Università degli Studi di Catania

Facoltà di Scienze Matematiche, Fisiche e Naturali

Dipartimento di Matematica e Informatica

Dottorato di ricerca in Matematica Applicata

XXV Ciclo

Optimization of homogeneous emitter and thin-film solar cells

Angelo Greco

Advisor
Prof. Vittorio Romano

Anno Accademico 2012-2013

UNIVERSITÁ DI CATANIA

Optimization of homogeneous emitter and thin-film solar cells

by

Angelo Greco

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

at the Dipartimento di Matematica ed Informatica

December 2013

"O frati," dissi, "che per cento milia
perigli siete giunti a l'occidente,
a questa tanto picciola vigilia

d'i nostri sensi ch'è del rimanente
non vogliate negar l'esperïenza,
di retro al sol, del mondo sanza gente.

Considerate la vostra semenza:
fatti non foste a viver come bruti,
ma per seguir virtute e canoscenza".

Dante, Inferno, Canto XXVI, vv. 112-120

"Appena a sud del carbonio c'è il silicio. Come accade spesso tra i vicini si tratta di una prossimità ambigua, che crea un po' di disagio. Come il carbonio, ma in misura minore, il silicio ha la capacità di formare alcune tra le lunghe molecole a catena necessarie per i processi complessi come la vita. E tuttavia il silicio non ha dato origine a un proprio tipo di vita; forse però da questo punto di vista è solamente dormiente. I prodotti principali del carbonio, gli organismi viventi, hanno impiegato miliardi di anni per mettere a punto meccanismi di accumulo e dispersione dell'informazioni (una definizione austera e sintetica di ciò che intendiamo per «vita»); nel frattempo il silicio è rimasto in attesa. La recente alleanza tra le due regioni, che ha visto organismi basati sul carbonio sviluppare utensili basati sul silicio per la tecnologia dell'informazione, ha portato alla schiavizzazione del silicio. Però gli organismi basati sul carbonio sono ricchi d'inventiva e stanno sviluppando sempre più le potenzialità nascoste del silicio, tanto che forse un giorno il silicio capovolgerà i rapporti di forza con il suo vicino settentrionale e assumerà il ruolo dominante. Sicuramente sui tempi lunghi il silicio ha grandi potenzialità, perché il suo metabolismo e la sua replicazione possono essere meno complessi di quelli del carbonio. Questo potrebbe rivelarsi uno dei più astuti giochi di alleanze di tutto il Regno" […]

Peter Atkins, "*The Periodic Kingdom: A Journey Into the Land of the Chemical Elements*".
New York, BasicBooks, 1995

Contents

Acknowledgements

First of all, I want to thank my advisor, Prof. Vittorio Romano, for his scientific competence, for his ability to teach and, above all, for his patience. Without his trust in me, I would not have ever begun my PhD studies.
Moreover, I would like to thank Prof. Giuseppe Nicosia, from the Dipartimento di Matematica e Informatica, Catania University, for all the scientific support he was willing to give me as for the Genetic Algorithm based Optimization part of this work.
Thanks also to Dr. Giovanni Carapezza, former Temporary Research Fellow at the Department, in cooperation with Prof. Nicosia.
I would like to remember that this thesis has been developed within the Project *ENIAC Joint Undertaking*, *Energy for a green society: from sustainable harvesting to smart distribution, equipment, materials, design solutions and their applications*.

Finally, I thank my parents. They have always provided me with the moral support I needed when facing the tough challenge to complete my PhD studies while working as an engineer at the same time.

Preface

Over the last decades the world has been experimenting an increasing pressure to find solutions to energy crisis issues. As a result, the scientific research has been boosted towards the development of solutions related to alternative energy sources.
For the future decades, we can only imagine either to face the current standard energy demand or to face an increased one.
Since life quality levels are getting continuously better in most of the world, energy needs in such a scenario could be satisfied only by photovoltaic energy if we think to meaningfully cut the consumption of both traditional fossil fuels and nuclear energy.
This story comes from the past, the need for the development of renewable energy sources put itself under the world spotlight for the first time in the 70s, when the western countries experimented a serious energy crisis sponsored by the Middle-East OPEC countries that decided to dramatically reduce their crude oil export as a way to politically press over the western diplomacies after the Yom Kippur War between Israel and its Arab neighbors.
In the last twenty years, the Climate Change issues have gained enough popularity as well among developed countries' public opinions to become a further motivation to invest in the research and development of renewable sources in order to improve their efficiency.
In this geopolitical context, several research programs have been supported. Among them, the *ENIAC* Project *Joint Undertaking*, *Energy for a green society: from sustainable harvesting to smart distribution, equipment, materials, design solutions and their applications*, within which the present work has been performed.
In particular, this thesis is focused on the optimization of photovoltaic cells, through the use of mathematics tools and optimization techniques based on new theories like the Genetic Algorithm ones.
In the first chapter, the solar cell physics is briefly introduced, with special attention to light absorption phenomenon, to the main sources of loss in photovoltaic conversion and to the most important geometric parameters involving the cell efficiency, that will be the object of the analysis in the following chapters.
In the second chapter, the global optimization problem and main techniques are introduced: the deterministic methods, the direction, the tunneling and the probabilistic methods, showing their advantages and drawbacks.
Since we will usually deal with more than one objective function to optimize in our analysis, the multiobjective programming techniques are introduced in the third chapter, taking into account Pareto optimality theory and the most important multiobjective techniques, ranging from the ex-ante to the ex-post methods, to the interactive ones, suggesting to us the importance of the solver opinions in order to give importance to results and search directions.
In the fourth chapter, we deal the direct search algorithm for optimization, ranging from the line search methods, to the linear approximation to the quadratic approximation ones.
Then, in the fifth chapter, a numerical simulation of a monocrystalline selective emitter solar cell is presented. As for the model implementation, the tool used has been the Technology Computer Aided Design (TCAD) Sentaurus. Moreover, a homogeneous emitter cell is presented and simulated in the same section, taking into account the dependence of efficiency on both LDOP (lowly doped) and HDOP (heavily doped) profiles. Finally, it has been performed an analysis of the loss

mechanisms involving the photovoltaic cells, bounding the Internal Quantic Efficiency of both HE (Homogeneous Emitter) and SE (Selective Emitter) cells, getting to the conclusions of the advantages of the SE cell over the HE one.
Anyway, also the number and the way the contacts are placed within the cell play a key role in optimizing the device's efficiency. That is why in the sixth chapter, it has been performed a simulation of a rear contacted solar cell with special attention to the dependence of the cell's output on the metallization fraction as for the short circuit current, the fill factor, the open circuit voltage and the device's internal efficiency.
In the seventh chapter, a homogeneous emitter cell has been simulated, writing a Matlab input code to perform the simulation with given parameters by the TCAD Sentaurus. A genetic algorithm has been used, gaining an improvement in Fill Factor and Efficiency of the simulated cell with regard to the HE cell of the seventh chapter. Since we are trying to optimize both fill factor and efficiency of the device, we deal with a multiobjective problem and some trade-offs solutions have been introduced among the points of the Pareto front. The simulation has been launched twice, using a maximum number of generations of 300 and 1700 respectively. By the combined use of both a genetic algorithm and a such a powerful tool like TCAD Sentaurus, the results obtained in this part are innovative and represent a quantitative improvement of the results previously obtained over this argument in literature. The comparison against the reference structures, together with the improvements gained is summarized in the tables and figures at the end of the chapter.
In the eighth chapter, the thin-film cells are briefly introduced. Since the pressure towards the efficiency increase has got strong and stronger in last two decades, this gives a simultaneous answer to the increasing material and manufacturing cost of photovoltaic devices.
In the ninth chapter, the thin-film physics is briefly described, together with the mathematical methods that will be used in the following chapter to calculate the absorption and the light's diffused component. So, this chapter is dedicated to the thin-film optical model, but it also deals with the Monte Carlo method, a powerful stochastic method that will be used to model in a stochastic way the photons behavior in the photovoltaic device.
In the tenth and last chapter, the thin-film silicon cell is optimized through the use of a Genetic Algorithm. Since the thin-film technology is intimately bound to the reduction of production costs, the optimization problem takes into account the balance between the thickness of the cell layers and the manufacturing cost, becoming a profit maximization problem. Also this chapter deals with the thin-film structures from an innovative standpoint. The obtained results are new and they represent a noticeable improvement in the optimization of a tandem cell.

1 Introduction to solar cells

Semiconductor solar cells are fundamentally quite simple devices. Semiconductors have the capacity to absorb light and to deliver a portion of the energy of the absorbed photons to carriers of electrical current – electrons and holes. A semiconductor diode separates and collects the carriers and conducts the generated electrical current preferentially in a specific direction. Thus, a solar cell is simply a semiconductor diode that has been carefully designed and constructed to efficiently absorb and convert light energy from the sun into electrical energy. *[13]*
A solar cell, so, is a device able to convert part of the energy coming from sunlight into electrical power, by the exploitation of the photovoltaic effect, that takes place when the light on a double leyer of semiconductive material produces a potential difference between the layers.

1.1 Solar radiation

All electromagnetic radiation, including sunlight, is composed of particles called **photons**, which carry specific amounts of energy determined by the spectral properties of their source. Photons also exhibit a wavelike character with the wavelength, λ, being related to the photon energy, E_λ, by the equation *[1]*

$$E_\lambda = \frac{hc}{\lambda}$$

where h is Plank's constant and c is the speed of light.
The sun has a surface temperature of 5762 K and its radiation spectrum can be approximated by a black-body radiator at that temperature. Emission of radiation from the sun, as with all black-body radiators, is isotropic. However, the Earth's great distance from the sun means that only those photons emitted directly in the direction of the Earth contribute to the solar spectrum as observed from Earth. Therefore, for practical purposes, the light falling on the Earth can be thought of as parallel streams of photons. Just above the Earth's atmosphere, the radiation intensity, or Solar Constant, is about **1.353 kW/m^2** and the spectral distribution is referred to as an ***air mass zero*** (AM0) radiation spectrum. The Air Mass is a measure of how absorption in the atmosphere affects the spectral content and intensity of the solar radiation reaching the Earth's surface. The Air Mass number is given by *[2]*

$$Air\ Mass = \frac{1}{cos\theta}$$

where θ is the angle of incidence (θ = 0 when the sun is directly overhead). The Air Mass number is always greater than or equal to one at the Earth's surface. An easy way to estimate the Air Mass has been given by Green as

$$Air\ Mass = \sqrt{1 + \left(\frac{S}{H}\right)^2}$$

where S is the length of a shadow cast by an object of height H. A widely used standard for comparing solar cell performance is the AM1.5 spectrum normalized to a total power density of 1 kW/m^2. The spectral content of sunlight at the Earth's surface also has a diffuse (indirect) component owing to scattering and reflection in the atmosphere and surrounding landscape and can account for up to 20% of the light incident on a solar cell. The Air Mass number is therefore further

defined by whether or not the measured spectrum includes the diffuse component. An AM1.5g (global) spectrum includes the diffuse component, while an AM1.5d (direct) does not. *[1]*
Black body (T = 5762 K), AM0, and AM1.5g radiation spectrums are shown in the picture below

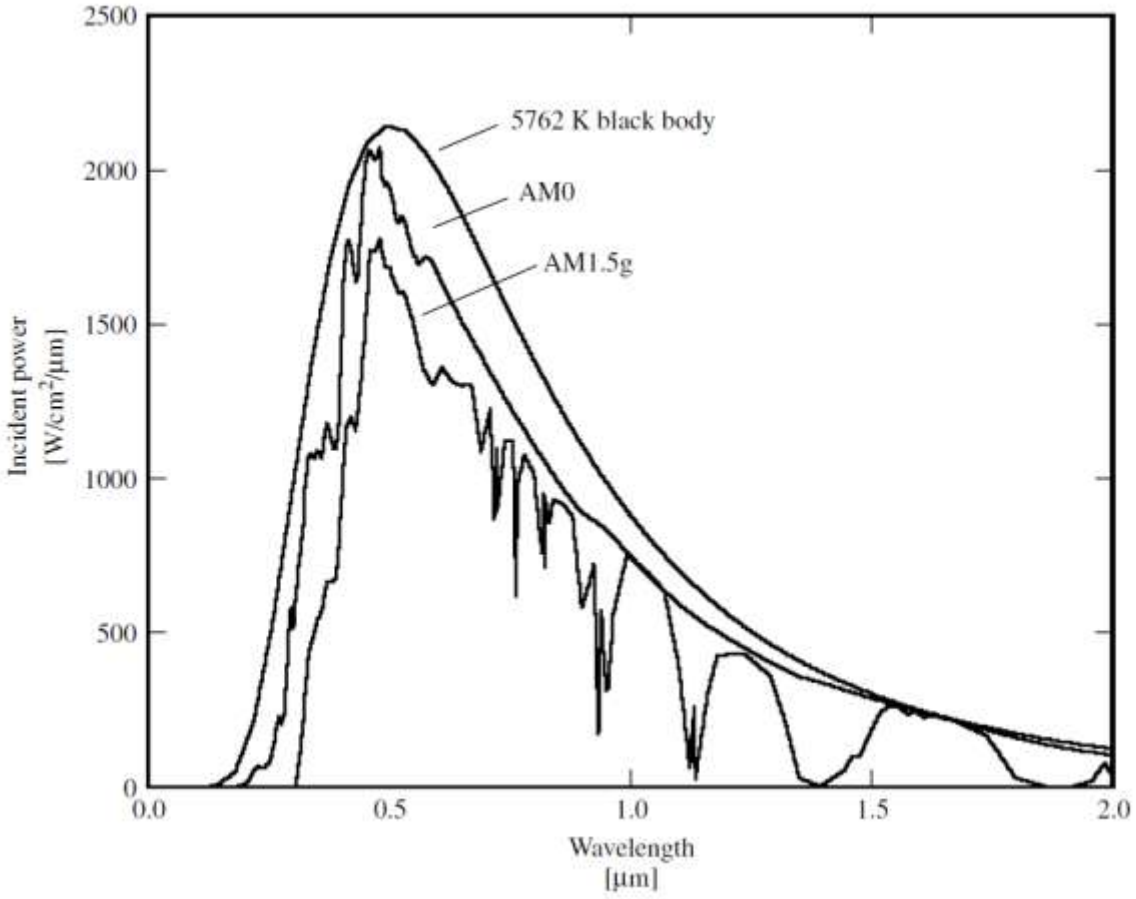


Figure 1.1 - The radiation spectrum for a black body at 5762 K, an AM0 spectrum, and an AM1.5 global spectrum

The specific amount of energy carried by photons is related to the spectral properties of the source they come from. Below, it is possible to get a look on the light spectrum, related to the wavelength.

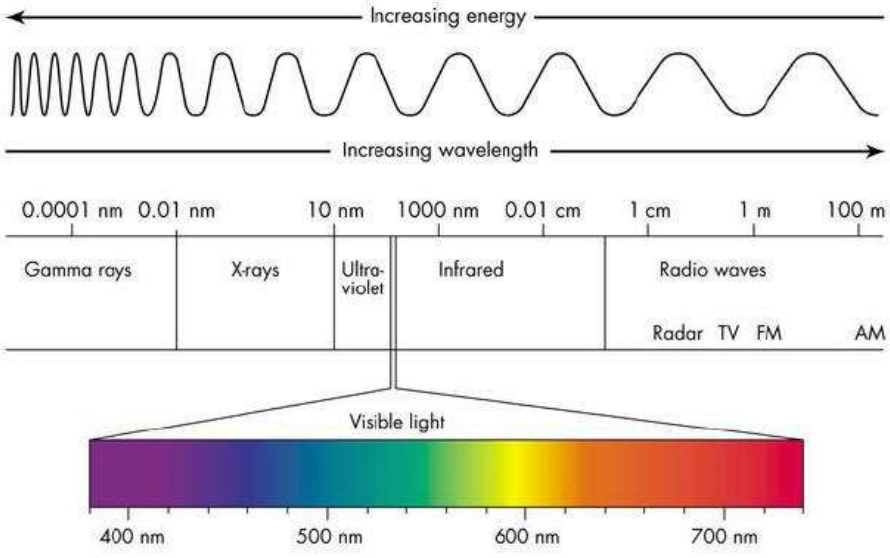


Figure 1.2 - The light spectrum with reference to wavelength and special attention to the visible light range

As we have already seen when introducing the Air Mass parameters, the atmosphere is responsible for alterations of the electromegnatic spectrum. This is, essentially, due to two reasons. The first one is that the atmosphere is divided into different layers and each of them is responsible for the absorption of radiations with a specific wavelength. The second one is that the atmosphere is responsible for the phenomenon of the Rayleigh's scattering, related to the collision among different light wavelength and the air, leading to a radiations' alteration. *[22]*

1.2 Silicon in photovoltaic technology

Solar cells can be fabricated from a number of semiconductor materials, most commonly **silicon** (Si) – crystalline, polycrystalline, and amorphous. Materials are chosen largely on the basis of how well their absorption characteristics match the solar spectrum and their cost of fabrication. Silicon has been a common choice due to the fact that its absorption characteristics are a fairly good match to the solar spectrum, and silicon fabrication technology is well developed as a result of its pervasiveness in the semiconductor electronics industry. *[21]*

Electronic grade semiconductors are very pure crystalline materials. Their **crystalline nature** means that their atoms are aligned in a regular periodic array. This periodicity, coupled with the atomic properties of the component elements, is what gives semiconductors their very useful electronic properties. Below, you can see an abbreviated periodic table of elements.

I	II	III	IV	V	VI
		B	C	N	O
		Al	Si	P	S
Cu	Zn	Ga	Ge	As	Se
Ag	Cd	In	Sn	Sb	Te

Figure 1.3 – Part of the periodic table of elements

Note that silicon is in column IV, meaning that it has four valence electrons, that is, four electrons that can be shared with neighboring atoms to form covalent bonds with those neighbors.

In the case of Silicon, each atom forms a covalent bond with four more atoms, and these are all valence atoms. That is the way the lattice in the molecular structure of silicon is created. *[18]*

The dual behavior of semiconductor, insulant at low temperatures and conductive at higher ones, can be related to the behavior of electrons between valence and conduction bands. When temperature increases, the thermal energy somministrated to the lattice, leads to the breaking of some covalent bonds created among valence atoms. These atoms, thermally excited, can jump away from the valence band to the conduction band and they will be responsible for the conduction phenomena in photovoltaic.

The amount of energy needed to break this kind of bond is called **band gap** ($\mathbf{E_g}$) and can be determined as follows in the case of silicon *[1]*

$$E_g = 1.21 - 3.60 \cdot 10^{-4} T \quad eV$$

So, for a temperature of 0 K, the band gap for silicon is 1.21 eV, and, at room temperature (300 K), it has decreased to 1.1 eV. That is why, at temperatures below zero, semiconductors are usually considered as insulant materials. *[22]*

When temperature increases, the probability for a valence electron to break its covalent bonds and jump to the conduction band increases too. This phenomenon involves not only the valence electrons, because, when one of more of them leaves the valence band, it is no more completely occupied by the electrons, so, electrons belonging to lower states of the band can move to fill the free states in the valence band if properly excited by an electric camp.

A material with the previously underlined features is called **intrinsic semiconductor**. In this case, the density of electrons is equal to the density of holes. *[1]*

As shown in the picture below, electron near the maxima in valence band have been thermally excited to the empty states near the conduction-band minima, leaving behind holes. The excited electrons and remaining holes are the negative and positive mobile charges that give semiconductors their unique transport properties.

