field $E(t)$. We will neglect amplitude fluctuations in this case.

$$E(t) = E_0 \exp i[2\pi\nu_0 t + \phi(t)] \tag{3.3}$$

Autocorrelation function looks like:

$$R_E(\tau) = E_0^2 \exp[i2\pi\nu_0\tau]\langle \exp i[\phi(t+\tau) - \phi(\tau)]\rangle\,. \tag{3.4}$$

The mean value $\langle \exp i[\phi(t+\tau) - \phi(\tau)]\rangle$ can be calculated via spectral phase noise density $S_\phi(f)$. First of all, let us assume that these fluctuations are ergodic (averaging over time is equivalent to averaging over ensemble):

$$\overline{\exp[i\Phi(t,\tau)]} = \langle \exp[i\Phi(t,\tau)]\rangle = \int_{-\infty}^{\infty} p(\Phi)\exp(i\Phi)\,d\Phi\,, \tag{3.5}$$

where

$$\Phi(t,\tau) = \phi(t+\tau) - \phi(t) \tag{3.6}$$

– is the phase increment over time τ. In the right part of exp. (3.5) we use a regular definition of mathematical expected value (mean) of $\exp[i\Phi(t,\tau)]$ for the given probability distribution of $p(\Phi)$. For a large number of uncorrelated events the central limiting theorem allows to use the Gaussian probability distribution

$$p(\Phi) = \frac{1}{\sigma\sqrt{2\pi}}\,\exp\left(-\frac{\Phi^2}{2\sigma^2}\right) \tag{3.7}$$

with the regular dispersion of σ^2. Since the function $p(\Phi)$ is purely real, only the real (cosine) part will be left in the integral (3.5). Substitution of (3.5) in (3.7), taking into account $\int_{-\infty}^{\infty}\exp(-a^2x^2)\cos x\,dx = \sqrt{\pi}/a\exp(-1/4a^2)$ will give us:

$$\langle \exp[i\Phi(t,\tau)]\rangle = \exp\left(-\frac{\sigma^2}{2}\right)\,. \tag{3.8}$$

Using (2.11) for $\langle\Phi\rangle = 0$ (3.6), we get:

$$\begin{aligned}\sigma^2(\Phi) &= \langle\Phi^2\rangle = \langle[\phi(t+\tau) - \phi(\tau)]^2\rangle \\ &= \langle[\phi(t+\tau)]^2\rangle - 2\langle[\phi(t+\tau)\phi(\tau)]\rangle + \langle[\phi(\tau)]^2\rangle\,.\end{aligned} \tag{3.9}$$

From (3.2) we get:

$$\langle[\phi(t+\tau)\phi(\tau)]\rangle = \int_0^\infty S_\phi(f)\cos(2\pi f\tau)\,df = R_\phi(f)\,, \tag{3.10}$$

$$\langle[\phi(t+\tau)]^2\rangle = \langle[\phi(\tau)]^2\rangle = \int_0^\infty S_\phi(f)\,df = R_\phi(0)\,. \tag{3.11}$$

Substitution of (3.10) and (3.11) into (3.9) will give us:

$$\sigma^2 = 2\int_0^\infty S_\phi(f)[1-\cos 2\pi f\tau]\,df\,, \tag{3.12}$$

which allows to calculate the autocorrelation function (3.4):

$$R_E(\tau) = E_0^2 \exp[i2\pi\nu_0\tau] \exp\left(-\int_0^\infty S_\phi(f)[1-\cos 2\pi f\tau]\,df\right). \tag{3.13}$$

Equations (3.1) and (3.13) allow to calculate the power spectral density from the given spectral density of phase fluctuations $S_\phi(f)$ (see (2.34)):

$$S_E(\nu-\nu_0) = E_0^2 \int_{-\infty}^\infty \exp -[i2\pi(\nu-\nu_0)\tau] \exp\left(-\int_0^\infty S_\phi(f)[1-\cos 2\pi f\tau]\,df\right) d\tau \tag{3.14}$$

under condition that the integral (3.14) converges.

3.1.1 Spectrum with shallow high-frequency fluctuations

Let us consider a case of shallow high-frequency fluctuations. For simplicity let us re-write the expression (3.14) for circular frequencies

$$S_E(\omega) = E_0^2 \int_{-\infty}^\infty \exp -[i(\omega-\omega_0)\tau] \exp\left(-\int_0^\infty S_\omega(\omega')\frac{[1-\cos\omega'\tau]}{\omega'^2}\,d\omega'\right) d\tau \tag{3.15}$$

and define the function $F(\tau)$ as

$$F(\tau) \equiv \int_0^\infty S_\omega(\omega')\frac{[1-\cos\omega'\tau]}{\omega'^2}\,d\omega'\,. \tag{3.16}$$

Let us also introduce the dispersion of frequency fluctuations according to

$$\sigma_\omega^2 = \overline{\Omega}^2 = \int_0^\infty S_\omega(\omega')d\omega' \tag{3.17}$$

with $S_\omega(\omega')$ being a one-sided spectral density. The fluctuation process $\omega(t)$ possesses some typical correlation time τ_Ω (see Fig. 3.1). For a white noise the correlation time equals to zero which provides flat and infinitely broad noise spectrum. In real processes the spectrum is always finite and the correlation time is non-zero $\tau_\Omega > 0$. The spectrum width is proportional to $1/\tau_\Omega$ which directly follows from the properties of Fourier transformation (or the Heisenberg inequality).

Let us consider the case when

$$\overline{\Omega}^2\tau_\Omega^2 \ll 1\,. \tag{3.18}$$

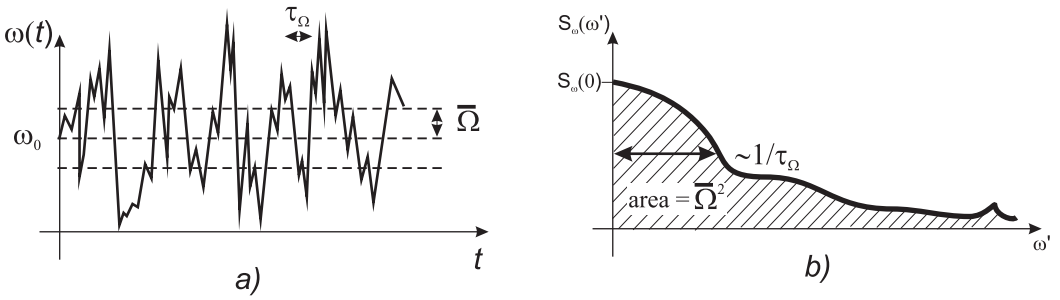


Figure 3.1: a) Fluctuating frequency. The mean value ω_0, the dispersion $\overline{\Omega}$ and the correlation time τ_Ω. b)Relations between $S_\omega(0)$, τ_Ω and $\overline{\Omega}^2$ derived from the typical spectrum shape.

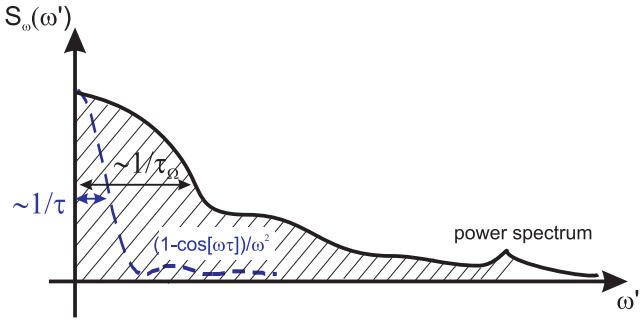


Figure 3.2: Relation between the spectral width of shallow high-frequency fluctuations and function $\frac{[1-\cos\omega'\tau]}{\omega'^2}$.

It corresponds to a case of a short-correlated (fast) and weak (shallow) frequency noise. For example, such noise may dominate in the case when some weak but fast process in the physical system plays a dominant role. In the case of lasers the good example would be spontaneous emission which results is fast small phase jumps of the resulting electric field due to contribution of spontaneous photons to the laser mode. Exactly this process dominates in the emission of semiconductor lasers. The resulting spectral line shape will be derived further in this section.

Let us consider the product $S_\omega(\omega')\frac{[1-\cos\omega'\tau]}{\omega'^2}$ under integral (3.16). The function

$$\frac{[1-\cos\omega'\tau]}{\omega'^2} \tag{3.19}$$

has a main maximum of the width of $1/\tau$ (see Fig. 3.2). Assuming the case $\tau_\Omega \ll \tau$ one can approximate

$$F(\tau) = S_\omega(0)\int_0^\infty \frac{[1-\cos\omega'\tau]}{\omega'^2}\,d\omega' = \frac{\pi}{2}S_\omega(0)\tau \tag{3.20}$$

since

$$\int_0^\infty \frac{1-\cos bx}{x^2} = \frac{\pi|b|}{2}\,. \tag{3.21}$$

We obtain a diffusion process with the diffusion coefficient of

$$D = \frac{\pi S_\omega(0)}{2}\,. \tag{3.22}$$

Thus, we get for the (3.15):

$$S_E(\omega) = E_0^2 \int_{-\infty}^{\infty} \exp[-D\tau] \exp -[i(\omega-\omega_0)\tau]\, d\tau\,. \tag{3.23}$$

The exponent $\exp[-D\tau]$ cuts the expression under integral at $\tau \approx 1/D$. Let us remember that we considered the case $\tau_\Omega \ll \tau$, so we get $D\tau \approx 1$ and $D\tau_\Omega \ll 1$. Substituting D from (3.22) we get

$$\frac{\pi}{2} S_\omega(0)\tau_\Omega \ll 1\,. \tag{3.24}$$

From Fig. 3.2 one can approximate

$$\overline{\Omega}^2 = \int_0^\infty S_\omega(\omega')d\omega' \approx \frac{S_\omega(0)}{\tau_\Omega}\,, \tag{3.25}$$

which gives us $S_\omega(0) \approx \overline{\Omega}^2\tau_\Omega$. Combining this result with (3.24) we get $\frac{\pi}{2}\overline{\Omega}^2\tau_\Omega^2 \ll 1$ which is compatible with the initial assumption (3.18).

Finally, we can easily calculate the integral (3.16)

$$S_E(\omega) = 2E_0^2 \frac{D}{D^2 + (\omega-\omega_0)^2}. \tag{3.26}$$

This is the Lorentzian line shape.

Conclusions: for the case of shallow high-frequency noise one should expect the Lorenzian line shape according to (3.26).

3.1.2 Spectrum with slow and deep frequency fluctuations

In this section we will consider different case compared to 3.1.1, namely, the case of slow and deep fluctuations. This case corresponds mathematically to

$$\overline{\Omega}^2\tau_\Omega^2 \gg 1 \tag{3.27}$$

(compare to (3.18)). This case corresponds to the situation when the oscillator frequency is perturbed by some slow but intense process. In case of the laser

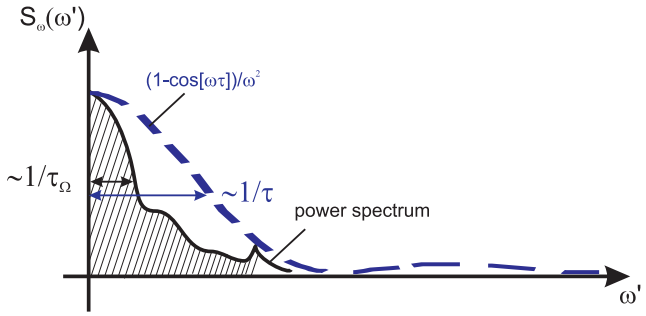


Figure 3.3: Relation between the spectral width of deep low-frequency fluctuations and a function $\frac{[1-\cos\omega'\tau]}{\omega'^2}$.

it can be acoustic noise or temperature fluctuations, or density perturbations in gas lasers. In the intuitive picture, the oscillator frequency will randomly fluctuate around some value and the resulting line shape will be a sum of some "instant prints" of the narrow lines with individually shifted frequencies. According to the Central limiting theorem one should expect Gaussian distribution. Let us show it mathematically. Let us assume that in this case only time intervals $\tau = 1/\omega \ll \tau_\Omega$ are important in our analysis. In this case

$$1 - \cos(\omega\tau) \approx \frac{\omega^2\tau^2}{2} \tag{3.28}$$

and the integral 3.16 is given as

$$F(\tau) = \tau^2 \int_0^\infty S_\omega(\omega')d\omega' = \overline{\Omega}^2\tau^2\,. \tag{3.29}$$

Correspondingly,

$$S_E(\omega) = E_0^2 \int_{-\infty}^{\infty} \exp -[i(\omega-\omega_0)\tau]\exp[-\overline{\Omega}^2\tau^2]\,d\tau\,. \tag{3.30}$$

The second exponent cuts the integral at the typical time of $\tau^2 = 1/\overline{\Omega}^2$ and, together with assumption $\tau \ll \tau_\Omega$ we get $\tau^2\overline{\Omega}^2 \gg 1$ which is compatible with (3.27).

Taking the integral 3.30 one gets

$$S_E(\omega) = \frac{\sqrt{\pi}E_0^2}{\overline{\Omega}} e^{-(\omega-\omega_0)^2/4\overline{\Omega}^2}\,, \tag{3.31}$$

which is the Gauusian line shape as expected from intuitive considerations.

3.1.3 Spectrum with a weak phase noise

Expression (3.14) can be transformed using (3.10) (3.11) as following:

$$S_E(\nu-\nu_0) = E_0^2 \int_{-\infty}^{\infty} \exp[-R_\phi(0)]\exp[R_\phi(\tau)]\exp[-i2\pi(\nu-\nu_0)\tau]\,d\tau\,. \tag{3.32}$$

If phase fluctuations are weak $\int_0^\infty S_\phi(f)\,df \ll 1$, we can expand two first exponents in the power series, leaving the leading order:

$$S_E(\nu-\nu_0) = E_0^2 \int_{-\infty}^{\infty} [1 - R_\phi(0) + R_\phi(\tau)] \exp[-i2\pi(\nu-\nu_0)\tau]\,d\tau\,. \tag{3.33}$$

Using definition of the Dirac-δ function and Wiener-Khinchin theorem (2.28) we get

$$S_E(\nu-\nu_0) = E_0^2[1 - R_\phi(0)]\delta(\nu-\nu_0) + E_0^2 S_\phi^{\text{2-sided}}(\nu-\nu_0)\,. \tag{3.34}$$

Power spectrum consists of the carrier frequency (δ-function) at $\nu = \nu_0$ and two symmetrical sidebands, proportional to the spectral density of the pase noise S_ϕ for $f = |\nu - \nu_0|$.

For commercial devices the useful parameter is the spectral purity $\mathcal{L}(f)$ which corresponds to noise level in the side band measured by spectrum analyzer:

$$\mathcal{L}(f) = \frac{S_\phi^{\text{2-sided}}(\nu-\nu_0)}{1/2E_0^2}\,. \tag{3.35}$$

3.1.4 Spectrum with phase noise: power in the carrier and carrier collapse

Returning back to the equation (3.32)

$$S_E(\nu-\nu_0) = E_0^2 \int_{-\infty}^{\infty} \exp[-R_\phi(0)] \exp[R_\phi(\tau)] \exp[-i2\pi(\nu-\nu_0)\tau]\,d\tau\,. \tag{3.36}$$

can leave the exponent $\exp[-R_\phi(0)]$ without expanding it to power series and only expand the second one $\exp[R_\phi(\tau)]$. In this case the result (3.34) will be transformed to

$$S_E(\nu-\nu_0) = E_0^2 \left(e^{-R_\phi(0)}\delta(\nu-\nu_0) + e^{-R_\phi(0)} S_\phi^{\text{2-sided}}(\nu-\nu_0)\right)\,. \tag{3.37}$$

We see that fraction of power in the carrier ($\delta(\nu-\nu_0)$) is proportional to

$$e^{-R_\phi(0)} = e^{-\phi_{\text{rms}}^2}\,, \tag{3.38}$$

where $-\phi_{\text{rms}}^2$ is the dispersion of the phase fluctuations (r.m.s. phase deviation squared). For e.g. $\phi_{\text{rms}}^2 = 0$ we see that the power fraction in the carrier $P_{\text{carrier}} = 1$ as expected for a noise-free signal. For phase fluctuations on the order of $\phi_{\text{rms}}^2 = 1$ the fraction promptly drops $P_{\text{carrier}} = 1/e$.

Let us consider a noisy signal with phase fluctuations of ϕ_{rms}. If the signal is transformed in the second harmonic generation process, the phase deviation

doubles $\phi'_{\text{rms}} = 2\phi_{\text{rms}}$ which will result in the fact, that the dispersion of the phase fluctuations will be quadrupled.

$$\phi'^2_{\text{rms}} = 4\phi^2_{\text{rms}} \,. \tag{3.39}$$

Accordingly, the power in the carrier will be reduced accordingly to (3.38). In general case, if the signal is converted in the nth harmonic, the power in the carrier will change as

$$P'_{\text{carrier}} = e^{-\phi'^2_{\text{rms}}} = e^{-n^2\phi^2_{\text{rms}}} = (P_{\text{carrier}})^{n^2} \,. \tag{3.40}$$

For example, if the laser light at 972 nm contains $P_{\text{carrier}} = 0.97$ power in the carrier, transformation in the 8th harmonic at 121 nm will result in $P'_{\text{carrier}} = (0.99)^{64} = 0.61$. Even for the best oscillators, frequency multiplication will result in the prompt growth of the phase noise contribution which can result in a so-called *carrier collapse*. Intuitively one can explain such feature by the fact that the noise spectral components will multiply with the carrier in the non-linear process of harmonic transformation which result in relative increase of the phase noise contribution.

3.2 Measurement methods

The spectral density of frequency (or phase) fluctuations can be measured by different means. Function $S_\nu(f)$ can be measured with the help of spectrum analyzer which can be modelled as a set of narrow-band filters and detectors measuring power at the output of each of the filter. Other method uses digital spectrum analyzers with built-in FFT transformation function:

$$\Delta\phi(f) = \mathcal{F}(\Delta\phi(t)) \,. \tag{3.41}$$

Spectral density of phase fluctuations will be given as :

$$S_\phi(f) = \frac{[\Delta\phi(f)]^2}{\text{BW}} \,, \tag{3.42}$$

where the bandwidth BW should satisfy and inequality $\text{BW} \ll f$.

Frequency and phase fluctuations can be transformed into amplitude fluctuations by using a discriminator. It can be slopes of frequency responses of different electronic filters, Fabri-Perot cavities of absorption line. If the oscillator frequency is tuned to the slope the power at the filter output will be linearly dependent on the frequency:

$$V(\nu - \nu_S) = (\nu - \nu_S)k_d + V(\nu_S) \,, \tag{3.43}$$

where k_d – is the slope a the frequency n_S. A detector, installed after the filter will allow to transform frequency fluctuations in the power fluctuations (see fig. 3.4 a).

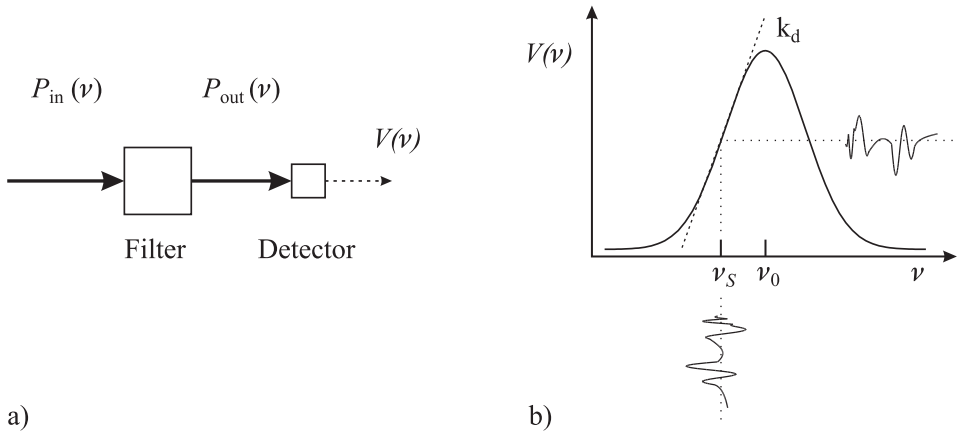


Figure 3.4: a) Spectral sensitivity of filter transmission can be used for transforming frequency fluctuations of input signal $P(\nu)$ into voltage fluctuations $V(\nu)$. b) At proper selection of the working point the filter serves as frequency discriminator: The voltage fluctuations will be approximately proportional to frequency fluctuations. (see (3.43))

3.2.1 Heterodyne measurements

Higher frequencies may be measured by implementation of a heterodyne technique. In the heterodyne the studied signal at frequency ν is mixed with a stable frequency ν_0 giving the residual at the output. Let us consider two harmonic signals at high frequency ν and ν_0 (see fig 3.5 a), b), e.g. two laser fields which are overlapped at a photodiode. A diode is sensitive not to electric field, but to power of the signal which is proportional to the square of sum of amplitudes. The photo-detector output will contain a signal with a frequency of

$$\nu_{\text{beat}} = |\nu - \nu_0| \tag{3.44}$$

which can be filtered by a low-pass filter.

More generally, the photo-detector plays a role of a non-linear element. In radio-electronics are broadly used balanced mixers which multiplies two signals - input signal (RF) withe the heterodyne signal (LO) producing at the output the signal of intermediate frequency (IF). If input signal are harmonic ones, one can write

$$\cos(\omega_{\text{RF}}t)\cos(\omega_{\text{LO}}t) = \frac{1}{2}\cos[(\omega_{\text{RF}} + \omega_{\text{LO}})t] + \frac{1}{2}\cos[(\omega_{\text{RF}} - \omega_{\text{LO}})t]\,. \tag{3.45}$$

The product contains only sum and differential frequencies. A balanced mixer is typically built from four diodes and two transformers which, based on diode non-linearity, multiplies two input signals.

For two signals at the same frequency $\omega_{\text{LO}} = \omega_{\text{RF}} = \omega$, but with some

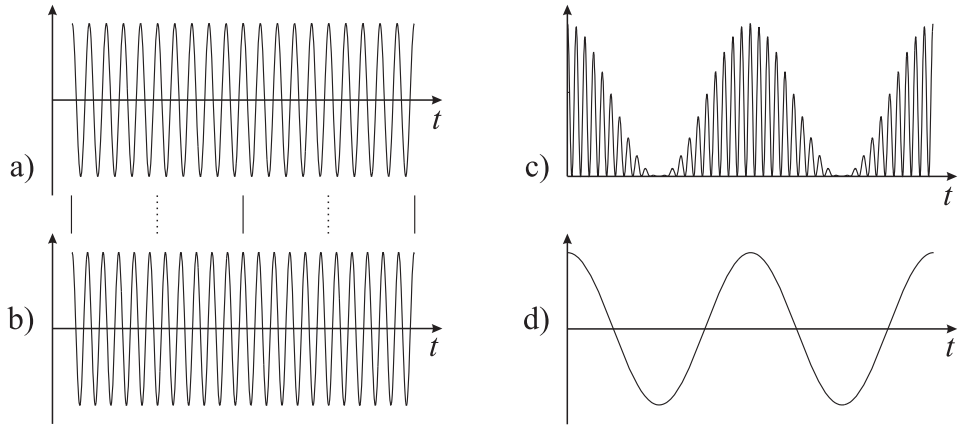


Figure 3.5: a, b) Signals different in frequency by 10%. c) Squared sum of signals a) b). d) Beatnote.

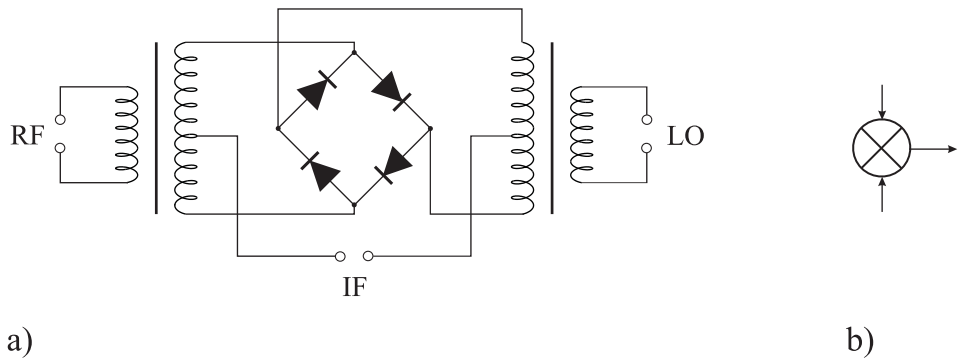


Figure 3.6: Balanced mixer: a) Schematics b) symbol.

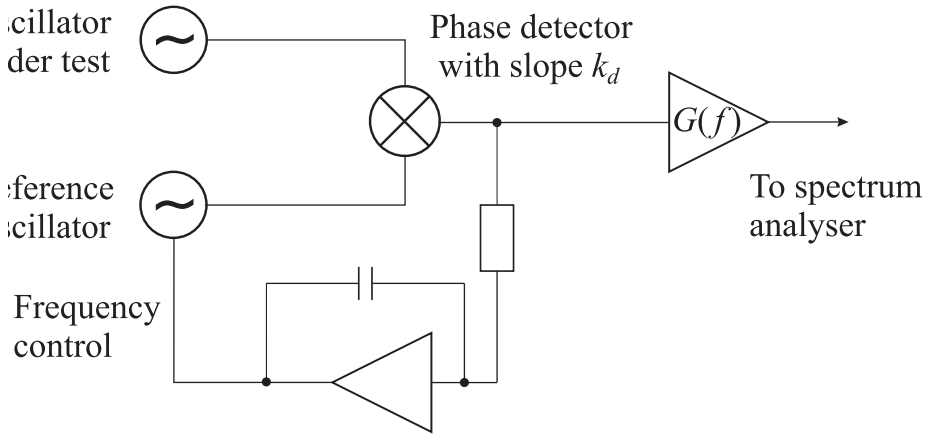


Figure 3.7: Scheme for measurement of oscillator phase noise.

phase difference ϕ we get (3.45)

$$\cos(\omega t + \phi)\cos(\omega t) = \frac{1}{2}[\cos(2\omega t + \phi) + \cos\phi]\,. \tag{3.46}$$

The low-frequency part of the output signal will depend on the phase difference $\frac{1}{2}\cos\phi$.

Using this setup one can measure phase fluctuations according to Fig. 3.7, where the phase of oscillator under study is compared with the phase of some reference high-quality synthesizer.

Exercise: Calculate the spectrum of signal with white frequency noise

Solution: Let us consider an oscillator which frequency fluctuations can be presented as a white (frequency-independent) noise S_ν^0 (see table 2.1). Since

$$S_\phi(f) = \frac{S_\nu^0}{f^2} = \frac{\nu_0^2 h_0}{f^2}\,, \tag{3.47}$$

we can calculate the integral in the exponent (3.14) analytically using the expression $\int_0^\infty [1-\cos(bx)]/x^2\,dx = \pi|b|/2$:

$$\begin{aligned} S_E(\nu-\nu_0) &= E_0^2 \int_{-\infty}^{\infty} \exp -[i2\pi(\nu-\nu_0)\tau]\exp(-\pi^2 h_0\nu_0^2|\tau|)\,d\tau \\ &= 2E_0^2 \int_0^{\infty} \exp\left(-\tau\left[i2\pi(\nu-\nu_0)+\pi^2 h_0\nu_0^2\right]\right)\,d\tau\,. \end{aligned} \tag{3.48}$$

Calculating the integral (3.48) and leaving only the real part we get the power spectrum :

$$S_E(\nu-\nu_0) = 2E_0^2\frac{h_0\pi^2\nu_0^2}{h_0^2\pi^4\nu_0^4 + 4\pi^2(\nu-\nu_0)^2} = 2E_0^2\frac{\gamma/2}{(\gamma/2)^2+4\pi^2(\nu-\nu_0)^2}\,, \tag{3.49}$$

where $\gamma \equiv 2h_0\pi^2\nu_0^2 = 2\pi(\pi h_0\nu_0^2) = 2\pi(\pi S_\nu^0)$. Hence, the spectrum of the oscillator which frequency fluctuations can be presented as white noise S_ν^0, will have a Lorentzian shape with the full width on a half maximum of

$$\Delta\nu_{1/2} = \pi S_\nu^0 \,. \tag{3.50}$$

Lecture 4: General relativity in applications to time and frequency transfer

Space and time in Einstein's theory of gravitation, basics of General relativity. Minkowski metric tensor. Time transformation in rotating frame, gravitational red shift, time dilation, Sagnac effect. Methods of time and frequency transfer, clock synchronization. One way and two way transfer. Transfer of optical frequencies.

Accurate frequency and time signals are extremely important for science and technology. Technologies which we are today considered as standard ones (navigation, geodesic measurements, global communication networks, high bit-rate data transfer) vastly use highly accurate time and frequency signals. For fundamental science these signals are demanded in satellite navigation, interferometry with very large base, measurements of fundamental constants and development of new standards for metrology.

All these methods are using (directly of indirectly) time and frequency transfer. Time and frequency information obtained at large distance from the source allows to set up or correct local time scales, control oscillators and measure the time delay between two sources. Taking into account that the time and frequency signals transferred by electromagnetic waves penetrate the space at the speed of light c, one can calculate coordinates from time delays. For accurate determination of local time one has to take into account all feasible time delays in cables and in space, etc. All these delays contribute to the final uncertainty of time transfer. The task of frequency transfer is somehow simpler – one need only that the delay does not change in time.

For comparison of modern highly accurate frequency signals one has to take into account limitations arising from the General Relativity. According to SI definition of second, each clock give the "true" second in their local reference frame. For an observer residing in another frame, the local time will be influenced by a gravitational potential, generally different for different systems. According to the General Relativity, time runs faster or slower depending of the gravitational potential. Also, the frequency is influenced by the relative

velocity and acceleration of two frames. For example, only due to the difference of gravitation potentials, the observer in the National metrology Institute in Germany (PTB) at a hight of 80 m over see level will see the relative difference of $2 \cdot 10^{-13}$ in respect to NIST clock (USA) at the hight of 1,6 km. The observer at PTB will think that the clock at NIST runs faster.

4.1 Basics of General Relativity

A clock at the Earth surface influence the gravitation field and acceleration in the rotation frame. Clock in the accelerated frame should be treated in frames of General Relativity of the curved time and space (time-space metric).

In this theory, the *interval*

$$ds^2 = g_{\alpha,\beta}(x^\mu)dx^\alpha dx^\beta \tag{4.1}$$

gives the relation between two infinitesimally close time-space events. Tensor $g_{\alpha,\beta}(x^\mu)$ is a metric tensor depending on coordinates, and $(x^\mu) \equiv (x^0 = ct,\, x^1,\, x^2,\, x^3)$ are time-space coordinates with the coordinate time t and the speed of light c. In equation (4.1) one uses the summation of the repeating indices according to Einstein. A time and space curvature in the Solar system is small due to the fact, that the gravitation field is small. Metric tensor components $g_{\alpha,\beta}(x^\mu)$ differ from Minkovsky tensor for Special Relativity $g_{00} = -1$, $g_{ij} = \delta_{ij}$ only by small corrections as a power series for the small parameter, namely the gravitational potential. Here we use the symbol $\delta_{ij} = 1$ for $i = j$ and $\delta_{ij} = 0$ for $i \neq j$. Around the Earth the potential is weak and can be approximated by a Newtonian potential U. Tensor components in an inertial non-rotating geocentric system will be equal to

$$g_{00} = -\left(1 - \frac{2U}{c^2}\right), \qquad g_{0j} = 0, \qquad g_{ij} = \left(1 + \frac{2U}{c^2}\right)\delta_{ij}\,. \tag{4.2}$$

The non-diagonal elements of this metrical tensor in this case equal 0. The relativistic interval can be approximated as

$$ds^2 = -\left(1 - \frac{2U}{c^2}\right)c^2dt^2 + \left(1 + \frac{2U}{c^2}\right)\left[(dx^1)^2 + (dx^2)^2 + (dx^3)^2\right], \tag{4.3}$$

where the gravitational potential $U = U_E + U_T$ is the sum of Newtonian gravitational potential of Earth U_E and tidal potential U_T, which is due to external bodies (Sun, Moon, etc.). For the approximation of the Earth gravitational potential one takes

$$U_E = \frac{GM_E}{r} + J_2 GM_E a_1^2 \frac{(1 - 3\sin^2\phi)}{2r^3}\,, \tag{4.4}$$

where the coordinate r is calculated from the center of the Earth. This equation takes into account the increase of Earth radius towards equator and the potential depends on the latitude which is given by the angle ϕ. The angle ϕ is calculated from the equatorial plane and is positive in the northern hemisphere. The equatorial Earth radius equals to $a_1 = 6\,378\,136.5\,\mathrm{m}$, while the value $GM_E = 3,986\,004\,418 \cdot 10^{14}\ \mathrm{m}^3/\mathrm{s}^2$ is the product of the gravitational constant and the Earth mass. The quadrupole coefficient for the Earth is equal to $J_2 = +1,082\,636 \cdot 10^{-3}$. The expression (4.4) for the gravitational potential provides the accuracy for the gravitational red shift and, correspondingly, for the clock comparison the the relative level of $\delta\nu/\nu < 10^{-14}$.

In the coordinate system rotating together with Earth it is necessary to perform the coordinate system transformation into the system rotating with the constant angular velocity of ω:

$$\begin{aligned} x &= x'\cos(\omega t') - y'\sin(\omega t') \\ y &= x'\sin(\omega t') + y'\cos(\omega t') \\ z &= z' \\ t &= t'\,. \end{aligned} \tag{4.5}$$

The angular speed of Earth equals to $\omega = 7,292\,115 \cdot 10^{-5}\,\mathrm{rad/s}$. We consider the case when $\omega(x'^2 + y'^2) \ll c^2$.

Result of transformation of the frame without gravity taken into account. Substitution of (4.5) and corresponding differentials into the expression for interval $ds^2 = -c^2dt^2 + dx^2 + dy^2 + dz^2$ in the inertial system will give us

$$\begin{aligned} ds^2 &= -\left[1 - \frac{\omega^2}{c^2}(x'^2 + y'^2)\right] c^2dt'^2 - 2\omega y'dx'dt' + 2\omega x'dy'dt' + dx'^2 + dy'^2 + dz'^2 \\ &= g'_{\alpha,\beta}(x'^\mu)dx'^\alpha dx'^\beta\,, \end{aligned} \tag{4.6}$$

here we do not yet take into account the potential U. In the right part of (4.6) we got an expression $\omega^2\rho^2$, which is resulting from the effective potential $U_{\mathrm{centr}} = \omega^2\rho^2/2$ in the frame rotating with the angular velocity ω at the distance $\rho = \sqrt{x'^2 + y'^2}$ from the rotation axis. Thus, in the rotating system *without* the gravitational potential we get

$$g_{00} = -\left(1 - \frac{2U_{\mathrm{centr}}}{c^2}\right)\,. \tag{4.7}$$

It has the same structure as for the interval (4.3). This shows equivalency of gravitational potential and potential coming from the acceleration. From (4.6) we see, that the tensor in the rotating frame has non-zero non-diagonal elements.

Rotating frame and gravitational potential For the spherical coordinates (r – distance to the origin, ϕ – latitude and L – angular longitude) the coordinate transformation looks like

$$\begin{aligned} x' &= r\cos\phi\cos L \\ y' &= r\cos\phi\sin L \\ z &= r\sin\phi \\ t' &= t\,, \end{aligned} \tag{4.8}$$

and we can get the following expression for the interval:

$$ds^2 = -c^2dt^2 + [dr^2 + r^2d\phi^2 + r^2\cos^2\phi\,(\omega^2dt^2 + 2\omega\,dL\,dt + dL^2)]\,. \tag{4.9}$$

Compared to (4.2), the metric in the rotating coordinate system with the gravitational potential taken into account can be written as

$$g_{00} = -\left(1 - \frac{2U}{c^2} - \frac{(\vec{\omega}\times\vec{r})^2}{c^2}\right), \qquad g_{0j} = \frac{(\vec{\omega}\times\vec{r})_j}{c}\,, \qquad g_{ij} = \left(1 + \frac{2U}{c^2}\right)\delta_{ij}\,, \tag{4.10}$$

where the vector product of the angular velocity $\vec{\omega}$ and the radius-vector $\vec{r}$ showing from the center of Earth towards the observer is equivalent to the centripetal potential. Non-diagonal elements are responsible for Sagnac effect, which will be considered later.