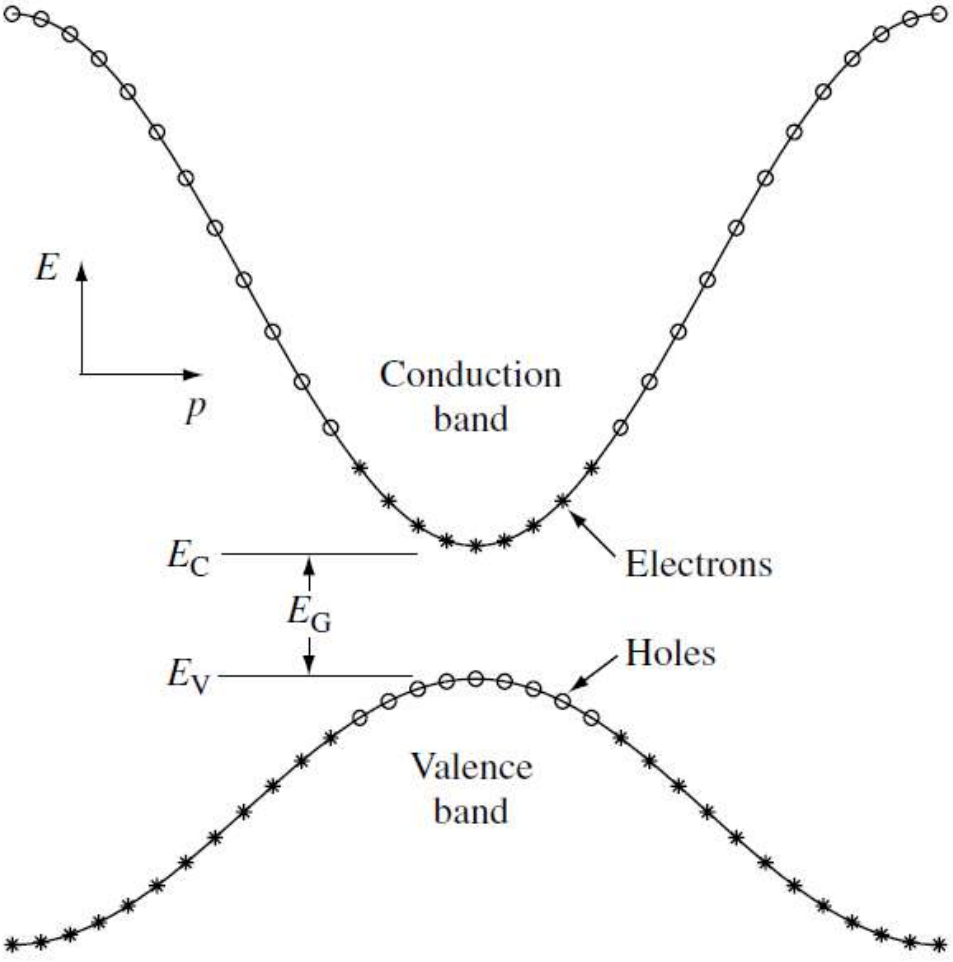


Figure 1.4 – A simplified energy band diagram at $T > 0$ K for a direct band gap (EG) semiconductor. Electrons near the maxima in valence band have been thermally excited to the empty states near the conduction-band minima, leaving behind holes. The excited electrons and remaining holes are the negative and positive mobile charges that give semiconductors their unique transport properties

1.3 Conduction band and valence band density of states and Fermi-Dirac distribution

The dynamic behavior of the electron can be established from the electron wave function, ψ, which is obtained by solving the time-independent **Schrodinger equation**:

$$\nabla^2 \psi + \frac{2m}{h^2}[E - U(r)]\psi = 0$$

where *m* is electron mass, *h* is the reduced Planck constant, *E* is the energy of the electron, and *U(*r*)* is the periodic potential energy inside the semiconductor. This equation states that the electron energy is quantized. If we consider a wave function *ψ(r)* and a sample made up by a cube of material through which *r* is the position vector, the huge number of energy levels allowed for the electron is very close one to another. *[2]*

So, if we consider the energy interval *(E, E+dE),* the number of states that own this level of energy is indicated as *n(E)dE,* with *n(E)* indicated as density of allowed states.

$$n(E) = \frac{8\sqrt{2}\pi m^{3/2} E^{1/2}}{h^3}$$

Not all of the allowed states are occupied. The density of occupied states will be

$$n_0(E) = n(E)p(E)$$

With $p(E)$ called **Fermi-Dirac function of probability**. It expresses the probability that a state at a given energy would be occupied as well.

$$p(E) = \frac{1}{e^{(E-E_F)/kT} + 1}$$

Where E_F is the Fermi's energy, the energy at which we have *p=1/2, k* is the Boltzmann's constant and *T* is the absolute temperature. This equation suggests that the parameter with a real importance is $\boldsymbol{E\text{-}E_F}$ that is the difference between the energy of the considered electron and the Fermi's energy, because only the electrons with an amount of energy close to the Fermi's one can play a role in the electrical conduction process. *[25]*

In the picture below, you can see the Fermi's energy at various temperatures. *[2]*

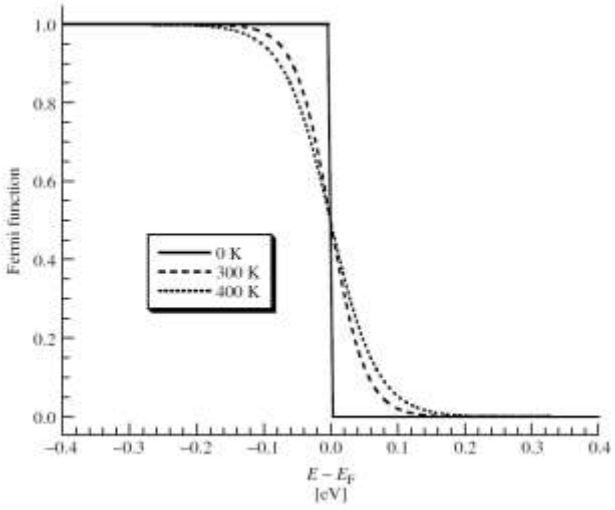


Figure 1.5 - The Fermi function at various temperatures

1.4 Donors, acceptors and doped semiconductors

Because of their low density, intrinsic semiconductors cannot produce a sufficient current for normal applications. Anyway, it is possible to alterate the normal properties of these materials through an adequate increase of carriers by doping the materials. Doping, in electronics, consists in contaminating the semiconductor with special impurities.
Let us consider a crystal of pure silicon, where we insert some atoms of elements of the 5^{th} group, like phosphorus. Four electrons belonging to these atoms will be shared with the closest silicon atoms to create covalent bonds, while the fifth electron, still available for an additional bond, will stay with its phosphorus atom. So, the thermal excite and the following jump to the conduction band would result easier for this last electron, rather for the others, now part of four covalent bonds.
As a result, an energy of only 0.05 eV, will be enough to free this electron and make it available for the electrical conduction, while it would be necessary an energy of 1.1 eV to take a silicon atom electron from the valence band to the conduction one. *[28]*
The phosphorus atom, within the silicon lattice, is called **donor**, because it lends an electron to the conduction band. So, adding donor atoms, it is possible to increase the electron density within the conduction band.
Semiconductors doped with donor atoms are called "**n type**" semiconductors, where n stands for negative, because the negative carriers (electrons) are much more than holes (positive carriers). In this type of semiconductors, electrons within the conduction band are called **majority carriers**, while the holes in the valence band are called **minority carriers**.
Let us consider now a pure silicon crystal doped with impurities belonging to elements from the third group of the periodic table, such as **boron**. Each boron atom is surrounded by four silicon atoms, but boron has three valence electrons, so one more electron is needed to complete the external atomic configuration (8 valence electrons). With a very low energy level, 0.05 eV it is possible to take an electron away from a silicon-silicon bond to use it to complete the valence configuration. The boron atom is so called **acceptor**, because it easily takes an electron from the valence band. Adding acceptor atoms, it is possible to dramatically increase the holes density within the valence band.
A semiconductor rich in this kind of impurities (acceptors) shows at room temperature an excess of positive carriers and it is called "**p type semiconductor**". In this last kind of semiconductors, majority carriers are the valence band holes, while the minority carriers are the conduction band electrons.

1.4.1 Drift and diffusion current

While the impurities are inserted in the semiconductor lattice, it is needed to understand how these carriers move inside the material. Electron speed without any electrical field can be estimated between 0 and v_f (Fermi's speed, the speed of an electron with a kinetic energy equal to the Fermi's one E_f). While, when applying an electrical field, electrons are accelerated, so that they have a small speed increase in the field direction, but towards the opposite versus, because electrons have a negative carriers, so that *[1]*

$$F = -eE$$

Electrons responsible for the conduction are the only ones with a speed close to v_f.
The parameter that measures the capacity of carriers to move freely within the material under an electrical field *E* is the following one

$$\mu = \frac{v}{E}$$

where *v* is the speed of the electrons in the direction opposite to the field's one. The current density of the carriers is called **drift current**. It is related to the carrier movement due to the electrical field over the semiconductor.
The drift current densities for holes and electrons can be written as

$$J_p^{drift} = qpv_{d,p} \quad \text{for holes}$$

$$J_n^{drift} = -qnv_{d,n} \quad \text{for electrons}$$

So, these currents depends on drift speed for holes and electrons, on the number of holes and electrons and, finally, on the electron's carrier.
Moreover, Electrons and holes in semiconductors tend, as a result of their random thermal motion, to move (diffuse) from regions of high concentration to regions of low concentration.
Much like how the air in a balloon is distributed evenly within the volume of the balloon, carriers, in the absence of any external forces, will also tend to distribute themselves evenly. This process is called ***diffusion*** and the diffusion current densities are given by

$$J_p^{diff} = -pD_p\nabla p$$

$$J_d^{diff} = qD_n\nabla n$$

So, the total current, for both holes and electrons, can be calculated as follows

$$J_p = J_{p,drift} + J_{p,diff}$$

$$J_n = J_{n,drift} + J_{n,diff}$$

1.4.2 p-n junctions

The development occurred to the photovoltaic technology started just from the studies over the p-n junction, elementary structure of the semiconductor devices physics. The p-n junction, essentially, is made up by a n-type and a p-type semiconductor put close one to another. When this two devices lay one along the other, the diffusion phenomenon has place, so that the holes move from the p-zone to the n-zone and the vice-versa happens for the electrons from the n-zone to the p-zone. This happens because of the distribution of holes and electrons in the semiconductor device that is not uniform, so that the diffusion current is generated. This phenomenon ends when the carriers are distributed uniformly. The diffusion of carriers, determines a region between the junction, called "**depletion region**", where finding carriers it is not possible and it is possible to measure an electric field not equal to zero, caused by the presence of ionized doping atoms, that is not counterbalanced by the lack of carriers within the region. *[3]*
The following picture represents the depletion region between the n-type and p-type region

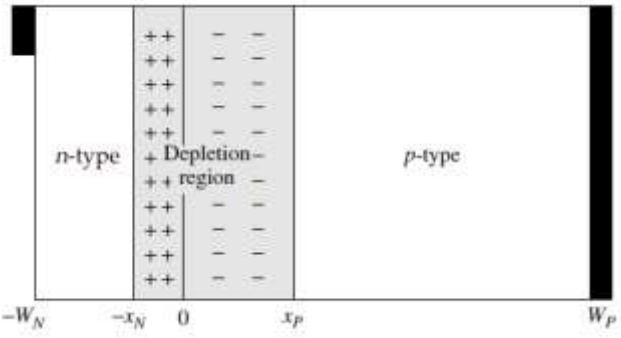


Figure 1.6 Simple solar cell structure used to analyze the operation of a solar cell. Free carriers have diffused across the junction ($x = 0$) leaving a space-charge or depletion region practically devoid of any free or mobile charges. The fixed charges in the depletion region are due to ionized donors on the *n*-side and ionized acceptors on the *p*-side

The difference of potential between the p-type and the n-type semiconductor is called "**built-in potential**", and it is equal to:

$$V_{bi} = \frac{KT}{q} ln\left(\frac{N_a N_d}{n_i^2}\right)$$

where N_a and N_d are, respectively, the p-type and n-type impurities concentrations introduced during the doping phase and n_i is the density of electrons in the conduction band.
The potential and the electrical field of the junction can be calculated using the *abrout change approssimation*. According to this equation, if we assume the density of spatial carriers in the *p* and *n* regions equal to, respectively $-qN_a$ and qN_d we will have that

$$x_n N_d = x_p N_a$$

where x_n and x_p are the width of the depletion regions in the n-type and p-type semiconductors. So, we can argue that, in equilibrium, in any point of the depletion region, the effect of the electrical field is counterbalanced by the effect of concentrations variation.
Now, we can consider the Poisson's equation:

$$\frac{d^2V}{dx^2} = -\frac{\rho}{\varepsilon} = \frac{qN_a}{\varepsilon}$$

where ρ is the density of carrier. So, integrating two times this equation, and assuming the boundary condition that the potential and the electrical field would be equal to zero for $x = -x_p$, we have

$$V = \frac{qN_a}{\varepsilon}\left(\frac{x^2}{2} + x_p x + \frac{x_p^2}{2}\right)$$

and it is now possible to express the potential for both sides of the junction:

$$V_p = \frac{qN_a x_p^2}{2\varepsilon} \quad \text{and} \quad V_n = \frac{qN_d x_n^2}{2\varepsilon}$$

So, if we sum this two values, we have the total value of the potential barrier on the junction.
The depletion region width can be calculated, approximately, considering the zone moving carriers free. In fact, when the intrinsec Fermi's level E_i is close to the real one E_f, n and p become almost

equal to each other. Assuming that N_a and N_d would be constant in the respective zones (**step junction**), the depletion region width W is equal to

$$W = x_p + x_n$$

1.5 Light absorption

The creation of electron–hole pairs via the absorption of sunlight is fundamental to the operation of solar cells. The excitation of an electron directly from the valence band (which leaves a hole behind) to the conduction band is called **fundamental absorption**. *[3]*

Both the total energy and momentum of all particles involved in the absorption process must be conserved. Since the photon momentum, $p\lambda = h/\lambda$, is very small compared to the range of the crystal momentum, $p = h/\_$, the photon absorption process must, for practical purposes, conserve the momentum of the electron.1 The absorption coefficient for a given photon energy, hv, is proportional to the probability, P_{12}, of the transition of an electron from the initial state E_1 to the final state E_2. So, the photon absorption process in a direct band semiconductor can be represented as in picture below

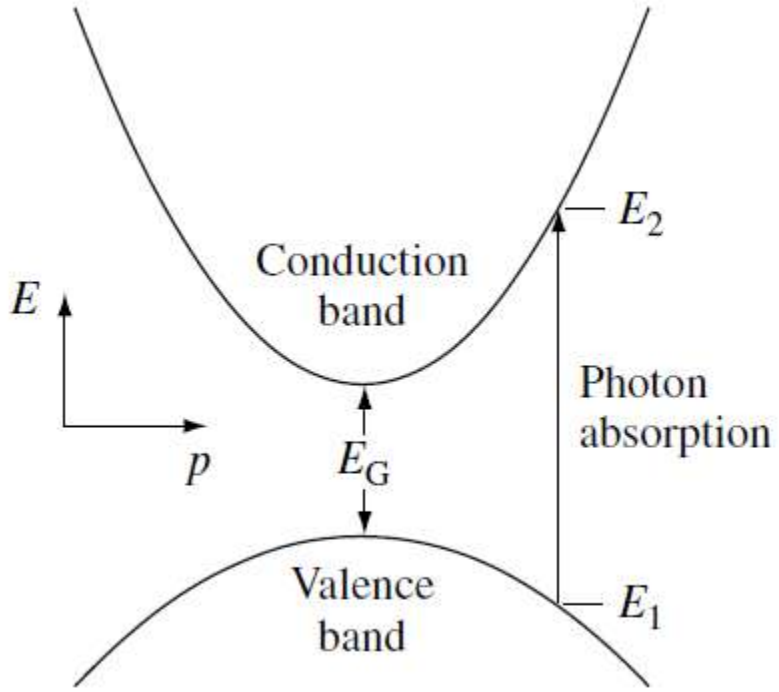


Figure 1.6 – Valence and conduction band in a direct band semiconductor

where the incident photon has an energy E_2-E_1>E_G.

Absorption results in creation of an electron-hole pair since a free electron is excited to the conduction band leaving a free hole in the valence band. *[18]*

While, in indirect band semiconductors, like silicon, where the valence-band maximum occurs at a different crystal momentum than the conduction-band minimum, conservation of electron momentum necessitates that the photon absorption process involve an additional particle. Phonons, the particle representation of lattice vibrations in the semiconductor, are suited to this process because they are low-energy particles with relatively high momentum. This phenomenon is represented in the following picture

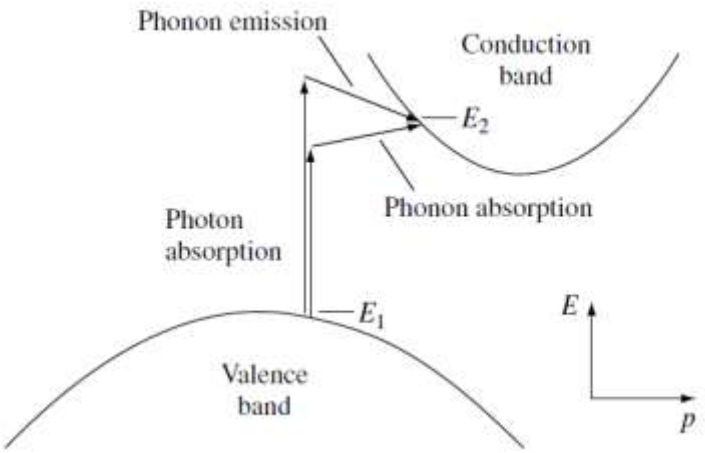


Figure 1.7 – Photon absorption in an indirect band gap semiconductor for a photon with energy $h\nu < E2 - E1$ and photon with energy $h\nu > E2 - E1$. Energy and momentum in each case are conserved by the absorption and emission of a phonon, respectively

Photon absorption in an indirect band gap semiconductor for a photon with energy $h\nu < E2 - E1$ and a photon with energy $h\nu > E2 - E1$. Energy and momentum in each case are conserved by the absorption and emission of a phonon, respectively.

In both direct band gap and indirect band gap materials, a number of photon absorption processes are involved, though the mechanisms described above are the dominant ones. A direct transition, without phonon assistance, is possible in indirect band gap materials if the photon energy is high enough. Conversely, in direct band gap materials, phonon-assisted absorption is also a possibility.

1.5.1 Recombination

When a semiconductor is taken out of thermal equilibrium, for instance by illumination and/or injection of current, the concentrations of electrons (n) and holes (p) tend to relax back toward their equilibrium values through a process called **recombination** in which an electron falls from the conduction band to the valence band, thereby eliminating a valence-band hole. There are several recombination mechanisms important to the operation of solar cells – recombination through **traps** (defects) in the forbidden gap, **radiative** (band-to-band) recombination, and **Auger** recombination.

The net recombination rate per unit volume per second through a single level trap (SLT) located at energy $E = E_T$ within the forbidden gap, also commonly referred to as **Shockley–Read–Hall recombination**, is given by

$$R_{SLT} = \frac{pn - n_i^2}{\tau_{SLT,n}\left(p + n_i e^{\frac{E_i - E_T}{kT}}\right) + \tau_{SLT,p}\left(n + n_i e^{\frac{E_T - E_i}{kT}}\right)}$$

where the carrier lifetimes τ_{SLT} are given by

$$\tau_{SLT} = \frac{1}{\sigma v_{th} N_T}$$

Where σ is the capture cross section, v_{th} is the thermal velocity of the carriers, and N_T is the concentration of traps. The capture cross section can be thought of as the size of the target present

to a carrier traveling through the semiconductor at velocity v_{th}. Small lifetimes correspond to high rates of recombination. If a trap presents a large target to the carrier, the recombination rate will be high (low carrier lifetime). When the velocity of the carrier is high, it has more opportunity within a given time period to encounter a trap and the carrier lifetime is low. Finally, the probability of interaction with a trap increases as the concentration of traps increases and the carrier lifetime is therefore inversely proportional to the trap concentration. *[25]*

Radiative (band-to-band) recombination is simply the inverse of the optical generation process and is much more efficient in direct band gap semiconductors than in indirect band gap semiconductors. When radiative recombination occurs, the energy of the electron is given to an emitted photon – this is how semiconductor lasers and light emitting diodes (LEDs) operate.

Auger recombination is somewhat similar to radiative recombination, except that the energy of transition is given to another carrier (in either the conduction band or the valence band). This electron (or hole) then relaxes thermally (releasing its excess energy and momentum to phonons). Just as radiative recombination is the inverse process to optical absorption, Auger recombination is the inverse process to *impact ionization*, where an energetic electron collides with a crystal atom, breaking the bond and creating an electron–hole pair.

Interfaces between two dissimilar materials, such as, those that occur at the front surface of a solar cell, have a high concentration defect due to the abrupt termination of the crystal lattice. These manifest themselves as a continuum of traps within the forbidden gap at the surface; electrons and holes can recombine through them just as with bulk traps. Rather than giving a recombination rate per unit volume per second, surface traps give a recombination rate per unit area per second. A general expression for surface recombination is

$$\int_{E_v}^{E_c} \frac{p_n - n_i^2}{\frac{p + n_i e^{\frac{E_i - E_t}{kT}}}{S_n} + \frac{n + n_i e^{\frac{E_i - E_t}{kT}}}{S_p}} D_{\Pi}(E_t)\, dE_t$$

where E_t is the trap energy, $D_{\Pi}(E_t)$ is the surface state concentration (the concentration of traps is probably dependent on the trap energy), and S_n and S_p are surface recombination velocities.

1.6 Theoretical limits to photovoltaic conversion

The conversion efficiency, maybe, is the most important parameter in photovoltaic tehcnology. The sun energy density is not as low as we cannot expect a generalized use of its energy, but, simultaneously, it is not so high that we can consider its exploitation simple. L'efficienza di conversione è forse il parametro più importante della tecnologia fotovoltaica. *[14]* La densità energetica del sole non è tanto bassa da non permetterci di avere aspettative su un uso generalizzato e efficiente della sua energia, ma, al tempo stesso, non è così alta da rendere ciò semplice. The solar cells efficiency is closely related to the hole-electron pairs due to the insulation and to the possibilità to avoid their recombination before they are conveyed into the outlet electrical circuit at evaluated the maximum efficiency to be expected from the solar cells, at 40.7% for the photonic spectrum approximated by the black body at the temperature of 6000K. *[15]* This value is not too high if we consider that the solar cells performs a inefficient use of photons, because most of them are not absorbed and their energy is not exploited in an optimal way. *[1]*

1.6.1 Recombination processes

Photons are absorbed in order to get conveyed towards the conduction band from the valence band, according the process known as electron-hole pair generation. Anyway, the same process can take place in the opposite versus, when an electron comes back to its valence band. As a result, the difference between the electrons pushed towards the conduction band by the sunlight absorption and the electrons that fall again into their valence band is equal to the net current extracted from the solar cell. This statement can also be presented as an equation *[1]*

$$\frac{I}{q} = \dot{N}_s - \dot{N}_r = \int_{\varepsilon_g}^{\infty} (\dot{n}_s - \dot{n}_r)\, d\varepsilon$$

where ε_g represents the band gap, while N_s and N_r are the inlet and outlet photons flows into and from the solar cell respectively, through any surface. In other words, if the cell is adequately contact equipped, the current is made up by the electrons that leave the conduction band through the n-type contacts, adequately doped. In the same way, in the valence band, the *I/q* ratio, represents the electrons that enter the valence band through the highly doped p-type contacts.

Finally, it is possible to estimate the solar cell theoretical reachable efficiency, as a function of the band gap, according to Shockley and Quesisser's assumptions, at 40.7%.

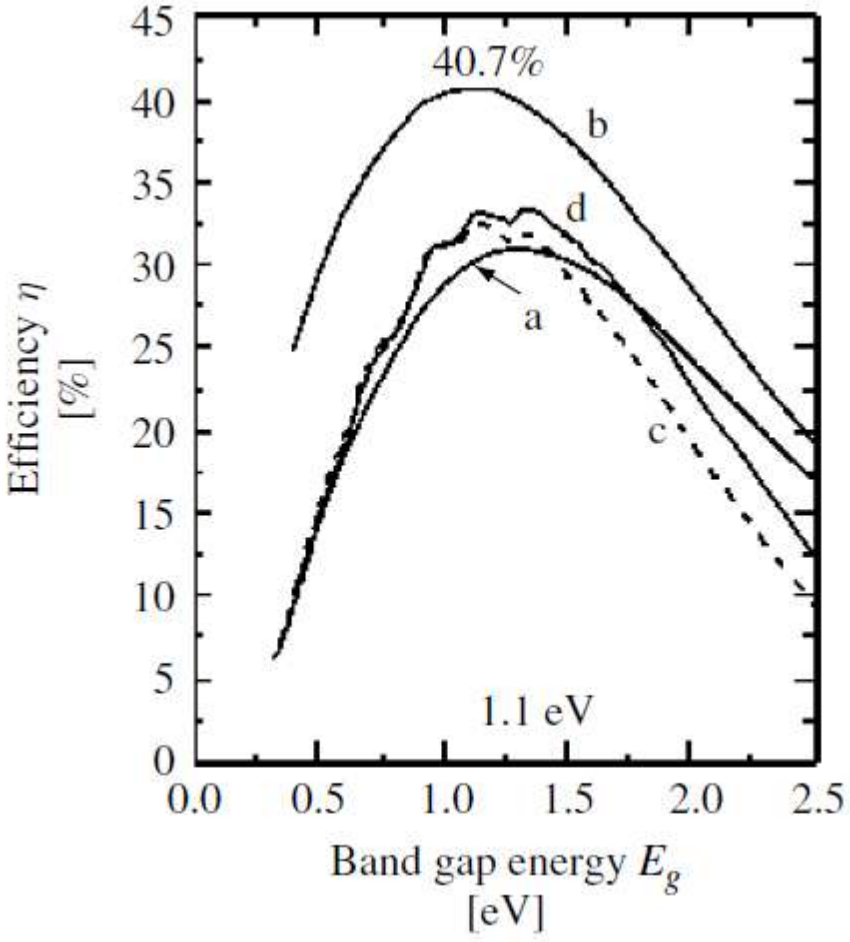


Figure 1.8 – Theoretical reachable efficiency for a solar cell as a function of the band gap

1.6.2 Solar cells thickness

If we consider the electrical perfomance, the optimal thickness depends on materials' structure and quality and it can involves several aspects. The thinnest cells can absorb less light, but this drawback can be counterbalanced by light trapping technologies. Moreover, the losses due to the shadow side of the device can decrease when reducing the cell thickness. It is also to be taken into account the economic pressure, from an industrial standpoint, towards the decrease of cell thickness, in order to reduce the product material costs, since a thinner cell is simply made up by less silicon.

1.6.3 Light trapping coatings

Silicon is featured by an high reflection index. It has been calculated that, when it is not coated, silicon reflects 30% of incident radiation. This is the way reflection losses are generated. A widely used solution, consists in the use of a material with a very low reflection index as a coating for silicon. This coating is generally an insulant material, designed to reduce reflection. Industrial trends are now positioned on the use of Titanium oxides through the process of chemical vapour deposition (CVD). *[1]*

2 Introduction to global optimization

Given a function $f: IR^n \rightarrow IR$, the global optimization methods tries to determine the global minum of the function $f(x)$, that is a point x^* in such a way that: *[5]*

$$f(x^*) \leq f(x) \; \forall \; x \in \mathbb{R}^n$$

These methods can be divided into the following groups:

1. Deterministic methods
2. Methods for Lipschitzian functions
3. Directions method
4. Tunneling methods
5. Probabilistic methods

2.1 Deterministic methods

A function $f: IR^n \rightarrow IR$ is called Lipschitzian if there exists a constant $L > 0$ (called Lipschitz's constant) exists in such a way that for any $x_1, x_2 \in \mathbb{R}^n$ it holds *[41]*

$$|f(x_1) - f(x_2)| \leq L||x_1 - x_2||$$

In other words a Lipschitzian function satisfyies the following statements:

$$f(x) \geq f(x_0) - L\big||x - x_0|\big|$$
$$f(x) \leq f(x_0) + L||x - x_0||$$

for any $x_0, x \in IR^n$.

The following function $f(x)$ is a Lipschitzian one

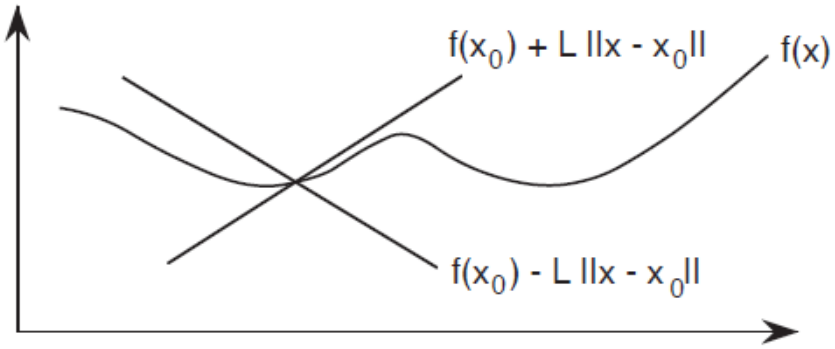


Figure 2.1 – A Lipschitzian function

The algorithms belonging to this group of optimization methods have in common the research for the value that minimizes the following problem:

$$\min_{x \in I_n} f(x)$$

with $I_n = \{x: A_i \leq x_i \leq B_i \ \ \forall\, i = 1,2, \dots, n\}$

We assume that:

1. The n-dimensional cube I_n should be in such a way that it contains a global minimum for $f(x)$
2. The function would be Lipschitizian over I_n
3. The L Lipschitz's constant would be known or it would be known its overestimation $\tilde{L}$

One of the most popular algorithm among these methods is the **Schumbert-Mladineo's** one. *[41]*

First step. Let it be $\tilde{L} > L$; given x_0 it is defined the function

$$F_0(x) = f(x_0) - \tilde{L}||x - x_0||$$

and x_1 is chosen so that:

$$F_0(x_1) = \min_{x \in I_n} F_0(x)$$

k[th] step. Once you get x_k the following function is defined

$$F_k(x) = \max_{j=0,\dots,k} \{f(x_j) - \tilde{L}||x - x_j||\}$$

and x_{k+1} is chosen in such a way that

$$F_k(x_k + 1) = \min_{x \in I_n} F_k(x)$$

the function $F_k(x)$ is featured by a very particular structure that can be used to define algorithms that in a finite number of steps solve the problem

$$\min_{x \in I_n} F_k(x)$$

So, if f^* is the minimum value of $f(x)$ over I_n, let they be F_k^* the minimum values of $F_k(x)$ over I_n. Let it be $\Phi \equiv \{x^* \in I_n,\ f(x^*) = f^*\}$ and let it be $\{x_k\}$ the sequence of points generated by the previous algorithm, then, it follows:

$$\lim_{k \to \infty} \inf_{x^* \in \Phi} ||x^* - x_k|| = 0$$

$$\lim_{k \to \infty} f(x_k) = f^*$$

and that the sequence $\{F_k^*\}$ is not decreasing and

$$\lim_{k \to \infty} F_k^* = f^*$$

Method's benefits

1. It does not require the solver to calculate the derivatives
2. It is possible to determine the point's sequence $\{x_k\}$ convergence, both from a theoretical and a computational standpoint
3. A stopping criterion exists for the algorithm. If $\{x_k\}$ and $\{F_k^*\}$ are the sequences generated, you have the following ones:

$$f(x_k) \geq f^* \geq F_k^*$$

$$f(x_k) \geq f^* \geq f(x_k) + r_k$$

where $r_k = F_k^* - f(x_k)$ and $\lim_{k \to \infty} r_k = 0$ because of the previous theorem. So, if $|r_k| < \varepsilon$ then x_k is a minimum for $f(x)$, far away from it for only an ε.

Method's drawbacks

1. it can be very difficult to define, ex-ante, an n-dimensional cube I_n containing at least a global minimum of $f(x)$
2. the method can be very heavy, from a computational point of view, because of the computing of $F_k(x)$ at each step
3. the hypothesis that $f(x)$ would be a Lipschitzian function is very restrictive
4. it is not always possible to get an overestimation of the Lipschitz's constant

2.2 Directions methods

The basis idea of the directions method is to define some directions, containing all the local optimals and to choose, among them, the one related to the lowest value of the objective function. These methods have been firstly proposed in the 70s, without getting good results at that time. Lately, through the implementation of a new, general approach, these methods have come again under the spotlight. The most popular direction method is the **Branin's method**. *[42]*

Branin's method

Let us suppose that $f(x)$ is continuous and $\nabla f(x)$ would be continuous as well. If you choose an x_0 it is possible to define some directions $x(t)$ in which $\nabla f(x(t))$ is parallel to $\nabla f(x_0)$. So, through the solution of the following system

$$\frac{d}{dt}\left[\nabla f\left(x(t)\right)\right] = \pm\nabla f\left(x(t)\right)$$

assuming the initial condition $x(0) = x_0$ it is possible to get some directions satisfying the following one

$$\nabla f\left(x(t)\right) = \nabla f(x_0)e^{\pm t}$$

The Branin's method is based on the following steps:

1. Determination of the solution $x(t)$ of the system
$$\frac{d}{dt}\left[\nabla f\left(x(t)\right)\right] = -\nabla f\left(x(t)\right), \quad x(0) = x_0$$

2. the direction $x(t)$ allows us to determine a stationary point x^* di $f(x)$. In fact, since
$$\lim_{t\to\infty} \nabla f\left(x(t)\right) = \lim_{t\to\infty} \nabla f(x_0)e^{-t} = 0$$
the direction $x(t)$ tends to x^*.

3. the stationary point x^* is slightly perturbated, gaining to the point $\tilde{x}_0 = x^* + \varepsilon$ and the following system is solved
$$\frac{d}{dt}\left[\nabla f\left(\tilde{x}(t)\right)\right] = \nabla f\left(\tilde{x}(t)\right), \quad \tilde{x}(0) = \tilde{x}_0$$
gaining the direction $\tilde{x}(t)$.

4. along the direction $\tilde{x}(t)$ we get away from the stationary point x^* because the norm of the gradient increases as t increases:
$$\nabla f\left(\tilde{x}(t)\right) = \nabla f(\tilde{x}_0)e^{t}$$

so, the direction $\tilde{x}(t)$ is followed until $\bar{t}$, when $\tilde{x}(\bar{t})$ is out of the stationary point "attraction zone".

5. the following system is now solved again

$$\frac{d}{dt}\left[\nabla f\left(x(t)\right)\right] = \nabla^2 f\left(x(t)\right)\frac{dx(t)}{dt} = \nabla f(x(t))$$

and the result is $\frac{d(x(t))}{dt} = \pm\nabla^2 f\left(x(t)\right)^{-1}\nabla f(x(t))$

these last equations define a Newton-like method. Because the Hessian matrix can become singular, the previous equations can lose meaning for some t. If $A(x)$ is the adjoint matrix of $\nabla^2 f(x)$, $A(x)$ always exists and it is true that $\nabla^2 f(x)^{-1} = \frac{A(x)}{\det[\nabla^2 f(x)]}$, so, the previous system can be replaced through a new parameterization with the following system

$$\frac{dx(t)}{dt} = \pm A\left(x(t)\right)\nabla f\left(x(t)\right)$$

Comments on the Brainin's method *[42]*

- It has never been proved that this method would be globally convergent, that is the curve defined reaches the global minimum.
- If the method is convergent, it is hard to know how many stationary points $f(x)$ has and, because of it, it is hard to state a stop criterion for the algorithm.
- The direction $x(t)$ is attracted by all the $f(x)$'s stationary points.
- The numerical solution of the differential equation systems defining the curve $x(t)$ is heavy from a computational standpoint.

2.3 Tunneling methods

The tunneling methods have been proposed in order to find in an efficient way the global minimum for functions with many local minima (in cases where the previous direction methods would not be adequate).

Tunneling algorithms' structure

Tunneling algorithms are made up by a sequence of cycles, each cycle is made up by two steps: a minimization phase, during which the objective function is minimized and a tunneling phase, where a "good" starting point is got for a following minimization phase.

Minimization phase

Given a starting point x_0, a local search is performed. It is equivalent to apply any convergent algorithm to a local minimum x_0^*.

Tunneling phase

The solver efforts are to find a point $x_1 \neq x_0$ in such a way that

$$f(x_1) = f(x_0^*)$$

so, theoretically, the tunneling methods generate a sequence in such a way that

$$f(x_k^*) \geq f(x_{k+1}^*)$$

and that the x_k points come close to the global minimum "passing under" the less important local minima, without taking into account how many and where they are.
This last feature is very important for problems with a lot of minima. The main drawback of these methods are the difficulties met when trying to find an x in such a way that $f(x) = f(x_k^*)$ and $x \neq x_k^*$. In order to avoid this situation, a new point is found, by the zero of the following:

$$T(x, x_k^*) = \frac{f(x) - f(x_k^*)}{[(x - x_k^*)^T (x - x_k^*)^\lambda]}$$

where λ is chosen iteratively, in such a way that the pole in x_k^* introduced in $T(x, x_k^*)$ makes $f(x) - f(x_k^*)$ equal to zero. Moreover, it must be taken into account that this method has no stop criterion. In fact, we could look for an $\bar{x}$ in such a way that $f(\bar{x}) = f(x_k^*)$, even when x_k^* is already the global minimum for $f(x)$.

2.4 Probabilistic methods

The probabilistic approaches define the global optimization problem over a limited region $D \subset \mathbb{R}^n$ that is:

$$\min_{x \in D} f(x)$$

Among the different probabilistic methods, we can refer to:

- Methods using random directions
- Multistart methods
- Chichinadze's methods
- Simulated annealing methods

Methods using random direction

When using this group of algorithms, at each iteration, the direction d_k is chosen randomly over an n-dimension sphere, with a unitary radius. These methods are based on the Gaviano's theorem, that states that if the sequence $\{x_k\}$ is in such a way that

$$x_{k+1} = x_k + \alpha_k d_k$$

where d_k is chosen randomly over the previous n-dimension sphere with a radius equal to 1 and α_k so that:

$$f(x_k + \alpha_k d_k) = \min_{\alpha} f(x_k + \alpha_k d_k)$$

then, $f(x_k) - f^* < \varepsilon$ happens with a probability that tends to one.

Multistart methods

These methods are based on the following considerations. *[43]*

Let $m(.)$ be the Lebesgue's measure over *D*. If *A* is a set with a measure *m(A)* in such a way that

$$1 \geq \frac{m(A)}{m(D)} = \alpha \geq 0$$

the probability *P(A,N)* that, taking into account N points randomly extracted (over *D*), at least one is inside *A* is given by:

$$P(A,N) = 1 - (1-\alpha)^N$$

and, from this equation, it follows:

$$\lim_{N\to\infty} P(A,N) = 1$$

so, we can conclude that, if we choose randomly many points, one of them, almost surely, is very close to the global minimum x^*.

2.5 Simulated annealing methods

These methods take inspiration from quantic mechanics theories. Let us consider a system made up by a very large number of particles of the same kind and let *s* stands for the system state and *E(s)* energy associated to this state. *[53]*. If the system is in thermal equilibrium, then the probability density that it would be in the state *s* is proportional to

$$e^{-\frac{E(s)}{KT}}$$

where, as defined in the previous chapters, *K* is the Boltzmann's constant and *T* is the temperature. It is generally known that, when lowering the temperature, states with low energy increase their own probability, up to the limit, when the temperature reaches the absolute zero and the only possible states are the ones with zero energy.

Now, let us consider a system, that associates at each state *x*, an energy amount:

$$E(x) = f(x) - f^* \geq 0$$

where, f^* is the global minimum for $f(x)$. Now, if the temperature would tend to zero, the states x^* would become more likely, in such a way that:

$$E(x^*) = f(x^*) - f^* = 0$$

In a more accurate way, we can underline the following theorem.

Let f(x) be a continuous function over a compact set $D \subset \mathbb{R}^n$. Let us assume that only one global minimum x exists for f(x) over D. Then, it is true that:*

$$x_i^* = \lim_{T\to 0} \frac{\int x_i e^{-(f(x)-f(x^*))/T} dx}{\int e^{-(f(x)-f(x^*))/T} dx} = \lim_{T\to 0} \frac{\int x_i e^{-f(x)/T} dx}{\int e^{-f(x)/T} dx} \qquad i = 1, \dots, n$$

that can be also expressed as:

$$x_i^* = \lim_{T\to 0} \int x_i P_T(x) dx = \lim_{T\to 0} \bar{x}_i(T)$$

where $P_T(x) = \frac{e^{-f(x)/T}}{\int e^{-f(x)/T} dx}$

it is a probability density where $\bar{x}_i(T)$ are the average values of some aleatory variables distributed according the density $P_T(x)$.
Then, the basic idea of these optimization methods is to simulate some aleatory arrays distributed according the probability density $P_T(x)$.
As *T* decreases, the arrays generated by the simulation come closer and closer, from a probabilistic standpoint, to the global minimum we are looking for.
The different algorithms belonging to this group, use different ways to perform this simulation.