4.2 Transformation of time: gravitational shift, time dilation, Sagnac effect

According to SI second definition, time indicated by the clock is so-called *local time* τ: The time measured in the coordinate system rigidly connected to the clock. Consider infinitely small transportation of the clock from one point to another which is described in some external frame by two coordinate points (x^0, x^1, x^2, x^3) and ($x^0 + dt$, $x^1 + dx^1$, $x^2 + dx^2$, $x^3 + dx^3$). The interval

$$d\tau = \frac{1}{c}\sqrt{-ds^2} \tag{4.11}$$

connects the increment of the *local time* $d\tau$ measured by the clock and the increment of *coordinate time* dt of time t, measured in some other, external frame. Time t is called as *coordinate time*. The increment of the coordinate time dt is connected with the increment of the local time by a simple expression

$$dt = d\tau\frac{dt}{d\tau}\,, \tag{4.12}$$

which can be calculated using (4.11) at some moment $(x^0,\, x^1,\, x^2,\, x^3)$. Integration of (4.12) along the *world's line* will give us the coordinate time $dt(t)$. The derivative $d\tau/dt$ can be calculated using (4.1) and (4.11) as

$$\frac{d\tau}{dt} = \sqrt{-g_{00}(x^0,\, x^1,\, x^2,\, x^3) - \frac{2}{c} g_{0i}(x^0,\, x^1,\, x^2,\, x^3)\frac{dx^i}{dt} - \frac{1}{c^2} g_{ij}(x^0,\, x^1,\, x^2,\, x^3)\frac{dx^i}{dt}\frac{dx^j}{dt}}\,. \tag{4.13}$$

Close to the Earth's surface the influence of the gravitational potential on the metric is small ($2U/c^2 \approx 1,4 \cdot 10^{-9} \ll 1$). Hence we will consider only small deviation from the flat space using some small parameter $h(t)$

$$\frac{d\tau}{dt} \equiv 1 - h(t)\,, \tag{4.14}$$

where $h(t)$ is the power series over $1/c$. The difference between the coordinate and local time is thus equals to

$$\Delta t \equiv t - \tau = \int_{t_0}^{t} h(t) dt\,. \tag{4.15}$$

The difference Δt can be calculated either using metric in the geocentric frame (4.3) or in the coordinate system rotating together with Earth (4.9). For the metric in the geocentric non-rotating frame (4.3) the non-diagonal elements equal to zero and the substitution into (4.13) will give us

$$h(t) = 1 - \sqrt{\left(1 - \frac{U}{2c^2}\right) - \frac{1}{c^2}\left(1 + \frac{U}{2c^2}\right) v^2}\,. \tag{4.16}$$

Expanding it in power series we will get

$$h(t) = \frac{U(t)}{c^2} + \frac{v^2}{2c^2} + \mathcal{O}\left(\frac{1}{c^4}\right)\,. \tag{4.17}$$

The second part in this expression is known as *time dilation* or the second order Doppler effect for the clock moving with velocity $\vec{v}$ in respect to the frame. The contribution of $\mathcal{O}\left(\frac{1}{c^4}\right)$ is typically less than 10^{-18} and will not be considered in the future.

For the frame, rotating together with Earth, one gets the following expression

$$h(t) = \frac{1}{c^2}\left[U_g + \Delta U(t) + \frac{V(t)^2}{2}\right] + \frac{2\omega}{c^2}\frac{dA_E}{dt}\,, \tag{4.18}$$

which one can get similar to (4.16) using the metric (4.9). Here $V(t)$ – is the modulus of the coordinate velocity in respect to the Earth. The last part appears due to the Sagnac effect:

$$\frac{1}{c^2}\int_{\mathcal{P}}^{\mathcal{Q}} (\vec{\omega} \times \vec{r}) \cdot d\vec{r} = \frac{1}{c^2}\int_{\mathcal{P}}^{\mathcal{Q}} \vec{\omega} \cdot (\vec{r} \times d\vec{r}) = 2\frac{1}{c^2}\int_{\mathcal{P}}^{\mathcal{Q}} \vec{\omega} \cdot d\vec{A}_E = \frac{2\omega A_E}{c^2}\,. \tag{4.19}$$

Here A_E is the area restricted by the projection on the equatorial plane of the vector originated from the Earth's center and pointing in the moving clocks as shown in fig.4.1. Potential $U_g = 6,263\,685\,75{\cdot}10^7\ \mathrm{m}^2/\mathrm{s}^2$ in (4.18) is the constant

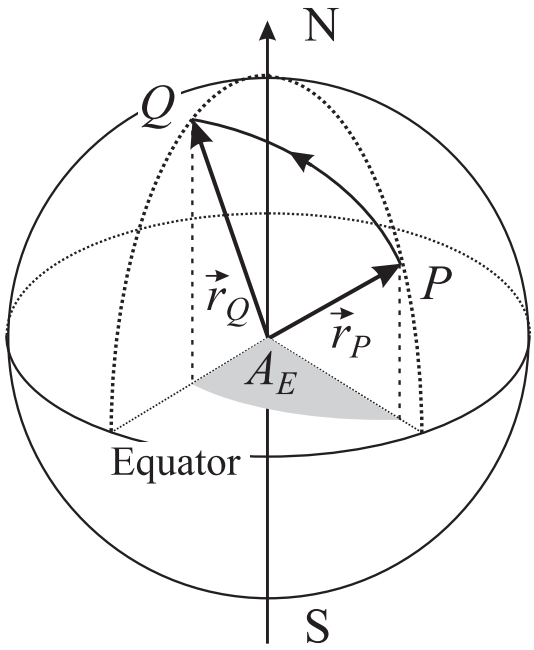


Figure 4.1: *Clock moving from the point $\mathcal{P}$ to $\mathcal{Q}$ on the Earth's surface accumulate a time shift due to Sagnac effect which is proportional to the area A_E.*

potential in the geocentric rotating coordinate frame on the *geoid's* surface which results from the equation (4.4) by adding the centripetal potential. The equation

$$\Delta U(\vec{r}) = \frac{GM_E}{r} + J_2 GM_E a_1^2 \frac{(1-3\sin^2\phi)}{2r^3} + (\omega^2 r^2 \cos^2\phi) - U_g \qquad (4.20)$$

gives the difference of gravitational potentials between the point with the coordinate $\vec{r}$ and the geoid's surface if the accuracy of 10^{-14} is enough.

Even better approximation is reached using expression

$$\frac{\Delta U(b,\phi)}{c^2} = (-1,08821\cdot 10^{-16} - 5,77\cdot 10^{-19}\sin^2\phi)\frac{b}{\mathrm{m}} + 1,716\cdot 10^{-23}\left(\frac{b}{\mathrm{m}}\right)^2, \qquad (4.21)$$

which depends on the hight from geoid b and the latitude ϕ. It is valid for heights $b < 15\,\mathrm{km}$ over the geoid, the relative uncertainty in this case is not larger than 10^{-15}.

4.3 Time and frequency comparison

First of all, we have to agree about what means *synchronization* of two clocks. We agree that *synchronized* clocks give the same reading at the same time. Today one uses the "coordinate synchronization" when two events described in some frame by full coordinate sets x_1^μ and x_2^μ correspondingly are considered as simultaneous if the time coordinates are equal ($x_1^0 = x_2^0$).

One can compare two clocks placed at different positions on the Earth's surface $\mathcal{P}$, $\mathcal{Q}$ by different means. Till some time the most regular method used the physical transportation, now the exchange of electromagnetic signals is widely used. Both these processes are described earlier mathematically in the geocentric frame. The frame can be chosen either using (i) the inertial frame with fixed direction of axes (e.g. pointing on distant quasars/stars) and the origin having the same instant velocity as Earth or (ii) the rotating frame. Equations will depend on the selected frame.

4.3.1 Comparing of transportable clock

If signal is transferred from point $\mathcal{P}$ to point $\mathcal{Q}$ using a transportable clock, the time difference in the non-rotating geocentric frame is equal to

$$\Delta t = \int_{\mathcal{P}}^{\mathcal{Q}} ds \left[1 + \frac{U(\vec{r}) - U_g}{c^2} + \frac{v^2}{2c^2}\right]. \tag{4.22}$$

Here $U(\vec{r})$ is the gravitational potential (only) at the clock position, v – the clock speed in a non-rotating geocentric frame, ds – the increment of the local distance in the clock frame.

In the rotating geocentric frame the time difference will be equal to

$$\Delta t = \int_{\mathcal{P}}^{\mathcal{Q}} ds \left[1 + \frac{\Delta U(\vec{r})}{c^2} + \frac{V^2}{2c^2}\right] + \frac{2\omega}{c^2} A_E \,, \tag{4.23}$$

where V is the clock speed in respect to the Earth surface. Vector $\vec{r}$ is pointing clocks during transportation from $\mathcal{P}$ to $\mathcal{Q}$. The vector projection $\vec{r}$ on the equatorial plane restricts the area A_E.

Three last terms in the (4.23) are the results of gravitation, time dilation and Sagnac effect. The latter results from the fact that clock and Earth are rotating with the same angular velocity. It means that the speed a non-rotation frame will depend on the latitude which defines the distance to the Earth's axis. The area A_E is considered as positive if clock are moving to the East. Situation shown in fig. 4.1 corresponds to the negative A_E.

4.3.2 Transfer using electromagnetic signals

For comparison of two separated clocks with the help of electromagnetic signals of radio- or optical frequency there exist basically three methods of different

complexity: (i) one-way transfer, (ii) differential method and (iii) differential method.

Time passing between emission and receiving of electromagnetic signal in a non-rotating geocentric frame equals to

$$\Delta t = \frac{1}{c}\int_{\mathcal{P}}^{\mathcal{Q}} d\sigma \left[1 + \frac{U(\vec{r}) - U_g}{c^2} + \frac{v^2}{2c^2}\right], \tag{4.24}$$

where $d\sigma$ is the increment of local distance between points $\mathcal{P}$ and $\mathcal{Q}$, all other values are the same as in (4.22).

In the rotating frame we will get

$$\Delta t = \frac{1}{c}\int_{\mathcal{P}}^{\mathcal{Q}} d\sigma \left[1 + \frac{\Delta U(\vec{r})}{c^2}\right] + \frac{2\omega}{c^2} A_E\,, \tag{4.25}$$

where $\Delta U(\vec{r})$ is the gravitational potential in the point $\vec{r}$ reduced by the geoid's potential in the coordinate system rotating with Earth and A_E – the equatorial projection area.

In the case if the signal is transmitted to the satellite at the geostationary orbit, the second term $\Delta U(\vec{r})$ results in the correction of about the 1 ns corresponding to the distance of $ct \simeq 30\,\mathrm{cm}$. The third term containing $2\omega/c^2 = 1,6227 \cdot 10^{-6}\,\mathrm{ns/km^2}$ can reach hundreds of nanoseconds.

One-way transfer

The simplest way of time and frequency transfer is the transmission of coded signals. The simplest examples is are the time got by phone, TV or the Internet. Radio-frequency transmitters of the short- and long wave range cover large areas where the receiver can get sufficient signal. On-board clocks at GPS satellites allow to receive accurate time signal over the whole globe.

The accuracy which can be achieved using this method depends on the propagation time and can reach a few tenths of a second if one uses the Internet line or satellite signal. This error can be significantly reduced if the client sends the signal back and the provider of time signal can measure the whole delay in the line client-server-client. Assuming that the delays in both propagation directions are the same, one can introduce the correction and compensate the significant part of the original error. For the satellite one can calculate the correction from the time distance to the satellite divided by speed of light

Differential method

Signal distributed by one source and received by two or more clients simultaneously can be used for client's clock synchronization. E.g. two users on the Earth's surface can use a signal from one satellite to synchronize their clocks. Consider two stations, A and B receiving the signal t_S propagating by two

paths S – A and S – B with time delays τ_{SA} and τ_{SB} correspondingly. After exchange later of measurement results (e.g. using regular Internet channel) $\Delta t_A = (t_S - \tau_{SA}) - t_A \quad \Delta t_B = (t_S - \tau_{SB}) - t_B$ one gets

$$\Delta t_B - \Delta t_B = (t_A - t_B) - (\tau_{SA} - \tau_{SB})\,, \tag{4.26}$$

which is the difference of times indicated by clocks $t_A - t_B$ and corresponding delays in the channels. This method is known as *differential* method and does not pose strict demands on the satellite clock accuracy since time t_S is cancelled after the subtraction. This method was very important till 2000 when the signal from GPS satellite was deliberately perturbed to reduce the accuracy in the public GPS channel.

Two-way transfer

The most accurate method in radio-frequency domain is the *two-way satellite time and frequency transfer*. Let us consider two stations A and B, each having its own clock, receiver and transmitter (fig. 4.2). Each of the stations sends the

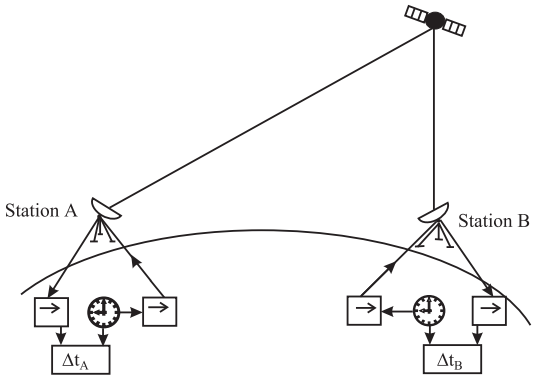


Figure 4.2: *Two-way transfer.*

signal to the satellite which re-transmits the signal to the other station. To reduce distortions of received signal by strong signal emitted by satellite the two signals are transmitted in different radio-frequency bands, e.g. 14 GHz for transmitting to the satellite and 12 GHz for back transmission.

At a moment t_A clocks at the station A give a time mark for beginning the signal transfer from A to B v via satellite and simultaneously trigger the time interval counter at station A. The very similar procedure is started at the station B at the moment t_B. Arriving signals from the satellite are used to stop time interval counters at station A and B. The result indicated by the time interval counters will be equal to

$$\Delta t_A = t_A - t_B + \delta_{B\to A} \tag{4.27}$$

$$\Delta t_B = t_B - t_A + \delta_{A\to B}\,. \tag{4.28}$$

If both directions are fully equivalent, the delays $\delta_{B\to A}$ $\delta_{A\to B}$ are the same. The time difference of the clocks at stations A and B ΔT can be calculated after both stations will exchange the measurement results. By subtracting (4.28) from (4.27), we get $\Delta T = (\Delta t_A - \Delta t_B)/2$.

Still, there are effects which can cause difference in propagation time for two directions. For that case the time difference will be equal to:

$$\begin{aligned} \Delta T &= \frac{\Delta t_A - \Delta t_B}{2} + \frac{(\tau_A^{\mathrm{up}} + \tau_B^{\mathrm{down}}) - (\tau_B^{\mathrm{up}} + \tau_A^{\mathrm{down}})}{2} + \frac{\tau_{A\to B} - \tau_{B\to A}}{2} \\ &+ \frac{(\tau_A^T - \tau_B^R) - (\tau_B^T - \tau_A^R)}{2} + \Delta\tau_R\,. \end{aligned} \tag{4.29}$$

The first term $(\Delta t_A - \Delta t_B)/2$ is the measured time difference, while the second term $[(\tau_A^{\mathrm{up}} + \tau_B^{\mathrm{down}}) - (\tau_B^{\mathrm{up}} + \tau_A^{\mathrm{down}})]/2$ is the contribution from delays in one and the other direction. If the signal transfer occurs approximately at the same moment, the second term can be neglected. The third term takes into account difference in the delays of the satellite re-translation device and is typically insignificant. The fourth contribution in (4.29) is the difference of time delays for signal transmission in the receiver and translator themselves. The last term $\Delta\tau_R$ is the Sagnac effect which takes into account the Earth rotation.

4.3.3 Transfer of optical frequencies

Lecture 5: Introduction to Global navigation systems

Global navigation system structure - space segment, ground segment, user segment. Satellites orbits, frequency shifts, accuracy. Data coding and decoding. CDMA, TDMA, FDMA methods. Atmospheric errors, corrections, clock synchronization. Atomic time scales TAI, UTC.

5.1 Global navigation system

Today space navigation systems become more powerful compared to ground-based systems. Among well-known satellite navigation systems are the American navigation system originally designed for military applications (NAVSTAR GPS) and Russian Global navigation system (GLONASS). Under development are European system GALILEO and Chinese BDS or COMPASS.

5.1.1 Principles of satellite navigation

One can distinguish three large segments in the The Global system for satellite navigation: (i) space segment, (ii) operational control segment and (iii) user segment. Space segment consists of a number of satellites which translate signals to users. The operational control segment consists of observation stations, ground antennas and the main control station. Observation stations track the satellites which are in their field of view and receive the navigation signals, transferring them to the main control station. Information is evaluated at this station for determining the actual orbits. The main station transfers information about individual orbit to each of the satellite thus updating navigation signals from the satellites.

On board of each of the satellite placed an atomic clock. Besides time signal, each of the satellites transfers information about its status and position on the orbit. The user determines his position using the data about distance to satellites positioned at some known coordinates in space. These distances are determined by corresponding time delays from satellites to the user.

For determining position on the Earth surface a GNSS receiver uses time marks from different satellites and compares them with local built-in clock. If a signal from some satellite "i" having coordinates $x_i,\ y_i,\ z_i$ was received by the user having coordinates $X,\ Y,\ Z$ (fig. 5.1), the time delay between the emission and receiving of the signal will define the distance from the satellite to the user.

If a clock in the receiver and clocks at satellites are perfectly synchronized, the distance to the satellite can be calculated form the propagation time delay $\Delta t_1\ \ R_1 = c \cdot \Delta t_1$. Measurement of distance to the second satellite will give the position of the receiver in the common plane. This position will be given by the intersection point of two circles with radii R_1 and R_2 as shown in fig. 5.1. For positioning in the 3D space one needs the third satellite. But, in general case clocks at receiver is not synchronized with on-board atomic clock. The error of time synchronization of $\delta t = 1\,\mu$s will result in the systematic positioning error of 300 m. In the two-dimensional case shown in fig. 5.1 it is implied that time T_U of receiver clock is faster than the navigation system time T_{GNSS} by $\delta t_u = T_U - T_{\mathrm{GNSS}}$. The measured distances will increase by $c \cdot \delta t_u$ which results in a wrong measurement of U' coordinate. Distances, calculated he signal including the clock uncertainty δt_u is called"pseudo-distance" $P_i = R_i + c \cdot \delta t_u$.

Composing 4 equations containing four different *pseudo-distances*, one can four unknown values: three space coordinates X, Y, Z and time difference δt_u:

$$\begin{aligned}
(x_1 - X)^2 + (y_1 - Y)^2 + (z_1 - Z)^2 &= (P_1 - c\,\delta t_u)^2\,, \\
(x_2 - X)^2 + (y_2 - Y)^2 + (z_2 - Z)^2 &= (P_2 - c\,\delta t_u)^2\,, \\
(x_3 - X)^2 + (y_3 - Y)^2 + (z_3 - Z)^2 &= (P_3 - c\,\delta t_u)^2\,, \\
(x_4 - X)^2 + (y_4 - Y)^2 + (z_4 - Z)^2 &= (P_4 - c\,\delta t_u)^2\,.
\end{aligned} \tag{5.1}$$

This non-linear system can be solved either by linearization, or in a closed form. The linearized system is obtained from the original one by Taylor expansion and is solved iteratively by substituting initial expected values for coordinates and time difference. Usually GNSS uses a reference ellipsoid for geocentric World system WGS84.

Further we give more detailed description of typical GPS characteristics, taking into account that other systems operate using similar principles.

5.1.2 GPS system operation

Clocks on board of GPS satellites are synchronized in respect to UTC(USNO) and the GPS systems distributes the time signal which is an approximation to UTC. GPS time scale is relied on readings of a number of atomic clock on board of the satellites and ground stations which are combined by a special procedure. This scale is corrected by an operational control segment in respect to UTC (USNO) of the US Naval observatory within maximal deviation of 1 μs.

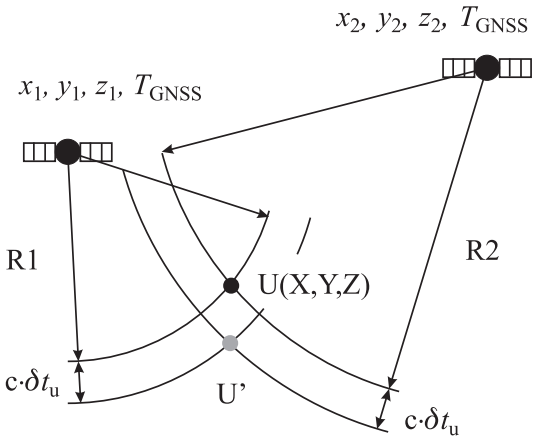


Figure 5.1: *Principles of GNSS operation and time evaluation.*

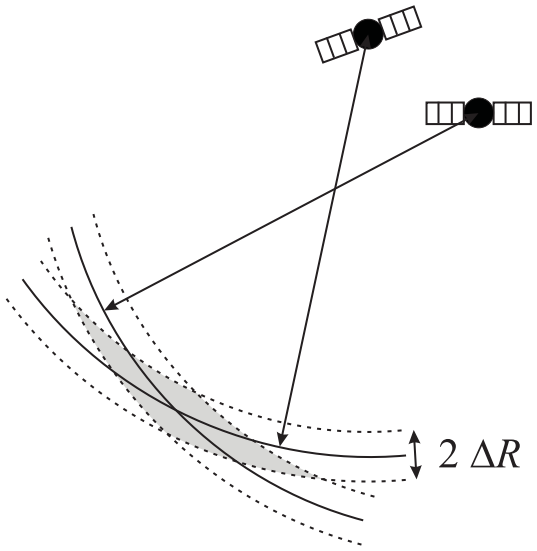


Figure 5.2: *Accumulated uncertainty due to geometric effects.*

Both scales were synchronized at midnight of January 6 1980, but now they differ because UTC(USNO) does not introduce *leap* seconds.

Satellite orbits

For a satellite on the stationary orbit the gravitational force provides the centripetal acceleration according to :

$$G\frac{M_E M_S}{R^2} = M_S \omega^2 R\,. \tag{5.2}$$

As we know, this law describes moving over a infinite number of closed Keplerian orbits. Here $GM_E = 3,986\,004\,418 \cdot 10^{14}\,\mathrm{m}^3/\mathrm{s}^2$ is the product of the gravitational and the Earth's mass. For covering the whole Earth's surface, from each of the ground points at least four satellites should be visible. The orbit should be selected that way, that the signal is strong enough and allows for the easiest description of the satellite position. For GPS satellites it is chosen as a half of the day, precisely 12 hours minus 2 minutes. This feature simplifies determination of the satellite coordinates to distant stars. For a selected rotation period, the radius of the orbit will be equal to 26 560 km. To make the second order Doppler effect and the gravitational red shift constant, the orbits are very close to circular with the eccentricity of $\epsilon = 0,02$. Eccentricity sets a relation between the big a and small b half-axes of the orbit as $b = a\sqrt{1-\epsilon^2}$.

To describe a satellite moving on an orbit one needs 6 coordinates: three space coordinates and three velocity components. This is very inconvenient. Usually for satellite moving on Keplerian orbit one selects so-called *Keplerian parameters*. The frame is "Earth equatorial frame" defined by the Earth's equator and the axis directed to the point of vernal equinox (the point when Sun crosses the Earth's equator plane in spring).

On-board clocks and satellite signals

Four independent clocks are installed on board of each of the satellites (Rb or Cs or both). In modern satellites the number of clocks may be larger. Clocks are used for on-board time synthesis. Since the on-board clocks are less accurate then from the main control station, satellites also transfer information about deviation from the true time scale.

To transport data signals, a suitable carrier frequency is required. The choice of the carrier frequency is submitted to the following requirements:

Frequencies should be chosen below 2 GHz, as frequencies above 2 GHz would require beam antennae for the signal reception. Ionospheric delays are enormous for frequency rages below 100 MHz and above 10 GHz. The speed of propagation of electromagnetic waves in media like air deviates from the speed of light (in vacuum) the more, the lower the frequency is. For low frequencies the runtime is falsified. The PRN-codes (explained below) require a high bandwidth for the code modulation on the carrier frequency. Therefore a range of high frequencies with the possibility of a high bandwidth has to be chosen. The chosen frequency should be in a range where the signal propagation is not influenced by weather phenomena like, rain, snow or clouds.

Based on these considerations, the choice of two frequencies proved to be advantageous. Each GPS satellite transmits two carrier signals in the microwave range, designated as L1 and L2 (frequencies located in the L-Band between 1000 and 2000 MHz). Civil GPS receivers use the L1 frequency with

1575.42 MHz (wavelength 19.05 cm). The L1 frequency carries the navigation data as well as the SPS code (standard positioning code). The L2 frequency (1227.60 MHz, wavelength 24.45 cm) only carries the P code and is only used by receivers which are designed for PPS (precision positioning code). Mostly this can be found in military receivers.

Modulation of the carrier signals C/A and P-Code

The carrier phases are modulated by three different binary codes: first there is the C/A code (coarse acquisition). This code is a 1023 chip long code, being transmitted with a frequency of 1.023 MHz. A chip is the same as a bit, and is described by the numbers one or zero. The name chip is used instead of bit because no information is carried by the signal. By this code the carrier signals are modulated and the bandwidth of the man frequency band is spread from 2 MHz to 20 MHz (spread spectrum). Thus the interference liability is reduced. The C/A code is a pseudo random code (PRN) which looks like a random code (see fig. 5.3) but is clearly defined for each satellite. It is repeated every 1023 bits or every millisecond. Therefore each second 1023000 chips are generated. Taking into account the speed of light the length of one chip can be calculated to be 300 m.

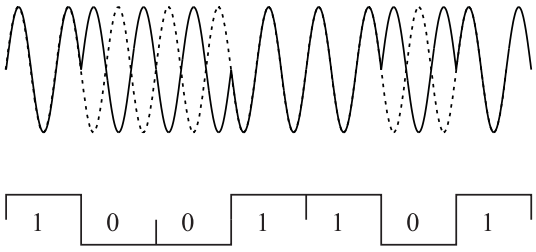


Figure 5.3: *Phase modulation by a pseudo-random code (PRN).*

Pseudo Random Numbers (PRNs)

The satellites are identified by the receiver by means of PRN-numbers. Real GPS satellites are numbered from 1 to 32. These PRN-numbers of the satellites appear on the satellite view screens of many GPS receivers. For simplification of the satellite network 32 different PRN-numbers are available, although only 24 satellites were necessary and planned in the beginning. For a couple of years, now more than 24 satellites are active, which optimizes the availability, reliability and accuracy of the network.

The mentioned PRN-codes are only pseudo random. If the codes were actually random, 21023 possibilities would exist. Of these many codes only few are suitable for the auto correlation or cross correlation which is necessary for the measurment of the signal propagation time. The 37 suitable codes

are referred to as GOLD-codes (names after a mathematician). For these GOLD-codes the correlation among each other is particularly weak, making an unequivocal identification possible.

The C/A code is the base for all civil GPS receivers. The P code (p = precise) modulates the L1 as well as the L2 carrier frequency and is a very long 10.23 MHz pseudo random code. The code would be 266 days long, but only 7 days are used. For protection against interfering signals transmitted by an possible enemy, the P-code can be transmitted encrypted. During this anti-spoofing (AS) mode the P-code is encrypted in a Y-code. The P- and Y-code are the base for the precise (military) position determination.

Transmission of data

In the GPS system data are modulated onto the carrier signal by means of phase modulations.

When a data signal shall be modulated onto a carrier signal by phase modulation, the sine oscillation of the carrier signal is interrupted and restarted with a phase shift of e.g. 180. This phase shift can be recognized by a suitable receiver and the data can be restored. Phase modulation leads to an extension of the frequency range of the carrier signal (leading to a spread spectrum) depending on how often the phase is shifted. When the phase changes, wave peaks are followed by wave minimums in a shorter distance than were in the original carrier signal (as can be seen in the graph). This kind of modulation can only be used for the transmission of digital data. The information (clock corrections, ephemerides, etc.) are transmitted together with PRN identification code with 50 Hz rate (see fig. 5.4).

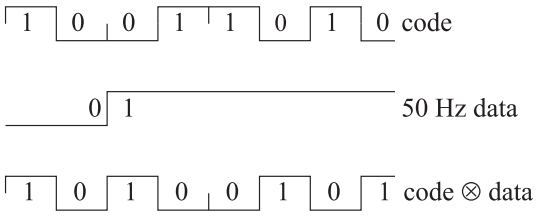


Figure 5.4: *Data coding in GPS signal.*

Uncertainties in GPS signals

Uncertainty for accurate measurement of coordinates with the help of GPS system first of all is given by effects which influence the pdeudo-distance to the satellite. It is called the “user uncertainty” UERE (User Equivalent Range Errors). It can increase by the geometric factor (fig. 5.2 which is given by so-called GDOP coefficient (Geometric Dilution of Precision). The coefficient

GDOP is derived by solving the set of equations using all satellites in the field of view. Analytical consideration shows that the GDOP coefficient is reversely proportional to the volume of the polyhedron with the vertexes at satellite positions. The uncertainty depends on a number of factors which cause the deviation of the "pseudo-distance" from true distances.

Ephemerides. For accurate determination of the GPS receiver one has to know the position of the satellite coordinates in respect to the globe. Because of perturbations the satellites move not exactly on Keplerian orbits. Perturbations can be of gravitational and non-gravitational nature. Deceleration of the satellite in the high atmospheric layers and the sun wind are the strongest non-gravitational perturbations. Gravitational perturbations result from the ellipsoidal Earth shape and the tidal potentials. The ellipsoidal earth shape causes the slow precession of the satellite's orbit. As a result of all these perturbations the satellite's orbits are not stationary and should be corrected once in a while by on-board engines. The satellite position is measured by ground stations with very well known coordinates on the Earth's surface. The main station analyzes received data and sends correction signals to satellites.

On-board clock uncertainties. According to the General relativity, the observed frequency of the on-board clock depends on ins gravitational potential and velocity (see (4.14), (4.16)). The effective potential for the clock orbiting around the Earth at the distance R and the angular velocity ω equals to

$$U = -\frac{GM_E}{R} - \frac{\omega^2 R^2}{2}\,. \tag{5.3}$$

For the clock on board of the satellite we get

$$U_{\text{satellite}} = -\frac{GM_E}{R} - \frac{GM_E}{2R} = -\frac{3}{2}\frac{GM_E}{R}\,, \tag{5.4}$$

using eqs. (5.3) (5.2).

For the clock resting on the geoid's surface we get $U_{\text{surface}} = -62,6\ (\text{km/s})^2$. The potential difference between two clocks result in the time difference of

$$\frac{\Delta\nu}{\nu} = \frac{\Delta U}{c^2} = \frac{1}{c^2}\left(-\frac{3}{2}\frac{GM_E}{R} + 62,6\cdot 10^6\,\frac{\text{m}^2}{\text{s}^2}\right)\,. \tag{5.5}$$

Using this equation one can calculate the time difference for different orbits as shown in fig. 5.5). For low-orbit satellites the difference is negative and becomes zero for the satellites orbiting at 3190 km over the geoid surface which is the half of the Erath's radius.

Time difference becomes positive for high orbits where GPS satellites or geostationary satellites are orbiting. For observation from the ground, GPS

clocks orbiting at the hight of $R = 26\,600$ km will be faster for 38,5 μs/day. To compensate this effect, the on-board clock signal is corrected for $-4,464\,733 \cdot 10^{-10}$ in relative units. Thus, the transmitted frequency is 10,229 999 995 432 6 MHz instead of 10,23 MHz. It does not take into account small eccentricity of the GPS satellite orbits. In the perigee the satellite is at the lower distance to the Earth and its speed is increased. Both these effects result in the reduction of the clock frequency if observed from the Earth. The maximal frequency deviation resulting from this effect is about 70 ns.

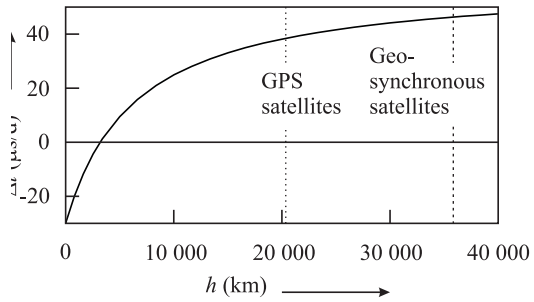


Figure 5.5: *Time difference accumulated during 24 hours between the clock on board of the satellite with the orbit's hight h over the geoid and the clock on the geoid calculated using (5.5).*

Atmospheric delays. Propagation of the electromagnetic waves emitted by the satellites through the Earth's atmosphere if different from propagation through vacuum. The strongest perturbations take place in the ionosphere. The refraction index of the ionosphere n_p for the phase velocity of the electromagnetic signal at frequency ν is given by the expression

$$n_p = 1 + \frac{c_2}{\nu^2}\,. \tag{5.6}$$

The coefficient $c_2 = -40.3 \times n_e$ Hz^2 depends on the free electron density n_e alon the propagation path to the satellite from the receiver. The integrated electron density is referred to as the total number of electrons TEC (Total Electron Content). TEC value is the number of free electrons in the cylinder with the base of 1 m^2. It can change in the range between 10^{16} m^{-2} to 10^{19} m^{-2} depending on the receiver position, high of the satellite Sun activity.

Since the GPS signal is modulated by information bits, it covers some frequency band. For the data transfer the *group* velocity defines how fast the pules (or bit) will propagate through a medium. The group velocity is defined by c/n_g where n_g corresponds to the refraction index. From a well known relation $n_g = n_p + \nu dn_p/d\nu$ and (5.6) we get

$$n_g = 1 - \frac{c_2}{\nu^2}\,. \tag{5.7}$$

Thus, the ionospheric delay for the data transfer can be given by

$$\Delta T = \frac{40,3 \cdot \mathrm{TEC}}{c\nu^2}\,. \tag{5.8}$$

If both bands L1 and L2 are used by the receiver, the difference of time delays will be equal to

$$\Delta\tilde{T} \equiv \Delta T(\mathrm{L1}) - \Delta T(\mathrm{L2}) = \frac{40,3 \cdot \mathrm{TEC}}{c}\left(\frac{1}{\nu_1^2} - \frac{1}{\nu_2^2}\right) = \Delta T(\mathrm{L1})\frac{\nu_2^2 - \nu_1^2}{\nu_2^2}\,. \tag{5.9}$$

One can thus calculate the delay ΔT_1 at the frequency L1 from the delay $\Delta\tilde{T}$ (5.9), which one can measure directly. Delay at the frequency L2 can be derived from ΔT_1 as $\nu_1^2/\nu_2^2 = (77/60)^2$.

In the case if receiver receives only L1 frequency band the ionospheric delay can be taken into account only from an empiric model. Parameters of this model are included in information sent by GPS satellites. The uncertainty can reach up to 50% from size of the effect itself.

The lower part of the atmosphere is called troposphere and basically does not possess dispersive properties for the frequencies lower than 15 GHz. The corresponding delay cannot be derived from L1 and L2 comparison. Change of the distance caused by tropospheric effects should be corrected using semi-empirical models. The correction typically corresponds to distance of few meters.

Navigation accuracy. From 1990 to 2000 the GPS signals were deliberately perturbed: a noise-like modulation was added to GPS signals. For a regular user it resulted in significant reduction of positioning accuracy to approx. 200 m. Authorized users (mainly military) possessed tools to decipher this modulation and avoid additional uncertainty.

Table 5.1 summarizes uncertainties for pseudo-range measurements coming from different sources.

For increasing the accuracy of coordinate and time measurements the "differential GPS" method is used. To implement this method, the on-ground stations with very well known coordinates are contributing to one-way GPS transfer. This method can considerably increase the accuracy of coordinate measurements with a user receiver.

Time and frequency transfer using GPS

Table 5.2 summarizes relative uncertainties corresponding to transfer of time and frequency by different methods using GPS satellites.

One-way method relies on direct time transfer form the GPS satellite. In the *differential* method two distant GPS receivers receive data from the same

Source of uncertainty	uncertainty
on-board clocks	3,0 m
satellite orbits	1,0 m
other perturbations	0,5 m
ephemerides prediction	4,2 m
other	0,9 m
ionospheric delay	2,3 m
tropospheric delay	2,0 m
receiver nose	1,5 m
propagation by different channels	1,2 m
others	0,5 m
sum	6,6 m

Table 5.1: Uncertainties for pseudo-range measurements (status 2006) for the space segment, control segment and user segment.

method	time uncertainty	relative frequencys uncertainty
one-way	<20 ns	$< 2 \cdot 10^{-13}$
one-channel differential	≈10 ns	$\approx 10^{-13}$
multi-channel differential	< 5 ns	$< 5 \cdot 10^{-14}$
differential with carrier phase measurement	< 500 ps	$< 5 \cdot 10^{-15}$

Table 5.2: Uncertainties at the level of 2σ for GPS measurements for 24 hr averaging time.

satellite simultaneously. For improving uncertainty one can implement *carrier phase* measurement. Usually the complicated geodesic receivers get all information about channels PA, P1, P2 as well as L1 and L2 phase.

For sub-nanosecond accuracies one needs to now the distance between antenna and the receiver with a very high accuracy (1 m is equivalent to 5 ns).

The accuracy of time transfer (status 2001) can be evaluated from the plot fig. 5.6. The main uncertainty comes from GPS (standard deviation approx. 2,6 ns) (TWSTFT) GPS.

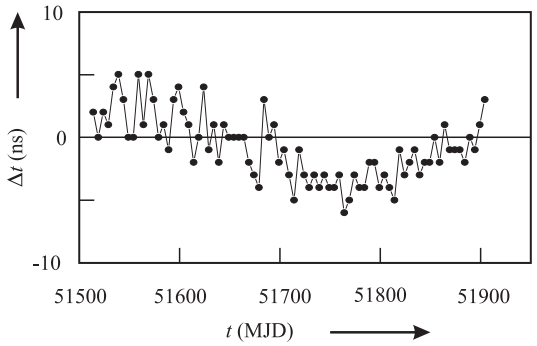


Figure 5.6: *The difference between time measurements at PTB and NPL for TW-STFT (two-way transfer) and differential GPS method (code C/A). The result shows full-day measurements for certain Julian days (MJD=0 corresponds to 0 o'clock 17 November 1858.)*

5.2 Code division multiplexing (Synchronous CDMA)

CDMA is a spread spectrum multiple access[6] technique. A spread spectrum technique spreads the bandwidth of the data uniformly for the same transmitted power. A spreading code is a pseudo-random code that has a narrow ambiguity function, unlike other narrow pulse codes.