Stop criteria

Within these probabilistic methods, a large number of different stop criteria have been proposed, but, the most interesting one is the one using a certain number of randomly chosen points over *D,* that tries to give to the solver an approximated value, as a probability $\tilde{P}_w$ of the function

$$P(w) = \frac{m(\{x: f(x) \le w\})}{m(D)}$$

where, as usual, *m(.)* is a set's Lebesgue's measure. After that, a point x^* can be considered a good estimation of the global minimum if

$$\tilde{P}(f(x^*)) \le \varepsilon << 1$$

3 Introduction to multiobjective programming

An optimization problem can be defined as both the minimization or maximization of a real function over a specified set. *[8]*
Its importance derives from the evidence that many real issues are formulated as an optimization probem. Anyway, almost any optimization problem is featured by the simultaneous presence of different objectives, that are real functions to be minimized or maximized, usually in conflict one against another. *[58]*
Let us consider the following optimization multiobjective problem:

$$\min (f_1(x) f_2(x) \dots f_k(x))^T \; with \; x \in F \subseteq \mathbb{R}^n \quad (1)$$

$$\text{where } k \geq 2 \; and \; f_i : \mathbb{R}^n \to \mathbb{R}$$

F is the set of feasible decision variables.
From now on we will refer to $\mathbb{R}^k$ as the **objective space** and $\mathbb{R}^n$ as the **decision variables space.**
An array $x \in \mathbb{R}^n$ so will be a **decision array** while $z \in \mathbb{R}^k$ is an **objective array.**
We will refer, moreover, to $f(x)$ as the objective function array $(f_1(x) f_2(x) \dots f_k(x))^T$ and to

$$Z = f(F) = \{z \epsilon \mathbb{R}^k : \exists x \in F, z = f(x)\}$$

as the imagine of the feasible region in the objective space
Especially, it is possible to say that an objective array $z \epsilon \mathbb{R}^k$ is feasible when $z \in Z$.
Moreover, it is possible to define the ***ideal*** objective array z^{id} as the array whose components are

$$z_i^{id} = \min_{x \in F} f_i(x)$$

The ideal situation represents the simultaneous optimization of all the objective functions. If there would not be conflicts among them, the trivial solution would be the one got by the separate solution of ***k different optimization problems*** (one for each objective function). This way we could just get the ideal array z^{id}. So, no particular solution technique would be needed. In order to avoid to treat this trivial case, it is requested to suppose that $z^{id} \notin Z$. This means that the functions $f_1(x), f_2(x), \dots f_k(x)$ are required to be, partly at least, in conflict one against another. *[58]*

3.1 Pareto optimality

The following optimality definition of a multiobjective problem has been first proposed by Edgeworth in 1881 and then redefined by Vilfredo Pareto in 1896 *[57]*

Given two arrays $z^1, z^2 \in \mathbb{R}^k$ we say that z^1 dominates z^2 according to Pareto ($z^1 \leq_P z^2$) when we have:

$$z_i^1 \leq z_i^2 \; \forall i = 1,2, \dots, k$$

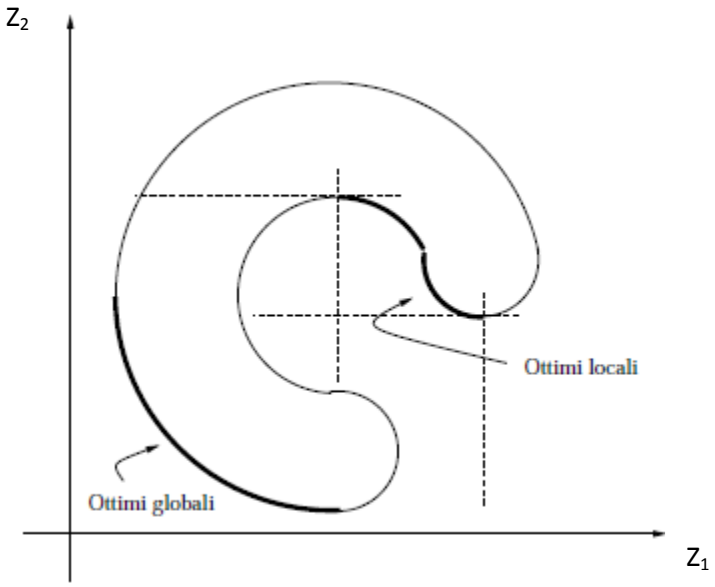
$$\text{and} \quad z_j^1 < z_j^2 \; for\; at\; least\; one\; j \in \{1, \dots, k\}$$

The binary relation $\leq_P$ is a **partial sort** in the set of k-tuples of real numbers. Through this relation, it is possible to define the Pareto optimality:

A decisions array $x^* \in F$ is optimal according to Pareto if there is no other array $x \in F$ in such a a way that

$$f(x) \leq_P f(x^*)$$

Below, a representation of local and global Pareto optimals:



Where Z_1 and Z_2 lay on the axes.
As a result, it is to say that an objective array $z^* \in Z$ is Pareto optimal when there is no other array $z \in Z$ in such a way that $z \leq_P z^*$.
So, if the solving method is already in a Pareto optimal point and the decision maker wishes to further reduce the value of one or more objective functions, it is needed to take into account a consequent increase in the value of some or all the objective functions. As a result, in the objective space, Pareto optimal points are to be considered as equilibrium points on the boundary of *Z*.
Now we define an **efficient boundary** the set of Pareto optimal points of the optimization problem. A Pareto optimum is therefore optimal, since it requests the satisfaction of the condition within the feasible set of the problem. It is also possible, moreover, to give a definition of Pareto local optimum:

a decision array $x^* \in F$ is a local optimum according to Pareto if exists a number $\delta > 0$ in such a way that x^* is a Pareto optimal within $F \cap B(x^*, \delta)$

where $B(x^*, \delta)$ is the neighborhood of center x^* and radius δ. Any global optimal point is a Pareto local optimal too, while the vice-versa is true only if some hypoteses are true:

1. The feasible set *F* is convex
2. All the objective functions $f_i(x)$ for *x = 1,2…k* are convex

In this case, it is possible to demonstrate that each Pareto local optimal is a global optimal point too. From this definition of Pareto optimum, it is possible to state the definition of **Pareto weak optimum** as follows:

an array $x^* \in F$ *is a Pareto weak optimal for the problem (1) if there is not any point* $x \in F$ *so that* $f(x) < f(x^*)$

where $f(x) < f(x^*)$ means $f_i(x) < f_i(x^*)$ $for\ each\ i = 1,2,\dots n$

It is possible to argue that the Pareto optimal set is a subset of the weak Pareto optimal and it is possible to define the local weak optimum:
a decision of array $x^* \in F$ *is a weak local optimum according to Pareto if exists a number* $\delta > 0$ *in such a way that* x^**is a Pareto weak optimum within* $F \cap B(x^*, \delta)$

Even for the weak optimality it is true that if the problem is convex, each local weak optimal is a Pareto global weak optimal too.

3.2 Efficient and dominated points

By the utilization of the concept of **cone** it is possible to generalize the definition of optimality and weak optimality according to Pareto: *[57]*

an array $y \in \mathbb{R}^n$ *is a conic combination of m arrays*$(x^1, x^2, \dots, x^m)$ *within* $\mathbb{R}^n$ *when it is possible to find m real numbers* $\lambda_1, \lambda_2, \dots, \lambda m$ *in such a way that:*

$$\sum_{i=1}^{m} \lambda_i x^i = y$$

where $\lambda_i \geq 0 \quad \forall i = 1, 2, \dots, m$

A set $D \subseteq \mathbb{R}^k$ *is a cone if the conic combination of the arrays of any finite subset of D belongs to D as well.*

Taken two arrays z^1 *e* z^2 *within* $\mathbb{R}^k$ *we can say that* z^1 *dominates* z^2 ($z^1 \leq_D z^2$) *if* $z^2 - z^1 \in D\backslash\{0\}$

Moreover, we can define an objectives array $z^* \epsilon Z$ ***efficient*** *in respect to a cone D if it is not possible to find any array* $z \epsilon Z$ *sothat* $z \leq_D z^*$*that is if and only if*

$$(z^* - D\backslash\{0\}) \cap Z = \emptyset.$$

Equally, an array of decisions $x^* \in F$ ***is efficient*** *respect to a cone $D \subset IR^k$ if and only if it does not exist any array $x \in F$ in such a way that $f(x) \leq_D f(x^*)$.*

1. **Optimality conditions**

Let us now consider a problem with regard to the set *F* defined by inequality constraints:

min f(x)
g(x) ≤ 0

where $f: \mathbb{R} \to \mathbb{R}^k$ for $k \geq 2$ and $g: \mathbb{R}^n \to \mathbb{R}^m$ are continuously derivable functions and F assumes the following structure:

$$F = \{x \in \mathbb{R}^n : g(x) \leq 0\}$$

We can indicate with the symbol $I_0(x) = \{i: g_i(x) = 0\}$ the set of valid constraints in the point x. It will be, moreover, $L: \mathbb{R}^{nxkxm} \to \mathbb{R}$ definied as follows $L(x, \lambda, \mu) = \lambda^T f(x) + \mu^T g(x)$ the Lagragian function coupled with the problem.

Let us remember, moreover, what the **Jordan's theorem**: one and only one of the two following systems has a solution,

$$Bz < 0, \begin{cases} B^T y = 0 \\ y \geq 0, \ y \neq 0 \end{cases}$$

where $B \in \mathbb{R}^{sxn}$, $z \in \mathbb{R}^n$ *and* $y \in \mathbb{R}^s$

Moreover, given a point $\bar{x} \in F$, an ***efficient direction*** *within $\bar{x}$ is an array $d \in \mathbb{R}^n$ with $d \neq 0$, for which exists a $\delta > 0$, in such a way that*

$$f(\bar{x} + \lambda d) \leq_P f(\bar{x}) \ \forall \ \lambda \in (0, \delta)$$

Thus, when moving from $\bar{x}$ along the direction of the *d* array and, for movements small enough, we are sure we will be finding points able to improve the value of at least n objective function, without, at the same time, worsening the value of the others. Moreover, we can indicate with

$$F(\bar{x}) = \{d \in \mathbb{R}^n \mid d \neq 0, f(\bar{x} + \lambda d) \leq_P f(\bar{x}) \ \forall \ \lambda \in (0, \delta) \ and \ some \ \delta > 0\}$$

the set of all the efficient directions within $\bar{x}$.
Evidently, if $\bar{x}$ is a Pareto local or global optimal, i twill result in $C(\bar{x}) \cap F(\bar{x}) = \emptyset$, that is no feasible direction can take us to points such that $f(x) \leq_P f(\bar{x})$.
To define from an analytical standpoint the Pareto otpimal points, it is needed to define these other sets:

$$F_0(\bar{x}) = \{d \in \mathbb{R}^n \mid \nabla f_i(\bar{x})^T d < 0, \forall\, i = 1, 2, \dots, k\}$$
$$\text{and} \quad G_0(\bar{x}) = \{d \in \mathbb{R}^n \mid \nabla g_i(\bar{x})^T d < 0, \forall\, j \in I_0(\bar{x})\}$$

For any x ∈ *F,* it results that $G_0(x) \subseteq C(x)$ and $F_0(x) \subseteq F(x)$ so,
necessary condition in order to the point $\bar{x} \in F$ *would be a Pareto optimal (local or global) is that*

$$G_0(\bar{x}) \cap F_0(\bar{x}) = \emptyset$$

So, the **Fritz-John's theorem for the multiobjective programming will be valid:**
necessary condition in order to get $\bar{x} \in F$ *as a Pareto optimal is that some arrays* $\lambda \in \mathbb{R}^k$ *will exist and* $\mu \in \mathbb{R}^m$ *in such a way that the following system would be satisfied:*

$$\sum_{i=1}^{k} \lambda_i \nabla f_i(\bar{x}) + \sum_{j=1}^{m} \mu_j \nabla g_j(\bar{x}) = 0$$

$$\mu^T g(\bar{x}) = 0, \qquad (\lambda, \mu) \geq 0, \qquad (\lambda, \mu) \neq (0,0)$$

3.3 Solution methods

The generation of the Pareto optimal solutions is an essential part of vectorial programming and, in most cases, from a mathematical standpoint, the problem (P) can be considered solved once the Pareto optimals set has been found. But, this result is not always sufficient. Sometimes, in fact, it is needed to sort all the solutions and then to select the best one respect to this sorting. That is why a decision maker is often needed, that is someone able to decide, according to its preferences, how to sort the Pareto optimals set of the problem (P).
On the basis of the role played by the decision maker over the problem solution, the multiobjective programming solving methods can be divided into four groups. *[58]*

a) **Methods without preferences** according to which the decision maker plays no role and it is considered satisfiying the finding of any Pareto optimal.

b) **Ex-post methods** according to which the whole Pareto optimal set is generated and the nit is presented to the decision maker, in order to let it choose the best optimal solution among the solutions presented.

c) **Ex-ante methods** according to which the decision maker specifies its preferences before the problem solving starts. On the basis of the information got from the decision maker, the best optimal solution is determined, avoiding the generation of all the Pareto optimal solutions.

d) **Interactive methods** according to which the decision maker specifies its preferences while the algorithm goes on. This is the way the solution process is guided towards the most satisfying possible solutions.

Beyond this distinction, all the multiobjective programming solution methods are based on the same idea, that is to shift the solution process from the original problem to another one **with only one objective function**. This technique, through which the mono-objective problem is got from the problem (P) is called ***scalarization***.

e) **Methods with no preferences**

While using the methods with no preferences, the solution process simply generate an optimal Pareto solution, whatever it is, without considering the decision maker preferences. The method analyzed is the so called **GOAL method**, that makes us look for the solution that minimizes, within the objectives set, the distance between the feasible region (*Z)* and any other reference point $z^{ref} \notin Z=f(F)$. The reference array will contain the desirable values for the single objective functions, especially, a possible choose of z^{ref} is $z^{ref} = z^{id}$. The problem got this way is the following one:

$$\min \ ||f(x) - z^{id}||_p$$
$$g(x) \leq 0$$

In this problem $\|\cdot\|_p$ stands for the **norma** of an array, that has a value between 1 and ∞ and, in particular, if p = ∞ the problem P_p is known as the **Tchebycheff's problem**. Let us suppose now we already know the global array of objectives. Under these hypotheses, the problem P_p always has a solution and the following properties are valid

Any global solution of the problem P_p with $1 \leq p \leq \infty$ is a Pareto global optimal for the problem P.

Any local optimal of the problem P_p with $1 \leq p \leq \infty$ is a Pareto local optimal for the problem P.

While, if we consider the case p = ∞, the following is valid:

Any local (global) Pareto optimal for the Tchebycheff's problem (P_∞) is a local (global) weak Pareto optimal for the problem P.

From this statement, the fact that at least an optimal Pareto solution of P_∞ exists for the problem P. Choosing p = 1 and p = ∞ is very convenient in the case the original multiobjective problem is linear, because through simple manipulations of the problem P_p it is possible to get a linear problem again and then to adopt the well known Linear Programming techniques for its solution. So, let us suppose that *P* is linear, that is:

$$min \ (c_1^T x, c_2^T x, \dots, c_k^T x)$$
$$Ax \leq b$$

- Case *Norm p = 1*

The scalarized problem is

$$min \sum_{i=1}^{k} \left| c_i^T x - z_i^{id} \right|$$
$$Ax \leq b$$

that can be simply transformed into a Linear Programming problem by adding k auxiliary variables, α_i for i=1, 2, … , k, getting:

$$\min \sum_{i=1}^{k} \alpha_i$$
$$\begin{cases} \left| c_i^T x - z_i^{id} \right| \leq \alpha_i \\ Ax \leq b \end{cases} \quad \text{with i=1, 2, …, k}$$

- Case *Norm p = ∞*

In this case, the scalarized problem is

$$\min \max_{i=1,\dots,k} \left\{ \left| c_i^T x - z_i^{id} \right| \right\}$$
$$Ax \leq b$$

and it can be easily transformed into a PL problem by adding only one auxiliary variable α, gaining:

$$\min \alpha$$
$$\begin{cases} \left| c_i^T x - z_i^{id} \right| \leq \alpha \\ Ax \leq b \end{cases} \quad \textit{with } i = 1, 2, \dots, k$$

f) **Ex-post methods**

The methods belonging to this group generate the Pareto solution set. In fact, the decision maker preferences are taken in account only after that the solution process ends, allowing the decision maker itself to choose the array or the arrays considered the best one(s) among the **Pareto optimals** generated.

The main drawback of these methods, consists in the fact that, often, the Pareto optimal generation process is hard from a computational standpoint and, if the solutions among which we have to choose are a large number, the decision maker choose is not easy. That is why, it is very important the way the solutions are presented to the decision maker.

1. Weights methods

Let us consider the following problem:

$$Min \sum_{i=1}^{k} w_i f_i(x)$$
$$g(x) \leq 0$$

In which $w \in \mathbb{R}_+^k$ and the w_i coefficients are normalized in such a way that:

$$\sum_{i=1}^{k} w_i = 1$$

A relation between the P_w problem solutions and the P problem Pareto points exists and this relation can be expressed as follows:
any local (global) solution for the P_w* problem is a local (global) weak Pareto optimal for the problem *P and, moreover, **if the problem P_w for a certain weights array $w \geq 0$, has only one solution, so, it is a Pareto optimal for the P problem.**

It is also possible to guess some hypoteses about the weights, if, in fact, all of them are positive, the following is valid:
if $w_i > 0$ for any i, each local (global) solution for the problem P_w is a local (global) Pareto optimal for the problem P.
If the multiobjective problem P is convex, it is possible to state the following existence statement:
Let it be x^* a Pareto optimal for the problem P. If P is convex, so the weights $w \in \mathbb{R}_+^k$ exists, with

$$\sum_{i=1}^{k} w_i = 1$$

and, in such a way that x^* is a solution for the problem P_w too

2. ε-Constraints method

It is requested to select an objective function $f_l(x)$ among P's objectives and then all other functions $f_i(x)$ with i=1, 2, … , k are transformed and i ≠ l, into constraints, imposing some upper bounds over their values. The problem obtained, so, is the following one:

$\min f_l(x)$
$f_i(x) \leq \varepsilon_i \quad \forall\, i = 1, 2, \dots, k \text{ and } i \neq l$
$g(x) \leq 0$

Where $l \in \{1, 2, \dots, k\}$

It is so valid the following, that **any solution for P_ε *is a weak Pareto optimal for the P problem.***
It is also valid the following, that **an array $x^* \in F$ *is a Pareto optimal for P, if and only if it is a solution for P_ε for any $l \in \{1, 2, \dots, k\}$ and being $\varepsilon_i = f_i(x^*)$ for any $i \neq l$.***

Finally **if the point $x^* \in F$ *is the only solution for the problem* P_ε *for any* $l \in \{1, 2, \dots, k\}$ *and with* $\varepsilon_j = f_j(x^*)$ *for any* $j \neq l$, *so, it is a Pareto optimal for the problem P.***

- ***Ex ante methods***

In the ex-ante methods the decision maker gives the needed information before the solution process starts. Then, the algorithm stops just when a Pareto optimal has been found. Some examples of these ex-ante methods are **the Value Function method** and the **lexicographic method**.

The Value Function method requie that the decision maker specifies an **analytic expression of a utility objectives function *U(z)***. Then, the following problem is solved:

$$\begin{aligned} \min \quad & U(f(x)) \\ & g(x) \leq 0 \end{aligned}$$

If the *U(z)* would be linear, we would get the weights' problem P_w again, while the weights' method is an ex-post one, in which the solver firstly has to generate all the Pareto solutions, by the modification of the weights time by time, the Value Function method is an ex-ante one, that the decision maker communicates its preferences through the weights and imposes this way the relative utilities of the objective functions. It is so needed to underline that the *U(z)* function, generally, could be not linear.
Using the lexicographic method, the decision maker sorts the objective functions according to their relative utility, and, at this point, the solution process starts with the minimization of the first objective function over the feasible original set *F*, that is the solution of the problem:

$$\begin{aligned} \min \quad & f_1(x) \\ & g(x) \leq 0 \end{aligned}$$

If the problem P_i has only one solution, then this one is a solution for the *P* problem too and the algorithm ends. Else, the second objective function is to be minimized according to the lexicographic sorting. But this time, in addition to the original constraints, an additional constraint is added, and its purpose is to guarantee that the optimal point value would not worsen the value of the first objective function evaluated. Then, the problem becomes:

$$\begin{aligned} \min \quad & f_2(x) \\ & f_1(x) \leq f_1(x^{1*}) \\ & g(x) \leq 0 \end{aligned}$$

where x^{1*} is a solution of the problem P_i.
It is true that **any solution got through the lexicographic method is a Pareto optimal one for the problem P.**

- **Interactive methods**

The general flow-chart for an interactive method is the following one: *[59]*

1. Find a starting feasible solution
2. Propose this starting solution to the decision maker
3. If the solution is good for the decision maker, then the algorithm stops
4. Otherwise, on the basis of the suggestions coming from the decision maker, a new solution is generated and the algorithm goes again to step 2

An interactive method is the **STEP method**.
Let us suppose that in a Pareto optimal point, the decision maker knows which objective functions have an acceptable value and which have not. Let us suppose, moreover, that all the objective functions would be limited over the feasible region.
At each iteration of this method, a Pareto optimal is presented to the decision maker, that, on the basis of what it knows about the objective functions' values in the interested point, specifies the objective functions for which it is acceptable an increase of the value, in order to further reduce the values of the other functions. This is equivalent to let that the indexes set $J = \{1, 2, \dots, k\}$ is divided at each step into two subsets:

1. $J_<$ set containing the indexes of the objective functions whose values are not satisfying for the decision maker at the present step

2. $J_> = J \setminus J_<$

If $J_< = \emptyset$ then the algorithm stops because the Pareto optimal that i salso the best solution for the decision maker has been found. Otherwise, the decision maker is requested to specify some bounds ε_i over the objective functions whose index is contained in the set $J_>$ and then is to be solved the problem:

$$\min \left(\sum_{i \in J_<} \left| f_i(x) - z_i^{id} \right|^p \right)^{\frac{1}{p}}$$

$$\begin{aligned} & f_i(x) \le \varepsilon_i && i \in J_> \\ & f_i(x) \le f_i(x^*) && i \in J_< \\ & g(x) \le 0 \end{aligned}$$

where $1 \le p \le \infty$ and x^* is the Pareto optimal point found during the previous step.
The algorithm has to stop, in order to avoid useless computing, even when the set $J_>$ is empty, that is when all the objective functions has unsatisfying values. So, the algorithm can be summarized as follows:

1. With any method, the first Pareto optimal point $x^1 \in F$ is found and then has to be set $h = 1$
2. The decision maker is requested to divide the set J into $J_>^h$ and $J_<^h$. if $J_>^h = \emptyset$ or $J_<^h = \emptyset$ the algorithm goes directly to step 4.
3. The following problem is solved:

$$\min \left(\sum_{i \in J_<^h} \left| f_i(x) - z_i^{id} \right|^p \right)^{\frac{1}{p}}$$
$$f_i(x) \leq \varepsilon_i^h \qquad i \in J_>^h$$
$$f_i(x) \leq f_i(x^h) \qquad i \in J_<^h$$
$$g(x) \leq 0$$

for $1 \leq p \leq \infty$. Let it be x^{h+1} the solution of the problem above, it is set $h = h+1$ and the algorithm goes again to step 2.

4. It is set $x^* = x^h$ and the algorithm stops.

4 Direct search algorithms for optimization calculations

4.1 Line search methods

Line search methods for unconstrained optimization are iterative. A starting vector of variables $\underline{x}_1 \in \mathbb{R}^n$ has to be given, and, for *k=1,2,3,…,* the k-th iteration derives $\underline{x}_{k+1}$ from $\underline{x}_k$ in the following way. A nonzero search direction $\underline{d}_k \in \mathbb{R}^n$ is chosen. Then the function of one variable $\phi(\alpha) = F(\underline{x}_k + \alpha \underline{d}_k),\ \alpha \in \mathbb{R}$, receives attention, in order to pick a new vector of variables of the form

$$\underline{x}_{k+1} = \underline{x}_k + \alpha \underline{d}_k \quad (1)$$

for example, an "exact line search" would set the step-length α_k to an α that minimizes $\phi(\alpha)$. In practice, however, one tries to choose α_k in a way that requires very few values of $F(\underline{x}_k + \alpha \underline{d}_k)$, $\alpha \in \mathbb{R}$, on each iteration, and it is unusual to satisfy the condition

$$F(\underline{x}_{k+1}) \leq F(\underline{x}_k),\ k = 1,2,3 \ldots \quad (2)$$

of course the search directions should be able to explore the full space of the variables. Therefore, line search methods should have the property that, for some integer, $l \leq n$, any consecutive search directions span $\mathbb{R}^n$ in a strict sense. If this condition failed, then a nonzero $v \in \mathbb{R}^n$, would be (nearly) orthogonal to the directions. Therefore a convenient form of the strict sense is that the bound

$$max\left\{|\underline{v}^T \underline{d}_j|/\|\underline{d}_j\|^2 : j = k-l+1, k-l+2, \ldots, k\right\} \geq c\|\underline{v}\|_2, \quad \underline{v} \in \mathbb{R}^n \quad (3)$$

is satisfied for $k \geq l$, where c is a positive constant. For example, a way of achieving this condition, which gives $l=n$ and $c=n^{-1/2}$ is to let each $\underline{d}_k$ be a coordinate direction in $\mathbb{R}^n$ and, to cycle round the n coordinate directions recursively as k increases. Rosenbrock provides an extension of this technique that is sometimes useful. His first n directions are also the coordinate directions, but, when k is any positive integer multiple of n, then, before starting the (k+1)-th iteration, he generates $\underline{d}_{k+1}, \underline{d}_{k+2}, \ldots, \underline{d}_{k+n}$ in sequence, by applying the Gram-Schmidt procedure to the differences $\underline{x}_{k+1} - \underline{x}_{k-n+j},\ j = 1,2,\ldots,n$. Further, he ensures that every step-length is nonzero, although condition (2) may have to fail.

Unfortunately, condition (3) and exact line searches do not guarantee that limit points of the sequence $\underline{x}_k,\ k = 1,2,3,\ldots$, are good estimates of optimal vectors of variables, even if the objective function is continuously differentiable, and the level set $\{x: F(\underline{x}) \leq F(\underline{x}_1)\}$ is bounded. Indeed, Powell gives an example of bad behavior, with $n=3$ and exacr line searches, where the sequence $\underline{d}_k,\ k = 1,2,3,\ldots$, is generated by cycling round the coordinate directions. Here, for each integer I in [1,6], the infinite sequence $\underline{x}_{6j+i},\ j = 1,2,3,\ldots$, tends to one vertex of a cube, and, in the path from $\underline{x}_k$ to $\underline{x}_{k+6}$ tends to be a cycle along six edges of the cube. Further, the objective function is constant

on each of these edges, which implies that two components of the gradient $\underline{\nabla}F$ are zero at each limiting vertex.

The other component of $\underline{\nabla}F(\underline{x}_k)$, however, is bounded away from zero for each stationary point of *F*. Therefore, it is easy to modify the algorithm so that the objective function becomes less than the actual limit of decreasing sequence $F(\underline{x}_k)$, $k \to \infty$. Specifically, we replace $\underline{d}_j$ by a difference approximation to $-\underline{\nabla}F(\underline{x}_j)$ for any integer *j* that is sufficiently large. Furthermore, there is another remedy that does not require an estimate of $\underline{\nabla}F$.

The kind of ingredient that avoids the bad behavior above is imposing the condition that, if $\|\underline{x}_{k+1} - \underline{x}_k\|$ is bounded away from zero, then $F(\underline{x}_k) - F(\underline{x}_{k+1})$ is bounded away from zero too. Hence, in the usual case when $F(\underline{x}_k)$, $k = 1,2,3, \ldots$, converges monotonically, we have the limit

$$\|\underline{x}_{k+1} - \underline{x}_k\| \to 0 \ \ as\ k \to \infty \quad (4)$$

which prevents the cycling round the edges of the cube.

Further, if the directions $\underline{d}_j$, $j = 1,2,3, \ldots$, satisfies inequality (3) and, if $\underline{x}^*$ is any limit point of the infinite sequence $\underline{x}_k$, $k = 1,2,3, \ldots$, then $\underline{\nabla}F(\underline{x}^*) = 0$ can be obtained by a suitable line search, provided that *F* is continuously differentiable and bounded below.

We are going to prove this assertion, not only because the method of proof provides a demonstration of the kind of analysis that can establish convergence properties. A way of achieving the restriction, due to Lucidi and Sciandrone, will be given after the proof.

We aim tom deduce a contradiction from the assumption $\|\underline{\nabla}F(\underline{x}^*)\|_2 = \eta$, where η is a positive constant and where $(\underline{x}^*)$ is a limit point of the sequence $\underline{x}_k$, $k = 1,2,3, \ldots$, as stated already. We seek some integers *j* such that $\underline{\nabla}F(\underline{x}_j)^T \underline{d}_j / \|\underline{d}_j\|_2$ is bounded away from zero, because then the step-length α_j of the equation $\underline{x}_{j+1} = \underline{x}_j + \alpha_j d_j$ can be choosen so that $F(\underline{x}_j) - F(\underline{x}_{j+1})$ is aso bounded away from zero, which gives the required contradiction if this happens an infinite number of times.

Now, by setting $\underline{v} = \underline{\nabla}F(\underline{x}^*)$ in expression (3), we deduce that the inequality

$$\left\|\underline{\nabla}F(\underline{x}^*)^T \underline{d}_j\right\| / \|\underline{d}_j\|_2 \geq c\|\underline{\nabla}F(\underline{x}^*)\|_2 = c\eta \quad (5)$$

is achieved at least once for every *l* consecutive positive integers *j*. Further, because $\underline{\nabla}F$ is continuous, this inequality implies $\left\|\underline{\nabla}F(\underline{x}^*)^T \underline{d}_j\right\| / \|\underline{d}_j\|_2 \geq 1/2c\eta$, provided that $\underline{x}_j$ is sufficiently close to $\underline{x}^*$. Specifically, $\underline{x}_j$ is close enough to $\underline{x}^*$ if it satisfies $\|\underline{x}_j - \underline{x}^*\|_2 < \varepsilon$, where ε is a positive constant that provides the property

$$\|\underline{\nabla}F(\underline{x}) - \underline{\nabla}F(\underline{x}^*)\|_2 \leq \frac{c\eta}{2} \quad if \ \ \|\underline{x} - \underline{x}^*\|_2 < \varepsilon \quad (6)$$

because then the Cauchy-Schwarz inequality and condition (5) give the bound

$$\frac{\left|\underline{\nabla}F(\underline{x}^*)^T\underline{d}_j\right|}{\|\underline{d}_j\|_2} \geq \frac{\left|\underline{\nabla}F(\underline{x}^*)^T\underline{d}_j\right| - \left|\{\underline{\nabla}F(\underline{x}_j) - \underline{\nabla}F(\underline{x}^*)\}T\underline{d}_j\right|}{\|\underline{d}_j\|_2} \geq c\eta - \frac{c\eta}{2} = \frac{c\eta}{2} \quad (7)$$

therefore it remains to show that, on an infinite number of occasions, l consecutive positive integers j satisfy $\|\underline{x}_j - \underline{x}^*\|_2 \leq \varepsilon$. The limit (4) is helpful, because it admits an integer $j_0 > 0$ such that $\|\underline{x}_{j+1} - \underline{x}_j\|_2 \leq \frac{\varepsilon}{2} / (l-1)$ holds for all $j > j_0$. Hence, if $\|\underline{x}_k - \underline{x}^*\|_2 \leq \frac{\varepsilon}{2}$ occurs for some integer $k > j_0$ then $\|\underline{x}_j - \underline{x}^*\|_2 \leq \varepsilon$ is obtained by every integer j in *[k, k+l-1]*. This does happen an infinite number of times, because $\underline{x}^*$ is limit point of $\underline{x}_k, k = 1,2,3, \ldots$, even if we require the differences between the chosen integers k to be at least l. the proof is complete.

The line search procedure of Lucidi and Sciandrone is suitable for the above analysis, although some parameters are required that may be difficult to choose well in practice. They are numbers γ and δ that satisfy $\gamma > 0$ and $0 < \delta < 1$ and a positive sequence $\{\beta_k : k = 1,2,3, \ldots\}$ that tends to zero as $k \to \infty$. Then, on each iteration, there is a search for a step-length $\alpha_k = \alpha$ that has the properties

$$\begin{cases} F(\underline{x}_k + \alpha d_k \leq F(\underline{x}_k) - \gamma\alpha^2\|\underline{d}_k\|_2^2 \quad and \\ \min\left[F(\underline{x}_k + \hat{\alpha}\underline{d}_k), F(\underline{x}_k + \hat{\alpha}\underline{d}_k)\right] \geq F(\underline{x}_k) - \gamma\alpha^2\|\underline{d}_k\|_2^2 \ \right] \end{cases} \quad (8)$$

where $\hat{\alpha} = \frac{\alpha}{\delta}$. If the first line of expression (8) holds for a trial $\alpha > 0$, then either α is acceptable or the second line shows that a step-length of larger modulus is allowed by the first line, namely $\frac{\alpha}{\delta}$ or $-\frac{\alpha}{\delta}$. Thus, the modulus of α is increased if necessary, and the second line is tested for a new α. This procedure is continued recursively until α is acceptable, which happens eventually because we are assuming that F is bounded below. Alternatively, if the first line of expression (8), not only for the initial α but also for $-\alpha$, then α is replaced by $\alpha\delta$ and these tests are tried again. Thus, the second inequality of expression (8) is achieved by the new α. Again recursion is applied, either until an acceptable step-length is found.

Moreover, the search directions $\underline{d}_k$, $k = 1,2,3, \ldots$, have to satisfy the strict linear independence condition (3). These constructions provide the conclusion $\underline{\nabla}F(\underline{x}^*) = 0$ as shown below. The first line of expression (8) and equation (1) imply the bound

$$F(\underline{x}_k) - F(\underline{x}_{k+1}) \geq \gamma\|\underline{x}_k - \underline{x}_{k+1}\|_2^2 \quad k = 1,2,3, \ldots, \quad (9)$$

when α_k is positive and the bound is trivial when α_k is zero. Therefore, the limit (4) at the beginning of the given analysis is valid and the conclusion $\underline{\nabla}F(\underline{x}^*) = 0$ of the analysis holds, provided that inequality (7) causes $F(\underline{x}_j) - F(\underline{x}_{j+1})$ to be bounded away from zero. Further, the method of analysis allows us to restrict attention to values of j that satisfy two more conditions. Firstly, we assume $j > j_0$, where j_0 is fixed positive integer, which may be larger than j_0 introduced earlier. Thus, we allow for the zero step-length in the line search procedure under consideration.

Secondly, we assume $\|\underline{x}_j - \underline{x}^*\|_2 \leq \varepsilon$, although our previous use of this bound was only to establish the existence of integers *j* that have the property (7). Thus, the uniform continuity of $\underline{\nabla} F$ in any neighbourhood of $\underline{x}^*$ provides the condition

$$\|\underline{\nabla} F(\underline{x}) - \underline{\nabla} F(\underline{x}_j)\|_2 \leq \frac{c\eta}{4} \quad if \quad \|\underline{x} - \underline{x}_j\|_2 \leq \hat{\varepsilon} \quad (10)$$

for all of the values of *j* that are obtained, where $\hat{\varepsilon}$ is a positive number that is independent of *j*.
now, it follows from expressions (7) and (10) that the gradient of the line search function $\phi(\alpha) = F(\underline{x}_k + \alpha \underline{d}_k)$, $\alpha \in \mathbb{R}$, is bounded by the inequality

$$|\phi'(\alpha)| = |\underline{d}_j^T \underline{\nabla} F(\underline{x}_j + \alpha \underline{d}_j)| \geq \|\underline{d}_j\|_2 \|\underline{\nabla} F(\underline{x}_j + \alpha \underline{d}_j) - \underline{\nabla} F(\underline{x}_j)\|_2 \geq \frac{c\eta}{4} \|\underline{d}_j\|_2 \quad if \quad \|\alpha \underline{d}_j\|_2 < \hat{\varepsilon} \quad (11)$$

Therefore, by choosing the sign of α to be opposite to the sign of $\phi'(0)$ and by applying $\phi(\alpha) = \int_0^\alpha \phi'(\theta) d\theta$, we find the relation

$$F(x_j + \alpha d_j) \leq F(x_j) - \frac{c\eta}{4} \|\alpha d_j\|_2 \quad if \quad \|\alpha d_j\|_2 \leq \hat{\varepsilon} \quad (12)$$

Thus, the first line of expression (8) is achieved by every α of the appropriate sign that satisfies $\|\alpha \underline{d}_j\|_2 < \hat{\varepsilon}$ and $\|\alpha d_j\|_2 \leq \frac{c\eta}{4}$. It follows that, if the parameter β_j of the line search procedure is at most $\delta \min\left[\hat{\varepsilon}, \frac{c\eta}{4\gamma}\right]$, and if the first trial value of α on the j-th iteration is at least β_j, then the procedure provides a step-length α_j that is positive. The first of these conditions is irrelevant, if j_0 is sufficiently large, as assumed before and, any sensible implementation observes the second condition. Therefore, both the inequalities (8) hold for $k = j$ with $\alpha = \alpha_j > 0$. We deduce form the second one and from the property (12) that $\|\hat{\alpha} \underline{d}_j\|_2 = \frac{\|\alpha_j \underline{d}_j\|_2}{\delta}$ is no less than $\min\left[\hat{\varepsilon}, \frac{c\eta}{4\gamma}\right]$, which gives the inequality

$$\|\underline{x}_{j+1} - \underline{x}_j\|_2 = \|\alpha_j \underline{d}_j\|_2 \geq \delta \, min\left[\hat{\varepsilon}, \frac{c\eta}{4\gamma}\right] \quad (13)$$

Thus condition (9) provides a positive lower bound on $F(\underline{x}_j) - F(\underline{x}_{j+1})$ as required. Therefore line search methods without derivatives can provide convergence properties of the kind that are acclaimed by theoreticians when $F(\underline{x}), x \in \mathbb{R}^n$, need not be convex.

4.2 Linear approximation methods

The changes to the variables in the simplex methods depend on the positions $\underline{v}_i, i = 1,2,\dots,n+1$ of the vertices of the current simplex and on an integer m in *[1, n+1]* that is usually defined by the conditions $F(\underline{v}_m) \geq F(\underline{v}_i), i = 1,2,\dots,n+1$. These methods make no other use of $F(\underline{v}_i), i = 1,2,\dots,n+1$, however, when choosing the next vector of variables for the calculation of the objective function, although the function values at the vertices can provide highly useful information when F is smooth. In particular, there is a unique linear polynomial from $\mathbb{R}^n \; to \; \mathbb{R}$, Φ say, that satisfies the interpolation conditions

$$\Phi(\underline{v}_i) = F(\underline{v}_i), \quad i = 1,2,\dots,n+1 \quad (1)$$

and often $\underline{\nabla}\Phi$ is very helpful for reducing the least calculated value of F. Therefore we will consider changes to the variables that are derived from Φ. The given procedures also allow constraints on the variables of the form

$$c_p(\underline{x}) \geq 0, \qquad p = 1,2,\dots,m \quad (2)$$

where m denotes the number of constraints from now until the end of the section. The constraints functions have to be specified by a subroutine that calculates $c_p(\underline{x}),\ p = 1,2,\dots,m$ at the points $\underline{x} \in \mathbb{R}^n$ that are generated automatically. These points include the vertices of the current simplex, in order that, for each p, we can let γ_p be the linear polynomial from $\mathbb{R}^n \; to \; \mathbb{R}$ whose coefficients are defined by the equations $\gamma_p(\underline{v}_i),\ i = 1,2,\dots,n+1$, which implies the conditions

$$\underline{x}_k \in \{\underline{v}_i : i = 1,2,\dots,n+1\} \; and \; F(\underline{x}_k) \leq F(\underline{v}_i),\ \ i = 1,2,\dots,n+1 \quad (3)$$

in the unconstrained case. Each iteration until termination generates a new vector of variables, $\underline{v}_{n+2}$ say, where the difference $\underline{v}_{n+2} - \underline{x}_k$ is either a "minimization step" or a "simplex step". The values of F and any constraint functions are calculated at $\underline{v}_{n+2}$. Then the $n+1$ vertices of the simplex of the next iteration are chosen by deleting one point from the set $\{\underline{v}_i : i = 1,2,\dots,n+2\}$. Further, $\underline{x}_{k+1}$ is defined in the way mentioned earlier, any ties being broken by retaining $\underline{x}_{k+1} = \underline{x}_k$, unless a change provides a strict improvement according to the criterion for the best vertex. An iteration also sets the parameters Δ_{k+1} and ρ_{k+1} before increasing k, where $\rho_1 = \Delta_1$. All of these operations receive further consideration below.