Each user in a CDMA system uses a different code to modulate their signal. Choosing the codes used to modulate the signal is very important in the performance of CDMA systems. The best performance will occur when there is good separation between the signal of a desired user and the signals of other users. The separation of the signals is made by correlating the received signal with the locally generated code of the desired user. If the signal matches the desired user's code then the correlation function will be high and the system can extract that signal. If the desired user's code has nothing in common with the signal the correlation should be as close to zero as possible (thus eliminating the signal); this is referred to as cross correlation. If the code is correlated with the signal at any time offset other than zero, the correlation should be as close to zero as possible. This is referred to as auto-correlation and is used to reject multi-path interference.

An analogy to the problem of multiple access is a room (channel) in which people wish to talk to each other simultaneously. To avoid confusion, people could take turns speaking (time division), speak at different pitches (frequency division), or speak in different languages (code division). CDMA is analogous to the last example where people speaking the same language can understand each other, but other languages are perceived as noise and rejected. Similarly,

in radio CDMA, each group of users is given a shared code. Many codes occupy the same channel, but only users associated with a particular code can communicate.

Synchronous CDMA exploits mathematical properties of orthogonality between vectors representing the data strings. For example, binary string 1011 is represented by the vector (1, 0, 1, 1). Vectors can be multiplied by taking their dot product, by summing the products of their respective components (for example, if $u = (a, b)$ and $v = (c, d)$, then their dot product $u \cdot v = ac + bd$). If the dot product is zero, the two vectors are said to be orthogonal to each other. Some properties of the dot product aid understanding of how W-CDMA works. If vectors a and b are orthogonal, then $\mathbf{a} \cdot \mathbf{b} = 0$.

5.2.1 Example

Start with a set of vectors that are mutually orthogonal. (Although mutual orthogonality is the only condition, these vectors are usually constructed for ease of decoding, for example columns or rows from Walsh matrices.) An example of orthogonal functions is shown in the picture on the left. These vectors will be assigned to individual users and are called the code, chip code, or chipping code. In the interest of brevity, the rest of this example uses codes,v, with only 2 bits.

Walsh matirces

$$H(2^1) = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix},$$

$$H(2^2) = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix},$$

and in general

$$H(2^k) = \begin{bmatrix} H(2^{k-1}) & H(2^{k-1}) \\ H(2^{k-1}) & -H(2^{k-1}) \end{bmatrix} = H(2) \otimes H(2^{k-1}).$$

Each user is associated with a different code, say v. A 1 bit is represented by transmitting a positive code,v, and a 0 bit is represented by a negative code, v. For example, if $v = (v0, v1) = (1, 1)$ and the data that the user wishes to transmit is $(1, 0, 1, 1)$, then the transmitted symbols would be $(v, v, v, v) = (v0, v1, v0, v1, v0, v1, v0, v1) = (1, 1, 1, 1, 1, 1, 1, 1)$.

Each sender has a different, unique vector v chosen from that set, but the construction method of the transmitted vector is identical.

Now, due to physical properties of interference, if two signals at a point are in phase, they add to give twice the amplitude of each signal, but if they are out of phase, they subtract and give a signal that is the difference of the amplitudes. Digitally, this behavior can be modelled by the addition of the transmission vectors, component by component.

If $sender0$ has code $(1, -1)$ and data $(1, 0, 1, 1)$, and $sender1$ has code $(1, 1)$ and data $(0, 0, 1, 1)$, and both senders transmit simultaneously, then this table describes the coding steps:

$signal0 = encode0 \cdot code0 = (1, 0, 1, 1) \cdot (1, -1) \equiv (1, -1, 1, 1) \cdot (1, -1) = (1, -1, -1, 1, 1, -1, 1, -1)$

$signal1 = encode1 \cdot code0 = (0, 0, 1, 1) \cdot (1, 1) \equiv (-1, -1, 1, 1) \cdot (1, 1) = (-1, -1, -1, -1, 1, 1, 1, 1)$

Because $signal0$ and $signal1$ are transmitted at the same time into the air, they add to produce the raw signal:

$(1, -1, -1, 1, 1, -1, 1, -1)+(-1, -1, -1, -1, 1, 1, 1, 1) = (0, -2, -2, 0, 2, 0, 2, 0)$

This raw signal is called an interference pattern. The receiver then extracts an intelligible signal for any known sender by combining the sender's code with the interference pattern, the receiver combines it with the codes of the senders. The following table explains how this works and shows that the signals do not interfere with one another: $code0 = (1, -1), signal = (0, -2, -2, 0, 2, 0, 2, 0)$

$decode0 = pattern \cdot vector0$

$decode0 = ((0, -2), (-2, 0), (2, 0), (2, 0)) \cdot (1, -1)$

$decode0 = ((0 + 2), (2 + 0), (2 + 0), (2 + 0))$

$data0 = (2, -2, 2, 2),$ **meaning** $(1, 0, 1, 1)$

$code1 = (1, 1), signal = (0, -2, -2, 0, 2, 0, 2, 0)$

$decode1 = pattern \cdot vector1$

$decode1 = ((0, -2), (-2, 0), (2, 0), (2, 0)) \cdot (1, 1)$

$decode1 = ((0 - 2), (-2 + 0), (2 + 0), (2 + 0))$

$data1 = (-2, -2, 2, 2),$ **meaning** $(0, 0, 1, 1)$

5.2.2 Asynchronous CDMA

When mobile-to-base links cannot be precisely coordinated, particularly due to the mobility of the handsets, a different approach is required. Since it is not mathematically possible to create signature sequences that are both orthogonal for arbitrarily random starting points and which make full use of the code space, unique "pseudo-random" or "pseudo-noise" (PN) sequences are used in asynchronous CDMA systems. A PN code is a binary sequence that appears random but can be reproduced in a deterministic manner by intended receivers. These PN codes are used to encode and decode a user's signal in Asynchronous CDMA in the same manner as the orthogonal codes in synchronous CDMA (shown in the example above). These PN sequences are statistically uncorrelated, and the sum of a large number of PN sequences results in multiple access interference (MAI) that is approximated by a Gaussian noise process (following the central limit theorem in statistics). Gold codes are an example of a PN suitable for this purpose, as there is low correlation between the codes. If all of the users are received with the same power level, then the variance

(e.g., the noise power) of the MAI increases in direct proportion to the number of users. In other words, unlike synchronous CDMA, the signals of other users will appear as noise to the signal of interest and interfere slightly with the desired signal in proportion to number of users.

5.2.3 Flexible allocation of resources

Asynchronous CDMA offers a key advantage in the flexible allocation of resources i.e. allocation of a PN codes to active users. In the case of CDM (synchronous CDMA), TDMA, and FDMA the number of simultaneous orthogonal codes, time slots and frequency slots respectively are fixed hence the capacity in terms of number of simultaneous users is limited. There are a fixed number of orthogonal codes, time slots or frequency bands that can be allocated for CDM, TDMA, and FDMA systems, which remain underutilized due to the bursty nature of telephony and packetized data transmissions. There is no strict limit to the number of users that can be supported in an asynchronous CDMA system, only a practical limit governed by the desired bit error probability, since the SIR (Signal to Interference Ratio) varies inversely with the number of users. In a bursty traffic environment like mobile telephony, the advantage afforded by asynchronous CDMA is that the performance (bit error rate) is allowed to fluctuate randomly, with an average value determined by the number of users times the percentage of utilization. Suppose there are 2N users that only talk half of the time, then 2N users can be accommodated with the same average bit error probability as N users that talk all of the time. The key difference here is that the bit error probability for N users talking all of the time is constant, whereas it is a random quantity (with the same mean) for 2N users talking half of the time.

In other words, asynchronous CDMA is ideally suited to a mobile network where large numbers of transmitters each generate a relatively small amount of traffic at irregular intervals. CDM (synchronous CDMA), TDMA, and FDMA systems cannot recover the underutilized resources inherent to bursty traffic due to the fixed number of orthogonal codes, time slots or frequency channels that can be assigned to individual transmitters. For instance, if there are N time slots in a TDMA system and 2N users that talk half of the time, then half of the time there will be more than N users needing to use more than N time slots. Furthermore, it would require significant overhead to continually allocate and deallocate the orthogonal code, time slot or frequency channel resources. By comparison, asynchronous CDMA transmitters simply send when they have something to say, and go off the air when they don't, keeping the same PN signature sequence as long as they are connected to the system.

5.2.4 Spread-spectrum characteristics of CDMA

Most modulation schemes try to minimize the bandwidth of this signal since bandwidth is a limited resource. However, spread spectrum techniques use a transmission bandwidth that is several orders of magnitude greater than the minimum required signal bandwidth. One of the initial reasons for doing this was military applications including guidance and communication systems. These systems were designed using spread spectrum because of its security and resistance to jamming. Asynchronous CDMA has some level of privacy built in because the signal is spread using a pseudo-random code; this code makes the spread spectrum signals appear random or have noise-like properties. A receiver cannot demodulate this transmission without knowledge of the pseudo-random sequence used to encode the data. CDMA is also resistant to jamming. A jamming signal only has a finite amount of power available to jam the signal. The jammer can either spread its energy over the entire bandwidth of the signal or jam only part of the entire signal.

CDMA can also effectively reject narrow band interference. Since narrow band interference affects only a small portion of the spread spectrum signal, it can easily be removed through notch filtering without much loss of information. Convolution encoding and interleaving can be used to assist in recovering this lost data. CDMA signals are also resistant to multipath fading. Since the spread spectrum signal occupies a large bandwidth only a small portion of this will undergo fading due to multipath at any given time. Like the narrow band interference this will result in only a small loss of data and can be overcome.

Another reason CDMA is resistant to multipath interference is because the delayed versions of the transmitted pseudo-random codes will have poor correlation with the original pseudo-random code, and will thus appear as another user, which is ignored at the receiver. In other words, as long as the multipath channel induces at least one chip of delay, the multipath signals will arrive at the receiver such that they are shifted in time by at least one chip from the intended signal. The correlation properties of the pseudo-random codes are such that this slight delay causes the multipath to appear uncorrelated with the intended signal, and it is thus ignored.

Lecture 6: Precision measurements in astrophysics

Pulsars as astrophysical sources of periodic pulses. Physics of pulsars. Pulsars in double star systems. Drift of perimetrium and General relativity tests. Radiation of gravitational waves. Search for drift of the fine structure constant.

Methods of precision time and frequency measurements and clock synchronization open new opportunities for cosmological studies, physics of intrastellar media, study of exaplanetory orbits, etc. In this lecture we will consider a few implementations and examples.

6.1 Pulsars and Frequency Standards

In 1967 researchers discovered cosmic objects emitting periodic radio-signals which attracted a lot of interest from the community. These objects were called "pulsars ". They emit very broad (from radio frequencies to γ-rays) periodic pulses with intervals from one milliseconds to a few seconds. The number of discovered objects reached more than 1000 in 1998. It was found out that the time interval τ between pulses was very stable ($\Delta\tau/\tau \approx 10^{-3}$), the pulses should been emitted by solid bodies. One can assume a spinning body with a radiofrequency source fixed on its surface with a narrow angular emission pattern. The emitting cone periodically scans Earth similar to the light projector in lighthouse Fig. 6.1.

For a rigid spinning body on can get a restriction coming from the fact, that the linear speed on it's surface cannot exceed the speed of light c. If a pulsar rotates with a period of 1 ms, its radius R cannot exceed 50 km. Pulsar PSR B1937+21 has a rotation period of 1.6 ms. It is very improbable that pulsars with shorter period will be discovered. It is due to another relation, connecting the centripetal acceleration and gravitational force at its surface. Using the relation one can set a restriction to the highest angular velocity $\Omega = \sqrt{GM/R^3}$, where G is the gravitational constant, R is the radius and $M = 4\pi R^3\rho/3$ is the mass of the body. If we substitute the highest known density ρ which is the neutron star density $\rho \approx 10^{17}\,\mathrm{kg/m^3}$ we get the minimal rotation period of 1,2 ms. It is considered, that pulsars are the rotating neutron

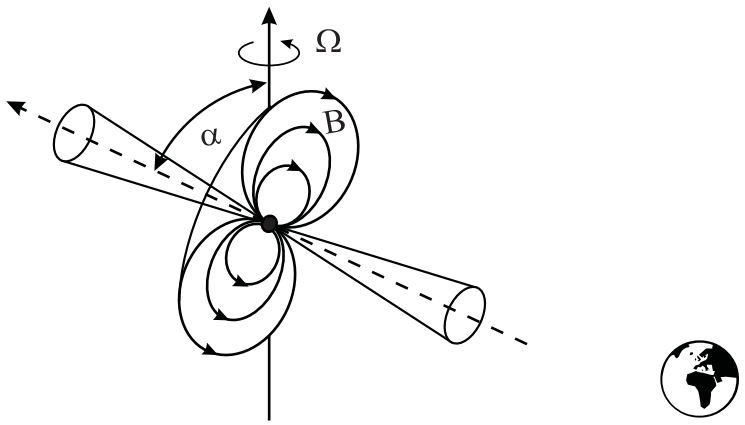


Figure 6.1: The model of a pulsar. The cone of radiation crosses Earth with a given period.

stars.

Neutron star is the star which has burnt its nuclear fuel. Stars with masses in the range $5M_\odot \leq M \leq 10M_\odot$ ($M_\odot$ the solar mass) can turn in neutron stars. Typically, there is an equilibrium between the gravitational forces compressing a star and the radiation pressure. When star is burnt out, the radiation pressure reduces and it turns into the supernova with corona expanding over a large volume. Temperature of the remaining matter becomes so high, that proton react with electrons turning into neutrons and neutrinos ($p^+ + e^- \rightarrow n + \nu$). The remaining matter consists of neutrons n forming the neutron star. If the original star had a magnetic field, its strength will significantly grow after the collapse. Assume, that the the radius of initial star was $R_i \approx 7 \cdot 10^8$ m before collapse and become $R_f \approx 5 \cdot 10^4$ m after the collapse. Due to the conservation of magnetic flow we get $B_i 4\pi R_i^2 = B_f 4\pi R_f^2$, which corresponds to the growth of magnetic field strength for 8 orders of magnitude. It can reach the level of $B_f = 10^8$ T. There are pulsars with the magnetic field on its surface of up to $8 \cdot 10^{10}$ T called "magnetars".

Periodicity of the radiation emitted by pulsars can be explained by the model described in Fig. 6.1. Since a neutron star is rotating with the angular velocity Ω, charged particles accelerate along the magnetic field lines in the star's magnitosphere. Radiation is preferably emitted close to the magnetic poles of the neutron star in narrow conic space volumes with the axes coinciding with the magnetic axis of the star.

The magnetic axis of the star does not coincide with its rotation axis, which means that the radiation of the pulsar periodically crosses the observer's position. It means that the period of pulses detected from the neutron star should coincide with its rotation period. Such pulse sources are typically observed in a radio-frequency region from a few hundreds of megahertz to a few gi-

gahertz. Although the total power emitted by the pulsar is huge, the power which reaches the Earth is very small and can be detected only with very sensitive devices. Usually, the spectral density of the pulsar reaching the Earth in in the range of $10^{-29}\,\mathrm{W\,m^{-2}\,Hz^{-1}}$ to $10^{-27}\,\mathrm{W\,m^{-2}\,Hz^{-1}}$ for the detection band around 400 MHz. Typically, it is not possible to detect the pulsed directly because of poor signal/noise ratio. But, knowing that the pulses are emitted periodically, one can use regular methods of phase detection which allows to measure signals coming from pulsars. Digitized signal from the telescope is accumulated in different time windows corresponding to the expected signal period. It was discovered, that each pulsar possesses its own characteristic pulse envelope (averaged over many periods) as shown in fig. 6.2.

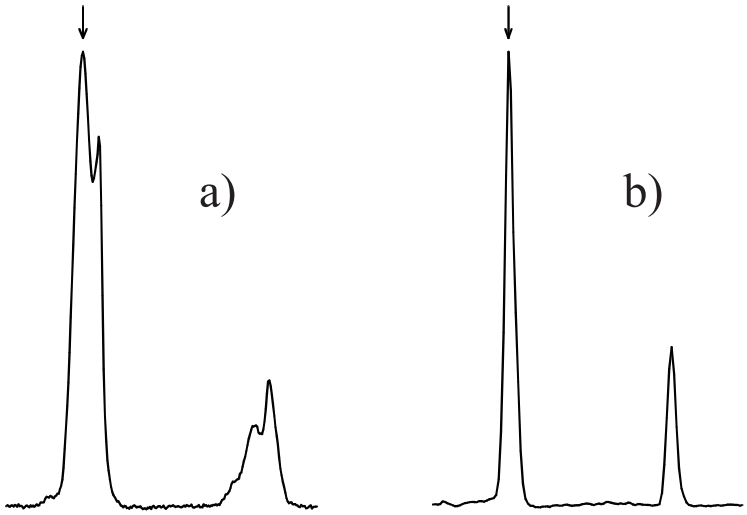


Figure 6.2: *Averaged envelopes for the pulsars PSR B1855+09 and PSR B1937+21 measured at frequencies 1.4 GHz and 2.4 GHz. Such envelopes are the "fingerprints" of each of the pulsar.*

About 3% of pulsars have an additional small pulse which falls approximately to the center of the period of the main pulse (see fig. 6.2). Such signal structure may be explained by the fact, that the observer on the Earth may receive the signal from both poles of the pulsar.

Pulsars can be distinguished in two main groups. The first group of "slow" or "regular" pulsars contains most of the pulsars. Pulsars, belonging to this group have the period P in the range (33 ms $< P <$ 5 s). The period of these pulsars is continuously growing with the typical rate of $\dot{P} \approx 10^{-15}\,\mathrm{s/s}$. The second group is referred to as "millisecond pulsars" and includes pulsars with the period from 1,5 ms to 30 ms. The period of these pulsars changes much slower, down to $\dot{P} \approx 10^{-19}\,\mathrm{s/s}$.

There is another difference between "slow" and "millisecond" pulsars - they have different age (10^5 years $< \tau < 10^9$ years and 10^9 years respectively)

and different strengthes of magnetic field on their surface ($B \approx 10^8\,\mathrm{T}$ and $B \approx 10^4\,\mathrm{T}$). About 80% of millisecond pulsars have orbital twins, while for the slow pulsars the fraction of double-stars (pusar+twin) is much less (1%).

Relying on the accepted model describing a pulsar for interpretation of experimental data, the pulsar can be treated as a huge classical magnetic dipole M. The magnetic moment is rotating with the angular frequency Ω and the angle between the magnetic moment and the rotation axis is α. In frames of classical electro-dynamics, the rotating magnetic dipole emits radiation with the power equal to

$$\frac{dE}{dt} = \frac{2(M\sin\alpha)^2\Omega^4}{3c^2}\,. \tag{6.1}$$

Emitted power results in deceleration of rotation (the power is taken from rotation energy)

$$E_{\mathrm{rot}} = \frac{1}{2}\Theta\Omega^2\,, \tag{6.2}$$

where Θ is a moment of inertia of the neutron star. For the sphere of radius $R \approx 15\,\mathrm{km}$ and density of $\rho \approx 10^{17}\,\mathrm{kg/m^3}$ it equals to $\Theta = 2/5MR^5 = 8/15\pi\rho R^5 \approx 1,3\cdot 10^{38}\,\mathrm{kg\,m^3}$. The rate of the energy loss can be calculated from the angular velocity of the pulsar $\Omega = 2\pi/P$ and its derivative $\dot{\omega} = -2\pi\dot{P}/P^2$:

$$\frac{dE_{\mathrm{rot}}}{dt} = \Theta\Omega\dot{\Omega} = -4\pi^2\Theta\frac{\dot{P}}{P^2}\,. \tag{6.3}$$

For most of the slow pulsars the rate is in the range $10^{23}\ \mathrm{W} \leq \dot{E}_{\mathrm{rot}} \leq 10^{26}\ \mathrm{W}$. The upper limit corresponds to the power emitted by our Sun due to nuclear fusion. Comparing the rotational energy losses (6.3) and full energy emitted by magnetic dipole (6.1) we get

$$\dot{\Omega} = \frac{2(M\sin\alpha)^2}{3\Theta c^3}\Omega^3\,. \tag{6.4}$$

Using this equation, and evaluating the pulsar's magnetic moment one can evaluate the magnetic field on it's surface $B \propto \sqrt{P\dot{P}}$.

6.1.1 Pulsar chronometry

For measuring parameters which define the pulsar properties one should take into account parameters influencing the signals received by antennas on the Earth. First of all, one should take into account the Earth's rotation. Typically, the baricentric (with the Sun at origin) system is used. There are a lot of corrections which should be taken into account if the signal is measured on the Earth: interstellar medium dispersion, General relativity corrections including time dilation, gravitational red shift, Saphiro delay, etc. Pulsar timing us used for pulsar time scale which possesses a very good long-term stability down to 10^{-15} in 3 years averaging time.

The quality of pulsar time scale is deteriorated by some unexpected frequency jumps which take place occasionally - so called *glitches*. After the glitch the pulsar period changes. The glitch can be treated as an neutron star "earthquake" changing the moment of inertia of the star and its angular velocity.

A pulsar is a unique object, since it can be considered as a distant clock in a strong gravitational field. Even more interesting is the case of so-called "binary pulsar", when the pulsar rotates in the system of another star. Such an object allows to make sensitive tests of General relativity theory. For this research Hulse and Taylor were awarded a Nobel Prize in 1993. They studied a double pulsar 1913+16 consisting of a neutron star and its twin.

Their research allowed to make a high-sensitive test of relativity theory which is 4 orders of magnitude more sensitive than the prominent test based on the perithelium precession of Mercury. As will be shown in the next section, the rotation period changes with the rate of $\dot{P} \approx -3 \cdot 10^{-12}$ which may be due to the emission of gravitational waves.

6.2 Binary pulsars

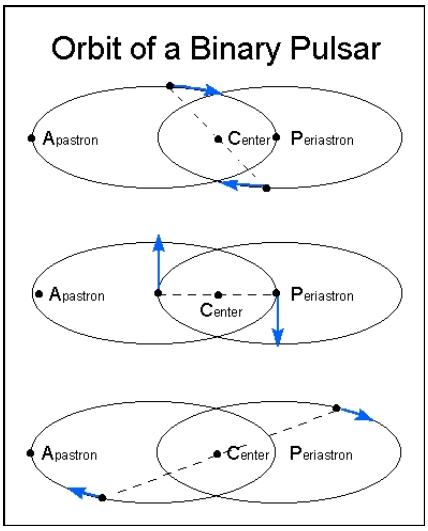


Figure 6.3: *Orbit of binary pulsar*

The pulsar and its companion both follow elliptical orbits (Fig. 6.3) around their common center of mass. Each star moves in its orbit according to Kepler's Laws; at all times the two stars are found on opposite sides of a line passing through the center of mass. The period of the orbital motion is 7.75 hours, and the stars are believed to be nearly equal in mass, about 1.4 solar masses.

As shown in the figure here, the orbits are quite eccentric. The minimum separation at periastron is about 1.1 solar radii; the maximum separation at apastron is 4.8 solar radii. In the case of PSR 1913+16, the orbit is inclined at about 45 degrees with respect to the plane of the sky, and it is oriented such that periastron occurs nearly perpendicular to our line of sight.

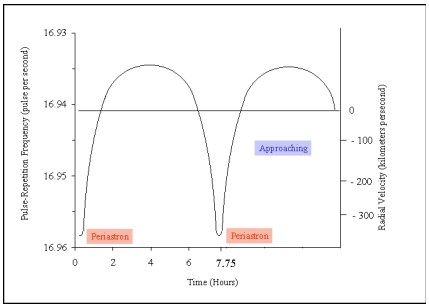


Figure 6.4: *Pulse repetition frequency of a binary pulsar*

The pulse repetition frequency, that is, the number of pulses received each second, can be used to infer the radial velocity of the pulsar as it moves through its orbit. When the pulsar is moving towards us and is close to its periastron, the pulses should come closer together; therefore, more will be received per second and the pulse repetition rate will be highest (Fig. 6.4). When it is moving away from us near its apastron, the pulses should be more spread out and fewer should be detected per second.

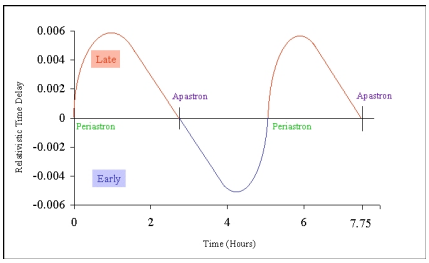


Figure 6.5: *Time ticking in a binary pulsar system.*

When they are closer together, near apastron, the gravitational field is stronger, so that the passage of time is slowed down – the time between pulses (ticks) lengthens just as Einstein predicted. The pulsar clock is slowed down when it is travelling fastest and in the strongest part of the gravitational field; it regains time when it is travelling more slowly and in the weakest part of the field (Fig. 6.5). The orbit of the pulsar appears to rotate with time; in the diagram (Fig. 6.6), notice that the orbit is not a closed ellipse, but a continuous elliptical arc whose point of closest approach (periastron) rotates with each

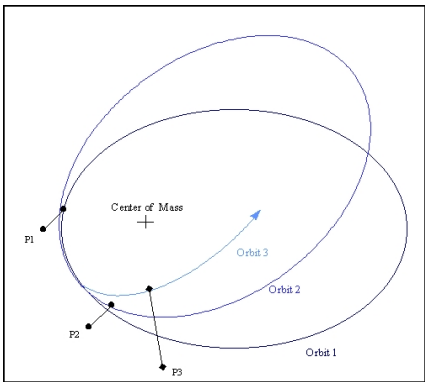


Figure 6.6: *Precession of the pulsar orbit.*

orbit. The rotation of the pulsar's periastron is analogous to the advance of the perihelion of Mercury in its orbit. The observed advance for PSR 1913+16 is about 4.2 degrees per year; the pulsar's periastron advances in a single day by the same amount as Mercury's perihelion advances in a century.

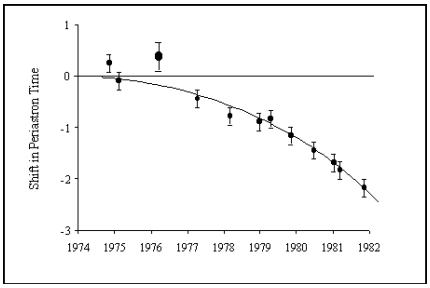


Figure 6.7: *Precession of the pulsar orbit.*

In 1983, Taylor and collaborators reported that there was a systematic shift in the observed time of periastron relative to that expected if the orbital separation remained constant (Fig. 6.7). In the diagram shown here, data taken in the first decade after the discovery showed a decrease in the orbital period as reported by Taylor and his colleagues of about 76 millionths of a second per year. By 1982, the pulsar was arriving at its periastron more than a second earlier than would have been expected if the orbit had remained constant since 1974

6.3 White dwarfs

Masses of white dwarfs are on the order of Sun mass, but their size is much smaller $R << 0.01 R_{\odot}$, which means, that their density is very high and each

cubic centimeter of the white dwarf's matter weights many tons $\rho \sim 10^5 - 10^9\,\mathrm{g/cm^3}$. At such densities electron shells in atoms are destroyed and the matter consists of electron-nuclei plasma. Sine electrons are fermions, the electronic component is the degenerate electronic gas. Pressure P of such gas depends on the density:

$$P = K_1 \rho^{5/3}\,, \tag{6.5}$$

where K_1 is the constant and ρ is the gas density. Contrary to the Clapeyron's equation (equation of state of ideal gas), the degenerative electronic gas pressure does not depend on temperature (Fig. 6.8).

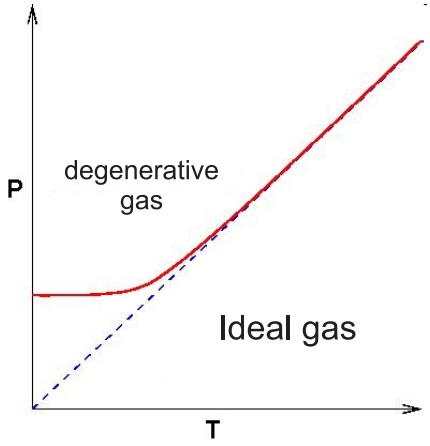


Figure 6.8: *Equations of states for ideal and degenerative gases.*

Equation (6.5) is valid only for a non-relativistic electron gas. Since the Fermi energy is very large, and the relation $kT << E_F$ is valid, the gas remains degenerative even at very high temperatures. Since two electrons cannot be at the same quantum state according to Pauli's principle (energy and the momentum cannot be the same), electrons in the white dwarf grow so much that the gas becomes a relativistic. For the relativistic electron gas the dependency differs from (6.5):

$$P = K_2 \rho^{4/3}\,. \tag{6.6}$$

The averaged density of the white dwarf equals $\rho \sim M/R^3$, where M is its mass and R – its radius. Pressure will be proportional to $P \sim M^{4/3}/R^4$ and its gradient inside the star will be given by

$$\frac{P}{R} \sim \frac{M^{4/3}}{R^5} \tag{6.7}$$

Gravitational force, acting against this pressure can be written as :

$$\frac{\rho G M}{R^2} \sim \frac{M^2}{R^5}\,. \tag{6.8}$$

Although eqns. (6.7) and (6.8) show similar dependency on radius of the star, the forces differently depend on the mass: As $\sim M^{4/3}$ and $\sim M^2$ respectively. Due to this fact, there is a certain mass of the star when they balance and, since gravitational force stronger depends on the mass, than the electronic gas pressure difference, the radius of the white dwarf decreases with its mass (see Fig. 6.9). If the mass overcomes a certain limit, the star will collapse and will turn into a neutron star. The limit calls a "Chandrasekhar limit".

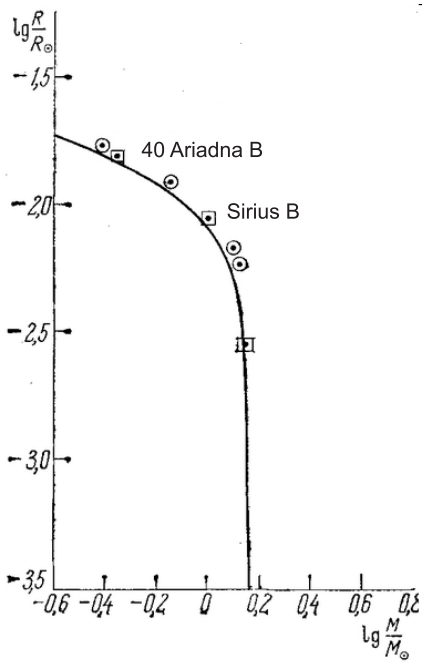


Figure 6.9: *Precession of the pulsar orbit.*

6.4 Introduction to gravitational waves

The effects of a passing gravitational wave can be visualized by imagining a perfectly flat region of spacetime with a group of motionless test particles lying in a plane. As a gravitational wave passes through the particles along a line perpendicular to the plane of the particles (i.e. following your line of vision into the screen Fig. 6.10), the particles will follow the distortion in spacetime, oscillating in a "cruciform" manner. The area enclosed by the test particles does not change and there is no motion along the direction of propagation.

Gravitational wave has a very small amplitude (as formulated in linearized gravity). However they enable us to visualize the kind of oscillations associated with gravitational waves as produced for example by a pair of masses in a circular orbit. In this case the amplitude of the gravitational wave is a constant,

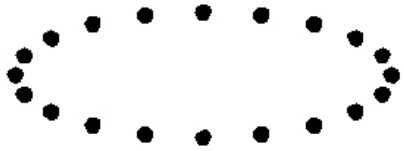


Figure 6.10: *Effect of gravitation wave on a ring of particles. The wave passes orthogonal to the screen.*

but its plane of polarization changes or rotates at twice the orbital rate and so the time-varying gravitational wave size (or 'periodic spacetime strain') exhibits a variation as shown in Fig. 6.10.

Power radiated by orbiting bodies. Gravitational waves carry energy away from their sources and, in the case of orbiting bodies, this is associated with an inspiral or decrease in orbit. Imagine for example a simple system of two masses such as the Earth-Sun system moving slowly compared to the speed of light in circular orbits. Assume that these two masses orbit each other in a circular orbit in the x-y plane. To a good approximation, the masses follow simple Keplerian orbits. However, such an orbit represents a changing quadrupole moment. That is, the system will give off gravitational waves.

Suppose that the two masses are m_1 and m_2, and they are separated by a distance r. The power given off (radiated) by this system is:

$$P = \frac{dE}{dt} = -\frac{32}{5}\,\frac{G^4}{c^5}\,\frac{(m_1 m_2)^2(m_1+m_2)}{r^5}\,, \tag{6.9}$$

where G is the gravitational constant, c is the speed of light in vacuum and where the negative sign means that power is being given off by the system, rather than received. For a system like the Sun and Earth, r is about 1.5×10^{11} m and m_1 and m_2 are about 2×10^{30} and 6×10^{24} kg respectively. In this case, the power is about 200 watts. This is truly tiny compared to the total electromagnetic radiation given off by the Sun (roughly 3.86×10^{26} W).

In theory, the loss of energy through gravitational radiation could eventually drop the Earth into the Sun. However, the total energy of the Earth orbiting the Sun (kinetic energy plus gravitational potential energy) is about

1.14×10^{36} joules of which only 200 joules per second is lost through gravitational radiation, leading to a decay in the orbit by about 10^{-15} meters per day or roughly the diameter of a proton. At this rate, it would take the Earth approximately 1×10^{13} times more than the current age of the Universe to spiral onto the Sun.

Wave amplitudes from the EarthSun system. We can also think in terms of the amplitude of the wave from a system in circular orbits. Let θ be the angle between the perpendicular to the plane of the orbit and the line of sight of the observer. Suppose that an observer is outside the system at a distance R from its center of mass. If R is much greater than a wavelength, the two polarizations of the wave will be

$$h_+ = -\frac{1}{R}\frac{G^2}{c^4}\frac{2m_1m_2}{r}(1+\cos^2\theta)\cos\left[2\omega(t-R)\right] \tag{6.10}$$

$$h_\times = -\frac{1}{R}\frac{G^2}{c^4}\frac{4m_1m_2}{r}(\cos\theta)\sin\left[2\omega(t-R)\right]. \tag{6.11}$$

Here, we use the constant angular velocity of a circular orbit in Newtonian physics: $\omega = \sqrt{G(m_1+m_2)/r^3}$.

For example, if the observer is in the x-y plane then $\theta = \pi/2$, and $\cos(\theta) = 0$, so the $h_\times$ polarization is always zero. We also see that the frequency of the wave given off is twice the rotation frequency. If we put in numbers for the Earth-Sun system, we find:

$$h_+ = -\frac{1}{R}\frac{G^2}{c^4}\frac{4m_1m_2}{r} = -\frac{1}{R}\,1.7\times10^{-10}\,\text{meters}. \tag{6.12}$$

In this case, the minimum distance to find waves is $R > 1$ light-year, so typical amplitudes will be $h \sim 10^{-26}$. That is, a ring of particles would stretch or squeeze by just one part in 10^{26}. This is well under the detectability limit of all conceivable detectors.

6.4.1 Very Large Baseline Interferometry

Radio-astronomy played an important role in the study of astrophysical objects providing an essential information about our Universe. The smallest angle between two objects θ resolvable by a telescope equals to

$$\theta = \alpha\frac{\lambda}{b}. \tag{6.13}$$

Diffraction on the aperture b limits he resolution of a telescope. The constant α is on the order of 1 and depends on the telescope's shape and its illumination.

To reach resolution demanded by today's goals, the size of radio-telescope (radio-telescopes typically work in the range from 1 cm to 1 m) should be so big that it cannot be implemented on practice.

Diffraction limit (6.13) results from interference of different partial waves on different parts of the telescope's aperture. A resolution can be increased if signals from two different telescopes are combined with proper phase relation. Mutual correlation of signals from two different receivers will give an interference pattern. After proper analysis it allows to recover an exact position of the astrophysical object and its shape. E.g. the system VLA (Very Large Array) combines 27 antennas withe the maximal separation od 36 km. Its resolution at 43 GHz reaches 0.04 arc. sec.

In other case, telescope can cover different continents (VLBI) as shown in Fig. 6.11. For such large distances the physical combination of signals in real

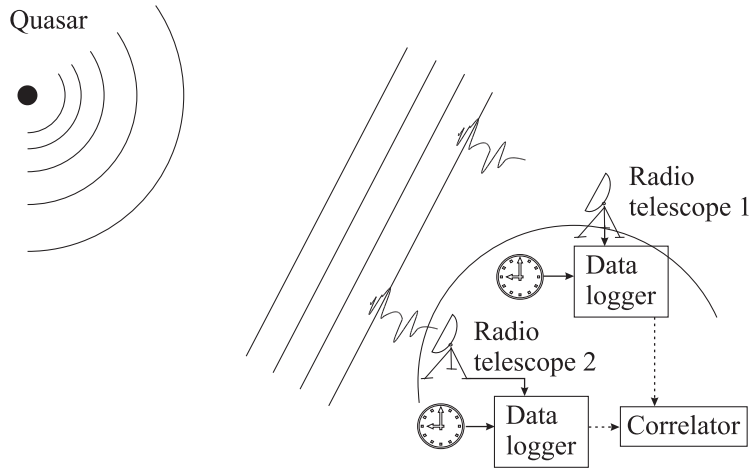


Figure 6.11: *Interferometry with a large baseline (VLBI).*

time is not possible. Instead, signals are recorded simultaneously together with time marks from synchronized clocks. After correction for the Doppler shift data are analyzed using correlator. One can consider VLBI operation principle as measuring time delay in signal arriving by two telescopes at large distance.

VLBI allows for a very accurate stellar coordinate system (1 arc sec using quasars as the reference). It is very useful for defining the position of astrophysical objects and description of Earth rotation. Measuring the relative position of antennas allows for measuring relative velocity of continents.