The minimization of $\Phi(\underline{x}), x \in \mathbb{R}^n$, subject to constraints (4) is a linear programming problem that usully fails to have a finite solution in the case $m < n$. Further, it is likely that the linear approximations are too inaccurate to be useful when $\underline{x}$ is far from the current simplex. Therefore we consider algorithms that employ trust region bounds. Specifically, the vector $\underline{v}_{n+2}$ is the vector of the k-th iteration has to satisfy the inequality

$$\|\underline{v}_{n+2} - \underline{x}_k\| \leq \rho_k \quad (4)$$

where ρ_k is a positive number that is available at the beginning of the iteration, but is may be reduced occasionally. On most iterations, $\underline{v}_{n+2}$ is the vector $\underline{x}$ that minimizes $\Phi(\underline{x})$ subject to $\|\underline{x} - \underline{x}_k\| \leq \rho_k$ and the conditions (4) and then $\underline{v}_{n+2} - \underline{x}_k$ is the "minimization step", provided that $\|\underline{v}_{n+2} - \underline{x}_k\|$ is as small as possible if the solution to this subproblem is not unique. It can happen, however, that the constraints of the subproblem are inconsistent, and then the "minimization step" is defined by minimizing the greatest violation of a linear constraint, namely $\max\{-\gamma_p(\underline{v}_{n+2}: p = 1,2,\dots,m\}$, subject to inequality (4) where again any nonuniqueness is taken up by reducing $\|\underline{v}_{n+2} - \underline{x}_k\|$. Powell addresses these calculations when the vector norm is Euclidean and recommends a procedure that generates the path $\underline{v}_{n+2}(\alpha),\ 0 < \alpha < \rho_k,\ in\ \mathbb{R}^n$, where $\underline{v}_{n+2}(\alpha)$ is the $\underline{v}_{n+2}$ that would be required if ρ_k were equal to α. This path begins at the point $\underline{v}_{n+2}(0) = \underline{x}_k$, and is continuous and piecewise linear. Further, the different pieces of the path correspond to different indices of critical constraints, the q-th constraint being critical if and only if the conditions $\gamma_q(x) \leq 0$ and $\gamma_q(x) \leq \gamma_p(x),\ p = 1,2,\dots,m$ hold. Sometimes the length $\|\underline{v}_{n+2} - \underline{x}_k\|$ of the "minimization step" is too small and then it is usual to replace $\underline{v}_{n+2} - \underline{x}_k$ by a simple step. There are also iterations that calculate only a "simplex step". The reasons for these alternatives are as follows.

We consider the case when there are no given constraints on the variables, when $\underline{v}_{n+2} - \underline{x}_k$ is a "minimization step", when $\|\underline{v}_{n+2} - \underline{x}_k\|$ is large enough for $F(\underline{v}_{n+2})$ to be calculated and when the new function value has the property

$$F(\underline{v}_{n+2}) \geq F(\underline{x}_k) \quad (5)$$

Then, because the definition of the minimization step implies $\Phi(\underline{v}_{n+2}) \geq F(\underline{x}_k)$, the approximation $\Phi(\underline{v}_{n+2}) \approx F(\underline{v}_{n+2})$ is inadequate. There are two main causes of the inadequacy, and it is important to distinguish between them. Firstly, $\underline{v}_{n+2}$ may be so far from $\underline{x}_k$ that very god linear approximations to F in a neighbourhood of $\underline{x}_k$ may be unsuitable at $\underline{v}_{n+2}$, due to second and higher order terms of lack of smoothness of the objective function. Secondly, although the bound (4) may ensure that any one of these very good approximations provides a minimization step that is successful at reducing the least calculated value of F, the interpolation condition (1) may define a linear polynomial Φ that is unhelpful. This can happen if one or more of the distances $\|\underline{v}_i - \underline{x}_k\|$, $i = 1,2,\dots,n+1$, is much greater than ρ_k or if the current simplex is nearly degenerate. The appropriate remedy in the first case is so shorten the length of the minimization step on the next iteration by choosing $\rho_{k+1} < \rho_k$, which is a standard technique in trust region algorithms. In the second case, however, the remedy is to choose a better simplex. When $\underline{v}_{n+2}$ is calculated for the latter purpose, we call $\underline{v}_{n+2} - \underline{x}_k$ a "simplex step".

The choice of independent Φ, except for a plus or minus sign, and $\underline{v}_{n+2}$ becomes one of the vertices of the simplex of the next iteration. The need for such steps is clear if a given constraint on the variables is linear and, if, $\underline{v}_{n+2}$ satisfies the constraint as an equation for all minimization steps. Indeed, if there were no simplex steps in this case, and if all of the vertices of the initial simplex have been removed from the current simplex by earlier iterations, then all of the current vertices $\underline{v}_i,\ i = 1,2,\dots,n+1$ are on the boundary of the linear constraint. Thus the equations (1) fail to

define the coefficients of Φ, because the matrix of the equations is singular. Therefore a reason for the "simplex steps" is to oppose any tendencies for the current simplex to become degenerate.

It has been mentioned that $\Delta_1 > 0$ is a prescribed parameter that controls the size of the initial simplex. Most of the later iterations employ $\Delta_k = \Delta_{k-1}$, and each Δ_k is an acceptable length for the edges of the current simplex, the length being relevant to the suitability of the linear polynomial Φ defined by the equations (1). Specifically, it is assumed that the nonlinearities of the objective function may damage the usefulness of the approximation $\Phi \approx F$, if any of the distances $\|\underline{v}_i - \underline{x}_k\|$, $i = 1,2,\dots,n+1$, is much greater than Δ_k. On the other hand, when "minimization steps" are successful at improving the best vector of variables so far, then there is no need for any "simplex steps". Thus, 20 consecutive iterations, say, may make changes to the variables that are minimization steps, and all of the changes may be roughly in the same direction in $\mathbb{R}^n$, which causes $\max\{\|\underline{v}_i - \underline{x}_k\| : i = 1,2,\dots,n+1\}$ to become large. Eventually, however, we expect the sequence of successful iterations to be interrupted by a minimization step that makes $\underline{v}_{n+2}$ no better than $\underline{x}_k$, which means that inequality (5) occurs in the unconstrained case. Then, the next iteration employs a "simplex step". When the k-th iteration tries to take a simplex step, an integer $l \; in \; [1,n]$ is calculated that has the property

$$\|\underline{v}_l - \underline{x}_k\| = \max\{\|\underline{v}_i - \underline{x}_k\|, \qquad i = 1,2,\dots,n+1\} \quad (6)$$

Further, the condition $\|\underline{v}_l - \underline{x}_k\| \leq \beta\Delta_k$ is tested, where $\beta > 1$ is a prescribed constant that has the value $\beta = 2.1$ in the work of Powell (1994). If the test fails, then $\underline{v}_{n+2}$ is chosen in a way that makes it suitable to delete $\underline{v}_l$ from the set $\{\underline{v}_i,\ i = 1,2,\dots,n+2\}$, when generating the vertices of the simplex of the next iteration. Specifically, letting $\underline{w}_l \in \mathbb{R}^n$ be a vector of unit length that is orthogonal to the face of the current simplex that is without $\underline{v}_l$, we let $\underline{v}_{n+2}$ be the point

$$\underline{v}_{n+2} = \underline{x}_k \pm \Delta_k \underline{w}_l \quad (7)$$

where the $\pm$ sign is negative if and only if $\underline{x}_k - \Delta_k \underline{w}_l$ is better than $\underline{x}_k \pm \Delta_k \underline{w}_l$. Otherwise, if $\|\underline{v}_l - \underline{x}_k\| \leq \beta\Delta_k$ is achieved, the algorithm seeks a different integer $l \; in \; [1,n+1]$. Indeed, letting σ_i be the distance from $\underline{v}_i$ to the plane in $\mathbb{R}^n$ that contains the vertices of the current simplex that are different from $\underline{v}_i$, the new l minimizes σ_l subject to $\underline{v}_l \neq \underline{x}_k$. Therefore a very small value of σ_l indicates that the simplex is nearly degenerate. The inequality $\sigma_l \geq \alpha\Delta_k$ is tried, where $\alpha < 1$ is another positive constant, for instance $\alpha = \frac{1}{4}$.

If the inequality fails, then $\underline{v}_{n+2}$ is defined by formula (7) for the new l, where the $\pm$ sign and $\underline{w}_l$ are as before. Further, the new simplex is generated by replacing $\underline{v}_l$ by $\underline{v}_{n+2}$ in the list of vertices, which increases the volume of current simplex by the factor Δ_k/σ_l. If $\sigma_l \geq \alpha\Delta_k$ holds, however, then the positions of the vertices $\underline{v}_i,\ i = 1,2,\dots,n+1$, are assumed to be adequate for the equations (1) that define Φ, and, we say that simplex is "acceptable". Then the iteration tries to generate $\underline{v}_{n+2}$ by a "minimization step" instead of by a "simplex step".

We are now ready to consider the choices between the minimization and simplex step alternatives, the values of Δ_k and ρ_k, $k = 1,2,3,\dots$, and a condition for terminating the calculation. Simple rules are recommended for adjusting Δ_k and for termination. Specifically, Δ_1 is given, and, until

termination, the k-th iteration sets $\Delta_{k+1}=\Delta_k$, where k is still the iteration number. The value of Δ_k at the start of the k.th iteration is provisional, however, in order that a few iterations can reduce Δ_k, although no increase are allowed. A positive parameter Δ_* say, has to be prescribed that satisfies $\Delta_* \leq \Delta_1$ because it is a lower bound on every Δ_k. the changes in Δ_k have to be such that $\Delta_k = \Delta_*$ occurs after a finite number of reductions. The calculation terminates when this situation occurs and Δ_k has already reached the value Δ_*. These rules afford the following useful properties. Every iteration until termination picks a vector of variables $\underline{v}_{n+2}$ that satisfies the inequality

$$\left\|\underline{v}_{n+2}-\underline{x}_k\right\| \geq \Delta_k \quad (8)$$

and Δ_k is not reduced until this condition seems to prevent further improvements to the variables. The user pick a value of Δ_1 that causes substantial adjustments to the variables to be tried at the beginning of the calculation, which can alleviate the damage from any random noise in the function values. Then the bound (8) can be refined gradually by the decreases in Δ_k. Further, when the given functions are smooth, good accuracy can usually be achieved at termination by letting Δ_* be sufficiently small.

Powell in 1994 sets $\rho_k = \Delta_k$ throughout the calculation, but changes to the variables that are much greater than Δ_k are sometimes necessary for efficiency. Indeed, there are unconstrained calculations with quadratic objective functions such that, when Δ_k is reduced, the distance from $\underline{x}_k$ to the optimal vector of variables is of magnitude $M\Delta_k$, where M is the condition number of second derivative matrix $\nabla^2 F$. Therefore it may be helpful to allow ρ_k to be much larger than Δ_k. Moreover, the initial choice $\rho_1 = \Delta_1$ has been already mentioned and, it is reasonable to set ρ_k to the new value of Δ_k when Δ_k is decreased, because ρ_k should become less than old value of Δ_k for the moment, but the condition (8) excludes $\rho_k < \Delta_k$. These remarks suggest the following guidelines for the choice of $\rho_k, k = 1,2,3,\ldots$ we pick $\rho_k = \Delta_k$, which causes ρ_k to be less than its value at the beginning of any iteration that decreases Δ_k, but there are no other changes to ρ_k during an iteration. The value $\rho_{k+1} = \rho_k$ is often set at the end of the k-th iteration, and it always occurs when $\underline{v}_{n+2} - \underline{x}_k$ is a "minimization step" that provides $\underline{x}_{k+1} \neq \underline{x}_k$. If a minimization step fails to improve the best vector of variables so far, however, than the next minimization step is required to be substantially shorter than the present one, except that the bound (8) is preserved. Therefore the value

$$\rho_{k+1} = \max\left[\Delta_k, \frac{1}{2}\left\|\underline{v}_{n+2}-\underline{x}_k\right\|\right] \quad (9)$$

for example, may be suitable.

Each iteration until termination has to choose a "minimization step" or a "simplex step". The method that fixes the choice is specified below using the nomenclature that a minimization step is "long enough" if it satisfies inequality (8) and is "questionable" unless its length is exactly Δ_k and the current simplex is "acceptable". When the iteration does not reduce Δ_k and the choice between the alternatives is determined by the following four rules, which are given in order of priority.

1. A minimization step is preferred if it is long enough and if either *k=1* or the previous iteration improved the best vector of variables so far.

2. A minimization step is preferred if it is long enough and if the previous iteration applied a simplex step.
3. A simplex step is preferred neither rule 1 nor rule 2 apply and if the current simplex is not acceptable.
4. A minimization step is preferred if it is long enough, if the current simplex is acceptable and if the previous iteration employed a minimization step that is questionable. Thus the remaining possibilities are the following two situations.

5. The current simplex is acceptable and the minimization step is not long enough.
6. The current simplex is acceptable, the minimization step is long enough, the previous iteration applied a minimization step that is not questionable, but that iteration did not improve the best vector of variables.

In these cases the time has to come to reduce Δ_k and ρ_k. Therefore termination occurs if Δ_k has attained the value Δ_*. Otherwise, after reducing Δ_k and ρ_k, the required choice is determined by three more rules.

7. The minimization step is preferred if it is long enough for the new Δ_k.
8. The simplex step is chosen if rule 7 fails and if the current simplex is not acceptable for the new Δ_k.
9. In all other cases, Δ_k is still too large, so we introduce a recursion by branching back to the part of the algorithm that either causes termination or reduces Δ_k. Thus, each iteration before termination picks just one vector $\underline{v}_{n+2}$ at which the values of the given functions from $\mathbb{R}^n \; to \; \mathbb{R}$ are calculated.

Another question that requires an answer is the choice of the *n+1* vertices of the simplex for the next iteration from $\{\underline{v}_i : i = 1,2,\dots,n+2\}$. We let $\underline{v}_l$ be the point that is not retained, which agrees with equation (7) when $\underline{v}_{n+2} - \underline{x}_k$ is a "simples step". We propose a new choice of *l* when $\underline{v}_{n+2} - \underline{x}_k$ is a "minimization step", however, because the technique in Powell (1994) assumes ρ_k=Δ_k for every *k*. Let $\underline{x}_{k+1} \in \{\underline{v}_i : i = 1,2,\dots,n+2\}$ be determined before *l* is selected, which is possible because we require the best vector of variables so far. Further, let the real multipliers θ_i, $i = 1,2,\dots,n+2$, satisfy the equation

$$\sum_{i=1}^{n+2} \theta_i (\underline{v}_i - \underline{x}_{k+1}) = 0 \quad (10)$$

where θ_i is zero for the integer i_* that is defined by $\underline{x}_{k+1} = v_{i*}$, but some of the other multipliers are nonzero. It follows from the nondegeneracy of the current simplex that the values of the multipliers are determined uniquely except for a scaling factor. Now, if *i* and *j* are different integers in *[1, n+2]* such that θ_i and θ_j are nonzero, and if S_i and S_j are the new simplices for *l*=*i* and *l*=*j*, respectively, the equation (10) implies the property

$$\left|(Vol\, S_i)/Vol\, S_j\right| = \left|\theta_i/\theta_j\right| \quad (11)$$

therefore it may be suitable to pick *l* by satisfying the condition $|\theta_l| = \max\{|\theta_i| : i = 1,2,\dots,n+2\}$. this method, however, would favour the retention of any points $\underline{v}_i$ that are far from $\underline{x}_{k+1}$, and we

do not want the new simplex to have a large volume because the lengths of some of its sides are much greater than Δ_k. Instead we take the view for the moment that, for every *i in [1, n+2]* such that $\|\underline{v}_i - \underline{x}_{k+1}\|$ exceeds Δ_k, the point $\underline{v}_i$ is replaced by the point on the line segment from $\underline{x}_{k+1}$ to $\underline{v}_i$ that is distance Δ_k from $\underline{x}_{k+1}$, but $\underline{v}_i$ is unchanged for the other values of *i*.
Then we choose *l* by applying the procedure just described to these new points. Specifically, for each integer *i* in *[1, n+2]*, we find that θ_i in equation (10) has to be scaled by max $[1, \|v_i - \underline{x}_{k+1}\|/\Delta_k]$, because of the temporary change to $\underline{v}_i$. Therefore we let *l* be an integer in *[1, n+2]* that has the property

$$|\theta_l| \max\left[\Delta_k, \|v_i - \underline{x}_{k+1}\|\right] \geq |\theta_i| \max[\Delta_k, \|v_i - \underline{x}_{k+1}\|], \quad i = 1,2,\dots,n+2 \quad (12)$$

Thus, if the current simplex has a vertex that is far from the best vector of variables, there is a tendency to exclude it from the simplex of the next iteration. The merit function, Ψ say, of the calculation provides a balance between the value of the objective function and any constraint violations, in order to determine the best vertex of the current simplex. Specifically, Ψ is the same as *F* when there are no constraints, and, for $m \geq 1$, the form

$$\Psi(\underline{x}_k) = F(\underline{x}_k) + \mu\,[\max\{-c_p(x) : p = 1,2,\dots,m\}]_+, \quad x \in \mathbb{R}^n \quad (13)$$

is taken from Powell (1994). Here μ is a parameter that is zero initially and that may be increased automatically as described below. Further, the subscript "+" indicates that the expression in square brackets is replaced by zero if and only if its value is negative. Thus $\Psi(\underline{x}) = F(\underline{x})$ occurs whenever $\underline{x}$ is feasible, and it is helpful to scale the constraint functions so that the values $-c_p(\underline{x})$, $p = 1,2,\dots,m$, have similar magnitudes for typical vectors $\underline{x}$. Expression (3) is extended to $m \geq 0$ by requiring the best vertex $\underline{x}$ to satisfy the conditions

$$\underline{x}_k \in \{v_i : i = 1,2,\dots,n+1\} \text{ and } \Psi(\underline{x}_k) \leq \Psi(\underline{v}_i), \quad i = 1,2,\dots,n+1 \quad (14)$$

after choosing $\underline{x}_k$, both the minimization and the simplex steps are independent of Ψ and μ, but Ψ is usually important to what happens next. Indeed, if the new vector of variables of the k-th iteration, namely v_{n+2} is generated by a minimization step, then usually another minimization step is chosen by the (k+1)-th iteration if and only if the strict inequality $\Psi(\underline{v}_{n+2}) \leq \Psi(\underline{x}_k)$ holds. Further, this inequality should be achieved if all linear approximations discussed before are exact. Therefore, we require the value of μ to provide the property

$$\Upsilon(\underline{v}_{n+2}) < \Upsilon(\underline{x}_k), \quad if \ \underline{v}_{n+2} - \underline{x}_k \text{ is a minimization step} \quad (15)$$

where Υ is the piecewise linear approximation

$$\Upsilon(\underline{x}) = \Phi(\underline{x}) + \mu\,[\max\{-\gamma_p(\underline{x}) : p = 1,2,\dots,m\}, \quad x \in \mathbb{R}^n \quad (16)$$

to the merit function. Now a minimization step either reduces the contribution from the constraints to expression (16), or the contribution is zero and $\Phi(\underline{v}_{n+2}) < \Phi(\underline{x}_k)$ occurs, when we are excluding steps that are zero, because they are abandoned automatically, due to the failure of inequality (8). It follows that condition (15) can be achieved whenever it is required by choosing a sufficiently large value of μ. therefore Powell (1994) propose the following technique for increasing μ.

Whenever a minimization step is calculated that has the property (8), we let $\overline{\mu}$ be the least nonnegative value of μ that provides $\Upsilon(\underline{v}_{n+2}) \leq \Upsilon(\underline{x}_k)$. Further, μ is unchanged in the case $\mu \geq \frac{3}{2}\overline{\mu}$, but, otherwise it is increased to $2\overline{\mu}$. A possible consequence of an increase in μ is that $\underline{x}_k$ is no longer the optimal vertex, and then the calculated minimization step would be incorrect. Therefore $\underline{x}_k$ is changed if necessary to another vertex that satisfies the condition (14). Then the minimization step is recalculated, so μ may have to be increased again, which may cause a further change to the optimal vertex. Fortunately, this procedure does not cycle, because each change to $\underline{x}_k$ causes a strict reduction in $\{-\gamma_p(\underline{x}) : p = 1,2,\dots,m\}$. Another use of Υ is that the $\pm$ sign of expression (7) is negative is and only if $\Upsilon(\underline{x}_k - \Delta_k w_l)$ is less than $\Upsilon(\underline{x}_k + \Delta_k w_l)$.