The largest on Earth interferometer base is limited by the Earth diameter 12 750 km. In 1997 Japan launched a satellite HALCA with 8-m on-board antenna. The orbit allowed for the base line of 30 000 km. Mission VSOP allowed to measure objects with resolution of 10^{-3} arc. sec. at 5 GHz frequency.

6.5 Search for drift of the fine structure constant

Lecture 7: Two levels atomic system and frequency standards

Optical Bloch equations. Pseudospin. Rabi oscillations. Excitation by sequence of coherent pulses. Ramsey method. Atomic interferometry. Microwave frequency standards. Hydrogen maser. Cesium beam apparatus. Allan deviation, stability, accuracy.

7.1 Two-level system

Description of any atomic system is usually started from two-level model. Most of the important processes like atomic level excitation, laser cooling, etc., are adequately described by this simplified model. Although most of the real atomic systems are very complex and have multiple levels, the two-level atom is still a very important model which significantly helps understanding of physical processes.

Consider a system with two levels with the energies E_1 and E_2, $E_2 > E_1$. The lower state is usually called *the ground state* and the upper state – *excited state*.

If two levels are coupled by an external field, the Hamiltonian can be written as

$$\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_{\text{int}} \,. \tag{7.1}$$

Here $\mathcal{H}_0$ presents the system itself, while interaction is given by $\mathcal{H}_{\text{int}}$. For the Hamiltonian $\mathcal{H}_0$ one can write the time-independent Schrödinger equation

$$\mathcal{H}_0 \phi_k(\vec{r}) = E_k \phi_k(\vec{r}) \,, \tag{7.2}$$

without taking into account the spontaneous emission. Here $k = 1, 2$, and $\vec{r}$ describes all internal degrees of freedom (electron positions, spins). External field cases perturbation $\mathcal{H}_{\text{int}}$. Since eigenfunctions $\phi_k(\vec{r})$ form a full set , the general solution $\psi(\vec{r}, t)$ of (7.8) can be given by their linear combination:

$$\psi(\vec{r}, t) = c_1(t) e^{-iE_1 t/\hbar} \phi_1(\vec{r}) + c_2(t) e^{-iE_2 t/\hbar} \phi_2(\vec{r}) \,. \tag{7.3}$$

The coefficients $c_k(t)$ are time dependent which results from interaction $\mathcal{H}_{\text{int}}$. The explicit form of the Hamiltonian describing interaction with an electromagnetic field can be obtained from power series of interaction of charged particle with an external field.

If a charged particle with mass m, charge q and coordinate $\vec{r}$ is placed in the field with a vector potential $A(\vec{r},t)$. If the wavelength is much larger compared to the atom size, the interaction can be presented as power series. The most important interaction Hamiltonian forms are:

$$\mathcal{H}_{\text{int}} = -\vec{d}\cdot\vec{E}(\vec{r}_0,t) = +q\vec{r}\cdot\vec{E} \text{ (electric dipole interaction)}, \quad (7.4)$$

$$\mathcal{H}_{\text{int}} = -\vec{\mu}\cdot\vec{B}(\vec{r}_0,t) \quad \text{(magnetic dipole interaction)}, \quad (7.5)$$

$$\mathcal{H}_{\text{int}} = \frac{q}{2}\vec{r}\cdot\vec{r}\cdot\vec{\nabla}_{r0}\vec{E}(\vec{r}_0,t) \text{ (electric quadruple interaction)}, \quad (7.6)$$

where $\vec{d}$ and $\vec{\mu}$ are the quantum mechanical electric dipole and magnetic operators, correspondingly.

The electric dipole moment equals $\vec{d} = q\vec{r} = e\vec{r}$ ($e = 1,602\cdot10^{-19}$ A s. The electric dipole interaction couples atomic levels with opposite parity (e.g. S-P, P-D) if corresponding selection rules are fulfilled.

The magnetic dipole interaction describes interaction of atomic magnetic moment $\vec{\mu}$ with magnetic field $\vec{B}(\vec{r}_0,t)$ of the electromagnetic field. Magnetic dipole transitions can be excited between levels of the same parity, most frequently used are magnetic transitions between ground-state sublevels in alkali atoms (Cs and Rb microwave standards) and H-maser.

Weak electric quadrupole transitions are widely used in optical frequency standards and can couple, e.g. S and D levels.

Here we will consider electric dipole interaction. External electric field tries to separate positive and negative charges, polarizing a particle. We assume that the induced dipole is parallel to the electric field of the plane-polarized wave $\vec{E}(\vec{r}_0,t) = E_0\vec{\epsilon}\cos(\omega t)$.

7.2 Optical Bloch equations

Evolution of a two-level system can be nicely presented by a geometric picture suggested by Feinmann, Wernon and Hellwarth. One can re-write equation (7.3) introducing coefficients $C_{1,2}(t)$:

$$\psi(\vec{r},t) = C_1(t)\phi_1(\vec{r}) + C_2(t)\phi_2(\vec{r})\,. \quad (7.7)$$

Here we merged the fast oscillating part directly in the coefficients. Substituting this expression in time-dependent Schrödinger equation

$$\mathcal{H}\psi = i\hbar\frac{\partial\psi(t)}{\partial t} \quad (7.8)$$

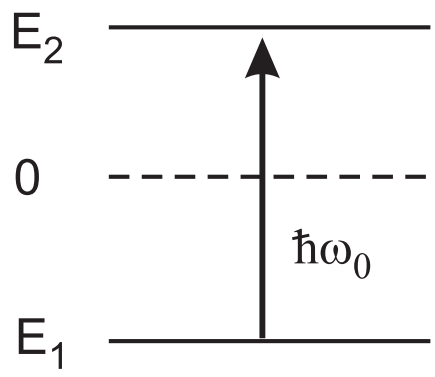


Figure 7.1: A two-level system.

will give is the following set of equations:

$$\frac{dC_1(t)}{dt} = +i\frac{\omega_0}{2}C_1(t) - \frac{i}{\hbar}C_2(t)H_{12}(t) \tag{7.9}$$

$$\frac{dC_2(t)}{dt} = -i\frac{\omega_0}{2}C_2(t) - \frac{i}{\hbar}C_1(t)H_{21}(t)\,. \tag{7.10}$$

Here $H_{21} \equiv \int \phi_2^* \mathcal{H}_{\rm int}\phi_1 d^3r$, $H_{12} \equiv \int \phi_1^* \mathcal{H}_{\rm int}\phi_2 d^3r = H_{21}^*$ and $\hbar\omega_0 = E_2 - E_1$. The zero energy is taken as $(E_1 + E_2)/2 = 0$, which will give us $E_2 = \hbar\omega_0/2$ and $E_2 = -\hbar\omega_0/2$. The sketch is shown in Fig. 7.1.

Feinmann used three real functions using coefficients $C_1(t)$ and $C_2(t)$:

$$R_1'(t) \equiv C_2(t)C_1^*(t) + C_2^*(t)C_1(t) \tag{7.11}$$

$$R_2'(t) \equiv i[C_2(t)C_1^*(t) - C_2^*(t)C_1(t)] \tag{7.12}$$

$$R_3'(t) \equiv C_2(t)C_2^*(t) - C_1(t)C_1^*(t)\,, \tag{7.13}$$

which and form a vector in the 3D space: $\vec{R}'(t) = (R_1'(t), R_2'(t), R_3'(t))$. This vector is usually formally treated as a *pseudospin* vector. From (7.11) – (7.13) one can get a relation (the total population of the system equals 1)

$$R_1'^{\,2}(t)+R_2'^{\,2}(t)+R_3'^{\,2}(t) = [C_2(t)C_2^*(t)+C_1(t)C_1^*(t)]^2 = \left(|c_2(t)|^2 + |c_1(t)|^2\right)^2 = 1\,. \tag{7.14}$$

The pseudospin length is constant and equals 1: $|\vec{R}'(t)|^2 = 1$. Its end moves along some trajectory on so-called *Bloch* sphere of unit radius.

To understand the pseudospin components we will write equations describing its evolution. For example, the derivative $dR_1'(t)/dt$ can be calculated form (7.11):

$$\frac{dR_1'(t)}{dt} = \frac{dC_2(t)}{dt}\,C_1^*(t) + C_2(t)\frac{dC_1^*(t)}{dt} + \frac{dC_2^*(t)}{dt}\,C_1(t) + C_2^*(t)\frac{dC_1(t)}{dt}\,. \tag{7.15}$$

Using (7.9) and (7.10), as well as (7.12), (7.13) we get:

$$\begin{aligned}\frac{dR_1'(t)}{dt} &= \frac{1}{i\hbar}\frac{\hbar\omega_0}{2}C_2C_1^* + \frac{1}{i\hbar}C_1C_1^*H_{21} + \frac{1}{i\hbar}\frac{\hbar\omega_0}{2}C_2C_1^* - \frac{1}{i\hbar}C_2C_2^*H_{12}^* \\ &- \frac{1}{i\hbar}\frac{\hbar\omega_0}{2}C_2^*C_1 - \frac{1}{i\hbar}C_1C_1^*H_{21}^* - \frac{1}{i\hbar}\frac{\hbar\omega_0}{2}C_2^*C_1 + \frac{1}{i\hbar}C_2C_2^*H_{12} \\ &= \frac{2\omega_0}{2i}(C_2C_1^* - C_2^*C_1) + \frac{1}{i\hbar}C_1C_1^*(H_{21} - H_{21}^*) + \frac{1}{i\hbar}C_2C_2^*(H_{12} - H_{12}^*) \\ &= -\omega_0 R_2'(t) - \frac{2}{\hbar}\mathrm{Im}(H_{21})R_3'(t)\,. \qquad (7.16)\end{aligned}$$

One can get similar equations for $R_2(t)$ and $R_3(t)$. The full equation set called *optical Bloch equations* is written here:

$$\frac{dR_1'(t)}{dt} = -\omega_0 R_2'(t) - \frac{2}{\hbar}\mathrm{Im}(H_{21})R_3'(t) \qquad (7.17)$$

$$\frac{dR_2'(t)}{dt} = +\omega_0 R_1'(t) - \frac{2}{\hbar}\mathrm{Re}(H_{21})R_3'(t) \qquad (7.18)$$

$$\frac{dR_3'(t)}{dt} = +\frac{2}{\hbar}\mathrm{Re}(H_{21})R_2'(t) + \frac{2}{\hbar}\mathrm{Im}(H_{21})R_1'(t)\,. \qquad (7.19)$$

It can be compactly written as:

$$\frac{\vec{R}'(t)}{dt} = \vec{\Omega}' \times \vec{R}'(t)\,, \qquad (7.20)$$

where $\vec{\Omega}'$ is some vector with three real components

$$\vec{\Omega}' \equiv \left(\frac{2}{\hbar}\mathrm{Re}(H_{21}), -\frac{2}{\hbar}\mathrm{Im}(H_{21}), \omega_0\right)\,. \qquad (7.21)$$

Equation (7.20) is very similar to the equation describing dynamics of a spinning body or precession of the spin-1/2 particle in magnetic field. Exactly this fact caused the name *pseudospin*.

The optical Bloch equations describe how a two-level system interacts with an external electromagnetic field. The components R_1' R_2' correspond to the real and imaginary parts of atomic polarization, while the component R_3' is the probability difference to find the system in the upper ϕ_2 or lower ϕ_1 state. In other words it is the population inversion of the system. For an atom in the ground state the Bloch vector is directed down, for the upper state - to the upper pole of the sphere.

This simple picture adequately describes the system evolution only in the case if the precession rate $\vec{R}'$ is much faster compared to the change of vector $\vec{\Omega}'$. For example we consider a regular π-pulse: the resonant pulse of electromagnetic field is applied for the time τ selected such way, that its product to the Rabi frequency equals π: $\Omega_R\tau = \pi$, Fig. 7.2. In usual application the pulse

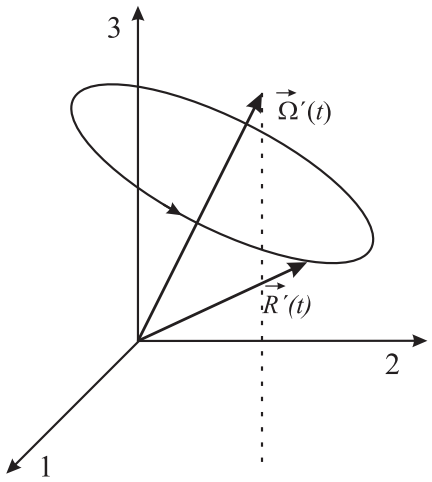


Figure 7.2: The pseudospin vector is precessing around vector $\vec{\Omega}'$.

duration is much larger compared to the period of electromagnetic field and pseudospin rotates many time around the vertical axis of the Bloch sphere.

To get rid of these not informative multiple rotations, the pseudospin vector is usually treated in the rotating frame u, v, w. It rotates with the frequency of electromagnetic field ω around axis 3, the axis w coincides with axis 3.

We also assume, that interaction is electric dipole interaction and $\mathcal{H}_{12} = \mathcal{H}_{21}^* = -\vec{d}\cdot\vec{E}$ (7.4). In this case the equations (7.17) – (7.19) can be re-written as:

$$\frac{dR_1'(t)}{dt} = -\omega_0 R_2'(t) \tag{7.22}$$

$$\frac{dR_2'(t)}{dt} = +\omega_0 R_1'(t) + \frac{2d_r}{\hbar} E_0 \cos\omega t\, R_3'(t) \tag{7.23}$$

$$\frac{dR_3'(t)}{dt} = -\frac{2d_r}{\hbar} E_0 \cos\omega t\, R_2'(t)\,. \tag{7.24}$$

Now we can make the coordinate transformation according to

$$R_1'(t) = u\cos\omega t - v\sin\omega t \tag{7.25}$$

$$R_2'(t) = u\sin\omega t + v\cos\omega t \tag{7.26}$$

$$R_3'(t) = w\,. \tag{7.27}$$

Substitution of (7.26) into (7.22) and replacing the derivative from (7.25) will give us

$$\dot{u}\cos\omega t - \dot{v}\sin\omega t = (\omega-\omega_0)u\sin\omega t + (\omega-\omega_0)v\cos\omega t \tag{7.28}$$

$$\dot{u}\sin\omega t + \dot{v}\cos\omega t = -(\omega-\omega_0)u\cos\omega t + (\omega-\omega_0)v\sin\omega t + \frac{2d_r}{\hbar}E_0\cos\omega t\, w$$

$$\dot{w} = -\frac{2d_r}{\hbar}E_0\cos\omega t\sin\omega t\, u - \frac{2d_r}{\hbar}E_0\cos^2\omega t\, v\,.$$

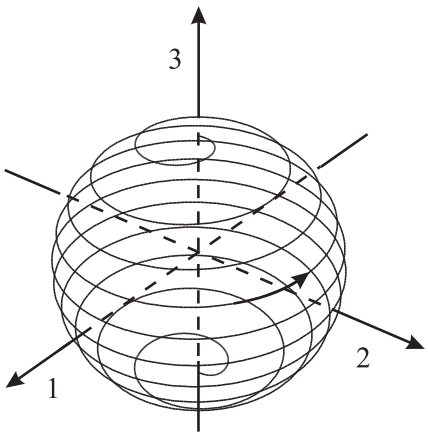


Figure 7.3: Evolution of the Bloch vector $\vec{R}'$ under a resonance π-pulse, applied to an atom initially found in the ground state.

Then, we can multiply (7.28) and (7.29) by $\cos\omega t$ and $\sin\omega t$ correspondingly and add the results. We will get the Bloch equations in the rotating frame:

$$\dot{u} = (\omega - \omega_0)v + \frac{d_r}{\hbar}E_0 \sin 2\omega t\, w \tag{7.29}$$

$$\dot{v} = -(\omega - \omega_0)u + \frac{d_r}{\hbar}E_0(1 + \cos 2\omega t)\, w \tag{7.30}$$

$$\dot{w} = -\frac{d_r}{\hbar}E_0 \sin 2\omega t\, u - \frac{d_r}{\hbar}E_0(1 + \cos 2\omega t)\, v\,. \tag{7.31}$$

It is clear, that equations contain two types of terms - slowly varying a the frequency of $\omega - \omega_0$ and rapidly oscillating at the frequency 2ω. Usually, in optical and radio-frequency regions the detuning is much smaller compared to the frequency and rabidly oscillating terms can be neglected. If we neglect these terms (which is called a *rotating wave approximation*) and introduced the Rabi frequency Ω_R

$$\Omega_R = \frac{eE_0}{\hbar}\int \phi_1^*(\vec{r})\,\vec{r}\cdot\vec{\epsilon}\,\phi_2(\vec{r})d^3r = \frac{E_0 d_r}{\hbar} \tag{7.32}$$

we will finally get the simplified Bloch equations in the rotating frame:

$$\dot{u} = (\omega - \omega_0)v \tag{7.33}$$

$$\dot{v} = -(\omega - \omega_0)u + \Omega_R w \tag{7.34}$$

$$\dot{w} = -\Omega_R v\,. \tag{7.35}$$

Since the frame transformation is the unitary transformation, the length of the pseudospin vector $\vec{R} = (u, v, w)$ remains 1. Similar to (7.20) one can re-write

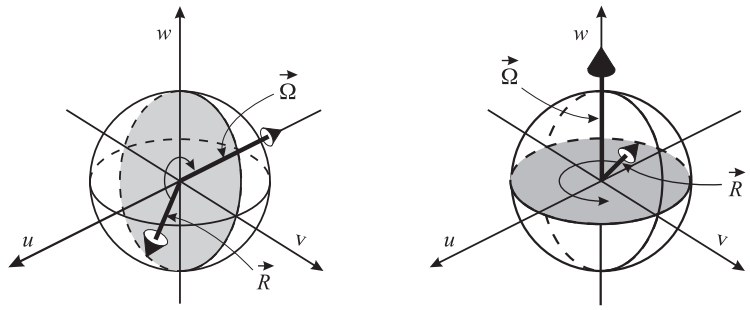


Figure 7.4: a) For zero detuning the pseudo-spin vector precesses in the $v-w$ plane. b) Evolution of pseudo-spin vector if the external field has zero intensity Ω_R, non-zero detuning $\omega_0 - \omega > 0$ and initially the system is prepared in the coherent superposition of two states by $\pi/2$ pulse.

equations in the form of one vector equation

$$\frac{d\vec{R}(t)}{dt} = \vec{\Omega} \times \vec{R}(t) \tag{7.36}$$

with vector

$$\vec{\Omega} = (-\Omega_R, 0, \omega_0 - \omega)\,. \tag{7.37}$$

Representation of excitation using pseudospin vector is very useful if one considers pulse excitations. Figure 7.4 shows two examples for illustration.

The first example shows the situation when the field is tuned exactly with resonance with zero detuning. In this case vector Ω will be directed along $-u$. The pseudospin vector will rotate in the $v-w$ crossing south and north poles. The projection of the vector on w plane will describe regular Rabi oscillations.

Another case shown in Figure is the free evolution of atom excited in the coherent superposition of two states $(\phi_1 + \phi_2)/\sqrt{2}$. In this case $\omega - \omega_0 \neq 0$, but the field intensity equals zero $\Omega_R = 0$. Now the Bloch vector is oscillating in the plane $u - v$ and distribution of population does not change in time.

7.3 Ramsey method

One of the very efficient methods for excitation of atoms, ions and atomic ensembles is a *Ramsey method*. It was suggested by Norman Ramsey. The principle of this method is illustrated in Fig. 7.5. Ramsey method is used in Cs primary frequency standards, atomic fountains, atomic interferometry as well as in many other applications.

The excitation probability resulting from Ramsey excitation can be obtained by different methods which give the same result (i) consideration using

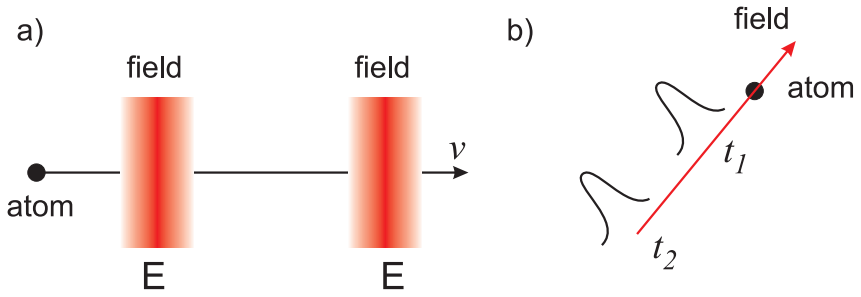


Figure 7.5: Excitation of atom/ion by the Ramsey method. a) Atom with velocity v flies through two consequent interaction zones with the same field. b) Atom at rest is illuminated by two consecutive field pulses at t_1 and t_2.

Bloch equations and Bloch vector representation (ii) using spectral representation (iii) considering atomic interferometry.

7.3.1 Bloch sphere representation

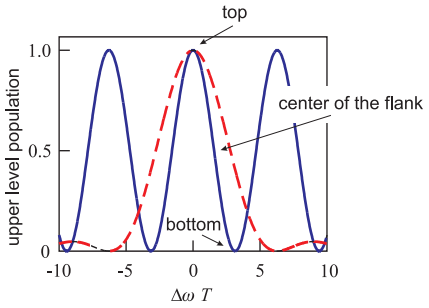


Figure 7.6: *The probability to find atom in the excited state after interaction with two short pulses $\tau \ll T$ following after each other in interval T (solid curve (7.40)). For comparison result of continuous excitation for time T is shown (dashed line).*

Ramsey considered the general case of evolution of atomic system illuminated by two pulses. After the first excitation atom turns in the coherent superposition of states which will freely evolute between pulses. The second pulse will again excite the atom and, dependent on the relative phase of atom itself and the external field the population of the upper state will either further increase or decrease. The probability to fine an atom in the excited state after the second interaction equals :

$$p(\tau+T+\tau) \equiv |c_2(\tau+T+\tau)|^2 \tag{7.38}$$
$$= 4\frac{\Omega_R^2}{\Omega_R'^2}\sin^2\frac{\Omega_R'\tau}{2}\left(\cos\frac{\Omega_R'\tau}{2}\cos\frac{\Delta\omega T}{2} - \frac{\Delta\omega}{\Omega_R'}\sin\frac{\Omega_R'\tau}{2}\sin\frac{\Delta\omega T}{2}\right)^2,$$

where Ω_R is the Rabi frequency. The Ω'_R is the generalized Rabi frequency

$$\Omega'_R \equiv \sqrt{\Omega_R^2 + \Delta\omega^2}\,, \tag{7.39}$$

and $\Delta\omega$ is the frequency detuning. Initially atom is in the ground state. Here T is the time interval between pulses and τ is the pulse duration. Close to the

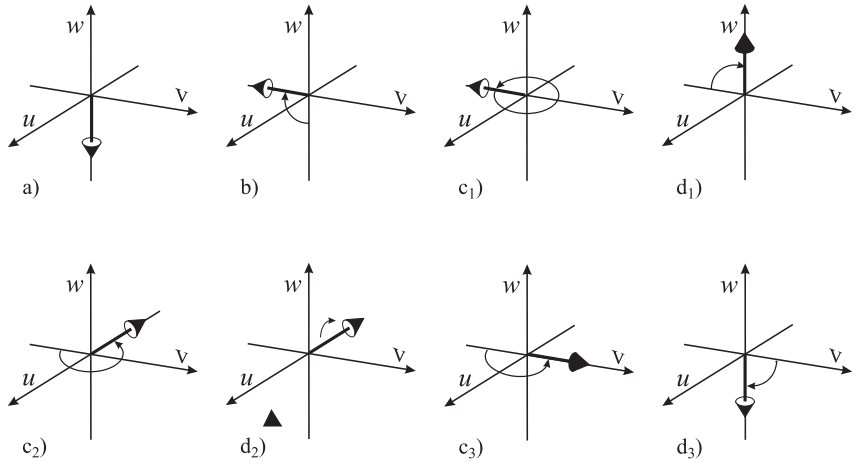


Figure 7.7: *Evolution of the pseudospin vector under excitation by two short $\pi/2$ pulses $\Omega_R\tau \ll \Delta\omega T$ for different T.*
a)-b) Excitation by the first $\pi/2$ pulse.
c_1) free evolution for $\Delta\omega T = 2\pi$ with the d_1) consecutive excitation by the second $\pi/2$ pulse. Atom is excited to the upper state (top of the fringe).
c_2) free evolution for $\Delta\omega T = 3/2\pi$ with the d_2) consecutive excitation by the second $\pi/2$ pulse. Atom is excited to the coherent superposition of states upper state (center of the fringe wing).
c_3) free evolution for $\Delta\omega T = \pi$ with the d_3) consecutive excitation by the second $\pi/2$ pulse. Atom is not excited (bottom of the fringe).

resonance ($\Delta\omega \ll \Omega_R$) one can approximate $\Omega_R \approx \Omega'_R$ and equation (7.38) is simplified:

$$p(\tau + T + \tau) \approx \frac{1}{2}\sin^2\Omega_R\tau[1 + \cos 2\pi(\nu - \nu_0)T]\,. \tag{7.40}$$

The maximal excitation of atom reaches at $\Omega_R\tau = \pi/2$, i.e. by excitation of atom by two consecutive $\pi/2$ pulses. The FWHM of the fringe equals :

$$\Delta\nu = \frac{1}{2T}\,. \tag{7.41}$$

The resolution of the Ramsey method is given by the time between pulses T.

If field in two interaction zones has a certain phase difference $\Delta\Phi$, the interference pattern will be shifted by :

$$\frac{\Delta\nu_\Phi}{\nu_0} = -\frac{\Delta\Phi}{2\pi\nu_0 T}\,. \tag{7.42}$$

For the frequency standards this shift is undesirable, but for many other applications it can be very helpful. This method allows sensitive measurement of the phase shift which can be caused by some external fields.

For simple representation of the Ramsey scheme we can implement the Bloch vector representation shown in Fig. 7.7. Three cases are shown - excitation to the top of the fringe, excitation to the center of the fringe flank and no excitation (fringe bottom) as shown in Fig. 7.6.

7.3.2 Spectral representation

The very similar excitation picture can be obtained in the spectral domain. If we look at the excitation field spectrum, it will consist of narrow fringes fit under the envelope. The fringe width will be given by time interval T, while the envelope width will be reversely proportional to the pulse width τ. An example for two Gaussian pulses is shown in Fig. 7.8. For square pulses the envelope function will be different.

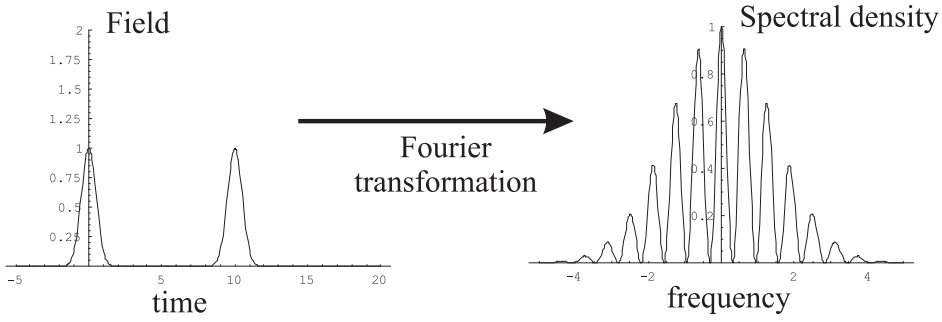


Figure 7.8: *Two Gaussian pulses and their Fourier transformation.*

7.3.3 Atomic interferometry

One can consider the Ramsey scheme also as and atomic interference. The first interaction will split the incoming atomic wave packet in two as shown in Fig.7.9. After free propagation each of the atomic packets will be split again and the interference between packets in the ground state (solid lines) and in the excited state (dashed lines) takes place. Depending on the relative phase of atom and the field either constructive or destructive interference pattern will appear.

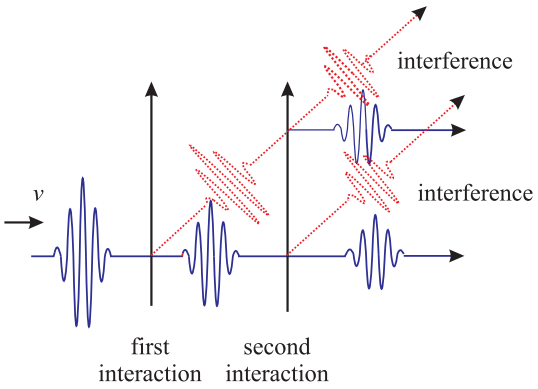


Figure 7.9: *Ramsey scheme treatment from the position of atomic interferometry.*

Similar interferometric methods can be applied in optical region, where atomic interferometers are widely used as sensitive gravimeters and allow tests of fundamental theories. In optical domain the Ramsey setup should be modified because the wave packets will be significantly separated in space due to large optical photon recoil.

All three mentioned treatments are equivalent.

7.4 Microwave frequency standards

7.4.1 Cesium beam clock

Atomic Cs beam clock is the most robust and till now most common primary frequency standard. They are compact, robust and e.g. are installed in the GPS satellites.

The operation principle of the Cs beam clocks is shown in Fig. 7.10. Atoms are emitted from the oven have equal population of all magnetic sublevels. To make Ramsey method feasible one should provide population difference by strong gradient magnetic field. The polarizer magnet deflects atoms in the desired state (e.g. $(F = 3,\, m_F = 0)$) shown in Fig. 7.11.

To address the clock transition ($m_F = 0 \leftrightarrow m'_F = 0$), the magnetic sublevels are split in the external homogenous magnetic field B orthogonal to the drawing plane. Atoms fly through the first interaction region with the microwave field which is excited in the U-shaped resonator. After the free evolution (in commercial clock 20 cm in length, in state-of art clocks up to a few meters) atoms fly through the next interaction zone of the same resonator. Atoms in different states are separated by the polarizer magnet and then detected. The error signal allow to lock external synthesizer (usually a stable quartz oscillator) to the transition.

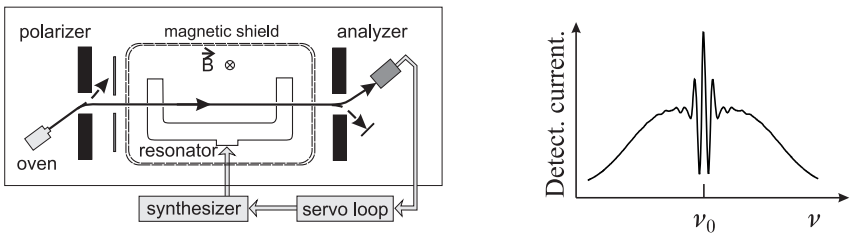


Figure 7.10: *Left: Cs beam clock schematics. Right: signal on the detector when the synthesizer is scanned.*

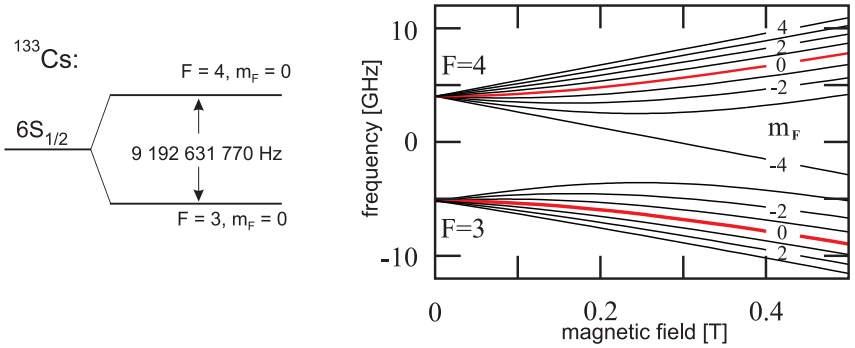


Figure 7.11: *Ground state splitting in Cs and magnetic sublevels shift in external magnetic field.*

Typical short-time instability of the commercial Cs clocks HP5071A ("Agilent") equals 5×10^{-12} for 1 s. The accuracy is limited by magnetic field instability at the level of $10^{-12} - 10^{-13}$.

7.4.2 Cs fountain clock

Laser cooling of atoms allow to significantly reduce velocity of atoms interacting with microwave field. Atom ^{133}Cs has strong cycling transition $6^2S_{1/2}(F = 4) \leftrightarrow 6^2P_{3/2}(F' = 5)$ at 852 nm which allows efficient laser cooling to a few μK regime.

Fig. 7.12 shows schematics of a Cs fountain clock. In contrast to a beam clock, atoms pass the same interaction in the same microwave cavity after a ballistic flight which lasts for approx. 1 s. Up to 10^7 atoms are laser cooled by 6 laser beams. Atoms are prepared in the proper initial state and then are launched by the same laser beams at the velocity of 4 m/s. The radial velocity corresponds to less then 1 cm/s.

Atoms pass two times through the same microwave cavity feeded by signal from the most stable tunable oscillator (special quartz oscillator). After the Ramsey excitation the population of the ground state components is detected

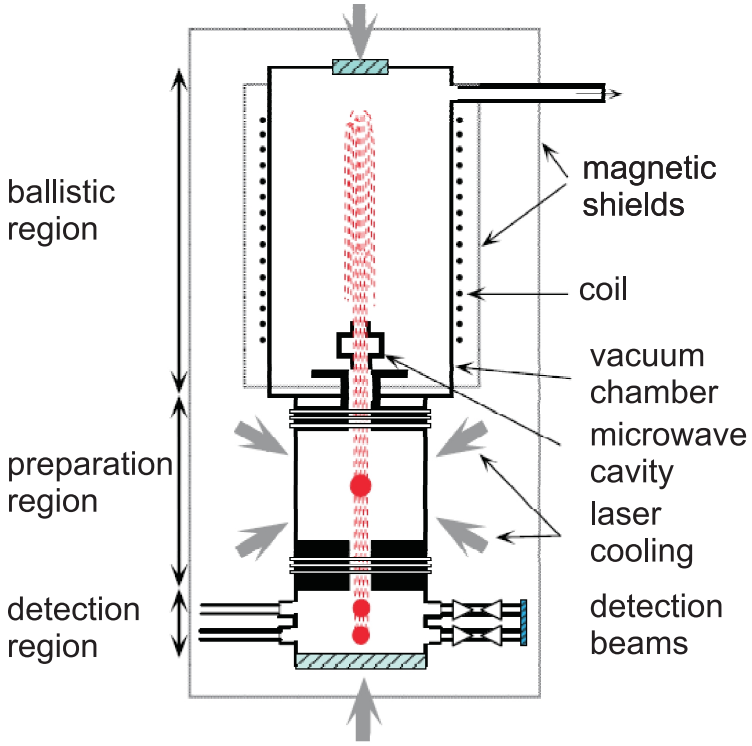


Figure 7.12: *Cs fountain clock schematics.*

by laser methods.

Typical ballistic flight time is of 1 s which corresponds to the fringe width of approx. 1 Hz as shown in Fig. 7.13.

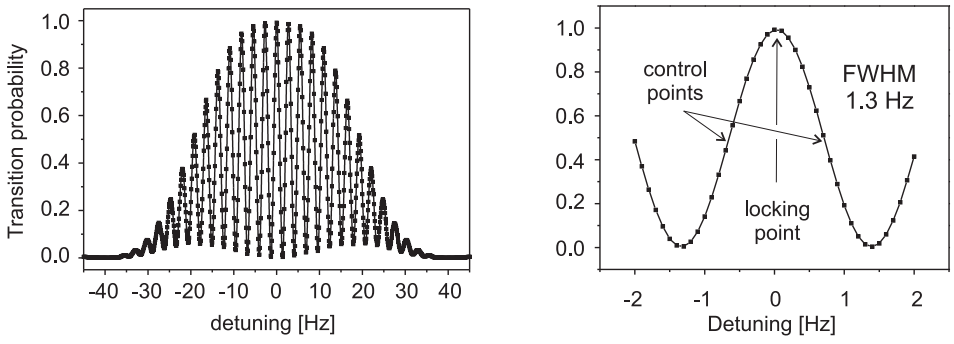


Figure 7.13: *Left: interference fringes of FOM fountain clock for the transition* $6S_{1/2}(F = 3, m_F = 0) \leftrightarrow 6S_{1/2}(F' = 4, m'_F = 0)$ *in* ^{133}Cs. *Right: zoom in the central fringe. Interrogating oscillator is locked to the flank of the fringe.*

The ballistic region is isolated from magnetic fields and weak homogeneous magnetic field is applied to select the $6S_{1/2}(F = 3, m_F = 0)$ and $6S_{1/2}(F' =$

$4, m'_F = 0$) sublevels (similar to the beam apparatus). The measurement cycle of the atomic fountain is about 1 s which means that the quartz oscillator frequency can be corrected in respect to the fringe center each second. The servo loop controls population on sublevels and correct oscillator frequency correspondingly.

7.4.3 Stability of Cs clocks

The short time instability of the beam and fountain clock is defined by the quartz oscillator. For times longer than 1 s the signal from Cs atoms will define the stability of the apparatus.

As seen from Fig. 7.14, the stability of commercial beam clock is significantly less stable compared to Cs fountain. For Cs beam clock HP5071A the Allan deviation is at the level of 5×10^{-13}. For fountain clock the instability drops as $1.3 \times \tau^{-1/2}$ and reaches the flicker noise level of $2 - 3 \times 10^{-16}$.

Most of the systematic effects are well studied. They are the Zeeman shift in magnetic field, black body radiation and collisional shift. The best fountain clock accuracy now reached 2×10^{-16} and is limited by the quantum projection noise. They are widely used for synthesis of SI second and in many fundamental applications.