4.3 Quadratic approximation methods

Now, we let the approximation $\Phi(\underline{x})$, $\underline{x} \in \mathbb{R}^n$, to the objective function $F(\underline{x}), \underline{x} \in \mathbb{R}^n$, be a quadratic polynomial instead of a linear polynomial. Therefore Φ has $\hat{n} = \frac{1}{2}(n+1)(n+2)$, say, independent coefficients, that may be defined by the interpolation conditions

$$\Phi(\underline{v}_i) = F(\underline{v}_i), \quad i = 1,2,\dots,\hat{n} \quad (1)$$

where the vectors $\underline{v}_i$, $i = 1,2,\dots,\hat{n}$ are the points in $\mathbb{R}^n$. These points should have the property that, if expression (1) is written as a system of linear equations, the unknowns being the coefficients, then the matrix of the system is nonsingular. The Lagrange functions of the interpolation problem will be useful later. Therefore we reserve the notation χ_i, $i = 1,2,\dots,\hat{n}$, for the quadratic polynomials from $\mathbb{R}^n$ to $\mathbb{R}$ that satisfy the equations

$$\chi_i(v_j) = \delta_{ij}, \quad 1 \leq i,j \leq \hat{n} \quad (2)$$

where δ_{ij} is the Kronecker delta. It follows that Φ is the function

$$\Phi(\underline{x}) = \sum_{i=1}^{n} F(\underline{v}_i)\chi_i(\underline{x}), \quad \underline{x} \in \mathbb{R}^n \quad (3)$$

the main advantage of the quadratic over linear polynomials is that quadratics include some second derivative information, which allows the development of algorithms that have useful superlinear convergence properties. We are going to consider some of the ideas that have been proposed for

constructing and applying quadratic approximations to F when there are no constraints on the variables.

The algorithm of Winfield, developed in 1973, not only employs the interpolation equations (1) to define Φ, but also it includes some of the earliest work of the objective function calculated so far: $\hat{n}$ of them being obtained before the first iteration. Let these values be $F(\underline{v}_i),\ \ i = 1,2,\dots,\hat{n}$, where $\hat{n} \geq \tilde{n}$, let $\underline{x}_k$ be a best vector of variables which means that it satisfies the conditions

$$\underline{x}_k \in \{\underline{v}_i : i = 1,2,\dots,\hat{n}\} \text{ and } F(\underline{x}_k) \leq F(\underline{v}_i),\ \ i = 1,2,\dots,\hat{n} \quad (4)$$

and let the current data be ordered so that the sequence of distances $\|v_i - \underline{x}_k\|,\ i = 1,2,\dots,\hat{n}$, increases monotonically. Then the k-th iteration generates the quadratic polynomial Φ by trying to interpolate the function values of only the first $\hat{n}$ terms of the sequence, in accordance with the notation (1). Further, the iteration calculates the vector $\underline{x} \in \mathbb{R}^n$ that minimizes $\Phi(\underline{x})$ subject to the bound $\|\underline{x} - \underline{x}_k\| \leq \rho_k$, where the trust region radius is chosen automatically and satisfies $\rho_k \leq 0.99\|\underline{v}_{\hat{n}} - \underline{x}_k\|$, in order that the value of F at the new point will be included in the interpolation conditions of the (k+1)-th iteration. One reason for mentioning the algorithm is that it acts in an enterprising way when the system (1) is degenerate. Specifically, the degeneracy is ignored, it is assumed that the calculation of Φ is sufficiently robust to provide a quadratic function that allows the trust region subproblem to be solved and the resultant $\underline{x}$ receives no special treatment. Thus some unpredictable changes to the variables occur that may remove the degeneracy after a few iterations. Indeed, Winfield (1973) states that "this natural cure of ill-conditioning is more efficient than restarting the algorithm by evaluating $F(\underline{x})$ at the points of a grid". The other methods that we study, however, ensure that each Φ is well defined.

The Lagrange functions that have been mentioned provide a convenient way of avoiding singularity in the equations (1). The technique suggests itself if one tries to modify the algorithm of Powell for unconstrained optimization described in the previous paragraph, so that the linear polynomial Φ is replaced by the quadratic one that is defined by the equations (1). We retain from the previous section the parameters Δ_k and $\rho_k,\ k = 1,2,3,\dots$, and the rules that give their values. Moreover, in the quadratic case, the points $\underline{v}_i,\ i = 1,2,\dots,\tilde{n},$ for the first iteration can be the vertices and the mid-points of the edges of a nondegenerate simplex in $\mathbb{R}^n$, where the lengths of the edges are still of magnitude Δ_1. Otherwise, for $k \geq 2$, these points are chosen by the previous iteration, and $\underline{x}_k$ satisfies the conditions

$$\underline{x}_k \in \{\underline{v}_i : i = 1,2,\dots,\hat{n}\} \text{ and } F(\underline{x}_k) \leq F(\underline{v}_i),\ \ i = 1,2,\dots,\tilde{n} \quad (5)$$

further, $\underline{v}_{n+1} - \underline{x}_k$ is still a "minimization step" if $\underline{v}_{n+1}$ is the vector $\underline{x} \in \mathbb{R}^n$ that minimizes $\Phi(\underline{x})$ subject to $\|\underline{x} - \underline{x}_k\| \leq \rho_k$ which is the trust region subproblem of the previous paragraph. On the other hand, a "simplex step" is usually required if the previous iteration generated a minimization step that failed to reduce the least calculated value of F. In this case, we let l be and integer in $[1, \tilde{n}]$ that maximizes $\|\underline{v}_l - \underline{x}_k\|$. If this distance is unacceptably large, then we have to pick a point $\underline{v}_{n+1}$ that will replace $\underline{v}_l$ in the system (1) on the next iteration. Therefore we require a formula that is suitable when Φ is a quadratic polynomial.

Now we are going to maximize the volume of the simplex of the next iteration subject to $\|\underline{v}_{n+2} - \underline{x}_k\| \leq \Delta_k$. Further, the volume of the simplex is a constant multiple of the modulus of the determinant of the matrix of the system (1) of the previous paragraph, when the usual basis of the space of linear polynomials is employed. Therefore an analogous choice of the "simplex step" when Φ is quadratic would maximize the modulus of the determinant of the $\hat{n}x\tilde{n}$ system (1), after $\underline{v}_l$ is replaced by $\underline{v}_{\hat{n}+1}$, where $\underline{v}_{\hat{n}+1}$ has to satisfy $\|\underline{v}_{\hat{n}+1} - \underline{x}_k\| \leq \Delta_k$.

We write $\underline{x} = \underline{v}_{\hat{n}+1}$ for the moment, we regard the new quadratic polynomial in $\underline{x}$. Further, the determinant must vanish if $\underline{x}$ is any point of the set $\{\underline{v}_i : i = 1,2,\dots,\hat{n}\}$ that is different from $\underline{v}_l$. Thus, an elementary normalization provides the identity

$$\frac{New\ determinant}{Old\ determinant} = \chi_l(\underline{x}), \qquad \underline{x} \in \mathbb{R}^n \quad (6)$$

therefore we define $\underline{v}_{\hat{n}+1} - \underline{x}_k$ to be a "simples step" for the chosen integer $l \in [1, \hat{n}]$ if and only if $\underline{v}_{\hat{n}+1}$ is a vector of variables $\underline{x}$ that maximizes $|\chi_l(\underline{x})|$ subject to $\|\underline{x} - \underline{x}_k\| \leq \Delta_k$. This definition has the advantage of being independent of the choice of basis of the space of quadratic polynomials. Further, the simplex step can be calculated by solving two trust region subproblems of the type that has been encountered already. Indeed, if two vectors of variable are generated by minimizing the quadratic functions χ_l and $-\chi_l$ subject to the trust region bound, then the required $\underline{v}_{\hat{n}+1}$ is the vector that gives the larger value of $|\chi_l|$.

When the k-th iteration tries to take a "simplex step", the algorithm may find that all the points $\underline{v}_i,\ i = 1,2,\dots,\hat{n}$, are sufficiently close to $\underline{x}_k$, which corresponds to the condition $\|\underline{v}_l - \underline{x}_k\| \leq \beta\Delta_k$. Then a test for neardegeneracy of the system (1) is required.

Therefore, we continue to let $\alpha < 1$ be a positive constant, for instance $\alpha = 1/4$ and we seek an integer l in $[1, \hat{n}]$ such that the replacement of $\underline{v}_l$ by $\underline{v}_{\hat{n}+1}$ increases the modulus of the determinant of the system (1) by a factor of more than $1/\alpha$, where $\underline{v}_{\hat{n}+1}$ is defined at the end of the previous paragraph, because this choice maximizes the modulus of the new determinant. Specifically, the test for near-degeneracy in the quadratic case is as follows. The integer l runs through the set $\{1,2,\dots,\hat{n}\}$, but similar tests on recent iterations may make it advantageous not to begin with *l=1*. For each *l,* the maximum value of $|\chi_l(\underline{x})|$, $\|\underline{x} - \underline{x}_k\| \leq \Delta_k$, is calculated. If $|\chi_l(\underline{x})| > 1/\alpha$ occurs, the task of searching for a suitable l is complete, because the replacement of $\underline{v}_l$ by the vector $\underline{v}_{\hat{n}+1}$ that has been mentioned provides a substantial improvement to the positions of the interpolation points. Then $\underline{v}_{\hat{n}+1} - \underline{x}_k$ is a "simplex step" and the functions value $F(\underline{v}_{\hat{n}+1})$ is required for the system (1) of the next iteration. Otherwise, if no current interpolation points are "acceptable" and the iteration may generate $\underline{v}_{\hat{n}+1}$ by a "minimization step". We also retain the rule that the minimization step is abandoned if it fails to satisfy $\|\underline{v}_{n+1} - \underline{x}_k\| \leq \Delta_k$, which is important to the criteria for reducing Δ_k and for termination.

We let each choice between a "minimization" and a "simplex step" in the quadratic case be the same as in the previous paragraph. A modification is needed, however, to the technique that selects the interpolation points for the *(k+1)-th* iteration, after $F(\underline{v}_{\hat{n}+1})$ has been calculated and $\underline{v}_{n+1} - \underline{x}_k$ is a minimization step. These points are all but one of the vectors $\underline{v}_i : i = 1,2,\dots,\hat{n}+1$ and again we let $\underline{v}_l$ denote the point that is rejected. Here it is important to note that, in contrast to the previous

paragraph, $\underline{v}_{n+1}$ is now independent of *l,* because it is generated by the minimization step before *l* is chosen. In order to retain a best vector of variable so far, we let *i* be an integer in [1, n̂ + 1] such that $F(\underline{v}_i)$ is the least of the function values $F(\underline{v}_i)$, $i = 1,2,\dots,\hat{n}+1$. Then the value $l = i_*$ is prohibited because $\underline{x}_{k+1}$ is going to be the point $\underline{v}_i$. It would be straightforward to pick the l that maximizes the modulus of the determinant of the system (1) on the next iteration if we wished to do so. Indeed, if $l \in [1, \hat{n}]$, then it follows from the identity (6) that the determinant of the new system is the determinant of the present one multiplied by $\chi_l(\underline{v}_{n+1})$. Therefore, after defining $\theta_{n+1} = 1$ and $\theta_i = \chi_i(\underline{v}_{n+1})$, $i = 1,2,\dots,\hat{n}$ and then replacing θ_{i*} by zero, we would let l satisfy the equation $|\theta_l| = max\{|\theta_i|: i = 1,2,\dots,\hat{n}+1\}$. Again, however, we prefer to take the distances $\|\underline{v}_i - \underline{x}_{k+1}\|$, $i = 1,2,\dots,\hat{n}$ into account.

Specifically, if *I* is any integer in [1, $\hat{n}$] such that $\|\underline{v}_i - \underline{x}_{k+1}\| > \Delta_k$ occurs, we make a notional shift of $\underline{v}_i$ to $\hat{\underline{v}}_i$, say, which is a point on the line segment from $\underline{x}_{k+1}$ to $\underline{v}_i$ that is within the distance Δ_k of $\underline{x}_{k+1}$. Further, we let χ_i be the quadratic polynomial that satisfies the Lagrange conditions $\hat{\chi}_i(\hat{\underline{v}}_i) = 1$ and $\hat{\chi}_i(\hat{\underline{v}}_j) = 0$, for very integer *j* in [1, $\hat{n}$] that is different from *i*. Hence, $\hat{\chi}_i$ is the function

$$\hat{\chi}_i(\underline{x}) = \frac{\chi_i(\underline{x})}{\chi_i(\hat{\underline{v}}_i)}, \quad \underline{x} \in \mathbb{R}^n \qquad (7)$$

Now, because of the inequality $\|\underline{v}_i - \underline{x}_{k+1}\| > \Delta_k$ we assume that the temporary replacement of $\underline{v}_i$ by $\hat{\underline{v}}_i$ would make the determinant of the system (1) more relevant to our consideration of possible near-degeneracy. Therefore we change the value of θ_i, given the previous paragraph, to the number $\hat{\chi}_i(\underline{v}_{n+1}) = \frac{\chi_i(v_{\hat{n}+1})}{\chi_i(\hat{\underline{v}}_i)}$, but there is still some freedom in the position of $\hat{\underline{v}}_i$. We have to avoid positions that are too close to other interpolation points, and it is easy to make $|\chi_i(\hat{v}_i)|$ as large as possible, because χ_i is a quadratic function of one variable on the line segment from $\underline{x}_{k+1}$ to $\underline{v}_i$. On the other hand, it would be unsuitable to allow $|\chi_i(\hat{v}_i)|$ to exceed one, because then $\|\underline{v}_i - \underline{x}_{k+1}\| > \Delta_k$ would assist the retention of $\underline{v}_i$ in the set of interpolation points. These remarks lead to the formula

$$\theta_i = \frac{\chi_i(v_{\hat{n}+1})}{\min[1, max\{|\chi_i(\underline{x}_{k+1}+\alpha[\underline{v}_i-\underline{x}_{k+1}])|: 0\le\alpha\le\bar{\alpha}\}]} \qquad (8)$$

where $\bar{\alpha} = \frac{\Delta_k}{\|\underline{v}_i-\underline{x}_{k+1}\|}$. Further, this choice is just $\theta_i = \chi_i(\underline{v}_{n+1})$ as before, when *i* is an integer in [1, $\hat{n}$] that satisfies $i \neq i_*$ and $\|\underline{v}_i - \underline{x}_{k+1}\| > \Delta_k$. Moreover, $\theta_{n+1} = 1$ is the most reasonable scaling factor to apply to the determinant when there is no change to the interpolation points. Therefore we recommend these values of θ_i and after replacing θ_{i*} by zero, we let *l* be an integer in [1, $\hat{n}+1$] that maximizes $|\theta_l|$.

We have found that, due to the identity (6), Lagrange functions are highly useful for selecting points $\underline{v}_i$, $i = 1,2,\dots,\hat{n}$, such that the quadratic polynomial Φ is well defined by the equations (1).

Another description of the use of Lagrange functions is given by Conn, Scheinberg and Toint. This work also addressed the idea of employing "Newton fundamental polynomials" instead of Lagrange functions, where these polynomials in the quadratic case are a constant Lagrange polynomial, *n*

linear Lagrange polynomials, and $\frac{1}{2}n(n+1)$ quadratic Lagrange polynomials that are derived from one, *n+1* and all of the interpolation points, $\underline{v}_i,\ i = 1,2,\dots,\hat{n}$ respectively. They provide a different basis of the $\hat{n} = \frac{1}{2}(n+1)(n+2)$ dimensional space of quadratic polynomials, which is helpful when fewer than $\hat{n}$ values of *F* are available to determine Φ. An outline of a trust region algorithm for unconstrained minimization without derivatives is given too. A major departure from the work of this section is that, the k-th iteration takes a "minimization step" that reduces *F* by an amount that compares favourably with the corresponding reduction in Φ, then Δ_{k+1} is allowed to be larger than Δ_k. Nevertheless, this trust region radius is reduced only when the positions of the interpolation points satisfy acceptability conditions that are similar to the ones specified in the complete paragraph following the equation (6). Therefore, in comparison with the technique of Jones that employs both Δ_k and ρ_k, several extra function values may occur if the larger trust region radius is successful for only a small number of iterations. An earlier paper by Conn, Scheinberg and Toint in 1997 also considers Newton fundamental polynomials and presents an outline of a similar trust region algorithm. Further, the convergence of the algorithm is studied under certain assumptions, including the uniform boundedness of the second derivative matrices $\nabla^2\Phi$. It is proved that, if the number of iterations is infinite, then the property $\lim inf_{k\to\infty} \left\|\underline{\nabla} F(\underline{x}_k)\right\| = 0$ is achieved.

The last topic of this section is the algorithm of Elster and Neumaier (1995) which is designed for optimization calculations. The algorithm is remarkable, because it combines quadratic approximations to *F* and trust regions with some of the properties of discrete grids that are considered before. Thus, termination is achieved, even if the values of the objective function are distorted by noise. There is a close analogy with the two trust region idea of Evan Jones, because it is appropriate to let ρ_k be the trust region radius and Δ_k be the grid size. The algorithm retains all the calculated values of *F.* Then, each quadratic approximation Φ is formed by least squares fitting to some of them, using a technique that is interesting, because it begins by generating a Hessian approximation *G,* and then it restricts attention to only about *2n+2* function values, in order to fit the parameters $a \in \mathbb{R},\ g \in \mathbb{R}^n$ and $k \in \mathbb{R}$ of the approximation

$$\Phi(\underline{x}) = a + \underline{g}^T(\underline{x} - \underline{x}_k) + \frac{k}{2}(\underline{x} - \underline{x}_k)^T G(\underline{x} - \underline{x}_k), \quad \underline{x} \in \mathbb{R}^n \qquad (9)$$

where $\underline{x}_k$ is still the vector of variables that provides the least value of *F* so far. The algorithm requires Φ, ρ_k and $\underline{x}_k$ for the calculation of a "minimization step". Then, the new vector of variables at the end of this step is shifted to the nearest grid point, $\underline{x}_+$ say. The use of grids ensures that, after only a finite number of iterations, the function value $F(\underline{x}_+)$ will have been found by an earlier iteration. When this happens, or when three consecutive minimization steps fail to achieve $F(\underline{x}_+) < F(\underline{x}_k)$, a procedure is involved that is similar to a "simplex step". The procedure derives and may apply a linear polynomial approximation to *F,* using values of the objective function at grid points that have to be neighbors of $\underline{x}_k$. Thus, the decision is taken whether or not to have to reduce Δ_k before resuming the minimization steps. Alternatively, termination occurs if a reduction in Δ_k is required but Δ_k is already at a prescribed lower bound. Several numerical experiments in Elster and Neumaier show that this algorithm compares favourably with the method of Nelder and Mead (1965) and with a finite difference implementation of a quasi-Newton algorithm.

5 Numerical simulation and modeling of monocrystalline selective emitter solar cells

Selective emitter (SE) solar cells in contrast with homogeneous emitter cells are characterized by having different doping profiles under the metal-contacted highly-doped region and the passivated one between contacts (lowly-doped region). *[10]*
The SE solar cell is obtained by a light diffusion (LDOP) followed by a heavy phosphorus diffusion (HDOP) in the contacted regions. The advantages of the SE cell consists in reduced recombination effects in the passivated LDOP surface region and in an enhanced spectral response in the blue region. A trade-off exists between the advantages listed above and the increase of the emitter resistance that affects the SE cell due to the lower doping concentration of LDOP. Two-dimensional (2-D) numerical simulations can be used to gain insight on the loss mechanism in SE cells and to aid the design of the devices. *[60]*
Here, we compare SE cells with a baseline homogeneous emitter cell. All the considered cells feature a p-type base region and a wafer thickness D_{sub}=180 μm. Numerical simulations are carried out by Synopsys-Sentaurus.
A n^+ p c-Si HE solar cell and a SE device have been simulated. The rear surface is fully contacted by the base electrode. We consider a 100 μm wide front metal electrode and a total lateral width of the heavy diffused region W_{se}=130 μm under the metal electrode. The top surface is coated with a 70 nm thick silicon nitride antireflective coating layer.