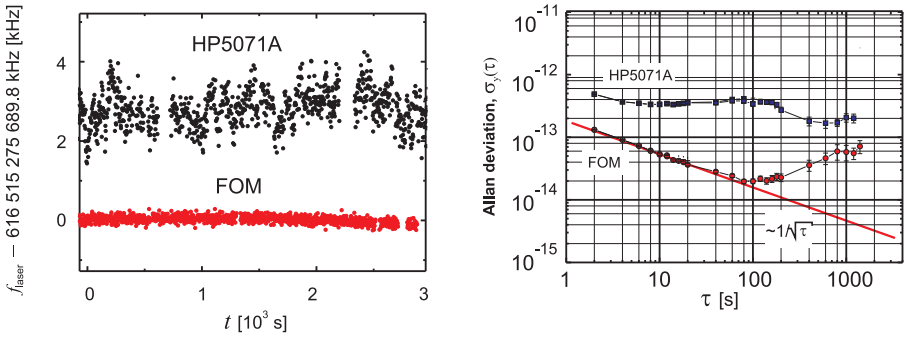


Figure 7.14: *Left: Frequency measurement of an ultra-stable laser using commercial Cs beam clock HP5071A and mobile fountain clock FOM. Right: Allan deviation corresponding to the left plot. Growing of Allan deviation at higher averaging times (¿100 s) is due to the thermal drift of laser frequency.*

Lecture 8: Laser cooling of atoms

Optical molasses. Doppler theory, Doppler limit. Subdoppler laser cooling: Sisyphus method, polarization gradient cooling. Recoil limit. Evaporative cooling. Applications. Bose-Einstein condensation of atomic gases.

Both spectral line width and the frequency of atomic transition depend on coordinates and velocities of particles interacting with electromagnetic field. It is very important for many applications to prepare particles in some definite initial state. Reduction of atomic velocity and localization of atoms in a small finite volume allow to suppress the Doppler effect, to increase interaction time with radiation and control external fields. Depending on how atoms are prepared in the initial state and what interrogation method is implemented, detected spectral line width can change by a few orders of magnitude as shown in Fig. 8.1.

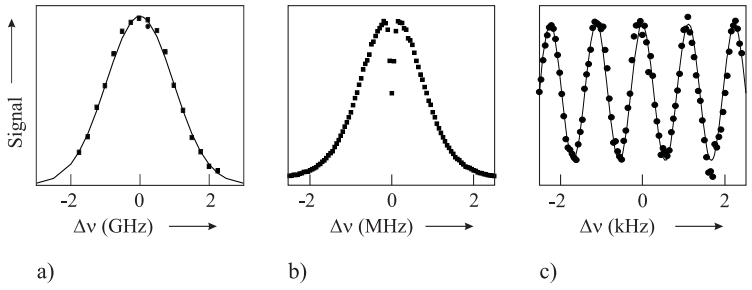


Figure 8.1: *Optical transition in Ca atoms ($\lambda = 657\,nm$) with the natural line width of $\Delta\nu \approx 0.37\,kHz$ measured using different methods a) Doppler broadened transition in the gas cell $\Delta\nu \approx 2\,GHz$. b) By saturation absorption spectroscopy $\Delta\nu \approx 150\,kHz$, the line width is limitd by time-of-flight broadening c) Ramsey spectroscopy on laser-cooled Ca atoms. One can measure the line width close to natural one.*

Laser cooling is the most efficient method to suppress Doppler effect and allows to increase the interaction time with particles.

Laser cooling of neutral atoms was first suggested by Hänsch and Shawlow (1975), while Weinland and Demelt suggested laser cooling of ions (1975).

8.1 Optical molasses

Let us take a two-level atom with the ground and excited state energies of E_g and E_e correspondingly. It absorbs a photon from a laser field with the wave vector $\vec{k}$ and the momentum of $\hbar\vec{k}$. After absorption it spontaneously emits a photon. Laser frequency is red detuned from the transition frequency $(E_e - E_g)/h$. The photon momentum is transferred to the atom changing atom's momentum $\vec{p} = m\vec{v}$. Assume that the Doppler shift $\Delta\nu = p/(m\lambda)$ is small in respect to the natural line width $\gamma = 1/(2\pi\tau)$, where τ is the life time of the excited state. It means that $\Delta\nu \ll \gamma$. Under this assumption the transferred momentum $\Delta\vec{p} = \hbar\vec{k}$ can be averaged out by a large number of absorption and emission processes which results in a classical force $\vec{F}$ applied to an atom. Spontaneously emitted photons will not contribute to the force $\vec{F}$, because they are emitted isotropically. The averaged force applied to an atom from absorbed photons equals to :

$$\vec{F} = \frac{N_e}{N}\frac{\hbar\vec{k}}{\tau}\,, \tag{8.1}$$

where N_e is the averaged number of atoms in the excited state, and $N = N_e + N_g$ is the total number of atoms (the sum of excited state atoms N_e and ground state atoms N_g). The ratio N_e/N can be derived using the saturation parameter S_0 which will give us :

$$\vec{F} = \frac{\hbar\vec{k}}{2\tau}\frac{S_0}{1 + S_0 + \left(\frac{\delta\nu}{\gamma/2}\right)^2}\,. \tag{8.2}$$

Here

$$S_0 \equiv \frac{I}{I_{\text{sat}}}\,, \tag{8.3}$$

where I_{sat} is the saturation intensity corresponding to the case when the resonant radiation transfers 1/4 of the population to the excited level. It can be written as

$$I_{\text{sat}} = \frac{2\pi^2 hc\gamma}{3\lambda^3}\,. \tag{8.4}$$

Theory of a two-level system describes the population of the upper level as

$$\frac{N_e}{N} = \frac{S_0}{2}\frac{(\gamma/2)^2}{(1 + S_0)(\gamma/2)^2 + \delta\nu^2}\,, \tag{8.5}$$

which, after substitution, will give the result (8.2).

If the intensity of a laser field is much lower compared to the saturation intensity $S_0 \ll 1$, force (8.2) is described by a Lorentzian line shape. The width of the Lorentzian approaches the natural line width.

For an atom moving with some velocity $\vec{v}$, the frequency detuning from the resonance will depend on velocity due to the Doppler effect. In the atomic frame, the detuning is equal to $\delta\nu = \nu - \nu_0 - \vec{k}\cdot\vec{v}/(2\pi)$. Let us consider an atom moving with velocity $\vec{v}$ placed in the field formed by two counter-propagating laser beams of equal intensity (e.g. a laser beam retro-reflected by a mirror). If the field is weak and $S_0 \ll 1$, forces caused by two laser beams may be added as following:

$$\begin{aligned}\vec{F}_{om} &= \frac{\hbar\vec{k}}{2\tau}\left(\frac{S_0}{1+S_0+4\left(\nu-\nu_0-\frac{\vec{k}\cdot\vec{v}}{2\pi}\right)^2/\gamma^2} - \frac{S_0}{1+S_0+4\left(\nu-\nu_0+\frac{\vec{k}\cdot\vec{v}}{2\pi}\right)^2/\gamma^2}\right) \\ &= \frac{\hbar\vec{k}}{2\tau}S_0\frac{16(\nu-\nu_0)\frac{\vec{k}\cdot\vec{v}}{2\pi\gamma^2}}{\left[1+S_0+\frac{4(\nu-\nu_0)^2}{\gamma^2}+\left(\frac{k^2v^2}{\pi^2\gamma^2}\right)\right]^2-\left[8(\nu-\nu_0)\frac{\vec{k}\cdot\vec{v}}{2\pi\gamma^2}\right]^2}\,. \qquad (8.6)\end{aligned}$$

Fig. 8.2 shows the dependency of force acting on an atom in the case of $S_0 = 0.3$ if the laser frequency ν is shifted from the transition frequency ν_0 by one spectral line width γ o the red: $\nu - \nu_0 = -\gamma$.

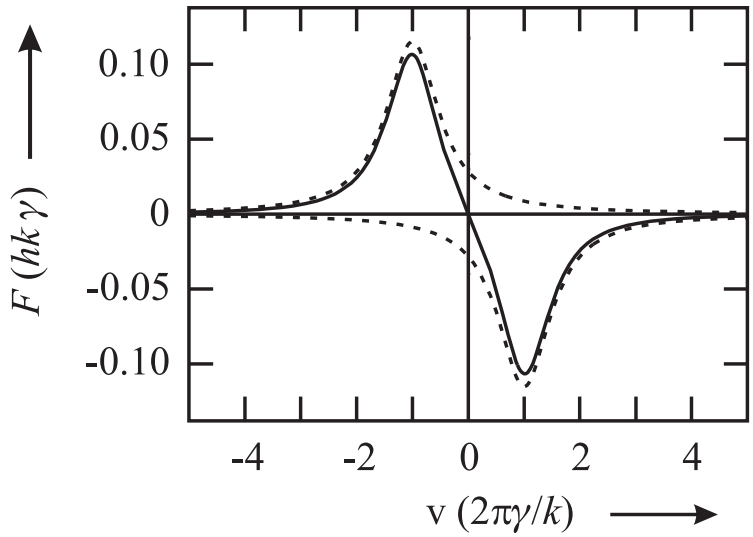


Figure 8.2: *A force acing on an atom depending on is velocity. Force is caused by absorption of photons from two counter-propagating laser fields of equal intensity. The curve is given by(8.6) at* $S_0 = 0.3 \;\; \nu - \nu_0 = -\gamma$.

At low velocity limit ($v < \gamma\lambda$) one can neglect higher order terms $((\vec{k}\cdot\vec{v}/\gamma^2)^2$

and higher) in (8.6). After reduction we get:

$$\vec{F}_{\text{om}} = \frac{8\hbar k^2 S_0(\nu - \nu_0)}{\gamma\left(1 + S_0 + \frac{4(\nu-\nu_0)^2}{\gamma^2}\right)^2}\vec{v} = \alpha\vec{v}\,. \tag{8.7}$$

The resulting force acting on the atom linearly depends on velocity at low velocity limit. If the laser frequency is red detuned in respect to the atomic resonance ($\nu - \nu_0 < 0$), one can find correspondence of the force $F_{\text{om}} = -\alpha\vec{v}$ with a viscose friction. In other words, for atoms with velocity $\vec{v}$ the frequency of the counter propagating beam is closer to the resonance due to the Doppler effect. Atoms, moving in the electrical field of given configuration will be decelerated by this viscose force. Due to this analogy, the expression "optical molasses" is used to describe this process.

8.2 The Doppler limit

One can think, that in the optical molasses atoms will be continuously decelerated and they will stop reaching $T = 0$ limit. In this case one neglects the fact, that even atoms at rest will absorb and emit photons. A recoil energy, which will be transferred to the each atom in the absorption process is equal to $(+\hbar^2k^2/2m)$, while in the emission process $(-\hbar^2k^2/2m)$. It will result in heating which will correspond to the increase of the kinetic energy of each of the atoms by $2\hbar^2k^2/2m$ (at average).

If the system reaches equilibrium, heating and cooling rates should be equal to each other :

$$\dot{E}_{\text{heat}} = -\dot{E}_{\text{cool}}\,. \tag{8.8}$$

The heating rate $\dot{E}_{\text{heat}}$ equivalent to the transferred energy per unit time, will be proportional to the fraction of atom in the excited state for each of the fields (8.5)and the decay rate $1/\tau = 2\pi\gamma$ of the excited state.Hence

$$\dot{E}_{\text{heat}} = 2\frac{(\hbar k)^2}{2m}\frac{2\pi\gamma}{2}\frac{2S_0}{1 + 2S_0 + 4(\nu - \nu_0)^2/\gamma^2}\,, \tag{8.9}$$

where we took into account that the saturation parameter equals $2S_0$. The cooling rate coming from the deceleration in optical molasses is equal to :

$$\dot{E}_{\text{cool}} = \frac{\partial}{\partial t}\frac{p^2}{2m} = \dot{p}\frac{p}{m} = F(v)v = -\alpha v^2\,. \tag{8.10}$$

Substituting (8.9), (8.10) and (8.7) in (8.8) and replacing v^2 by the averaged value $\langle v^2\rangle$, we get:

$$m\left\langle v^2\right\rangle = \frac{h\gamma}{4}\frac{[1 + 2S_0 + (2(\nu - \nu_0)/\gamma^2)^2]}{2(\nu - \nu_0)/\gamma}\,. \tag{8.11}$$

The right part of the equation (8.11) reaches minimum at $\nu - \nu_0 = \gamma/2$. Taking into account that $m\langle v^2\rangle/2 = k_B T/2$, we get, that the minimal temperature equals to:

$$T_D = \frac{h\gamma}{2k_B} = \frac{\hbar\Gamma}{2k_B} \qquad \text{(Doppler limit)} \tag{8.12}$$

under condition $S_0 \to 0$. Temperature T_D is the minimal temperature which can be achieved using considered cooling mechanism. Since the cooling results from the Doppler effect, the temperature limit is referred as to *Doppler limit*.

The Doppler limit in the three-dimensional case can be derived by similar considerations. Although the cooling rate is the same as in one-dimensional case, the heating rate will be three times higher, because in 3D case one should take into account 6 laser field instead of two.At the other hand side, the temperature is given by $m\langle v^2\rangle_{3D}/2 = 3k_B T/2$. As a result, the Doppler limit in the 3D case is the same, as in (8.12). E.g., for Cs atoms with the cooling transition $6\ ^2\mathrm{S}_{1/2} - 6\ ^2\mathrm{P}_{3/2}$ ($\lambda = 852\,\mathrm{nm}$, $\gamma = 5,18\,\mathrm{MHz}$), the Doppler limit equals $0,12\,\mathrm{mK}$. For Ca atoms and corresponding cooling transition $4^1\mathrm{S}_0 - 4^1\mathrm{P}_1$ ($\lambda = 423\,\mathrm{nm}$, $\gamma = 34,6$ Hz) – $0,83\,\mathrm{mK}$. Thermal velocity corresponding to Doppler limit can be obtained from the equation $1/2mv_D^2 = k_B T_D/2$:

$$v_D = \sqrt{\frac{h\gamma}{2m}}. \tag{8.13}$$

For the examples given above $v_{D,\mathrm{Cs}} = 8,82\,\mathrm{cm/s}$ $v_{D,\mathrm{Ca}} = 41,5\,\mathrm{cm/s}$.

8.3 Subdoppler cooling

Typical velocities of atoms cooled using the Doppler mechanism on strong resonance transitions are in the range from a few cm/s to few tens cm/s. Although it is much less compared to the thermal velocity of atoms (typically 300-500 m/s), they are still too high for many applications. For example, in the atomic fountain clock atoms should interact with radiation within 1 s. To load atoms in the shallow traps one also have to cool atoms to lower velocities.

For atoms possessing magnetic or hyperfine splitting of the ground state (e.g. Cs, Rb, Tm, etc.) there are other mechanisms which allow to laser cool atoms to even lower temperatures. Most frequency implemented is so called *polarization gradient* method were atoms are cooled in a laser field with spatially varying polarization. The gradient can be formed by two counter-propagating laser fields of opposite circular polarization or two counter-propagating laser fields of orthogonal linear polarizations which is called "Sisyphus cooling". We will consider this mechanism as an example.

For Sisyphus laser cooling the atom should move in two laser fields of equal amplitudes and frequencies, opposite wave vectors and perpendicular linear

polarizations:

$$\begin{aligned} \vec{E} &= E_0\hat{x}\cos(\omega t - kz) + E_0\hat{y}\cos(\omega t + kz) \\ &= E_0[(\hat{x}+\hat{y})\cos\omega t\cos kz + (\hat{x}-\hat{y})\sin\omega t\sin kz]\,. \end{aligned} \tag{8.14}$$

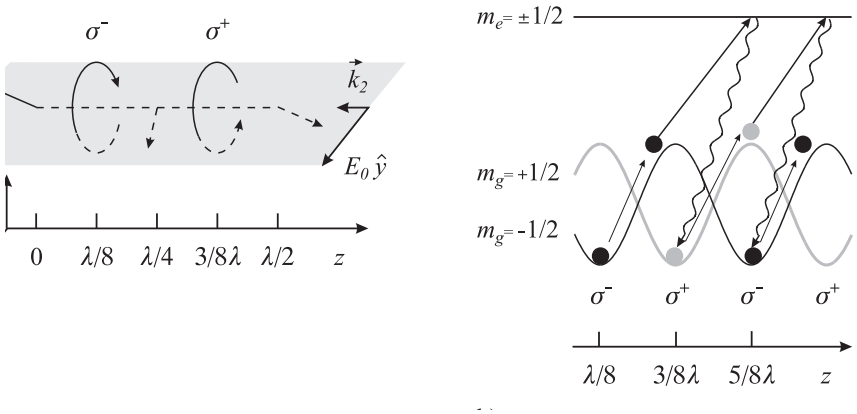


Figure 8.3: *Sisyphus cooling. a) Polarization distribution in the wave. b) Light shift in the wave modulates energies on the ground state magnetic sublevels ($m_g = +1/2$ $m_g = -1/2$), which modulates interaction with laser field.*

From (8.14) follows that at $kz = 0$ polarization of the field is linear and oriented by 45° to the x axis as shown in Fig. 8.3). Polarization changes to the orthogonal at $\lambda/4$ distance ($kz = \pi/2$), while at $kz = \pi/4$ ($z = \lambda/8$) it is circular.

The simplest model is the two-level atom with a resonance transition $J = 1/2 \to J = 3/2$. It can be other transition with different multiplicity. For the mentioned transition the magnetic sublevels $m_{\pm 1/2}$ of the ground state are periodically shifted which vary in space due to the polarization gradient as shown in Fig. 8.3 b). Assume that atom is initially in the ground state with $m_g = -1/2$, which has the lowest energy at $z = \lambda/8$. If atom is moving along z, it climbs the potential hill which results in losses of its kinetic energy. If laser field polarization at this point will change to σ^+, the atom will be optically pumped to the $m_g = +1/2$ state because of optical pumping via $m_e = +1/2$ state. Moving along z axis atoms again looses its kinetic energy climbing the next hill where is will be again optically pumped by σ^- radiation to $m_g = -1/2$ state (via $m_e = -1/2$ level). In analogy to the ancient Greek hero Sisyphus, who was punished by gods and had to roll the huge stone up to the hill again and again, this mechanism is called "Sisyphus". The best regime is reached in

the case if the averaged time of optical pumping necessary to transfer atoms to the next sublevel equals the travelling time of $\lambda/2$ distance.

Usually, in the experiment the polarizations are chosen to be σ^+/σ^- and slightly different mechanism plays a role. In this case the resulting polarization will be always linear, but rotating around the axis. Cooling will be reached by periodical re-distribution of population among the sublevels which will increase probability to absorb photon from the light field counter propagating to the atomic motion.

The minimal temperature which can be reached using subdoppler mechanism reaches the *recoil limit*

$$k_B T > E_r = (\hbar k)^2/2m. \tag{8.15}$$

For Cs atoms in 3D optical molasses the achievable temperature is $25\,\mu$K, which is much lower than the Doppler limit $0,12\,$mK, but still slightly higher compared to (8.15).

The recoil limit is difficult to overcome since it results from only one single emission of a resonance photon. Still, there are approaches allowing to laser cool atoms to even lower temperatures (e.g. by using EIT and Raman cooling), but in this case atom should not directly interact with the resonance laser field scattering photons. These exotic methods have not find applications in frequency standards.

One can reach low temperatures also by implementation of the second stage Doppler cooling, but using narrower transition. It is useful for atoms which do not have magnetic splitting of the ground state: ^{20}Mg, ^{40}Ca, ^{88}Sr all of them widely used in frequency standards. Using transitions of $\sim 1\,$Hz line width one can push the Doppler limit to microkelvin regime.

Lecture 9: Traps for neutral atoms

Magnetic dipole trap, optical dipole trap, optical lattices. Magneto-optical trap.

For many applications including precision measurements it is important not only to cool atoms to lower temperatures, but also to trap them at some local position in pace for longer time. For this purpose one can use electric, magnetic, gravitational and light forces which influence external degrees of freedom (coordinates and velocities) of the atom, ion or molecule and localize them at some position.

There are some limitations for stable traps. For some volume without charges $\Delta\Phi = 0$, where Φ is the electrostatic potential (if the charge density equals 0, the Maxwellian equations will give $div\vec{E} = \vec{\nabla} \cdot \vec{E} = \vec{\nabla} \cdot \vec{\nabla}\Phi = \Delta\Phi = \rho/\varepsilon_0 = 0$). From this relation one can conclude the it is impossible to configure electric charges such way, that in the free space between them a minimum or maximum of electric potential. This conclusion is usually called as the Irnshaw theorem. It means that is is impossible to build a stable electrostatic trap for ions.

It is also possible to prove, that in space free of charges and currents there are neither *maximum* of electric field, nor *maximum* of magnetic field. It means that it is not possible to build neither electro-static, nor magneto-static trap for atoms in the lowest energetic state, which always will move towards the maximum of electric of magnetic field. Ketterle and Pritchard proofed this statement for any combination of magnetic and electric fields.

Atomic and molecular ions can be trapped into ion traps since the Irnshaw theorem does not pose any limitation for using rotating field configurations when positive and negative field gradients are changing with high frequency. Since the electrostatic interactions are given by the electric field strength $\vec{F} = q\vec{E}$, the potential well is quite deep and can reach a few electronvolts. Recalling that $1\,\mathrm{eV} \hat{=} 11\,600\,\mathrm{K}$, it is clear that this depth is well enough to trap ions at room temperatures and at much higher temperatures as well. Ion traps will be considered in the next lectures.

Forces, acting on neutral atoms and molecules are much weaker compared

to electromagnetic (Coulomb) interactions. Most regularly these forces are due to interaction of electric field gradient with an induced dipole moment of the particle or due to interaction of magnetic field with the magnetic moment of the particle.

Unperturbed atoms cannot possess a permanent electric moment due to T-invariance. It is only possible to trap atoms using interaction with induced dipole moment of atoms.

9.1 Magnetic dipole trap

It is not difficult to prepare atoms in an internal state with a permanent magnetic moment which can interact with magnetic field. The force acting on the atom with the magnetic moment μ (projected on the magnetic field direction) will be given as

$$F = -\mu \overrightarrow{\nabla} B \,. \tag{9.1}$$

External magnetic field results in the shift of energy levels. If external magnetic field gradient is applied, a particle with magnetic moment will fill a force. The lowest energy level of any atom will always shift down in the external magnetic field. It means that any atom in the lowest energy state will move towards maximum of magnetic field. Atoms aiming for the maximum of magnetic field are called "high-field seekers".

Atoms in the excited states may move also to the minimum of the magnetic field (of course they can move to the maximum of the field too, depending on the state). Atoms aiming to the minimum of magnetic field are called as "low-field seekers". Since it is impossible to have a maximum of the magnetic field, only low filed seekers can be trapped in the minimum of the field.

One of the simplest configurations of the magnetic traps if the trap consisting of two coils in anti-Helmholtz configuration (Fig. 9.1). Coils build up a radial symmetric magnetic field in the plane $x-y$ if z axis is aligned along the coil axis. Close to the center the magnetic field vector changes linearly ($B_x = \{\partial B_x/\partial x\} \cdot x, B_y = \{\partial B_y/\partial y\} \cdot y, B_z = \{\partial B_z/\partial z\} \cdot z$). Taking into account the equation $div \cdot \vec{B} = \vec{\nabla} \cdot \vec{B}(\vec{r}) = 0$ we get $2\partial B_x/\partial x = 2\partial B_y/\partial y = -\partial B_z/\partial z$. It shows that the field gradient along z-axis is always twice as large as for x and y axis, but has an opposite sign.

Neutral atoms were trapped in such a trap in 1985 by H. Metcalf and it was the first demonstration of trapping neutral atoms. An important disadvantage of such trap is the possibility of a Mayorana spin-flip at trap center where magnetic field is essentially zero. Crossing the zero field point magnetic moment of the atom can flip and atoms will turn into high-field seekers which will result in the repulsive force from the trap center. Atoms will be lost from the trap and cannot be returned back.

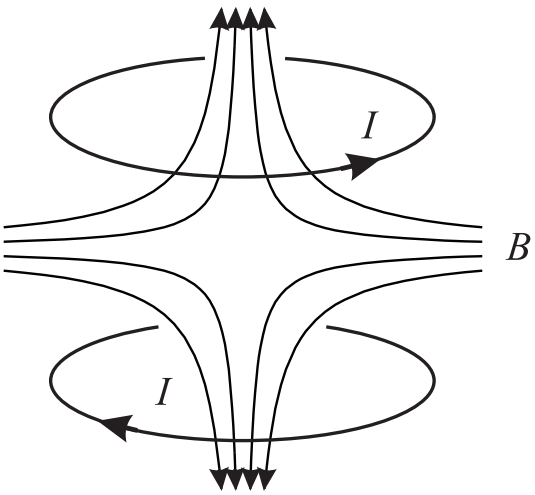


Figure 9.1: *Magnetic quadrupole trap consisting of two current loops in the anti-Helmholtz configuration.*

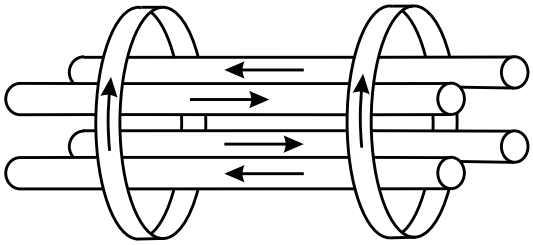


Figure 9.2: *Ioffe-Pritchard trap consisting of four current lines and two coils. Arrows show direction of current.*

One of the possible solutions is the implementation of the Ioffe-Proitchard trap shown in Fig. 9.2. Such trap possesses the minimum of the magnetic field as well, but the field does not turn to zero at its minimum. Such systems are widely used for production of quantum degenerative gases (Bose-Einstein and Fermi condensates).

Both these systems have very shallow potential gap and can be implemented only for trapping neutral particles at very low temperatures (less than 1 K).

9.2 Optical dipole trap

Atoms without magnetic moment can be trapped using an induced electric dipole moment in the external field. High field strength and gradients can be obtained in the focus of a laser beam.

If a two-level system undergoes interaction with the resonance laser field, the energy of its states will change due to the dynamic Stark shift. If the frequency of the field is red detuned from the atomic transition frequency, the ground state energy will be reduced and the energy of the excited state will grow as shown in Fig. 9.3. Accordingly, if the laser is blue detuned, the picture will change to the opposite. Depending of the frequency detuning, the atomic dipole moment will oscillate either in phase or out of phase with the external field. Thus, atom will be either pulled into the field maximum, or be pushed out of it.

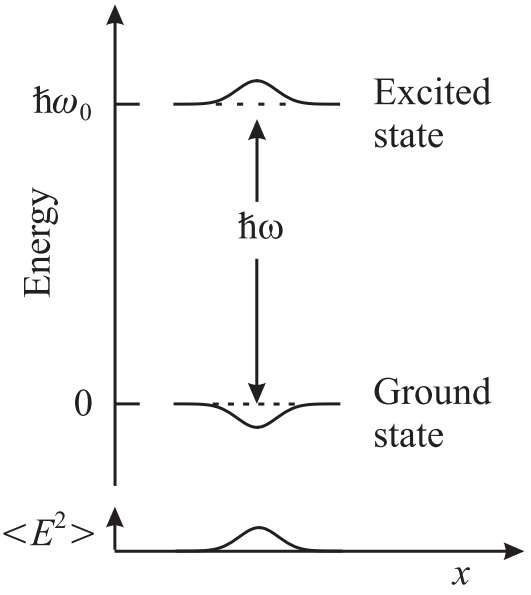


Figure 9.3: *Interaction of a two-level atom with a spatially inhomogeneous laser field tuned close to the resonance frequency. It results in the spatial-dependent shift of the atomic levels.*

The potential energy of an atom in the laser beam with the electric field amplitude E_0 is given by:

$$\begin{aligned} W_{\rm dip}(r,z) &= -\frac{6\pi\varepsilon_0 c^3}{\omega_0^2}\frac{\Gamma(\omega_0^2-\omega^2)}{(\omega_0^2-\omega^2)^2+\omega^6\Gamma^2/\omega_0^4}\frac{E_0^4}{4} \\ &\approx -\frac{3\pi\varepsilon_0 c^3}{4\omega_0^3}\left[\frac{\Gamma}{\omega_0-\omega}+\frac{\Gamma}{\omega_0+\omega}\right]E_0^2 \approx \frac{\hbar}{8}\frac{\Gamma^2}{\omega-\omega_0}\frac{I(r,z)}{I_{\rm sat}}\,. \end{aligned} \quad (9.2)$$

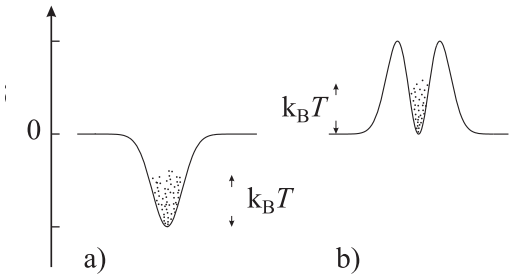


Figure 9.4: *Optical dipole traps with red (a) and blue (b) detuning. Trap with red detuning can be build by focussing a laser beam with a regular field distribution (fundamental Gaussian mode). Dipole trap with blue detuning can be built by e.g. focussing a radially symmetrical Laggerr-Gaussian mode LG_01 possessing a donut field distribution.*

Here we made an assumption that the frequency detuning is much larger than the natural line width ($\omega - \omega_0 \gg \Gamma$). We also made a rotating wave approximation and neglected the second term in the square brackets. We also used the expression for the laser field intensity $I(r,z) = (\varepsilon_0 c/2)E_0^2$ and the saturation intensity $I_{\rm sat}$ from (8.4).

The simplest optical dipole trap is the red detuned focused Gaussian laser beam (Fig. 9.4 a). Such a beam has a three-dimensional intensity maximum in focus. For the Gaussian beam we have

$$I(r,z) = \frac{2P}{\pi w_0^2(1+\frac{z^2}{z_R^2})} \exp\left[-\frac{2r^2}{w_0(1+\frac{z^2}{z_R^2})}\right] \approx \frac{2P}{\pi w_0^2}\left(1 - \frac{2r^2}{w_0^2} - \frac{z^2}{z_R^2}\right), \quad (9.3)$$

where P is the power of the Gaussian beam with radius w_0, while $z_R = \pi w_0/\lambda$ is the Rayleigh length. Approximation (9.3) is valid for small distances from the beam waist $zf < z_R \;\; r < w_0$. In this case both radial and axial directions are well approximated by harmonic potentials.

Contrary to the force in the optical molasses, the dipole force in the optical dipole trap does not saturate with the laser power. Spontaneous emission caused by absorption of photons in the dipole trap causes heating which is proportional to the number of scattered photons. A scattering rate $\Gamma_{\rm sc}$ which is the number of scattered photons per unit time is given by:

$$\Gamma_{\rm sc} = \frac{P_{\rm abs}}{\hbar\omega} = -\frac{2}{\hbar}\Im{\rm m}\{\alpha\}\frac{E_0^2}{4} \approx \frac{\Gamma^3}{8(\omega-\omega_0)^2}\left(\frac{\omega}{\omega_0}\right)^3 \frac{I}{I_{\rm sat}}. \quad (9.4)$$

The scattering rate $\Gamma_{\rm sc}$ becomes less for large detunings $\omega - \omega_0$ since it reduces as $(\omega - \omega_0)^{-2}$ (see (9.4)). In practice, most regularly used are optical

dipole traps with very large destining to the red from atomic resonance (Far of Resonance Traps or FORT).

Traps with the blue detuned laser field provide less scattering rate since atoms are accumulated in the area with low, close to zero intensity. A blue detuned optical dipole trap (e.g.a donut-shape Lagerr-Gaussian mode LG_{01} or an optical lattice) provides similar potential depth as the red detuned TEM_{00} trap. Blue detuned traps are more preferable for frequency standards since the dynamic Stark shift is less as in the red detuned traps where atoms are accumulated in the intensity maximum.

9.3 Magneto-optical trap

In optical molasses atoms are cooled to very low velocities. Force, acting on atoms is similar to the viscose force which decelerates atoms, but does not attract them (trap them) to some distinct point in space. The trapping force can be created by applying an inhomogeneous magnetic field.

Let us consider an atom with the ground state energy E_g and orbital momentum $J = 0$. For the excited state corresponding values are E_e and $J = 1$ as shown in Fig. 9.5. For example, such simple level scheme can be found in alkali-earth atoms (Ca, Sr, etc.) In this case the energy of the ground state can be treated as constant in magnetic field, while the excited state will be split in three magnetic sublevels ($m_J = 0, \pm 1$). Energies of $m_J = \pm 1$ magnetic sublevels linearly depend on magnetic field with the same coefficient (but with opposite sign). Assume, that the magnetic field B changes linearly if the coordinate z:

$$B_z(z) = bz. \tag{9.5}$$

Zeeman shift of the sublevels with $m_J \neq 0$ is given by

$$\Delta E(z) = \pm g_J \mu_B b z \,. \tag{9.6}$$

Due to this dependency the space-dependent detuning is formed:

$$\delta\nu = \nu - \nu_0 \mp \frac{v}{\lambda} \mp \frac{g_J \mu_B}{h} b z \,, \tag{9.7}$$

where g_J is the Landé factor of the excited state and μ_B – the Bohr magneton ($\mu_B/h = 1.4 \cdot 10^{10}\,\mathrm{Hz/T}$). Using the laser field propagating along z-axis one can excite transitions to the $m_J = 1$ $m_J = -1$ levels using σ^+ and σ^- circular polarization, correspondingly. Doing calculations similar to (8.7), but with the space-dependent term in the equation (9.7), we get:

$$F_z(z) = -Dz, \tag{9.8}$$

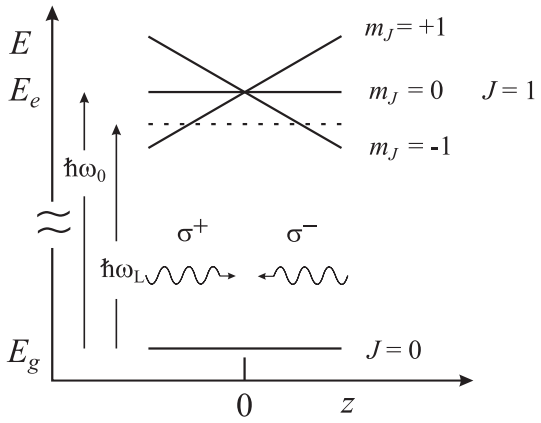


Figure 9.5: *Energy levels of atom in a magneto-optical trap.*

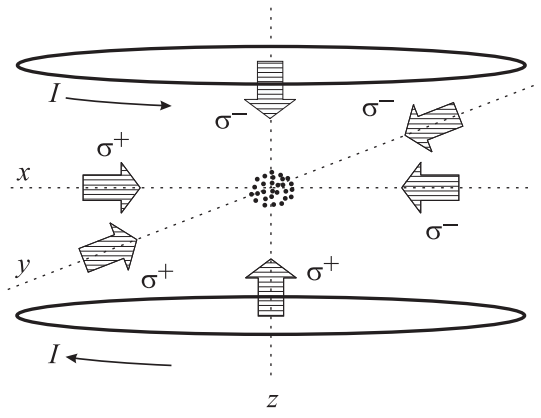


Figure 9.6: *Configuration of a magneto-optical trap.*

where the constant D is equal to :

$$D \approx \frac{8\mu_B bkS_0(\nu - \nu_0)}{\gamma\left(1 + S_0 + \frac{4(\nu-\nu_0)^2}{\gamma^2}\right)^2}. \tag{9.9}$$

Due to this force a parabolic potential $V(z) = Dz^2/2$ is formed which traps atoms. If both laser beams have the same intensity, the trap center will coincide with zero of magnetic field. The resulting force which takes into account both the viscose force of optical molasses and the retrieving force from quadratic potential (magnetic field inhomogeneity) equals to

$$F_z(z, v) = -Dz - \alpha v. \tag{9.10}$$

This is the equation of motion for the harmonic oscillator. Atom plays a role of a mass m oscillating with the eigenfrequency $\omega_0 = \sqrt{\frac{D}{m}}$ and the damping constant of $\Gamma = \frac{\alpha}{m}$.

To trap atoms in three-dimensional space on has to extend the given picture to all three coordinates maintaining proper polarization relations as shown in Fig. 9.6. Three-dimensional magnetic field minimum with linear dependency similar to (9.5) can be formed by two coils in anti-Helmholtz configuration as shown in Fig. 9.1. For regular MOTs the gradient is on the order of 0.05 T/m to 0.5 T/m.

Example. Calculate the eigenfrequency ω_0 and damping constant Γ in the trap for ^{40}Ca atoms. The magnetic field gradient equals $b = 0.1\,\mathrm{T/m}$, the frequency detuning equals $\omega - \omega_0 = \Gamma/2$, the cooling wavelength 423 nm.

Since the atomic mass equals $m = 40 \cdot 1.66 \times 10^{-27}\,\mathrm{kg}$ we get using (9.9) (8.7): $\omega_0 \approx 2\pi \cdot 2,4\,\mathrm{kHz}$ and $\Gamma \approx 1.56 \times 10^5\,\mathrm{s}^{-1}$. It is clear that the oscillation will be damped much faster as one oscillation period.

Loading of a magneto-optical trap. Maximal velocity of atoms which can be trapped in a magneto-optical trap equals $v_c \approx (2F_{\max} r/m)^{1/2} = (\hbar k \gamma r/m)^{1/2}$, where r – is the radius of the trapping light beam. Typically, it is around $v_c = 30\,\mathrm{m/s}$. Atoms with the velocities $v < v_c$ can be trapped directly from the low-velocity wing of Maxwellian distribution even without additional deceleration. More efficient method is to load atoms from a *Zeeman slower* where atoms are decelerated in one dimension using combination of resonant light and magnetic field. The equation describing the number of atoms in a MOT is:

$$\frac{dN}{dt} = R_c - \frac{N}{\tau_{\mathrm{MOT}}} - \beta N^2\,, \tag{9.11}$$

where R_c is the capturing rate and τ_{MOT} – an averaged life time of atom in a MOT. The second term is responsible for collisions with the background gas, while the third term describes collisions of atoms with each other. Equation (9.11) can be easily solved is one neglects the last term which is important only at high concentration of atoms. We will get:

$$N(t) = (N(0) - R_c\tau_{\mathrm{MOT}})e^{-t/\tau_{\mathrm{MOT}}} + R_c\tau_{\mathrm{MOT}}. \tag{9.12}$$

The loading curve shown in Fig. 9.7 approaches to the equilibrium state $N(t \to \infty) = R_c\tau_{\mathrm{MOT}}$ for the typical time τ_{MOT}. If initially the MOT is empty $N(0) = 0$, the loading curve is described by the following curve :

$$N(t) = R_c\tau_{\mathrm{MOT}}(1 - e^{-t/\tau_{\mathrm{MOT}}}). \tag{9.13}$$

Large MOT can contain up to $N \gg 10^7$ atoms at the density of $\rho > 10^{10}\,\mathrm{at/cm^3}$ which can be used in many different applications after switching of MOT beams.

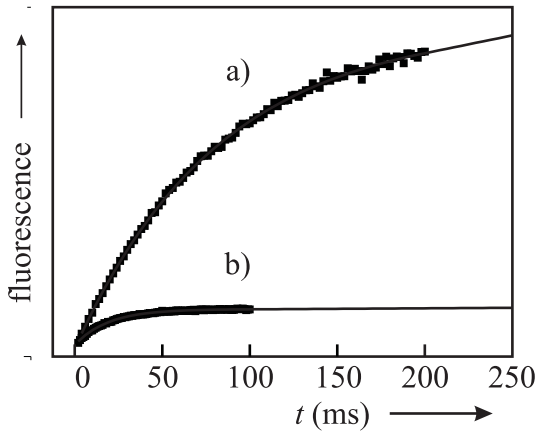


Figure 9.7: *Loading curves of Ca MOT and approximation according to (9.13). Using a repumper laser the life time increases from* $\tau_{MOT} = 19\,ms$ *(b) to* $\tau_{MOT} = 83\,ms$ *(a).*

9.4 Optical lattice

The optical lattice in an extension of the optical dipole trap. In contrast to the optical dipole trap formed by a focused laser beam with only one intensity maximum/minimum, the optical lattice is an interference pattern of two and more laser beams which possesses multiple intensity maxima/minima.

As example, consider two counter-propagating laser beams of similar polarization and intensity. The interference will result in a stationary distribution of intensity :

$$\begin{aligned} \vec{E} &= E_0\hat{\epsilon}\cos(\omega t - kz) + E_0\hat{\epsilon}\cos(\omega t + kz) \\ &= 2E_0\hat{\epsilon}\cos(kz)\cos(\omega t). \end{aligned} \tag{9.14}$$

The dynamic Stark shift is proportional to E^2 and, as follows from the expression is periodically varying with the coordinate z with node and antinodes separated by $\lambda/4$. Low-energy atoms can be trapped in potential wells and localize them in the volume much smaller compared to the wavelength. One can also build two- and three-dimensional lattice by intersection of a few standing waves, e.g. as shown in Fig. 9.8. In this case the potential wells look as shown in Fig. 9.9.

Typically, by loading the optical lattice by very cold atoms the population of a unit well turns to be very low. In applications to optical frequency standards optical lattices provide very long interaction times and very tight confinement of atoms which is important to reduce the Doppler effect. By proper tuning the wavelength of the optical lattice ("magic wavelength") the influence of the lattice potential on the clock transition can be turned to zero

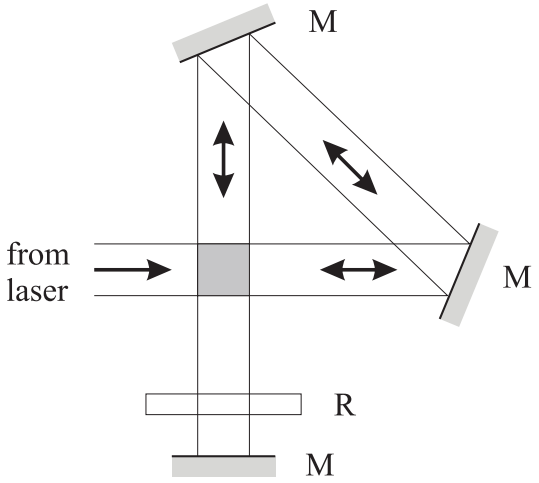


Figure 9.8: *Two-dimensional optical lattice. M – mirror, R – phase plate.*

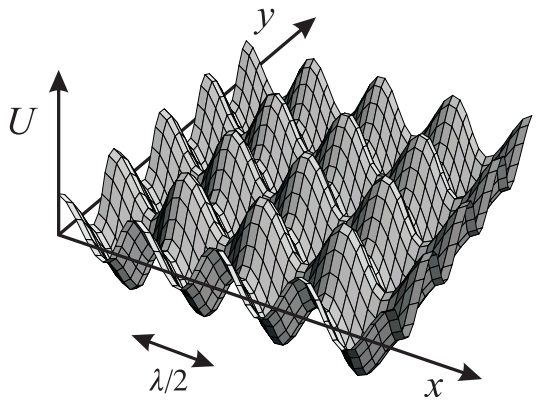


Figure 9.9: *Distribution of the potential in the two-dimensional optical lattice.*

in the first order. This happens in the case if the dynamic Stark shift of the lower and upper clock state become equal.

Lecture 10: Paul trap for ions

Traps for charged particles. Linear Paul trap. Equation of motion, Mathieu equations. Floquet-type solutions. Pseudopotential. Frequencies of micro- and macromotion. 3D Paul trap. Trap loading.

10.1 Traps for charged particles

The best reference for a frequency standard is an isolated medium at full rest, which possesses a strong absorption line with high Q-factor. Neutral atoms only weakly interact with external fields which mean that one needs high field intensities to tightly localize and trap atoms. High fields will, in turn, cause strong perturbation of the clock transition.

Ions, which are the charged particles, are easier to control, since the Coulomb interaction with electric field is very strong and allows to tightly localize atoms using so-called *ion traps*. Ions can be trapped for very long time (up to a few months) and they have extremely narrow transitions. One can study a single ion which is a perfect candidate for an isolated system since it does not interact with other particles. One can laser cool atoms using regular methods of laser cooling and reach very high localization and low velocities. It is also very important that ions can be *sympathetically* cooled, i.e. one atom can be cooled via Coloumb interaction by another, laser cooled atom.

One can localize charged particles by combination of electric and magnetic fields. As follows from the Irnshaw theorem, it is impossible to trap particles by static fields and alternating fields are necessary. There are two major types of traps:

- A Paul trap, where particles are trapped in inhomogeneous alternating electric field
- A Penning trap, where atoms are trapped in combination of static magnetic field and alternating electric field.

W. Paul was awarded a Nobel prise for invention of a trap configuration which is now referred as to the Paul trap.

10.2 Paul trap

Consider an electric field $\vec{E}(\vec{r})$ given by potential $\Phi(\vec{r})$ inside the trap volume. The field interacts with an ion with a charge $q = +e = 1{,}602 \cdot 10^{-19}\,\mathrm{A\,s}$. A force, acting on an ion will be given by

$$\vec{F}(\vec{r}) = e\vec{E}(\vec{r}) = -e \cdot \vec{\nabla}\Phi(\vec{r})\,, \tag{10.1}$$

acting towards the trap center. Here the operator $\vec{\nabla} = (\partial/\partial x, \partial/\partial y, \partial/\partial z)$ is used to find the field gradient. It is desirable, that the force will linearly depend on the distance form the trap center $\vec{r}$ as $\vec{F}(r) \propto \vec{r}$. In this case particles will undergo harmonic oscillations. In this case a scalar potential $\Phi(x, y, z)$ should have a quadratic dependency on coordinates

$$\Phi = const \cdot (ax^2 + by^2 + cz^2)\,, \tag{10.2}$$

where the constant is given by boundary conditions. Using the Laplace equation $\Delta\Phi = \nabla^2 = 0$ for a space free of charges, we get the restriction for the coefficients a, b c defining the potential in (10.2):

$$a + b + c = 0\,. \tag{10.3}$$

We will consider two cases for coefficients (10.3):

$$a = 1,\ b = -1,\ c = 0 \qquad \text{(linear quadrupole trap)} \tag{10.4}$$

and

$$a = b = 1,\ c = -2 \qquad \text{(three-dimensional trap).} \tag{10.5}$$

10.3 Linear quadrupole trap

The first combination of coefficients (10.4) describes the trap configuration with the potential independent on the z-coordinate:

$$\Phi = const \cdot (x^2 - y^2)\,. \tag{10.6}$$

It is a two-dimensional quadrupole potential shown in Fig. 10.1.

Such two-dimensional potential (10.6) can be formed by a system of four hyperbolic electrodes with a negative potential applied to upper and lower electrodes and positive - to the right and left electrodes (or vice versa) see Fig. 10.1. Assume, that he potential difference between the electrodes equals Φ_0. The constant from (10.2), (10.6) can be obtained from the boundary condition $\Phi(r_0) = \Phi_0/2 = const \cdot r_0^2$, where $2r_0$ is the distance between two opposite electrodes. We get $const = \Phi_0/2r_0^2$. The electric field is calculated

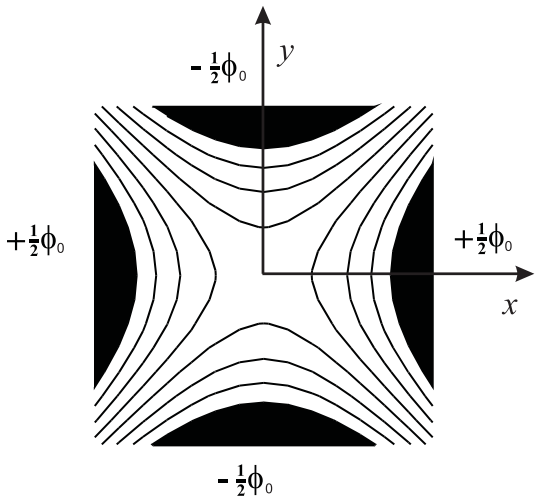


Figure 10.1: *Two-dimensional qudrupole potential in the $x-y$ plane can be produced by four hyperbolic electrodes.*

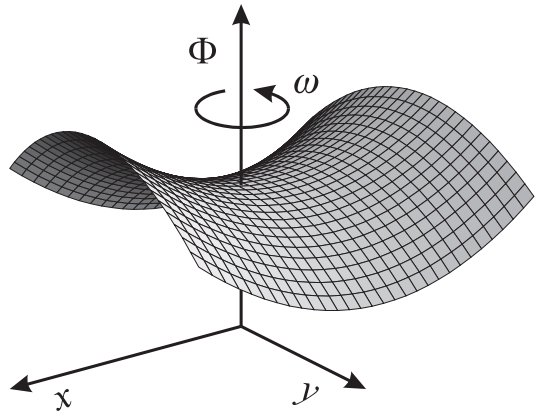


Figure 10.2: *The potential of the linear quadrupole trap has a saddle-like shape rotating around z-axis..*

from (10.1):

$$E_x = \frac{\Phi_0}{r_0^2}x,\ E_y = -\frac{\Phi_0}{r_0^2}y,\ E_z = 0. \tag{10.7}$$

In such a field the particle with the charge $+e$ will be repelled from the positive electrodes towards $x = 0$ position. Ion will harmonically oscillate along x-axis. At the other hand, it will be attracted to the negative electrode along the y-axis. According to (10.6) the potential will have a saddle-like shape. It has a minimum along x and maximum along y (10.2). Changing of polarity will reverse the picture: the ion will be repelled from y-electrodes and will be attracted to x ones.

To trap ions in both directions, the potential of the electrode pairs should

periodically change. It can be done by adding an alternating component $V_{\rm ac}$ at frequency ω to the constant bias voltage applied to electrodes $U_{\rm dc}$:

$$\Phi_0 = U_{\rm dc} - V_{\rm ac}\cos\omega t. \tag{10.8}$$

The potential shown in Fig. 10.2 will rotate at the frequency ω around z-axis. Although it is not obvious that periodical focussing and defocussing along x and y axes should result in trapping of an ion. Indeed, the averaged force seems to be zero.

As we will show later, it is not true: An efficient trapping force directed to the trap center appears. It comes from inhomogeneity of the trapping potential. Before deriving this force and corresponding *pseudopotential* let us consider the equation of motion for the trapped ion.

10.4 Mathieu equations

Consider an ion placed in the trap with the potential given by (10.8). Coordinates and velocities of an ion will be given by the following equations

$$\begin{aligned} F_x(t) &= m\ddot{x}(t) = eE(x)\cos\omega t = \frac{e}{r_0^2}(U_{\rm dc} - V_{\rm ac}\cos\omega t)x \\ F_y(t) &= m\ddot{y}(t) = eE(y)\cos\omega t = -\frac{e}{r_0^2}(U_{\rm dc} - V_{\rm ac}\cos\omega t)y\,, \end{aligned} \tag{10.9}$$

where $\ddot{x}(t)$ is d^2x/dt^2. Substituting (10.7) and introducing dimensionless parameters

$$\tau \equiv \frac{\omega}{2}t, \qquad a \equiv \frac{4eU_{dc}}{m\omega^2 r_0^2}, \qquad q \equiv \frac{2eV_{ac}}{m\omega^2 r_0^2}\,, \tag{10.10}$$

we get differential equations first analyzed by a French mathematician E. Mathieu:

$$\frac{d^2x(\tau)}{d\tau^2} + (a - 2q\cos 2\tau)x = 0 \tag{10.11}$$

and

$$\frac{d^2y(\tau)}{d\tau^2} - (a - 2q\cos 2\tau)y = 0. \tag{10.12}$$

The difficulty in this type of equations are the periodically changing coefficients. One can use this periodicity to analyze equations, namely, trying to find the solutions looking like that:

$$F_\mu(\tau) = e^{i\mu\tau}P(\tau)\,, \tag{10.13}$$

which are called Floquet-type solutions. Here $P(\tau)$ should be a periodical function with the same period as coefficients in (10.11) equal to π. If the solution is aperiodic, it can be represented as a combination of independent

Floquet solutions $F_\mu(\tau)$ and $F_\mu(-\tau)$. Parameter μ in the exponent depends only on coefficients a and q. In general case, presence of the exponent will result in an exponential growth of the amplitude which corresponds to non-stable regime. But, only in the case if the parameter μ becomes a real value $\mu = \beta \in$ Real, the solution will describe oscillations of the ions in a finite space around the equilibrium point. That will be the stable solution.

For applications, the oscillation amplitude should be smaller, that the inner size of the trap, otherwise ion will hit the wall. Since the characteristic exponent is the function of a and q, one has to calculate the dependency $a(q)$ for given $\beta = f(a, q)$ which can be done by different mathematical methods.

One of the examples for discussed dependency $a(q)$ is shown in Fig. 10.3. Shadowed areas correspond to stable areas for the following parameter range $0 \leq \beta \leq 1, 1 \leq \beta \leq 2, 2 \leq \beta \leq 3$. Mathematically, the integer numbers of β are the special cases. For example, the condition $\beta = 1$ defines the sharp threshold between two separate stability areas. It is not critical in practice, where the trap parameters are selected in such way, that parameters will lay safely within stability ranges.

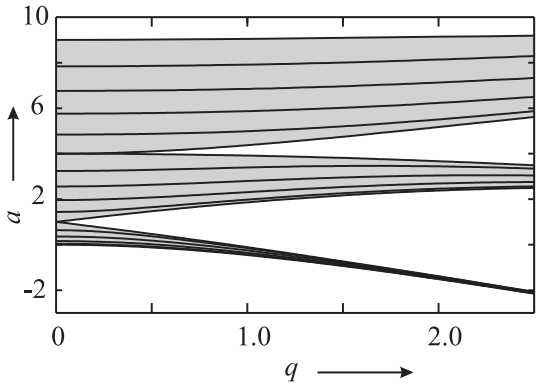


Figure 10.3: *Dependencies $a(q)$ for $\beta = f(a, q)$, calculated for $0 \leq \beta \leq 3$ with the step equal to 0.2 (lines). There are three stability ranges on the diagram (shadowed areas. Note, that the diagram is symmetrical, i.e. $a(q) = a(-q)$.*

Stable trapping of an ion is given only by trap parameters a q and does not depend on initial conditions. For stable trapping of a ion in two- and three-dimensional trap all parameters a_i and q_i $(i = x, y)$ should independently fall in stability regions. For two-dimensional trap the stability diagram is a joint plot shown in Fig. 10.4. For the x axis one should use $+a$, $+q$ and for the y axis – parameters $(-a)$, $(-q)$ using the fact that $a(q) = a(-q)$. The stable regime happens when both stability regions for x and y overlap. The first stability region is shown in details in Fig. 10.5 for different values of β.

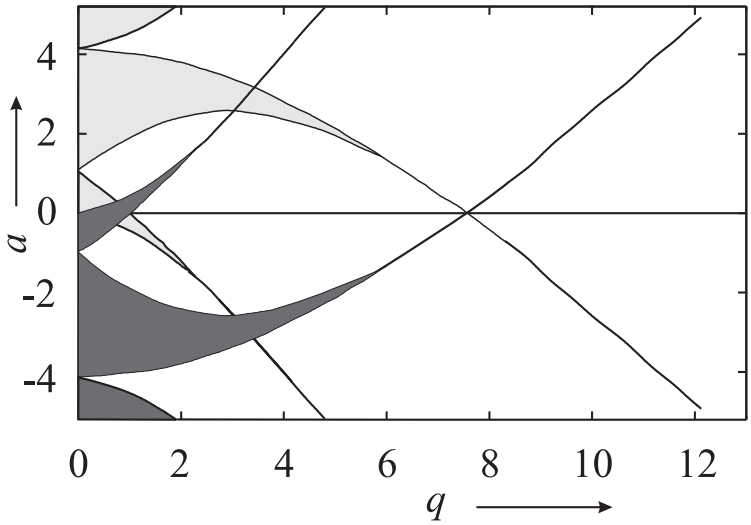


Figure 10.4: *Joint diagram for both coordinates x and y similar to Fig. 10.3. There are a few joint stability regions where individual diagrams overlap.*

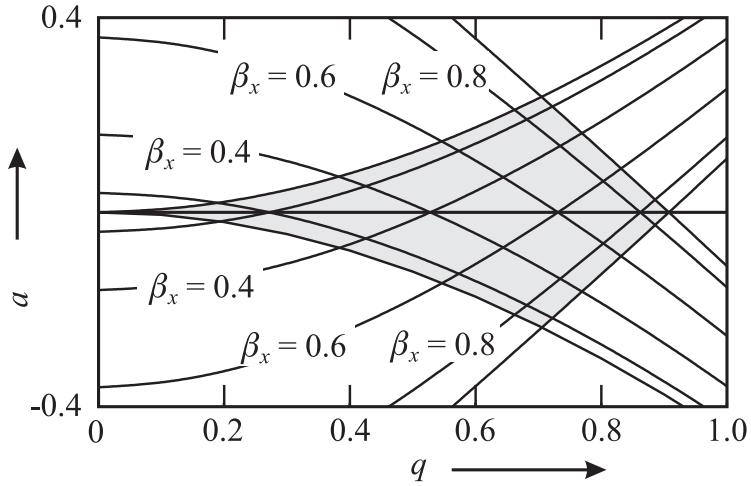


Figure 10.5: *The first joint stability region (shadowed area) in a two-dimensional Paul trap.*

10.5 Pseudopotential

Here we will study the origin of a trapping force appearing in the Paul trap. Assume, that initially ion is positioned away from the trap center at some position $\hat{x}$. First, for simplicity, consider the ion motion in a homogeneous oscillating electric field with the amplitude $\hat{E}$ and frequency ω. The equation of motion will look like

$$m\ddot{x}(t) = e\hat{E}(x)\cos\omega t \tag{10.14}$$

and (for convenience) assuming the initial condition $\dot{x}(0) = 0$ we will get the dependency

$$x(t) = \hat{x} - \frac{e\hat{E}}{m\omega^2}\cos\omega t. \tag{10.15}$$

The ion oscillates at the frequency of the applied field, but the phase of its oscillations differs for π from the applied field which is clear from minus sign. This process is called *micromotion*. The phase lag is very important and will result in the trapping force in an inhomogeneous field.

Now we will assume that the field is spatially inhomogeneous and is distributed according to Fig. 10.2. The ion is oscillating around some point $\hat{x} > 0$. If it is accelerated outwards the trap center, it should be closer to the trap center $x < \hat{x}$. In this region the field is less than at $\hat{x}$. Otherwise, when the ion is further away from the center $x > \hat{x}$ the field is stronger and the ion is accelerated towards the trap center. It means that there will be some non-zero effective force which will pull the ion towards the trap center. This force can defined via so-called *pseudopotential*. Assuming that $x(t) - \hat{x} << \hat{x}$ one can expand field in the power series:

$$\begin{aligned} F(t) &= eE(\hat{x})\cos\omega t + e\frac{dE(\hat{x})}{dx}(x-\hat{x})\cos\omega t + \cdots \\ &\approx eE(\hat{x})\cos\omega t - \frac{e^2E(\hat{x})}{m\omega^2}\frac{dE(\hat{x})}{dx}\cos^2\omega t\,. \end{aligned} \tag{10.16}$$

Here we used (10.15) for $x(t) - \hat{x}$ difference. After averaging of (10.16), its first term becomes zero, while the average of the second term is

$$F_{\rm av}(\hat{x}) = -\frac{e^2E(\hat{x})}{2m\omega^2}\frac{dE(\hat{x})}{dx}. \tag{10.17}$$

One can define a pseudopotential $\Psi_{\rm pseudo}$ corresponding to this force. In our 2D case it will be given by

$$\Psi_{\rm pseudo}(\hat{x},\hat{y}) = \frac{eE^2(\hat{x},\hat{y})}{4m\omega^2}\,. \tag{10.18}$$

The ion motion will consist of micromotion on the frequency of the driving field and much slower oscillations in the pseudopotential which are called as *secular motion*. The secular radial oscillation frequency ω_r can be calculated from the ion kinetic energy which should correspond to the potential (10.18):

$$e\Psi_{\rm pseudo} = \frac{1}{2}m\omega_r^2(x^2+y^2). \tag{10.19}$$

For simplicity assume $U_{\rm dc} = 0$. Now we will substitute $E^2(\hat{x},\hat{y}) = E_x^2 + E_y^2$ from (10.7) in (10.18), which will give us $\omega_r \approx eV_{\rm ac}/(\sqrt{2}m\omega r_0^2)$.

The described above quadrupole trap restricts the ion motion only on $x-y$ plane, ions can freely move along z-axis. There are different methods to restrict motion of the ion in z direction, for example, to place additional ring electrodes with positive repelling potential or to use segmented rods with the outer parts at a constant positive potential as shown in Fig. 10.6).

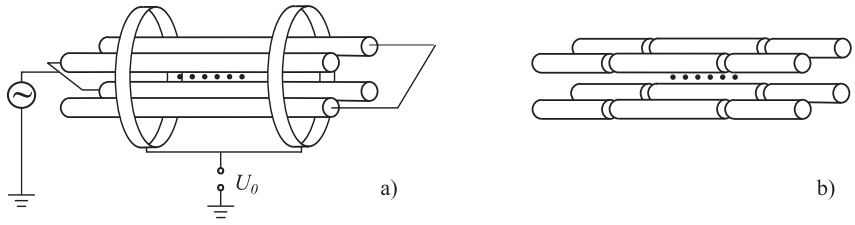


Figure 10.6: *Linear trap configurations with the radial potential similar to shown in Fig. 10.1. Traps have additional ring electrodes (a) or segmented rods (b) for axial confinement.*

10.6 Three-dimensional Paul trap.

Another solution of the equation (10.5) which is widely used in practice results in the three-dimensional potential:

$$\Phi = \frac{\Phi_0}{x_0^2 + y_0^2 + 2z_0^2} \cdot (x^2 + y^2 - 2z^2), \tag{10.20}$$

which can be produced with the potential surfaces of the following shape

$$x^2 + y^2 - 2z^2 \equiv r^2 - 2z^2 = \pm r_0^2. \tag{10.21}$$

The positive sign corresponds to an z-axially symmetrical hyperbolic surface which can be manufactured in practice as a ring electrode with the inner radius r_0 as shown in Fig. 10.7. The negative sign corresponds to a two hyperboloid branches separated by a distance $2z_0 = \sqrt{2r_0}$.

Electric field in radial direction (E_r) and axial direction (E_z) differ by the coefficient -2. The potential in cylindric coordinates will be given as

$$\Phi(r, z) = \frac{U_{\rm dc} + V_{\rm ac} \cos \omega t}{r_0^2 + 2z_0^2}(r^2 - 2z^2)\,, \tag{10.22}$$

where r_0 as z_0 are defined in Fig. 10.7. Parameters a and q (10.10) for the radial (a_r, q_r) and axial (a_z, q_z) directions will differ by factor -2:

$$a_z = -2a_r \equiv a, \qquad q_z = -2q_r \equiv q. \tag{10.23}$$

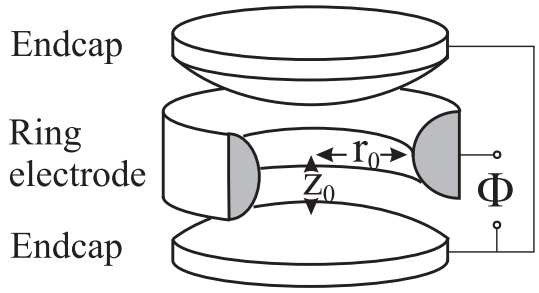


Figure 10.7: *Three-dimensional Paul trap.*

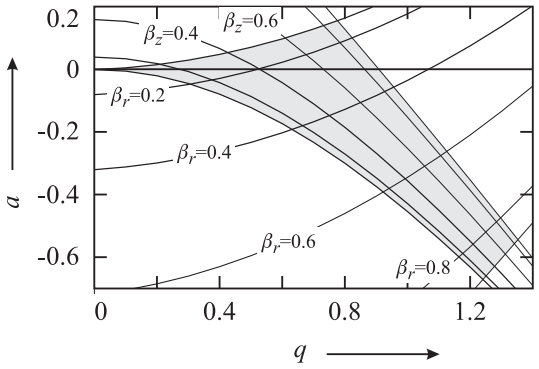


Figure 10.8: *The first stability region in the three-dimensional Paul trap.*

One can plot the stability diagram $a(q)$ by overlapping the axial and radial diagrams as shown in Fig. 10.8. The diagrams differ by a scaling factor of -2 (10.23). The first stability region becomes asymmetrical which differs from the two-dimensional case (Fig. 10.5).

In experiment only the first stability range is used. As example, an ion trap for trapping single $^{171}\mathrm{Yb}^+$ ions at PTB has a radius of $r_0 = 0,7\,\mathrm{mm}$ at the driving voltage $V_{\mathrm{ac}} = 500\,\mathrm{V}$ with a frequency of $\omega = 2\pi \cdot 16\,\mathrm{MHz}$. The trap parameters are $q_z = 0.11$ and $a_z \approx 2 \times 10^{-3}$ (10.10), (10.23)), which corresponds to the first stability region.

The pseudopotential $\Psi_{\mathrm{pseudo}}(\hat{r}, \hat{z})$ in the three-dimensional case can be calculated similar to (10.18):

$$\begin{aligned} \Psi_{\mathrm{pseudo}}(\hat{r}, \hat{z}) &= \frac{U_{\mathrm{dc}}}{2r_0^2}(\hat{r}^2 - 2\hat{z}^2) + \frac{eV_{\mathrm{ac}}^2}{4m\omega^2 r_0^4}(\hat{r}^2 + 4\hat{z}^2) \\ &= \frac{m\omega^2}{16e}[(q_r^2 + 2a_r)\hat{r}^2 + 4(q_r^2 - a_r)\hat{z}^2]\,. \end{aligned} \tag{10.24}$$

Parameters $e\Psi_{\mathrm{pseudo}}(r_0, 0)$ and $e\Psi_{\mathrm{pseudo}}(r_0/\sqrt{2}, 0)$ are the potential depths in the radial and axial directions. The depth in axial direction is two times larger as in the axial one. One can make the potential symmetrical fulfilling the condition $a_r = q_r^2/2$.

Lecture 11: Penning trap for ions and ion cooling

Penning trap for charged particles. Magnetron, cyclotron and axial frequencies. Precision mass comparison in the Penning trap. Synthesis of anti-hydrogen atoms in Penning traps. Cooling of ions. Doppler and sideband cooling. Sympathetic cooling. Lamb-Dicke regime.

11.1 Penning trap

Configuration of electrodes in the Penning trap is the same as in the 3D Paul trap, but the radio-frequency field is absent ($V_{\rm ac} = 0$). It means that ions will repel from the end cap electrodes (along z direction). In the $x-y$ plane the same potential will push ions out of the trap center. For trapping ions a strong magnetic field is applied along z-axis. Equation of electron motion can be written as

$$\begin{aligned} m\ddot{\vec{r}} &= e\vec{E}(\vec{r}) + e\dot{\vec{r}} \times \vec{B}, \qquad \text{which is equivalent} \\ m\ddot{x} &= e(E_r + \dot{y}B_z) \\ m\ddot{y} &= e(E_r - \dot{x}B_z) \\ m\ddot{z} &= eE_z. \end{aligned} \tag{11.1}$$

The electrical field components can be calculated from the potential Φ (10.22).

The last equation describes ion oscillations with the frequency of

$$\omega_z^2 = \frac{4eU_{\rm dc}}{m(r_0^2 + 2z_0^2)}\,, \tag{11.2}$$

which does not depend on B_z.

If only magnetic field is applied to a moving particle, it will perform a circular motion in the plane orthogonal to the field. The frequency of oscillations will be given by

$$\omega_c = \frac{e}{m}B_z \qquad \text{(cyclotron frequncy)}. \tag{11.3}$$

The cyclotron frequency can be obtained from the fact that the Lorentz forces provides the centripetal acceleration $evB_z = mv^2/r$ or $eB_z = m\omega_c$. In

the trap there is also the electric field $\vec{E}_r$ perpendicular to the magnetic field $\vec{B}_z$. Two fields will cause the motion of the particle along the ring orbit in the $x-y$ plane around z-axis which is called *magnetron* motion. A balance of electric and Lorentz force will define the magnetron frequency of this motion: $evB_z = e\omega_m r B_z = eE_r$. The magnetron frequency does not depend on the charge or mass of the particle, but depends only on electric and magnetic fields:

$$\omega_m = \frac{E_r}{r\,B_z} \qquad \text{(magnetron frequency)}. \tag{11.4}$$

For a typical Penning trap parameters $B_z \approx 1-5\,\mathrm{T}$ and $U_{ac} \approx 10-100\,\mathrm{V}$ the magnetron frequency will be of a few tens of kHz, the axial frequency ω_z – of a few hundred of kHz and the cyclotron frequency ω_c – a few MHz. Typically the following relation is fulfilled

$$\omega_c \gg \omega_z \gg \omega_m. \tag{11.5}$$

The ion trajectory in this case is a superposition of all three oscillation types as shown in Fig. 11.1. All oscillation are nearly independent and the trajectory consists of (i) fast cyclotron motion around the magnetic field axis (11.3), oscillations along the magnetic field axis (11.2) and a small drift along a circular trajectory calculated from (11.4).

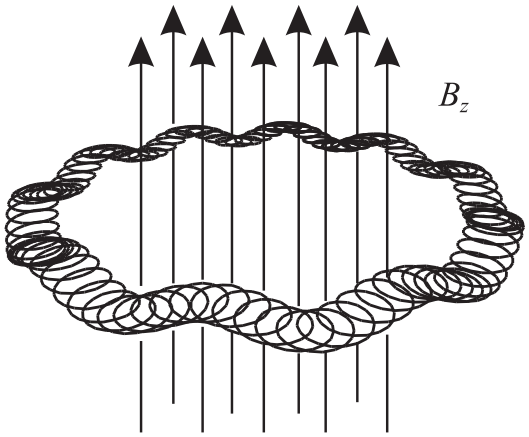


Figure 11.1: *Ion trajectory in the Penning trap. It is an orbit with epicycles in the $x-y$ plane with oscillations along the z axis. Here $\omega_c = 10\omega_z = 100\omega_m$.*

But, for other trap parameters the cyclotron motion can be of the same size as the magneton and picture will be different from the one shown in the Figure.

11.1.1 Rigorous solution

One can rigorously solve joint differential equations (11.1) for x and y:

$$\ddot{x} = \frac{e}{m}\left(\frac{2U_{\rm dc}}{r_0^2+2z_0^2}x + \dot{y}B_z\right) = \frac{\omega_z^2}{2}x + \omega_c\dot{y} \tag{11.6}$$

$$\ddot{y} = \frac{e}{m}\left(\frac{2U_{\rm dc}}{r_0^2+2z_0^2}y - \dot{x}B_z\right) = \frac{\omega_z^2}{2}x - \omega_c\dot{x}. \tag{11.7}$$

Let us add equation (11.6) to equation (11.7) multiplied by i. Also we introduce complex parameter $r = x + iy$. After this procedure we will get a simplified equation $\ddot{r} = \omega_z^2 r/2 - i\omega_c\dot{r}$. It can be solved by substitution $r = r_0\exp(i\omega t)$ which will result in a square equation $\omega^2 - \omega\omega_c - \omega_z^2/2 = 0$ for ω. Two roots of this equation will give us frequencies

$$\omega_c' = \frac{\omega_c}{2} + \sqrt{\frac{\omega_c^2}{4} - \frac{\omega_z^2}{2}} \qquad \text{(modified cyclotron frequency)} \tag{11.8}$$

$$\omega_m = \frac{\omega_c}{2} - \sqrt{\frac{\omega_c^2}{4} - \frac{\omega_z^2}{2}} \qquad \text{(magnetron frequency).} \tag{11.9}$$

If the value under the square root is positive ($\omega_c \geq \sqrt{2}\omega_z$), we will get two frequencies called as *modified cyclotron frequency* ω_c' and magnetron frequency ω_m. Modification of the original cyclotron frequency comes from the presence of electric field.

One can also get important relations by (i) adding the equations (11.8) and (11.9) or (ii) by squaring them and then adding them. As a result we get:

$$\omega_c = \omega_c' + \omega_m \tag{11.10}$$

$$\omega_c^2 = \omega_c'^2 + \omega_m^2 + \omega_z^2\,. \tag{11.11}$$

Both these frequencies are used to calculate the cyclotron frequency (11.3) which is very important for precision ion mass comparison.

For a cloud of ions in the Penning trap, fast magnetron motion along $(\vec{E}\times\vec{B})$ results in Doppler effect second order. For larger cloud size the effect is larger. Penning traps are not often used for optical or microwave frequency standards and are inferior to Paul traps. But they become very important tool for many fundamental applications like precision mass comparison, determination of anomalous magnetic moment of electron and producing antihydrogen atoms.

11.1.2 Ion energies in the Penning trap.

Consider an single-charged ion ($m = 100\,$a.u.) moving in the Paul trap with the potential difference of 10 V, $B = 5\,$T and $r = 1\,$mm. Which energy corresponds to each type of ion motion?

(i) The axial movement is a pure harmonic oscillatory motion and its energy is equally distributed between kinetic and potential energy. The maximal energy in the trap can reach $E_{\text{pot}} \approx 5\,\text{eV} = 8 \cdot 10^{-19}\,\text{J}$.

(ii) The cyclotron orbit has very small radius while the ion moves at very high speed. The energy is mostly kinetic and equals $e\omega_c B_z r^2 \approx 510^{-19}\,\text{J}$ for an orbit of $r = 0.5\,\text{mm}$.

(iii) The energy of magnetron motion is mostly potential. The ion mass equals $1,6 \cdot 10^{-25}\,\text{kg}$ the magnetron velocity will be equal $v \approx 1000\,\text{m/s}$ (11.4). The kinetic energy equals $E_{\text{kin}} = 1/2mv^2 \approx 8 \cdot 10^{-20}\,\text{J}$ which is much less compared to the potential energy (i).

The latter relation shows that the total energy reduces with the reduction of the radius of magnetron motion. Collisions will result to increasing the radius and, in the end, losses from the trap.

11.1.3 Interactions between trapped ions

All considered equations for Paul and Peening traps are valid only for the case if only one ion is trapped. If trap contains many atoms, the ion interaction will change the behavior of the ion motion. Ions will strongly interact by Coloumb interaction. If the ion energy is small compared to interaction energy, they will form crystalline structures. Very impressive structures can be obtained in linear Paul traps where field is nearly zero close to the axis. Cold ions will form a linear crystal which is similar to beads on the necklace as shown in Fig.11.2. More ions will form spirals or crystalline clouds which can contain up to 10^5 ions.

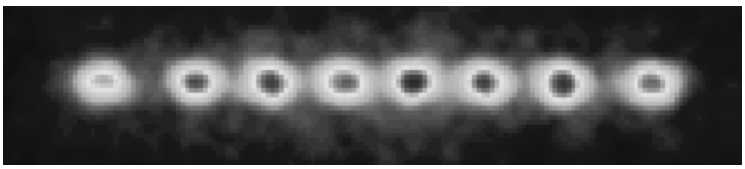


Figure 11.2: *Spontaneous radiation of 8 ions trapped in a linear quadrupole Paul trap.*

Because of joint oscillations, the spectrum of ion oscillation will contain new frequencies which will correspond to oscillatory eigenmodes. If the trap potential is purely harmonic, ions can not be heated by the driving field. But typically the potential is not perfectly harmonic and ions are heated by the field which is called *radio-frequency heating*. In this case the different degrees of freedom become dependent and the energy can be transferred between them. The spectrum of ion oscillations becomes complicated as shown in Fig. 11.3)

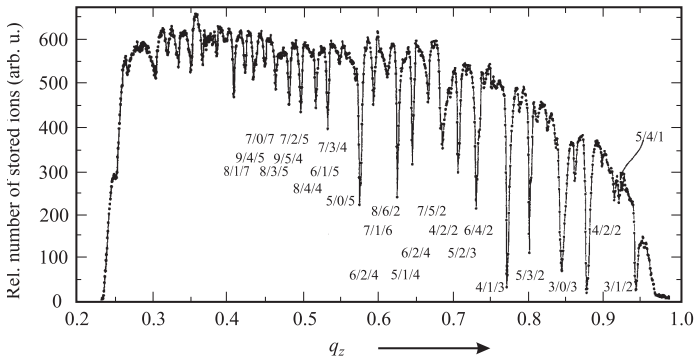


Figure 11.3: *Resonance losses of the ion trap dependent on driving field frequency. The losses result from ion heating.*

11.2 Lamb-Dicke regime

Excitation of optical transition in optical region has certain advantages for applications in frequency standard. But, motion of the atoms or ions result in the spectral line broadening due to the Doppler effect $\Delta\nu$, which is proportional to the frequency ν. For the particle temperature of only 1 mK the effect can reach a few MHz.

R.H. Dicke found out that if the particle is restricted in a small volume much smaller than the wavelength of the excitation field, contribution of Doppler effect vanishes. Let us show this on an example of ion in the trap.

In the trap ion oscillates. Assume that in the laboratory frame the ion is illuminated by a monochromatic field $E(t) = E_0 \sin \omega t$. In the frame of the ion the excitation field will turn into phase-modulated field due to the Doppler effect

$$E(t) = E_0 \sin(\omega t + \delta \sin \omega_m t)\,. \tag{11.12}$$

We know, that the spectrum of the phase-modulated field consists of a carrier frequency ω and a number of equidistant side bands $\omega \pm n\omega_m$ with $n = 1\,2, \ldots, \infty$. If the phase modulation is not deep ($\delta \equiv \Delta\omega/\omega_m \ll 1$) one can take into account only the carrier, the sidebands become negligibly small (1.27). Let us re-write this condition

$$\delta \equiv \frac{\Delta\omega}{\omega_m} = \frac{\omega v_{\max}}{\omega_m c} = \frac{\omega x_{\max}}{c} = \frac{2\pi x_{\max}}{\lambda} < 1\,, \tag{11.13}$$

where we use the expression for the Doppler shift $\Delta\omega = v_{\max}\omega/c$ and the energy relation for harmonic oscillations: $mv_{\max}^2/2 = Dx_{\max}^2/2$ or $v_{\max}^2 = \omega_m^2 x_{\max}^2$.

Consider an ion which oscillations are restricted in the volume of $d = 2x_{\max}$ size. We see, that if the Lamb-Dicke condition

$$d < \frac{\lambda}{\pi} \qquad \text{(Lamb-Dicke criterium)}, \tag{11.14}$$

is fulfilled, then according to (11.13) the condition $\delta < 1$ is also fulfilled.

The more strict condition (11.14) is fulfilled, the less the Doppler broadening of the spectral line. Radiation will be mostly absorbed ion the carrier and not in the sidebands. This criterium is also valid for any particle restricted in the volume of the size less than the wavelength of the excitation field. It is important in ion traps, optical lattices and hydrogen masers.

11.3 Trap loading

Loading of the ion trap significantly differs from loading of a magneto-optical trap because of absence of any dissipative force. It means, that the particle energy should be less than the energy barrier of the trap itself and the trap cannot be loaded from outside. It is similar to the case that the planet cannot be trapped to the stable orbit if it comes from infinity. To be trappable, the body should experience some collision or be born close to the potential well center.

There are a few methods how to load ions in the trap. The most usual one is the ionization of neutral atoms directly inside the trap, for example, by electron bombardment. Neutral particles easily penetrate to the trap center without interaction with magnetic and electric field. After impact with the electron, neutral particle becomes ionized, and if its energy is low enough, would be trapped in the ion trap. The similar method is photo-ionization which is isotope-selective and causes less induced charges on the trap walls.

This method cannot be used if one traps anti-particles or exotic isotopes from the accelerator ring. In this case one can lower the potential barrier at one side of the trap, then load the trap, and, as soon as ions are inside, lift up the barrier again. Of course, it should be done very fast, before ions are reflected from the other side of the trap.

11.4 Ion cooling

Ion traps can trap ions at high energies up to 2 eV which corresponds to temperatures of 20 000 K. Even the second order Doppler effect of such particles will be huge $\Delta\nu/\nu = -v^2/(2c^2) = -(mv^2/2)/(mc^2) \approx -2\,\mathrm{eV} \approx -10^{-11}$, which is too much for any applications in the frequency standards. Since ions in the ion trap are nearly completely isolated, thermalization is inefficient and ions should be cooled by some other methods. Isolation from the environment has also a lot of advantages: As soon as ions are cooled, they will not be heated for the long time.

We should agree about notation “temperature” which is appied in this case to a single ion. We will stick to the standard agreement that temperature T will define the energy of ion per degree of freedom : $E_i = kT/2$.

11.4.1 Energy dissipation by an electric circuit

Ions, oscillating in a trap, induce mirror charges and corresponding currents in the electrodes. If electrodes are shunted by a resistor, ion will loose energy and its oscillations will dissipate. In the limit, ion temperature will be given by shunt resistor temperature.

Dissipation rate of ion in the trap with the shunt circuit. Assume, that an ion with mass m and the charge q oscillates along z-axis at a distance of $2z_0$ from each other as shown in Fig. 10.7. Resistance of the shunt is R. Derive the energy dissipation rate of an ion.

If ion with velocity v is moving for the distance ds in the electric field E, the energy gain/loss will be equal to $dW_z = qEds$.

Corresponding power equals $dW_z/dt = qEds/dt \approx qUv/(2z_0)$. Current flowing between electrodes can be derived from the following relation $IU = qUv/(2z_0)$, which means $I = qv/(2z_0)$. This approximation corresponds to the case if the trap electrodes can be considered as plane capacitor.

The equivalent electric circuit is a current source connected to the plane capacitor via resistor R. The averaged power scattered on resistor equals $\langle I^2 R\rangle$. In the case if the capacitance is small $R \ll 1/(\omega_z C)$, the averaged power, dissipated by an ion can be calculated as

$$-\frac{dW_z}{dt} = \langle I^2 R\rangle = \frac{q^2 R W_z}{4mz_0^2}\,, \tag{11.15}$$

where we used the relation $W_z = m\langle v_z^2\rangle$ for the ion kinetic energy. Solution of this equation described the exponential power dissipation which the time constant of

$$t_0 = \frac{4mz_0^2}{q^2 R}\,. \tag{11.16}$$

This cooling method can be used for any ion. Usually, the trap electrodes are shunted by resistor which is cooled in liquid nitrogen or helium. It allows to cool ions to temperatures down to 1 K. For example, such cooling is used for positron cooling for production of anti-hydrogen.

11.4.2 Buffer gas cooling

The easiest way to cool atoms in the trap is to add some buffer gas in the vacuum volume. Since the trap depth is very high, collisions with a trapped ion will take energy from it without knocking it out of the trap. Each collision between the ion and buffer gas atom will reduce the energy of the ion by

$$\frac{\Delta E_{\rm kin}}{E_{\rm kin}} = \frac{m_{\rm buffer}}{m_{\rm ion}}\,. \tag{11.17}$$

One can cool ions to temperatures down to 4 K if cryogenic trap is used.

11.4.3 Doppler laser cooling

First who suggested laser cooling of ions in the trap were Weinland and Demelt in 1975. The Doppler cooling method is very similar to laser cooling of atoms in optical molasses. It relies on absorption of the photon which the energy lower than energy of emitted one (at average). The energy difference is taken from kinetic energy of a particle.

For ions, the typical resonance lines are quite strong (natural line width 10 MHz) and allow for efficient ion cooling. Laser cooling of an ion is technically less challenging compared to a free particle cooling, because cooling only from one direction is enough. Preferably ion will absorb photons when it moves opposite to a k-vector of a red-detuned cooling laser field and will loose energy correspondingly. Typically, only one retro-reflected laser beam is used with direction intersects with all degrees of freedom of moving ion.

Using this method one can reach the lowest temperature corresponding to the Doppler limit $kT_D \equiv h\gamma/2$ (see. (8.12)). The minimal temperature is achieved when the frequency detuning is half of the resonance line width $\gamma/2$. Typically, one can reach temperatures of $T_D \approx 1\,\mathrm{mK}$. This method can be used both in Paul and Penning traps.

If many ions are trapped, the heating due to Coloumb repulsion and corresponding increased potential inharmonicity will increase the heating rate which will result in higher temperatures. There are other heating mechanisms resulting from interaction of ions with patch charges on electrodes and influence of electrical noises of electrodes which may additionally heat ions.

The Doppler temperature is typically enough to obtain crystalline structures in the trap as shown in Fig. 11.2. Still, for many applications the temperature is not low enough.

Sideband cooling

The Doppler mechanism for ions and neutral particles are very similar. The sub-Doppler mechanisms, in turn, differ very much. For neutral particles one uses *Sisyphus* or *polarization gradient* methods where atom either climb the potential lattice formed due to the dynamical Stark shift or are cooled due to redistribution of population.

For ions sub-Doppler cooling relies on the fact, that an ion is trapped in the potential well and populates some vibrational levels of this trap. If potential is harmonic, its energy levels will be equidistant as shown in Fig. 11.4. Each of the electronic energy levels can be treated as a number of sublevels separated by the harmonic trap eigenfrequency. In other words, the oscillating ion will emit a phase-modulated field with the modulation frequency of the oscillations. The spectrum can be approximated by $\nu_0 \pm k\nu_m$, where ν_0 is the ion electronic transition frequency, ν_m is the oscillation frequency and k is the positive integer.

One can excite optical transitions between different vibrational sublevels. If a vibrational level number is the same for upper and lower electronic level ($k = 0$), the ion motion will not change after absorption of the photon. But, if ion will absorb or emit photon such that $k \neq 0$, the same number of quanta of motion will be absorbed or emitted (which means that the ion will vibrate stronger or weaker in the trap). Assume, that ion is illuminated by a red-detuned laser field resonant with the next vibrational sublevels with $k = -1$ (as shown in Fig. 11.4). The emission will preferably take place at the resonance frequency ν_0. It means, that each absorption-emission event will take one quantum of vibrational energy from the ion. After multiple processes, the ion will be cooled to the ground vibrational state. The emission pattern of an ion will change, as shown in the experimental plot of Fig. 11.5. If an ion is in the ground vibrational state, the absorption with state with $k = -1$ is not possible any more and the corresponding side band becomes strongly suppressed.

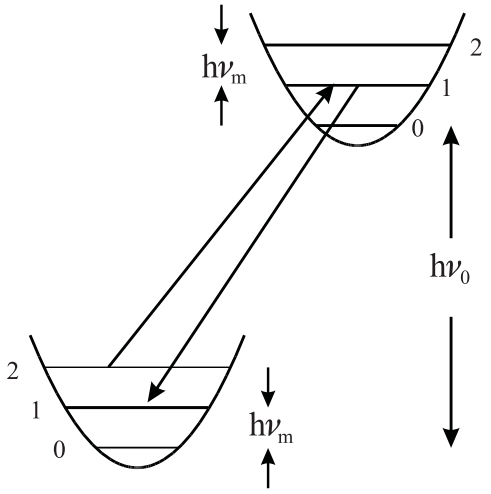


Figure 11.4: *Sideband cooling of an ion.*

Using this method one can cool ions to 10-100 μK. Sideband cooling is very important to reach the Lamb-Dicke regime in an ion trap, for quantum manipulations and quantum computation algorithms, where ion should populate a state with well-defined vibrational quantum number.

Sympathetic cooling.

One can cool ion using another *sparring* ion which is cooled by regular methods of laser cooling. Since both ions strongly interact by Coulomb interaction, taking energy from sparring ion will reduce kinetic energy or the *clock* ion as well. Typically, this method is used if the clock ion does not have strong transitions of they are not accessible by lasers. For example, one can cool

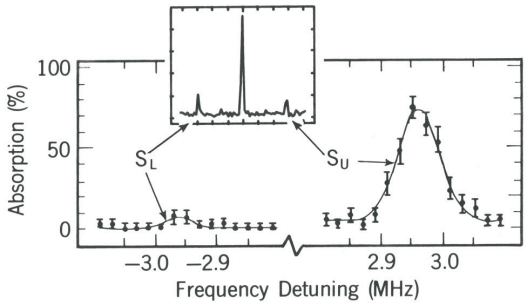


Figure 11.5: $^{198}Hg^{+}$ *ion absorption spectrum at 281.5 nm after sideband cooling. The red detuned sideband becomes strongly suppressed if the ion is cooled to the vibrational ground state.*

He^{+} ions by collisions with laser cooled Be^{+} ions. Recently, an Al^{+} clock ion was cooled by implementation of sympathetic cooling by Mg^{+} ion down to vibrational ground state and the world best result in optical clock accuracy was achieved (fractional uncertainty $< 10^{-17}$).

11.5 Detection of trapped ions

There are a few methods allowing detection of trapped ions in the trap. The most straightforward one is destructive and relies on opening the trap and detecting escaping ions by, e.g., electron channel detectors (*channeltrons*). Another method is to detect ion oscillations by sensitive measuring potential difference at electrodes. The mirror charge of moving ion will induce current in in the attach circuit which allows to detect the ion and measure its temperature.

11.5.1 Optical registration

Most regular way to detect ions in the trap (as well as neutral particles) is an optical detection. If ion is illuminated by a resonant laser field it will scatter photons which can be detected by a sensitive optical system (large NA objective lens and sensitive CCD camera). Typically, a single ion can scatter $10^6 - 10^7$ photons per second which can be readily detected with the detector having quantum efficiency of 10^{-2}. Once can image ions, study crystalline structures and ion chemistry.

The resonance laser is usually the same as the laser used for cooling.

Lecture 12: Methods of quantum logic in optical clocks

Precision measurements in the traps, electron shelving. Elements of quantum logic in ion traps. Motional degrees of freedom. CNOT gate. Cirac-Zoller gate. Information transfer between clock and cooling ions. Precision spectroscopy using quantum logic.

Ions are widely used as frequency standards because of the following reasons

- one can trap single ion which does not interact with other ions
- ion is trapped in the zero of electric potential: no external fields can perturb clock transitions
- ions can be laser cooled to the Lamb-Dicke regime
- ions possess very high-Q transitions in the microwave and optical regions

Ions, which are most successfully used in optical frequency standards are Hg^+, Al^+, Yb^+, Sr^+.

Trapped ions are also widely used as elements of quantum logic — the most successive and impressive results were obtained using ions. Here are the reasons why they are also very attractive for this field

- one can address individual ions by tight focussing of light on it
- the interaction between two ions in the trap is strong, while the isolation from the environment is nearly perfect. This Coulomb interaction is used for *quantum gates* which is the basic unit of the quantum circuit
- one can store quantum information and increase number of trapped ions (*scalability*)

For applications in quantum logic circuits one-electron Be^+, Mg^+ and Ca^+ ions are used.

Here we will consider some modern methods of precision measurements in the ion traps and some elements of quantum logic.

12.1 Electron shelving

Single trapped ions are widely used in optical clocks and demonstrate unprecedented stability and accuracy due to nearly perfect isolation from environment. But, for a clock transition with a typical life time of excited level of 1 s, the photon scattering rate cannot be higher than 0.5 photon per second which is lower than any detection level. The question arises: how to efficiently detect weak clock transition in a trapped ion?

To solve this problem the method of *electron shelving* is widely used. It can be used for ions which possess the V-scheme of levels with the common ground ground state for the weak clock and the strong cooling transition. An example of such scheme is shown in Fig. 12.1.

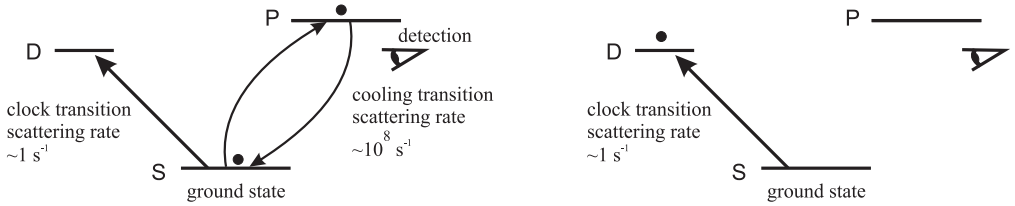


Figure 12.1: *Coupled atomic level V-scheme used for detection of quantum jumps. Left: population oscillates between S and P levels, the scattered photons are readily detected. Right: quantum jump occur. The population is transferred to D level. Ion does not scatter photons.*

Consider an ion which is illuminated simultaneously by two laser fields resonant with clock and cooling transition. Since the cooling S-P transition is very strong, the population will promptly oscillate between the ground state S and the upper state P scattering photons with high a rate up to $10^8\,\mathrm{s}^-1$. This radiation is readily detectable with sensitive optics as described in Sec.11.5.1. But, if the clock transition takes place, atom ceases scattering photons, because the ground state is depleted and the cooling laser cannot excite the strong transition.

It means, that for the time interval which lasts on the order of the D state life time (e.g. 1 s), the luminescence from the ion is not observed. A typical plot with the quantum jumps is shown in Fig. 12.2.

Using this method one can record the clock transition line. For larger detunings from the clock transition the quantum jump rate is low, it increases by approaching the resonance, reaches the maximum at exact resonance and then again decreases. Plotting the histogram one will record the clock transition spectral line shape.

The electron shelving (or quantum jumps) method is similar to quantum amplification, since the ion excitation can be detected with the nearly unit

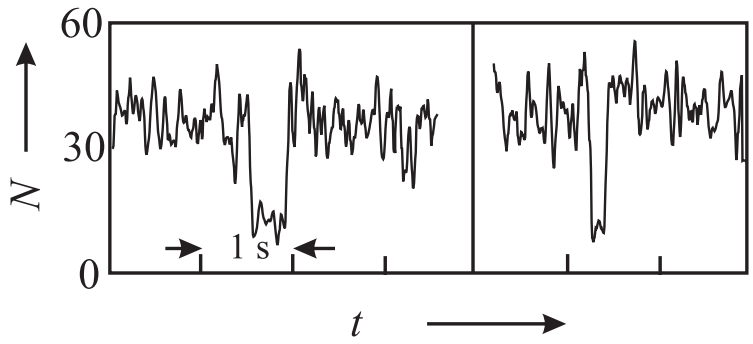


Figure 12.2: *Dark periods in the luminescence spectrum of In^+ ion in the trap corresponding to transitions in the long-living excited state.*

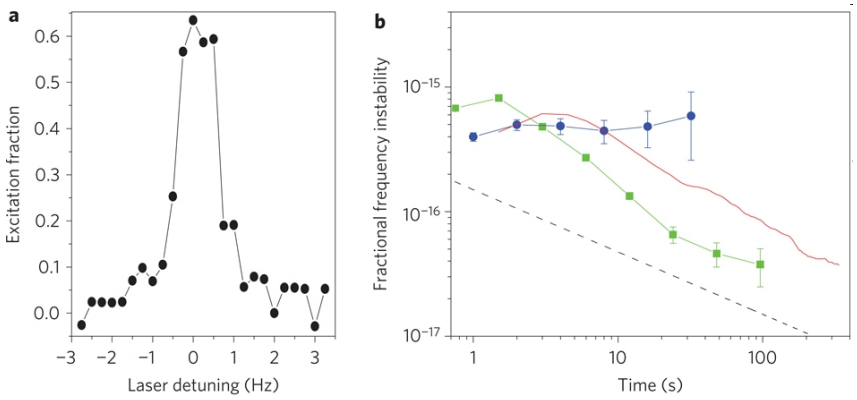


Figure 12.3: *Clock transition in an ion recorded by quantum jumps method. Corresponding frequency stability.*

probability. But, contrary to photomultiplier tubes and channel amplifiers, there amplification takes place directly in the ion.

12.2 Elements of quantum logic in ion traps

Ion in the ion trap with a long-lived (optical) transition can be considered as an isolated two-level two-level system which represents a nearly perfect *Q-bit*, i.e., and element of quantum information which can be represented as

$$|\psi\rangle = \alpha|1\rangle + \beta|2\rangle\,. \tag{12.1}$$

Using electromagnetic pulses which drive this transition one can perform different single Q-bit operations like, e.g. SWAP, by implementation a π-pulse:

$$\alpha|1\rangle + \beta|2\rangle \to \beta|1\rangle + \alpha|2\rangle\,. \tag{12.2}$$

One can implement infinite number of different one Q-bit operations on the vector $\left|\begin{array}{c}\alpha\\ \beta\end{array}\right|$ which can be represented by a unitary operator (gate) 2×2 matrix. E.g. the SWAP operation is represented by the matrix

$$X=\left|\begin{array}{cc}0 & 1\\ 1 & 0\end{array}\right|.$$

Indeed, $X\left|\begin{array}{c}\alpha\\ \beta\end{array}\right|=\left|\begin{array}{c}\beta\\ \alpha\end{array}\right|$.

A regularly implemented gate is the Hagamard gate which can be treated and a $\pi/2$-pulse acting on a two-level system:

$$H=\frac{1}{\sqrt{2}}\left|\begin{array}{cc}1 & 1\\ 1 & -1\end{array}\right|.$$

Still, similar to classical computation, it is not possible to perform any computation algorithm by using only one q-bit gate. We know, that in classical computation any logical element (AND, OR, NOT, etc.) can be implemented by different combinations of only one gate NAND.

Similar to that *any* quantum algorithm can be implemented if one can efficiently operate so-called CNOT gate (controlled NOT). Usually it is denoted as shown in Fig. 12.4.

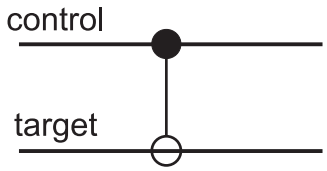


Figure 12.4: Two Q-bit CNOT gate.

CNOT gate consists of two Q-bits: one called control Q-bit which does not change during operation and one target Q-bit which may be changed depending on the state of the first Q-bit. The name "controlled NOT" describes the operation: one tries to make the NOT gate on the second target Q-bit. If the control Q-bit is in the state $|1\rangle$, the operation is successful, if it is in the state $|0\rangle$ nothing happens. This statement can be given by a raw of transformations

$$\begin{aligned}|00\rangle &\to |00\rangle\\ |01\rangle &\to |01\rangle\\ |10\rangle &\to |11\rangle\\ |11\rangle &\to |10\rangle\,,\end{aligned}$$

where the first Q-bit is the control one and the second - is target.

The gate can be also written as 4×4 matrix:

$$U_{\text{CNOT}} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{vmatrix}.$$

In 1995 Cirac and Zoller suggested how one can implement CNOT gate using ions in the linear Paul trap and then in 2002 the first CNOT gate was demonstrated in R. Blatt group using ultracold Ca^+ ions.

Another important gate is the *phase* gate

$$U_{\text{phase}} = \begin{vmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 \end{vmatrix}. \tag{12.3}$$

Transformation from the phase gate to the CNOT gate can be easily performed using two *Hagamard* gates on each of the Q-bits:

$$U_{\text{CNOT}} = \frac{1}{\sqrt{2}} \begin{vmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{vmatrix} \cdot U_{\text{phase}} \cdot \begin{vmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & -1 \end{vmatrix} \frac{1}{\sqrt{2}}, \tag{12.4}$$

which can be readily technically implemented.

12.3 Implementation of Cirac-Zoller gate

12.3.1 States of an ion

In the Cirac-Zoller gate two type of states are implemented: internal states (electronic or hyperfine transitions) or external motional states in the trap as shown in Fig. 12.5. We will denote internal logical states of an ion as $|g\rangle, |e\rangle$. These are the working states which we would like to perform a quantum gate at.

The second set of states which will be used is external motional states $|0\rangle, |0\rangle$.

For the Cirac-Zoller gate one also needs the auxiliary level $|aux\rangle$ as shown in Fig.12.6.

12.3.2 2π rotation of the spin-1/2 system

We know from the basics of quantum mechanics, that the rotation operation of the system possessing the spin S_z is given by

$$\mathcal{D}(\phi) = \exp(-\frac{i}{\hbar} S_z \phi). \tag{12.5}$$

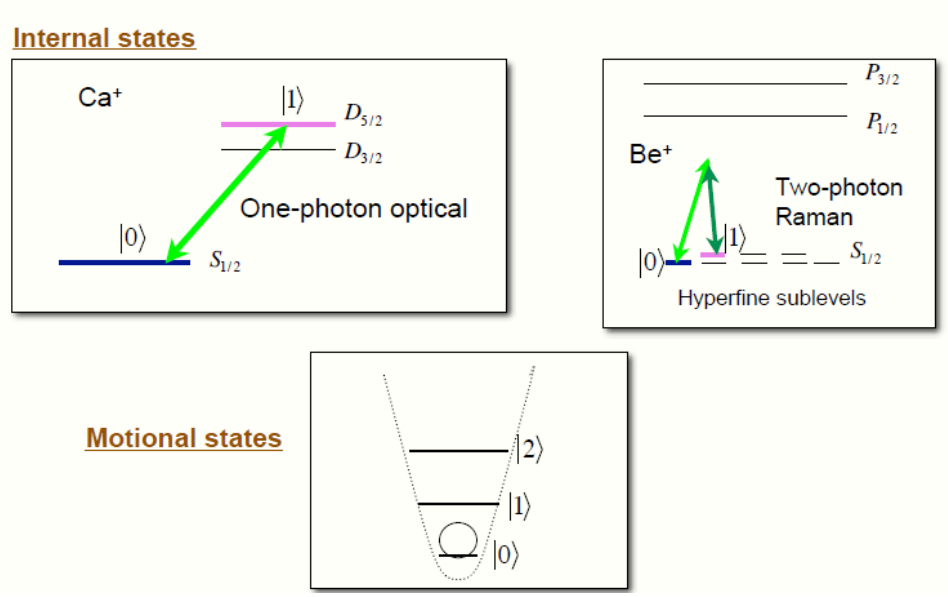


Figure 12.5: *Internal and motional states on a single ion in the trap.*

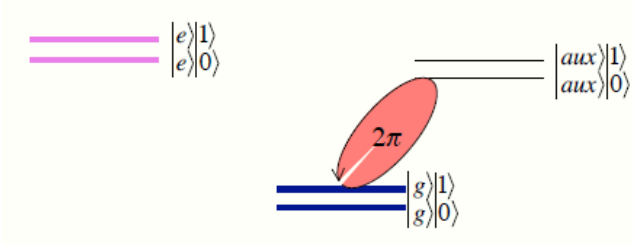


Figure 12.6: *Ion states necessary for implementation of the Cirac-Zoller gate.*

If we apply this operator with S_z (spin-1/2 system) to the state $|\alpha\rangle = |+\rangle\langle+|\alpha\rangle + |-\rangle\langle-|\alpha\rangle$ we will get

$$\exp(-\frac{i}{\hbar}S_z\phi)|\alpha\rangle = \exp\left(-\frac{i\phi}{2}\right)|+\rangle\langle+|\alpha\rangle + \exp\left(-\frac{i\phi}{2}\right)|-\rangle\langle-|\alpha\rangle \qquad (12.6)$$

we get the following important relation

$$|\alpha\rangle_{2\pi} = -|\alpha\rangle\,. \qquad (12.7)$$

This relation is illustrated in Fig.12.6. The 2π operation on the auxiliary level using the red vibrational sideband will result in

$$\begin{aligned} |g\rangle|0\rangle &\rightarrow |g\rangle|0\rangle \\ |e\rangle|0\rangle &\rightarrow |e\rangle|0\rangle \\ |g\rangle|1\rangle &\rightarrow -|g\rangle|1\rangle \\ |e\rangle|1\rangle &\rightarrow |e\rangle|1\rangle\,. \end{aligned} \tag{12.8}$$

12.3.3 Collective vibrational modes

If two ions are sitting in the potential well of the trap and interact with each other via Coulomb interaction, they experience collective vibrational modes. The number of modes along one of the coordinates coincide with the number of ions. An example is shown in Fig. 12.7.

It means, that if one of the ions will absorb a blue detuned photon resonance with one of the modes, the shared oscillations will start and for *both* ions the number of level in the potential well will increase (Fig. 12.5). And vice versa, if one shines red detuned laser, the number of shared quanta will decrease.

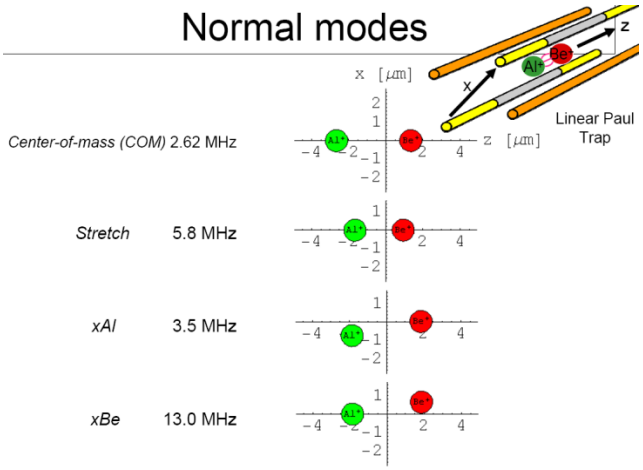


Figure 12.7: *Example of collective vibrational modes for two ions (Al^+, Be^+).*

An important message: if the ion(s) is(are) in the true ground state $|g\rangle|0\rangle$, it cannot absorb a red detuned laser since there is no $|e\rangle|-1\rangle$ level. This is widely used in ion logic schemes since it allows to address only one of the two lower levels in the ladder.

12.3.4 CNOT gate

We will consider how the CNOT gate can be implemented on two ions in the ion trap which are initially cooled to the vibrational ground state. Now we can

write the states and wavefuntions which correspond to operations illustrated in Fig. 12.8.

Initially the system is prepared in the state ψ_0:

$$\begin{aligned} &|g\rangle_c|g\rangle_t|0\rangle \\ &|g\rangle_c|e\rangle_t|0\rangle \\ &|e\rangle_c|g\rangle_t|0\rangle \\ &|e\rangle_c|e\rangle_t|0\rangle \end{aligned} \tag{12.9}$$

The first operation is a π-pulse on the red sideband applied to the control

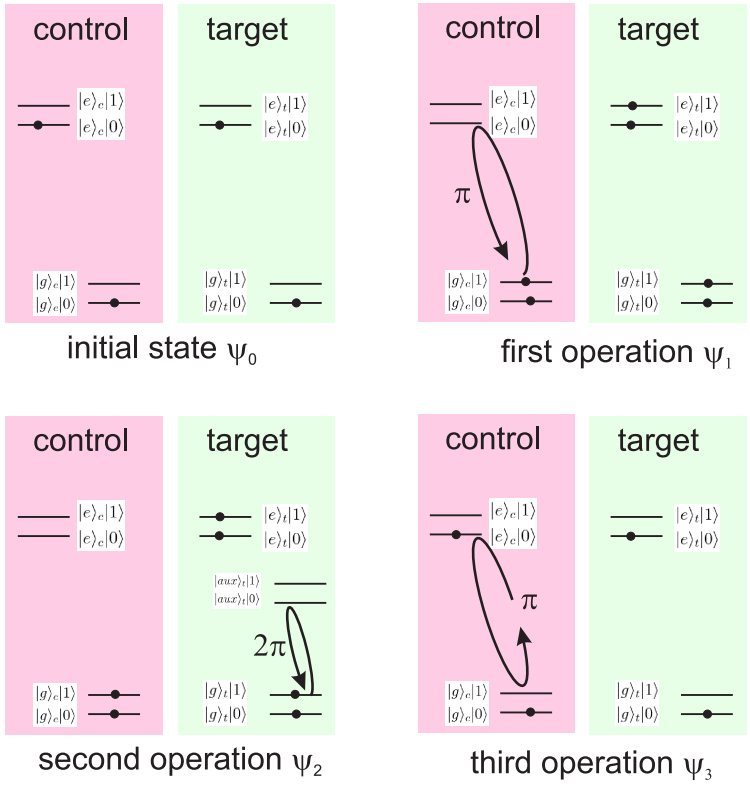


Figure 12.8: *Operation sequence for the Cirac-Zoller gate.*

gate. It will excite a collective oscillations of the both ions:

$$\begin{aligned} |g\rangle_c|g\rangle_t|0\rangle &\rightarrow |g\rangle_c|g\rangle_t|0\rangle \\ |g\rangle_c|e\rangle_t|0\rangle &\rightarrow |g\rangle_c|e\rangle_t|0\rangle \\ |e\rangle_c|g\rangle_t|0\rangle &\rightarrow -i|g\rangle_c|g\rangle_t|1\rangle \\ |e\rangle_c|e\rangle_t|0\rangle &\rightarrow -i|q\rangle_c|e\rangle_t|1\rangle \end{aligned} \tag{12.10}$$

Additionally, it will transfer population from the excited state of the control ion to its ground state with the corresponding phase which can be calculated from (12.5).

The second operation is the 2π-pulse on the target ion which will change the phase of the particular state $|g\rangle_t|1\rangle$:

$$\begin{aligned} |g\rangle_c|g\rangle_t|0\rangle &\rightarrow |g\rangle_c|g\rangle_t|0\rangle \\ |g\rangle_c|e\rangle_t|0\rangle &\rightarrow |g\rangle_c|e\rangle_t|0\rangle \\ -i|g\rangle_c|g\rangle_t|1\rangle &\rightarrow +i|g\rangle_c|g\rangle_t|1\rangle \\ -i|q\rangle_c|e\rangle_t|1\rangle &\rightarrow -i|q\rangle_c|e\rangle_t|1\rangle \end{aligned} \tag{12.11}$$

The last operation is again the π-pulse on the red sideband applied to the control ion:

$$\begin{aligned} |g\rangle_c|g\rangle_t|0\rangle &\rightarrow |g\rangle_c|g\rangle_t|0\rangle \\ |g\rangle_c|e\rangle_t|0\rangle &\rightarrow |g\rangle_c|e\rangle_t|0\rangle \\ +i|g\rangle_c|g\rangle_t|1\rangle &\rightarrow |e\rangle_c|g\rangle_t|0\rangle \\ -i|q\rangle_c|e\rangle_t|1\rangle &\rightarrow -|e\rangle_c|e\rangle_t|0\rangle \end{aligned} \tag{12.12}$$

We see that the system is returned back to the initial state ψ_0 for all states except the last one which changed the sign.

As we see from (12.3) it corresponds to the phase gate which can be converted to CNOT gate by two simple Hagamard operations (12.4).

12.4 Information transfer between clock and cooling ions. Precision spectroscopy using quantum logic.

One of the most successful implementation of quantum logic algorithms in practice is the recent demonstration of reading the clock transition in Al^+ ion with the help of an auxiliary sparring ion Be^+. Aluminum ion has very narrow and attractive clock transition at 267 nm. One of the most important advantages is that this transition is barely perturbed by the black body radiation.

The problem with Al^+ ion is that its strong cooling transition is at 167 nm which cannot be excited with modern lasers. It means that the electron shelving scheme cannot be implemented here (see Fig. 12.1). The problem of cooling can be overcome by implementation of the sparring ion (e.g. Be^+), but how then to read the information that the clock transition is exited?

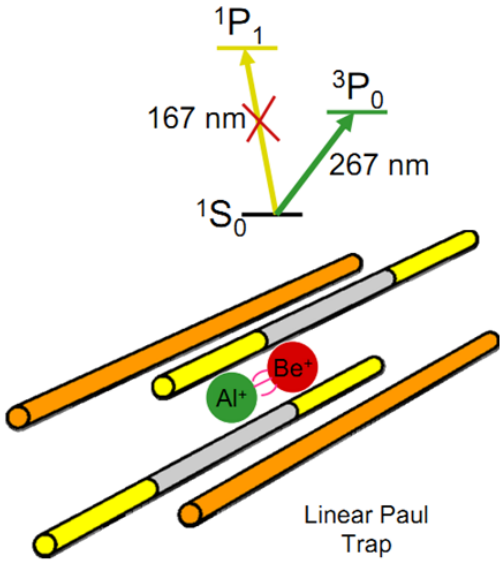


Figure 12.9: Clock transition in Al^+ and sympathetic cooling of this ion in the trap with Be^+

D. Weiland suggested to implement quantum algorithm to detect this transition which is illustrated in Fig. 12.10. Situation is very similar to the one described above when we considered a CNOT gate.

The initial state of the system is :

$$|\,\psi_0\rangle = |g\rangle_L\,|g\rangle_C\,|0\rangle\,. \tag{12.13}$$

which means that both a logical ion L and the clock ion C are in the ground states (electronic and vibrational).

After excitation of the clock transition the clock ion will be excited to the coherent superposition with amplitudes α and β (Fig. 12.10 b)):

$$\begin{aligned} |\psi_0\rangle \to |\psi_1\rangle &= |g\rangle_L\,[\alpha|g\rangle_C + \beta|e\rangle_C]\,|0\rangle \\ &= |g\rangle_L\,[\alpha|g\rangle_C\,|0\rangle + \beta|e\rangle_C\,|0\rangle]\,. \end{aligned} \tag{12.14}$$

Next, a blue-detunedπ-pulse is applied to the clock ion which will excited the common emotional excitation of both ions. It will influence only the ground state $|g\rangle_C$, since there is no resonance for $|e\rangle_C$ ($|g\rangle_C\,|-1\rangle$). ,

$$\begin{aligned} |\psi_1\rangle \to |\psi_2\rangle &= |\uparrow\rangle_L\,[\alpha|\uparrow\rangle_C\,|1\rangle_M + \beta|\uparrow\rangle_C\,|0\rangle_M] \\ &= |\downarrow\rangle_L\,|\uparrow\rangle_C\,[\alpha\,|1\rangle_M + \beta\,|0\rangle_M]\,. \end{aligned} \tag{12.15}$$

The π-pulse exciting the electronic state simultaneously transferred the quantum information to the logic ion via Coloumb interaction (common oscillations). The amplitude of the motional sates will be equal to the excitation

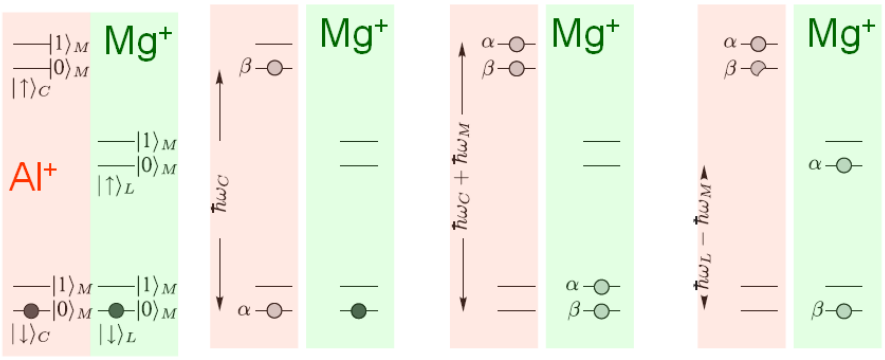


Figure 12.10: *Excitation of the clock ion C and reading by a logical ion L. a) – the initial state b) – excitation of the clock ion c) – π-pulse on the blue sideband applied to the clock ion d) – π-pulse at the red sideband .*

amplitudes in the clock ion. Now one can transfer these motional amplitudes to the amplitudes of the electronic states in the logic ion by applying a red-detuned π-pulse:

$$|\psi_2\rangle \rightarrow |\psi_{\text{final}}\rangle = [\alpha\,|e\rangle_L + \beta\,|g\rangle_L]\,|e\rangle_C\,|0\rangle\,. \tag{12.16}$$

Now one can readily read values α^2 and β^2 by the regular shelving method (projecting the states of the logic ion on the $|g\rangle\,, |e\rangle$) basis. It allows to measure the excitation probability of the clock ion.

Using this method, D. Wineland and co-workers reached extremely high frequency reproducibility of $< 10^{-17}$ and set the best known to date restriction to the possible variation of the fine structure constant α.

Lecture 13: Optical frequency measurements

Frequency conversions in optical domain. Second harmonic generation, phase modulation. Frequency dividers and frequency chains. Femtosecond mode-locked lasers. Time domain and frequency domain representation of femtosecond pulse train. Phase and group velocities in the laser cavities, carrier envelope offset frequency. Spectral broadening in nonlinear photonic crystal fiber. Nonlinear interferometer. Measuring absolute frequency of laser radiation.

As we know from the first lecture, increasing the carrier frequency ν_0 will in general case increase the accuracy of the clock which is defined by the resonance quality factor $Q = \nu/\Delta\nu$. Optical clocks possessing high carrier frequency of up to 10^{15} Hz demonstrate unprecedented stability and reproducibility as was explained in previous lectures. One can synthesize a highly stable electromagnetic wave at a few particular optical frequencies using different atomic samples.

Still, for the practical use this situation is not favorable. We know, that most of the communications, time and frequency signal distribution take place in radio- and microwave frequency domains covering 1 kHz-10 GHz. In this domain signals can be easily converted, counted, mixed and multiplied using regular semiconductors. From approx. 40 GHz the semiconductors stop working. It means that for optical clock the clockwork device, which will help to convert signals and transfer stability to the radio-frequency domain or or other optical frequencies was missing (see Fig. 1.7).

Since late 80-s there were a lot of attempts to built a system which will serve as a clockwork for optical clocks and a few so called *frequency chains* have been built in the number of leading scientific centers. The basics of any frequency conversion lays in non-linear processes (the second harmonic generation, sum frequency generation, the four-wave mixing, parametric downconversion) which allow to transfer stability from one frequency range to another.

The real breakthrough in this field was achieved by J. Hall and T.W. Hänsch by invention of a *frequency comb*. Here we will describe some of the processes involved and study the operation of frequency comb.

13.1 Introduction to some optical non-linear processes

Non-linearities in the optical domain are very weak and strong fields are necessary to reach strong effects. Non linear processes result from non-linear medium response — its polarization $P(E)$. In linear approximation $P \propto E$. In the general case $P(E)$ is the non-linear function and can be written as power series of E

$$P(E) = \epsilon_0 \left[\chi^{(1)}E + \chi^{(2)}E^2 + \chi^{(3)}E^3 + \ \cdots\right] . \tag{13.1}$$

Here $\chi^{(i)}$ are the non-linear susceptibility coefficients which are responsible for the process of the ith order. The square contribution in (**??**) is the tensor

$$P_i = \epsilon_0 \sum_{j,k=1}^{3} \chi^{(2)}_{i,j,k} E_j E_k \qquad i,j,k = 1,2,3\,. \tag{13.2}$$

If two waves with amplitudes E_1 and E_2 are superimposed the result is

$$\begin{aligned} (E_1 + E_2)^2 &= E_{01}^2 \cos^2\omega_1 t + 2E_{01}E_{02}\cos\omega_1 t\cos\omega_2 t + E_{02}^2\cos^2\omega_2 t \\ &= \frac{E_{01}^2}{2}(1-\cos 2\omega_1 t) + \frac{E_{02}^2}{2}(1-\cos 2\omega_2 t) \\ &+ E_{01}E_{02}[\cos(\omega_2-\omega_1)t - \cos(\omega_2+\omega_1)t]\,, \end{aligned} \tag{13.3}$$

where the terms with doubled frequencies, sum and differential frequencies appear.

The expression (13.3) presents three types of non-linear processes. One can treat this equation as if two photons with the frequencies ω_1 and ω_2 disappear, and one photon with the frequency ω_3 appear. If $\omega_1 = \omega_2$, the process is called *second harmonic generation*. If the frequencies are not equal it is called *sum frequency generation*.

If one reads (13.3), one photon ω_3 will born two photons which is called *optical parametric generation*.

The third process is *optical heterodyning* when the photon $\omega_1 - \omega_2$ is born. To satisfy the energy conservation law two other photons should appear: $\omega_1 + \omega_2 = \omega_1 + (-\omega_2 + \omega_2) + \omega_2 = (\omega_1 - \omega_2) + 2\omega_2$.

The four wave mixing involves three photons and is described by the cubic term in (13.1). If the condition of phase synchronism is fulfilled, three photons with frequencies $\omega_1, \omega_2, \omega_3$ will be converted to other frequencies. An important combination for this discussion is $\omega_4 = \omega_1 + \omega_2 - \omega_3$.

Please note, that due to parity and momentum conservation reasons the two-photon processes may take place only with anisotropic media (some crystals), while in anisotropic materials (glass, gas, liquid) they are very inefficient. On the other hand, the three-photon process does not have this restriction and can take place in isotropic media as well.

13.2 Ultrashort pulses and femtosecond laser basics

The heart of a frequency comb is a femtosecond laser which basics of operation will be briefly considered here.

Any laser is a device which has and amplification medium and the cavity around it. The laser can oscillate within the amplification spectrum of the medium at certain frequencies which are defined by a cavity length L. Separation between the cavity modes will be given as $c/2L$ for e.g. linear cavity. Formerly we considered single frequency lasers where one single laser mode is carefully selected by implementation of selective elements.

To build the pulsed laser, one has to reverse the task and try to excite as many modes in the laser as possible. It can be easily understood from the uncertainty relation $\delta\nu\tau \approx 1$. It indicates, that to reach short pulse regime one needs a very broad emission spectrum. For example, the spectrum of a femtosecond laser should spread over tens of nanometers.

Here is a simple illustration how one can get a pulse sequence from a set of continuous modes. Assume that the field consists of a number of monochromatic waves with the unit amplitude $E_n = e^{i(\omega_0 + n\Delta\omega)t}$, where n is the integer number. Here the freqeuncy ω_0 stands for the carrier frequency (e.g. optical frequency) and the freqeuncy ω will be given by the distance between the next cavity modes $\Delta\omega = 2\pi(c/2L)$.

Writing down the sum we will get

$$\begin{aligned} E(t) = \sum_{n=0}^{N-1} e^{i(\omega_0 + n\Delta\omega)t} = e^{i\omega_0 t} \sum_{n=0}^{N-1} e^{in\Delta\omega t} = \\ = e^{i\omega_0 t} \left[\sum_{n=0}^{\infty} e^{in\Delta\omega t} - \sum_{n=N}^{\infty} e^{in\Delta\omega t} \right] = \\ = e^{i\omega_0 t} \left[\frac{1}{1 - e^{i\Delta\omega t}} - e^{iN\Delta\omega t} \frac{1}{1 - e^{i\Delta\omega t}} \right] = \frac{1 - e^{iN\Delta\omega t}}{1 - e^{i\Delta\omega t}} e^{i\omega_0 t} , \end{aligned} \tag{13.4}$$

where we use the expression $\sum_{n=0}^{\infty} q^n = 1/(1-q)$ for $|q| < 1$. For the intensity we will get

$$I(t) \propto |E(t)|^2 = \frac{1 - \cos N\Delta\omega t}{1 - \cos \Delta\omega t} = \frac{\sin^2 N\Delta\omega t/2}{\sin^2 \Delta\omega t/2} . \tag{13.5}$$

An example for $N = 21$ is shown in Fig. 13.1. One can see that the intensity pattern looks like a periodic number of pulses which are separated by the interval $T = 2L/2$. The repetition rate of these pules is called the *repetition rate* and is equal to $f_{\text{rep}} = \Delta\omega/2\pi$. The width of the pulse, in turn, will be given by N: the pulse width τ is given by

$$\tau \approx \frac{2\pi}{\Delta\omega N} . \tag{13.6}$$

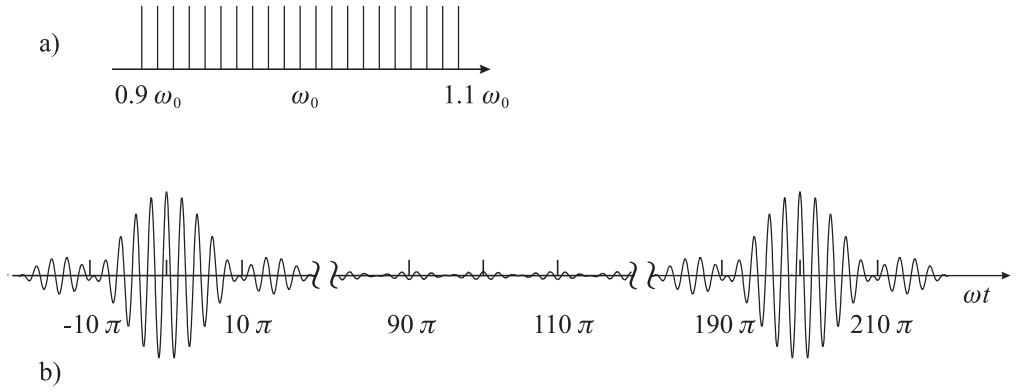


Figure 13.1: a) The set of $N = 21$ equidistant frequencies separated by $\Delta\omega - 0.01\omega_0$. b) The set of pulses for the case a) calculated according to (13.5).

It is clear that the pulse becomes shorter by increasing the number of contributing modes. For femtosecond lasers the number of modes reaches 10^6. It is necessary to point out that the calculations are valid only in the case if all phases of the modes in are the same (all of them equal zero $E_n = e^{i(\omega_0+n\Delta\omega)t}$). In more general case $E_n = e^{i(\omega_0+n\Delta\omega)t+\phi_n}$. If modes are not synchronized and all of the ϕ_n are different one will not obtain periodic pulse sequence.

13.3 EOM as the frequency shifting element

In the *electro-optical modulator* the phase of light field penetrating the birefringent crystal depends on the applied voltage. E.g the crystals which can be used as EOMs are $(NH_4)H_2PO_4$ (ADP), $LiTaO_3$, $LiNbO_3$). It is necessary that the crystal's optical axis was orthogonal to the light k-vector as shown in Fig. 13.2.

The light beam will be split in two — ordinary and extraordinary with corresponding refraction indexes n_o and n_e, the latter depends on the field E_z as

$$n = \left(n_e - \frac{1}{2} n_e^3 r_{zz} E_z \right) . \tag{13.7}$$

Here r_{zz} is the tensor describing the non-linear responde of the crystal

$$\delta\phi = \frac{2\pi n}{\lambda} L = \frac{n_e^3 r_{zz}}{\lambda} \frac{L}{d} \pi U_m \equiv \pi \frac{U_m}{V_\pi} , \tag{13.8}$$

where $U_m = dE_z$ is the voltage applied to the crystal. Voltage V_π, corresponding to $\delta\phi = \pi$:

$$V_\pi = \frac{\lambda}{n_e^3 r_{zz}} \frac{d}{L} . \tag{13.9}$$

EOM's are widely used for phase and frequency (intra cavity) modulation, for locking lasers and for optical frequency comb generation.

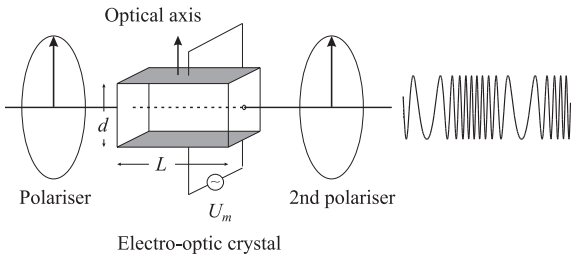


Figure 13.2: Phase-modulator based on a crystal.

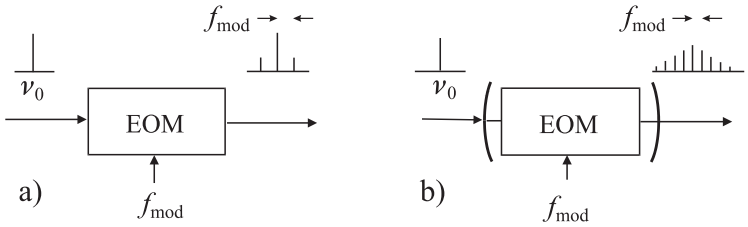


Figure 13.3: Left — EOM generating the sidebands. Right – EOM generating sidebands in the optical cavity.

13.3.1 EOM for the frequency comb synthesis

The sidebands which are generated by and EOM can be used for establishing a coherent "bridge" between and measurement the optical frequencies. To reach that one has to increase the modulation frequency and the depth of modulation.

To increase the modulation depth one can place the EOM in the optical resonator as shown in Fig. 13.3. The power in the k-th sideband will be given by (in respect to the carrier power P_c)

$$\frac{P_k}{P_c} = \exp\left(-\frac{\pi|k|}{F^*\ \delta}\right)\,, \tag{13.10}$$

where δ is the modulation depth and F^* is the finesse of the optical cavity. Typically, the power will decrease as 30 dB/THz by detuning from the carrier.

13.4 Kerr mode-locking

One of the very important mechanisms is the Kerr-lens mode-locking effect which is caused by the third-order susceptibility $\chi^{(3)}E^3$ in (13.1). If we leave only third order terms in (13.1), we get

$$D = \epsilon_0 E + P = \epsilon_0 \left(1 + \chi^{(1)}\right) E + \chi^{(3)} E^3 = \epsilon_0 \left[1 + \chi^{(1)} + \epsilon_0^{-1}\chi^{(3)} E^2\right] E\,. \tag{13.11}$$

The square brackets will give us non-linear susceptibility :

$$\epsilon' = \epsilon_1 + \epsilon_2 E^2 \tag{13.12}$$

with the linear part $\epsilon \equiv 1+\chi^{(1)}$ and the coefficient $\epsilon_2 \equiv \chi^{(3)}/\epsilon_0$. Since $n = \sqrt{\epsilon'}$ we get from (13.12)

$$n \approx n_0 + n_2 I\,. \tag{13.13}$$

The refraction index is proportional to the laser field intensity I. Typically around800 nm $n_2 \approx 3 \times 10^{-16}\,\mathrm{cm}^2/\mathrm{W}$.

If we consider a Gaussian beam, the intensity will be distributed radially which results in lensing effect. In the center of the beam the intensity is higher, the refraction index is higher which is analogy to the positive lens. This lens effect in the sapphire crystal or fiber the results in preferential soliton propagation (high intensity pulses).

Besides lensing effect, the kerr effect results in generation of new frequencies and pulses become chirped. If the pulse shape is $I(t) = I_0(t)[1-(t/\tau_p)^2+\cdots]$, the Kerr medium will transform it to :

$$E(t) \propto \exp[-(t/\tau_p)^2]\exp(i\omega_0 t)\exp(i\omega_0 L c^{-1}\{n_0 + n_2 I_0[1-(t/\tau_p)^2]\})\,. \tag{13.14}$$

Since

$$\Phi(t) = \omega_0 t + \omega_0 L c^{-1}\{n_0 + n_2 I_0[1-(t/\tau_p)^2]\}\,, \tag{13.15}$$

we get the instant frequency

$$\omega(t) \equiv \frac{d}{dt}\Phi(t) = \omega_0 - 2\omega_0 \frac{n_2 I_0 L}{c\tau_p^2} t\,. \tag{13.16}$$

13.4.1 Propagation of ultra short pulses

Propagation of short pulses in dispersive media have few new features compared to the monochromatic wave. Short pulse covers broad spectral interval and the pulse shape and delay can change significantly.

Consider the wave vector $k = 2\pi/\lambda = \omega n(\omega)/c$. The power series of k around ω_0:

$$k(\omega) = k(\omega_0) + (\omega-\omega_0)\left.\frac{dk}{d\omega}\right|_{\omega=\omega_0} + \frac{1}{2}(\omega-\omega_0)^2\left.\frac{d^2k}{d\omega^2}\right|_{\omega=\omega_0} + \cdots\,. \tag{13.17}$$

The first term

$$k(\omega_0) \equiv \frac{\omega}{v_\phi} \tag{13.18}$$

describes propagation of the sine carrier wave ω_0 inside the pulse envelope. The phase after the distance z equals $zk(\omega_0)$, and the time is given by $t_\phi = k(\omega_0)z/\omega_0 = z/v_\phi$.

The second term

$$\left.\frac{dk}{d\omega}\right|_{\omega=\omega_0} = \frac{1}{v_g} \tag{13.19}$$

gives the propagation of the envelope of the pulse v_g . The corresponding refraction index equals

$$n_g(\lambda) \equiv \frac{c}{v_g} = c\frac{dk}{d\omega} = c\frac{d}{d\omega}\frac{\omega\, n}{c} = n + \omega\frac{dn}{d\omega} = n(\lambda) - \lambda\frac{d}{d\lambda}n(\lambda)\,, \tag{13.20}$$

where we used $d\omega/d\lambda = -\omega/\lambda$.

The third term in (13.17) is so-called group velocity dispersion

$$\left.\frac{d^2k}{d\omega^2}\right|_{\omega=\omega_0} = \left.\frac{d}{d\omega}\frac{1}{v_g(\omega)}\right|_{\omega=\omega_0}. \tag{13.21}$$

The pulse shape changes after propagating in the medium. It is typical characterized by the parameter

$$D \equiv \frac{1}{L}\frac{dT}{d\lambda}\,, \tag{13.22}$$

where λ is the wavelength in vacuum, T is the propagation time for the pulse along the length L in the medium. Since $T = L/v_g$

$$D = \frac{d\frac{1}{v_g}}{d\lambda} = -\frac{\omega}{\lambda}\frac{d\frac{1}{v_g}}{d\lambda} = -\frac{2\pi c}{\lambda^2}\frac{d^2k}{d\omega^2}\,, \tag{13.23}$$

In the optical fibers the dispersion of the group velocity is caused also by the waveguide dispersion.

13.5 Precision optical spectroscopy and optical frequency measurements

The principle of modern optical frequency measurement is presented in fig. 13.4. A laser is tuned to the wavelength of a narrow metrological transition (usually referred to as a "clock transition") in an atomic, ionic or molecular sample. Most commonly, the laser frequency is stabilized by active feedback to a transmission peak of a well isolated optical cavity ("reference cavity") which allows to achieve sub-hertz spectral line width of the interrogating laser. Some recent advances in laser stabilization technique will be described in section **??**. The laser frequency is then scanned across the transition which allows to find the line center ω_0 using an appropriate line shape model. The measured transition quality factor can reach 10^{15} which provides extremely high resolution. To obtain the transition frequency the beat note ω_{beat} between the laser and one of the modes of the stabilized frequency comb is measured with the help

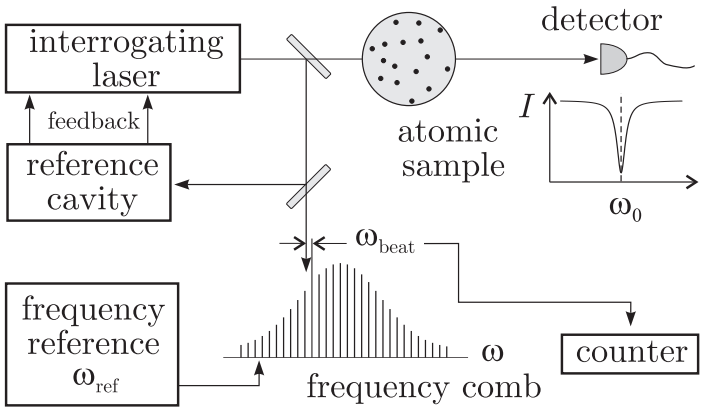


Figure 13.4: Setup for the measurement an optical transition frequency in an atomic sample with the help of an optical frequency comb.

of a frequency counter. Details of this type of measurement are presented in section 13.5.1. If the comb is stabilized to a primary frequency reference (i.e. a Cs atomic clock), the measurement presented in fig. 13.4 will yield the *absolute* frequency of the optical transition. Absolute frequency measurements allow a comparison of different results obtained at laboratories all over the world. On the other hand, if the comb is stabilized with the help of some other reference, which can be e.g. another optical frequency, the measurement will yield the ratio $\omega_0/\omega_{\mathrm{ref}}$. One can thus compare transition frequencies in different atomic samples avoiding time-consuming absolute frequency measurements.

13.5.1 Ultra-short pulse lasers and frequency combs

Frequency can be measured with by far the highest precision of all physical quantities. In the radio frequency domain (say up to 100 GHz), frequency counters have existed for a long time. Almost any of the most precise measurements in physics have been performed with such a counter that uses an atomic clock as a time base. To extend this accurate technique to higher frequencies, so called harmonic frequency chains have been constructed since the late 1960ies. Because of the large number of steps necessary to build a long harmonic frequency chain, it was not before 1995 when visible laser light was first referenced phase coherently to a cesium atomic clock using this method.

The disadvantage of these harmonic frequency chains was not only that they could easily fill several large laser laboratories at once, but that they could be used to measure a single optical frequency only. Even though mode locked lasers for optical frequency measurements have been used in rudimentary form in the late 1970ies, this method became only practical with the advent of femtosecond (fs) mode locked lasers. Such a laser necessarily emits a very

broad spectrum, comparable in width to the optical carrier frequency.

In the frequency domain a train of short pulses from a femtosecond mode locked laser is the result of a phase coherent superposition of many continuous wave (cw) longitudinal cavity modes. These modes at ω_n form a series of frequency spikes that is called a frequency comb. As has been shown, the modes are remarkably uniform, i.e. the separation between adjacent modes is constant across the frequency comb. This strictly regular arrangement is the most important feature used for optical frequency measurement and may be expressed as:

$$\omega_n = n\omega_r + \omega_{CE}\,. \tag{13.24}$$

Here the mode number n of some 10^5 may be enumerated such that the frequency offset ω_{CE} lies in between 0 and $\omega_r = 2\pi/T$. The mode spacing is thereby identified with the pulse repetition rate, i.e. the inverse pulse repetition time T. With the help of that equation two radio frequencies ω_r and ω_{CE} are linked to the optical frequencies ω_n of the laser. For this reason mode locked lasers are capable to replace the harmonic frequency chains of the past.

To derive the frequency comb properties as detailed by (13.24), it is useful to consider the electric field $E(t)$ of the emitted pulse train. We assume that the electric field $E(t)$, measured for example at the lasers output coupling mirror, can be written as the product of a periodic envelope function $A(t)$ and a carrier wave $C(t)$:

$$E(t) = A(t)C(t) + c.c.\,. \tag{13.25}$$

The envelope function defines the pulse repetition time $T = 2\pi/\omega_r$ by demanding $A(t) = A(t-T)$. The only thing about dispersion that should be added for this description, is that there might be a difference between the group velocity and the phase velocity inside the laser cavity. This will shift the carrier with respect to the envelope by a certain amount after each round trip. The electric field is therefore in general not periodic with T. To obtain the spectrum of $E(t)$ the Fourier integral has to be calculated:

$$\tilde{E}(\omega) = \int_{-\infty}^{+\infty} E(t)e^{i\omega t}dt\,. \tag{13.26}$$

Separate Fourier transforms of $A(t)$ and $C(t)$ are given by:

$$\tilde{A}(\omega) = \sum_{n=-\infty}^{+\infty} \delta\left(\omega - n\omega_r\right)\tilde{A}_n \qquad \text{and} \qquad \tilde{C}(\omega) = \int_{-\infty}^{+\infty} C(t)e^{i\omega t}dt\,. \tag{13.27}$$

A periodic frequency chirp imposed on the pulses is accounted for by allowing a complex envelope function $A(t)$. Thus the "carrier" $C(t)$ is defined to be whatever part of the electric field that is non-periodic with T. The convolution theorem allows us to calculate the Fourier transform of $E(t)$ from $\tilde{A}(\omega)$ and

$\tilde{C}(\omega)$:

$$\tilde{E}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} \tilde{A}(\omega')\tilde{C}(\omega-\omega')d\omega' + c.c. = \frac{1}{2\pi}\sum_{n=-\infty}^{+\infty} \tilde{A}_n\tilde{C}\left(\omega - n\omega_r\right) + c.c.\,. \tag{13.28}$$

The sum represents a periodic spectrum in frequency space. If the spectral width of the carrier wave $\Delta\omega_c$ is much smaller than the mode separation ω_r, it represents a regularly spaced comb of laser modes just like (13.24), with identical spectral line shapes. If $\tilde{C}(\omega)$ is centered at say ω_c, then the comb is shifted by ω_c from containing only exact harmonics of ω_r. The frequencies of the mode members are calculated from the mode number n:

$$\omega_n = n\omega_r + \omega_c\,. \tag{13.29}$$

The measurement of the ω_c as described below usually yields a value modulo ω_r, so that renumbering the modes will restrict the offset frequency to smaller values than the repetition frequency and (13.24) and (13.29) are identical.

If the carrier wave is monochromatic $C(t) = e^{-i\omega_c t - i\varphi}$, its spectrum will be δ-shaped and centered at the carrier frequency ω_c. The individual modes are also δ-functions $\tilde{C}(\omega) = \delta(\omega-\omega_c)e^{-i\varphi}$. The frequency offset (13.29) is identified with the carrier frequency. According to (13.25) each round trip will shift the carrier wave with respect to the envelope by $\Delta\varphi = \arg(C(t-T)) - \arg(C(t)) = \omega_c T$ so that the frequency offset may also be identified by $\omega_{CE} = \Delta\varphi/T$. In a typical laser cavity this pulse-to-pulse carrier-envelope phase shift is much larger than 2π, but measurements usually yield a value modulo 2π. The restriction $0 \le \Delta\varphi \le 2\pi$ is synonymous with the restriction $0 \le \omega_{CE} \le \omega_r$ introduced above. Figure 13.5 sketches this situation in the time domain for a chirp free pulse train.

Extending the frequency comb

The spectral width of a pulse train emitted by a fs laser can be significantly broadened in a single mode fiber by self phase modulation. Assuming a single mode carrier wave, a pulse that has propagated the length L acquires a self induced phase shift of

$$\Phi_{NL}(t) = -n_2 I(t)\omega_c L/c\,, \tag{13.30}$$

where the pulse intensity is given by $I(t) = \frac{1}{2}c\varepsilon_0|A(t)|^2$. For fused silica the non-linear Kerr coefficient n_2 is comparatively small but almost instantaneous even on the time scale of fs pulses. This means that different parts of the pulse travel at different speed. The result is a frequency chirp across the pulse without affecting its duration. The pulse is no longer at the Fourier limit so that the spectrum is much broader than the inverse pulse duration where the

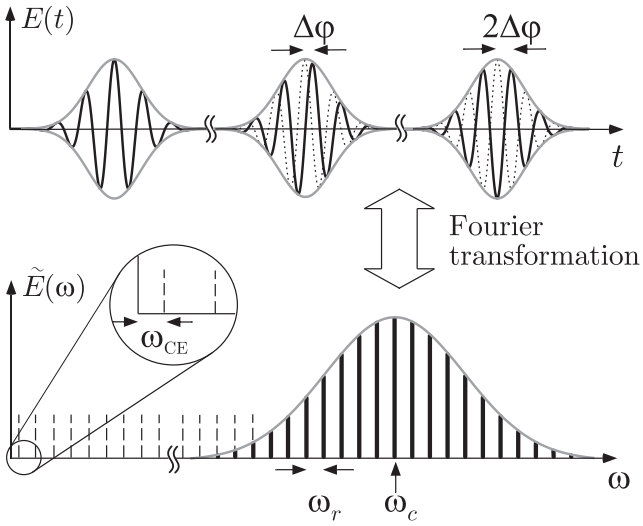


Figure 13.5: Consecutive un-chirped pulses ($A(t)$ is real) with carrier frequency ω_c and the corresponding spectrum (not to scale). Because the carrier propagates with a different velocity within the laser cavity than the envelope (with phase- and group velocity respectively), the electric field does not repeat itself after one round trip. A pulse-to-pulse phase shift $\Delta\varphi$ results in an offset frequency of $\omega_{CE} = \Delta\varphi/T$. The mode spacing is given by the repetition rate ω_r. The width of the spectral envelope is given by the inverse pulse duration up to a factor of order unity that depends on the pulse shape.

extra frequencies are determined by the time derivative of the self induced phase shift $\dot{\Phi}_{NL}(t)$. Therefore pure self-phase modulation would modify the envelope function in (13.25) according to

$$A(t) \longrightarrow A(t)e^{i\Phi_{NL}(t)}. \tag{13.31}$$

Because $\Phi_{NL}(t)$ has the same periodicity as $A(t)$ the comb structure of the spectrum is maintained and the derivations (13.28) remain valid because periodicity of $A(t)$ was the only assumption made. An optical fiber is most appropriate for this process because it can maintain the necessary small focus area over a virtually unlimited length. In practice, however, other pulse reshaping mechanism, both linear and non-linear, are present so that the above explanation might be too simple.

A microstructured fiber uses an array of submicron-sized air holes that surround the fiber core and run the length of a silica fiber to obtain a desired effective dispersion. This can be used to maintain the high peak power over an extended propagation length and to significantly increase the spectral broadening. With these fibers it became possible to broaden low peak power, high repetition rate lasers to beyond one optical octave as shown in fig. 13.6.

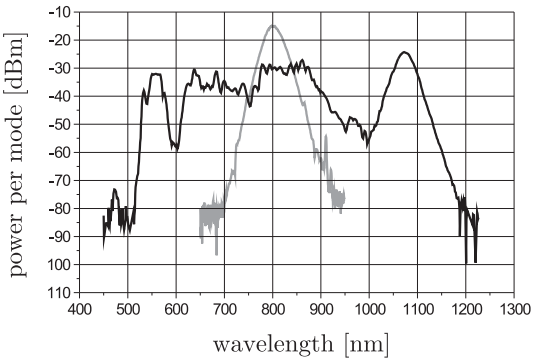


Figure 13.6: Power per mode of the frequency comb on a logarithmic scale (0 dBm = 1mW). The lighter 30 nm (14 THz at −3 dB) wide spectrum displays the laser intensity and the darker octave spanning spectrum (532 nm through 1064 nm) is observed after spectral broadening in a 30 cm microstructured fiber. The laser was operated at $\omega_r = 2\pi \times 750$ MHz (modes not resolved) with 25 fs pulse duration. An average power of 180 mW was coupled through the microstructure fiber.

Another class of frequency combs that can stay in lock for longer times are fs fiber lasers. The most common type is the erbium doped fiber laser that emits within the telecom band around 1550 nm. For this reason advanced and cheap optical components are available to build such a laser. The mode locking mechanism is similar to the Kerr lens method, except that non-linear polarization rotation is used to favor the pulsed high peak intensity operation. Up to a short free space section that can be build very stable, these lasers have no adjustable parts.

Self-referencing

The measurement of ω_{CE} fixes the position of the whole frequency comb and is called self-referencing. The method relies on measuring the frequency gap between *different* harmonics derived from the *same* laser or frequency comb. The simplest approach is to fix the absolute position of the frequency comb by measuring the gap between ω_n and ω_{2n} of modes taken directly from the frequency comb. In this case the carrier-envelope offset frequency ω_{CE} is directly produced by beating the frequency doubled red wing of the comb $2\omega_n$ with the blue side of the comb at ω_{2n}: $2\omega_n - \omega_{n'} = (2n - n')\omega_r + \omega_{CE} = \omega_{CE}$ where again the mode numbers n and n' are chosen such that $(2n - n') = 0$. This approach requires an octave spanning comb, i.e. a bandwidth of 375 THz if centered at the titanium-sapphire gain maximum at 800 nm.

Figure 13.7 sketches the $f - 2f$ self referencing method. The spectrum of a mode locked laser is first broadened to more than one optical octave with an

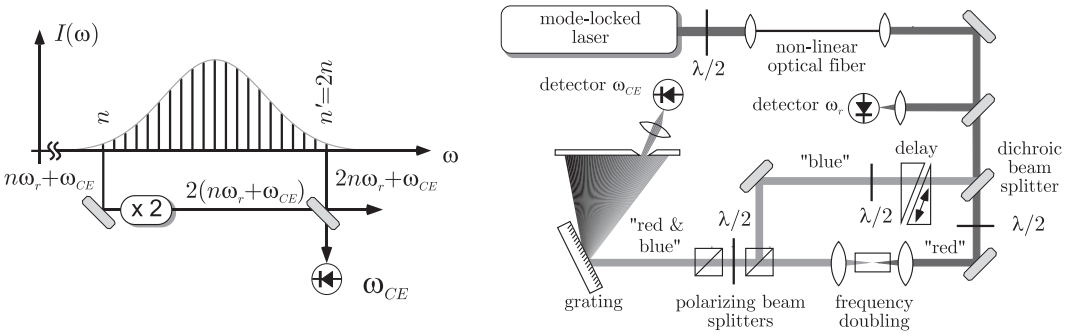


Figure 13.7: (left) — The principle of the $f - 2f$ self referencing relies on detecting a beat note at ω_{CE} between the frequency doubled "red" wing $2(n\omega_r + \omega_{CE})$ of the frequency comb and the "blue" modes at $2n\omega_r + \omega_{CE}$. (right) — More detailed layout of the self referencing scheme. See text for details.

optical fiber. A broad band $\lambda/2$ wave plate allows to choose the polarization with the most efficient spectral broadening. After the fiber a dichroic mirror separates the infrared ("red") part from the green ("blue"). The former is frequency doubled in a non-linear crystal and reunited with the green part to create a wealth of beat notes, all at ω_{CE}. These beat notes emerge as frequency difference between $2\omega_n - \omega_{2n}$ according to (13.24) for various values of n. The number of contributing modes is given by the phase matching bandwidth $\Delta\nu_{pm}$ of the doubling crystal and can easily exceed 1 THz.

As described, both degrees of freedom ω_r and ω_{CE} of the frequency comb can be measured up to a sign in ω_{CE} that will be discussed below. For stabilization of these frequencies, say relative to a radio frequency reference, it is also necessary to control them. Again the repetition rate turns out to be simpler. Mounting one of the laser's cavity mirrors on a piezo electric transducer allows to control the pulse round trip time. Controlling the carrier envelope frequency requires some effort. Any laser parameter that has a different influence on the cavity round trip phase delay and the cavity round trip group delay may be used to change ω_{CE}. Experimentally it turned out that the energy of the pulse stored inside the mode locked laser cavity has a strong influence on ω_{CE}. To phase lock the carrier envelope offset frequency ω_{CE}, one can therefore control the laser power through its energy source (pump laser).

Frequency conversions

Given the above we conclude that the frequency comb may serve as a frequency converter between the optical and radio frequency domains allowing to perform the following phase coherent operations:

- convert a radio frequency into an optical frequency. In this case both ω_r and ω_{CE} from (13.24) are directly locked to the radio frequency source.
- convert an optical frequency into a radio frequency. In this case the frequency of one of the comb modes ω_n is locked to a clock laser while the carrier envelope frequency ω_{CE} is phase locked to ω_r. The repetition rate will then be used as the countable clock output.
- convert an optical frequency to another optical frequency, i.e. measuring optical frequency ratios. In this case the comb is stabilized to one of the lasers as described in the second case, but instead of measuring ω_r one measures the beat note frequency between another laser and its closest comb mode ω'_n.