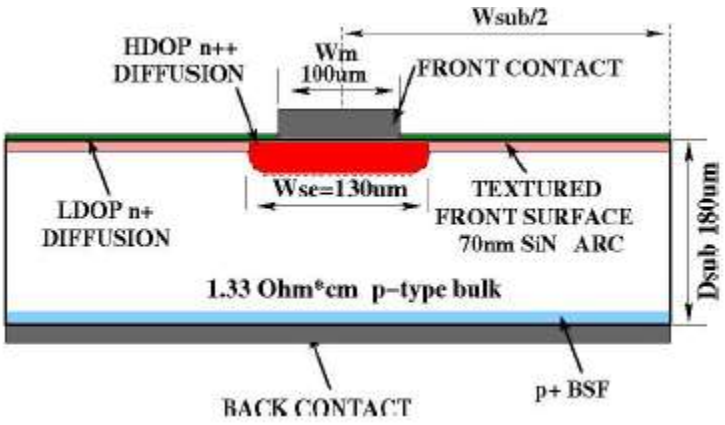


Figure 5.1 – Cross section of a selective emitter solar cell with wafer thickness D_{sub}=180 μm. The front metal contact width W_m is set to 100 μm and the lateral width of the diffusion under the metal grid W_{se} is set to 130 μm.

In the section above, representing a SE solar cell, it is possible to recognize the front metal contact width (W_m) set to 100 μm and the lateral width of the diffusion under the metal grid (W_{se}) set to 130 μm.
The simulated doping profiles are described by analytical functions. Both HE and SE solar cells feature a **back surface** (BSF) modeled by an error function boron profile with a peak doping set to 10^{20} cm^{-3} and 0.67 μm junction depth.

5.1 Simulation setup

The bulk doping concentration and base minority carrier lifetime are set to 10^{16} cm^{-3} (resistivity 1.33 Ωcm assuming a hole mobility of 470.5 cm^2/Vsec) and 200 μsec respectively, for both HE and SE. The emitter doping profiles are qualified by their sheet resistance, evaluated by using Arora model with **Sentaurus** default parameters. Surface recombination velocity is set to 10^5 cm/s for metal front and back contacts. The surface recombination velocity S at the front passivated interface is assumed to be dependent on the surface doping concentration N_{surf} of the emitter according to the following model *[10]*

$$S = S_0 \left[1 + S_{ref} \left(\frac{N_{surf}}{N_{ref}}\right)\right] \quad (1)$$

Where S_0=20 cm/s, N_{ref}=10^{16} cm^{-3} and S_{ref}=10^{-3} cm/s. So, according to equation (1), if N_{surf}=10^{20} cm^{-3}, then S=220 cm/s.

Electrical simulation takes into account Auger recombination, doping dependent Shockley-Read-Hall (SRH) bulk and surface recombination, radiative recombination, bandgap narrowing (del Alamo model), doping dependence of carrier mobility (Philips Unified mobility model) and mobility degradation at high fields (Canali Model). The standard Sentaurus model for intrinsic carrier concentration is adopted. Fermi statistics is enabled since heavy doping concentrations are considered.

Optical generation rate profiles are calculated assuming direct illumination with a standard AM1.5G spectrum and accounting for light trapping by a textured front surface. The light at the silicon top surface is assumed to be Lambertian distributed. The surrounding medium is air.

The multiple bounces of light inside the device are described analytically in terms of a geometric progression. External reflectivity, internal top and bottom reflectivity coefficients are calculated by using the Transfer Matrix Method (TMM) and they are wavelength dependent. Since the optical treatment is 1D, all coefficients are obtained by calculating their cosine-weighted average over the angle with respect to the normal direction. The shadowing under front grid fingers is assumed ideal.

The calculated parameters include **short circuit current density (J_{sc})**, **open circuit voltage (V_{oc})**, **fill factor (FF)** and **efficiency (η)**. The simulated electrical output power of the cell and FF accounts for power loss due to series resistance R according to the following expressions:

$$R_c = \frac{\rho_c}{W_m L} \quad (2)$$

$$R_m = \rho_m \frac{L}{3H_m W_m} \quad (3)$$

$$R = R_c + R_m \quad (4)$$

$$P_M = P_{M_0} - I_{MP}^2 R \quad (5)$$

where P_M is the effective maximum output power of the cell, I_{MP} is the current under maximum power condition, R_C is the metal-semiconductor-contact resistance and R_m the contact finger

resistance. The length and the thickness of the grid finger are set to L=3 cm and to H_m=12 μm respectively. The sheet resistivity of the metal and the contact resistivity are set to $\rho_M=6x10^{-6}$ Ωcm and $\rho_c=10^{-3}$ Ωcm, respectively.

5.2 Homogeneous emitter solar cell simulation

The HE solar cell features a 39 Ω/square emitter error function profile with peak doping concentration equal to $3.7x10^{20}$ cm^{-3} and a junction depth of 0.39 μm. *[10]*

For a given metal-grid width, increasing W_{sub} results in larger J_{sc} and to larger emitter resistance that degrades the FF, leading to an **optimum front contact pitch of approximately 2100 μm**.

5.2.1 Selective emitter: dependence of efficiency on LDOP profile

Several LDOP profiles are simulated at given HDOP (34 Ω/square with peak doping of $3.7x10^{20}$ cm^{-3} and junction depth set to 0.82 μm). for a fixed junction depth (0.27 μm) the peak doping of LDOP is changed from $5.0x10^{19}$ cm^{-3} to $3.0x10^{20}$ cm^{-3} corresponding to sheet resistances in the range 47 Ω/square – 215 Ω/square. The 34 Ω/square HDOP and the 109 Ω/square LDOP doping profiles of the HE solar cell are reported below

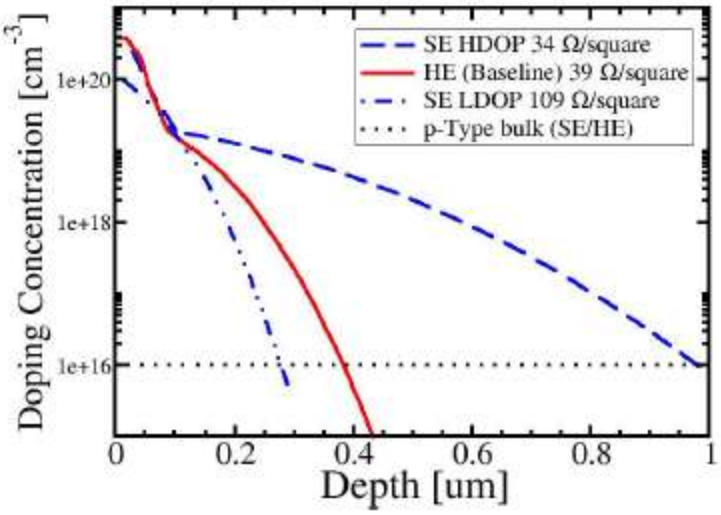


Figure 5.2 – Active doping profiles fot the HE (39 Ω/square) and for one SE solar cell (39 Ω/square HDOP-highly doped region – and 109 Ω/square LDOP-lightly doped region profiles).

In the picture, it is possible to look at the active doping profile for the HE (39 Ω/square) and for one SE solar cell (34 Ω/square HDOP – highly doped region and 109 Ω/square LDOP lightly doped region profiles). While the fill factor versus front contact pitch is reported in the following picture for the most significant LDOP profiles.

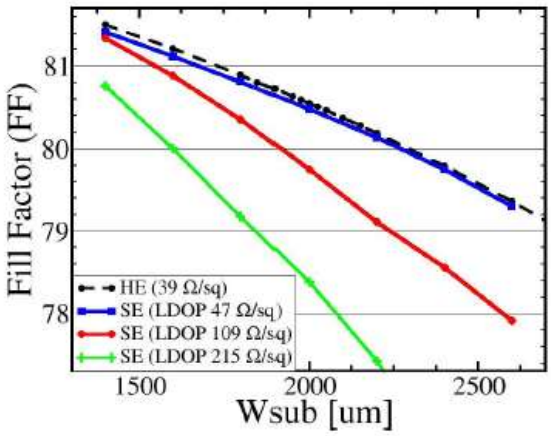


Figure 5.3 – Fill Factor (FF) of SE (HDOP 34 Ω/square) versus W_{sub} for different LDOP profiles and for the baseline (HE 39 Ω/square). FF for the baseline and the SE (47 Ω/square) are similar.

For a given front contact pitch value, the FF of the HE cells is larger than or equal to those of any considered SE cell thanks to its lower spreading resistance.

Peak Doping	Sheet Resistance
[cm^{-3}]	[Ohm/sq]
$3.00x10^{20}$	47
$2.00x10^{20}$	68
$1.50x10^{20}$	87
$1.15x10^{20}$	109
$1.00x10^{20}$	123
$9.00x10^{19}$	136
$5.00x10^{19}$	215

In the table above, the peak doping and sheet resistance values are reported for the considered lowly-doped region profiles for the SE solar cell. Simulated LDOP profiles are described by analytical error functions. For all LDOP profiles the junction depth is set to 0.27 μm.
So, simulation results highlight the dependence of FF on emitter resistance which is an increasing function of the contact pitch of the sheet resistance of the LDOP emitter profile. Efficiency versus front contact pitch W_{sub} obtained for different values of the LDOP sheet resistance is reported below together with results for the baseline cell:

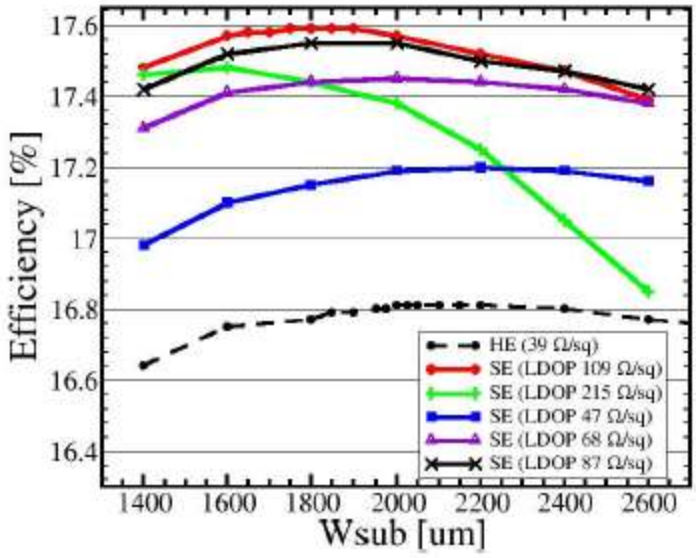


Figure 5.4 – Efficiency of SE (34 Ω/square) versus W_{sub} for different LDOP profiles and for the baseline (HE 39 Ω/square) solar cell.

For each considered W_{sub} value, the efficiency of the SE cell is larger than that of the baseline in spite of a lower fill-factor values because both J_{sc} and V_{oc} are larger. The maximum efficiency is provided by the 109 Ω/square profile (peak doping $1.15x10^{20}$ cm^{-3}) with front contact pitch equal to 1800 μm.
In the following table, it is possible to notice the difference in the main parameters between HE and SE cells:

	Wsub	Jsc	Voc	FF	Eff
	[μm]	[mA/cm^2]	[V]		[%]
HE 39 Ω/sq	2100	34.32	0.610	80.37	16.81
SE 34/109 Ω/sq	1800	35.17	0.623	80.35	17.59

Table 5.1 – Short circuit current (J_{sc}), Open Circuit Voltage (V_{oc}), Fill Factor (FF) and Efficiency for baseline (W_{sub}=2100 μm, HE 39 Ω/square) and SE (W_{sub}=1800 μm, HDOP 34 Ω/square, LDOP 109 Ω/square)

5.2.2 Selective emitter: dependence of efficiency on HDOP profile

The sheet resistance of the HDOP profile has been varied from 34 to 63 Ω/square, keeping the LDOP profile constant (109 Ω/square profile form previous analysis) and W_{sub}=1800 μm. The results of simulations show that there is only a slight dependence of efficiency on the HDOP parameters (peak doping, junction depth). The maximum value for conversion efficiency (17.61%) is obtained for the 46 Ω/square HDOP profile. *[10]*

5.3 Analysis of loss mechanisms

The SE (HDOP 46 Ω/square, LDOP 109 Ω/square) and the HE (39 Ω/square) solar cells are compared in terms of internal quantum efficiency (**IQE**).

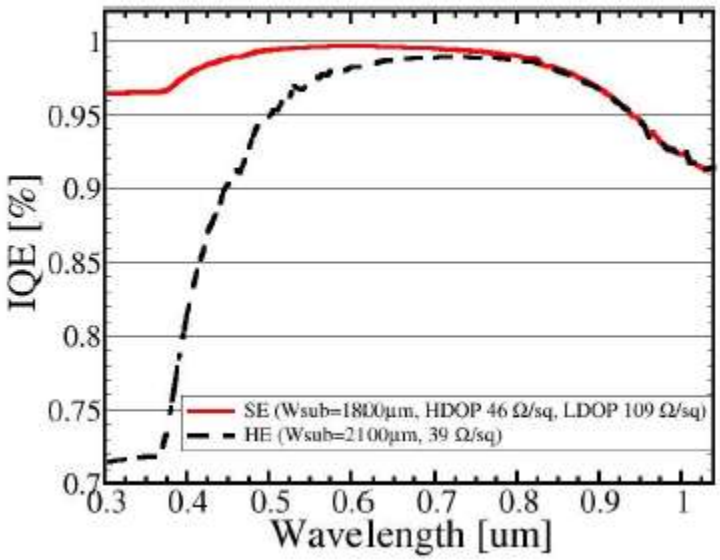


Figure 5.5 – Comparison of Internal Quantum Efficiency (IQE) between selective emitter SE (W_{sub} = 1800 µm, HDOP 46 Ω/square – ldop 109 Ω/square) and baseline HE (Wsub = 2100 µm, 39 Ω/square).

The SE cell features a better spectral response in the blue region resulting in higher short circuit current (35.17 mA/cm^2 against 34.32 mA/cm^2) even if the front contact pitch value is larger in the HE case, leading to reduced shadowing effect. The benefit in terms of spectral response is due to lower doping concentrations in the emitter leading to reduced Auger recombination and to a shallow junction in the passivated emitter region which improves separation for electron-hole pairs generated by photons at lower wavelengths (350-550 nm) that are absorbed close to the front surface due to a large absorption coefficient in c-Si. Furthermore, lower doping concentrations lead to reduced surface recombination rates at the front passivated interfaces. Simulations hightlight the influence of Auger recombination on efficiency as a major loss mechanism of a homogeneous solar cell with heavy and deep emitter diffusions. By selectively disabling Auger recombination effect, the SE and HE solar cells increase their efficiencies by an absolute 0.60% and 1.24 % respectively; this shows that reduced Auger recombination is the main reason for SE higher efficiency compared to the HE. The following figure compares the Auger recombination rates in the region close to the front surface of the device for SE and HE. The peak of the Auger recombination rate is 5.18x10^{22} cm^{-3}s^{-1} in the case of the SE and 2.15x10^{23} cm^{-3}s^{-1} for the baseline, confirming a reduced impact of Auger recombination in the SE cell.

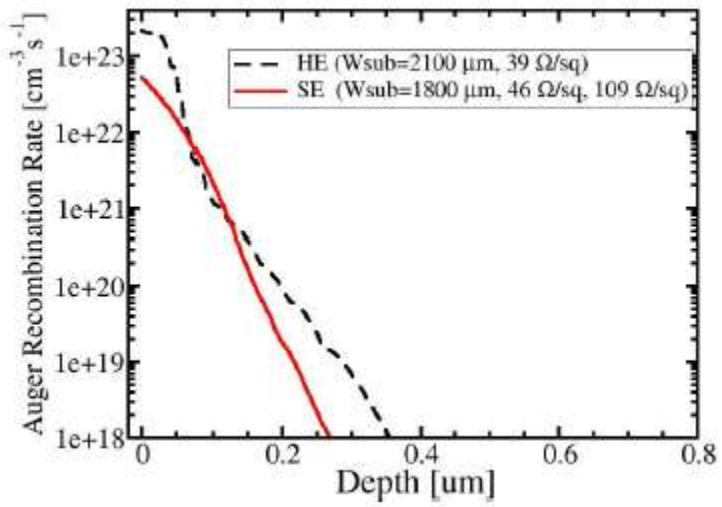


Figure 5.6 – Auger recombination rates for optimized SE (W_{sub}=1800 μm) and HE (W_{sub}=2100 μm) in the lowly-doped emitter region close to the top surface (Depth=0 μm).

5.4 Conclusions

It has been analyzed the efficiency, fill factor, short circuit current and open circuit voltage for a SE solar cell featuring a wafer thickness of 180 μm as a function of front contact pitch and emitter doping profiles. According to our simulations, a selective emitter solar cell may provide an efficiency up to 0.8% higher compared to an HE device. Advantages of double-diffused emitters arise from the enhancement of the collection efficiency, especially in the blue region of the spectrum, and from reduced Auger recombination due to a lighter emitter doping concentration.

6 Numerical simulation and modeling of rear point contact solar cells

The conversion efficiency of a solar cell is significantly limited by the recombination losses occurring at the rear contact. Conventional solar cells, which are uniformly contacted over the whole back silicon surface, are affected by significant recombination losses at the metal-semiconductor interface. High-efficiency silicon solar cells like PERC (Passivated Emitter and Rear Cell) and PERL (Passivated Emitter Rear Locally diffused) adopt local point contacts at the back surface, allowing the passivation of the uncontacted back silicon surface region to reduce the surface recombination rate and to increase the internal bottom reflectivity, leading to larger photocurrent densities.
The optimum design of rear point contact solar cells requires a trade-off between reduced recombination losses, light trapping properties and the series parasitic resistance. It is worth noting that the increase of parasitic resistance associated to 3-D conduction paths occurring when the extension of the contacted region is significantly smaller than the cell area. We will assume that the holes feature a circular shape. The metallization fraction *f* is expressed by the following expression *[11]*

$$f = \pi \left(\frac{s}{2p}\right)^2 \quad (1)$$

where *s* is the diameter of the contact holes and *p* represents the hole pitch. A sketch of the analyzed system is reported in the following figures:

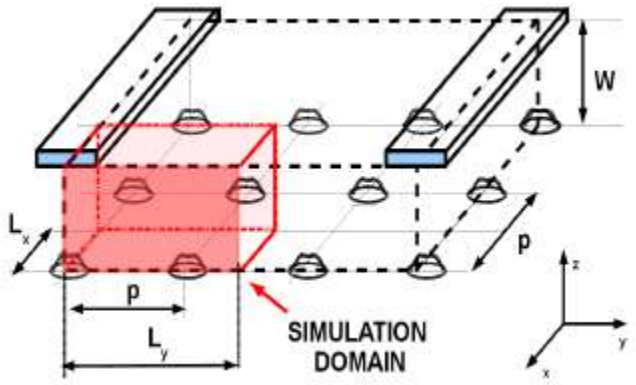


Figure 6.1 - 3-D Sketch of a rear point contact solar cell. W_{sub} denotes the front contact pitch and W_{met} the front contact finger width. The holes are equally distributedwith period *p*. The simulation domain is highlighted in red.

L_x and L_y are the width and the length of the simulation domain which are equal to half hole pitch and to half front contact pitch, respectively. The height of the simulation domain is equal to the wafer thickness *w*.

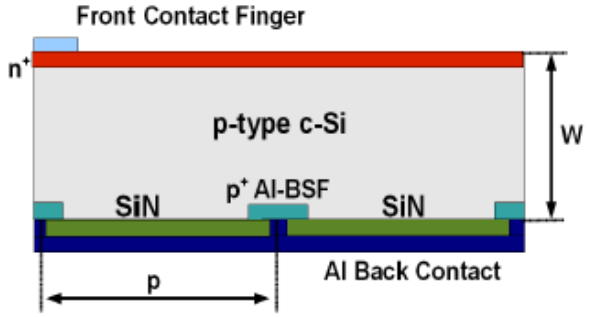


Figure 6.2 – A 2-D cross section of the rear point contact cell.

6.1 Simulation setup

Rear point contact solar cells featuring an aluminum **local back surface field** (**LBSF**) and **without LBSF** (**NOBSF)** have been simulated.

It has been investigated Czochralski (Cz) monocrystalline silicon (c-Si) devices with different values of substrate resistivity: $\rho_{sub} = 0.5\ \Omega$cm, and 1.0 Ωcm. For $\rho_{sub} = 1.0\ \Omega$cm and 10 Ωcm the Al/p-Si interface is rectifying, hence only the LBSF configuration is considered in this analysis. On the other hand, for $\rho_{sub} = 0.5\ \Omega$cm the rectifying action of the Al/p-Si system is negligible, therefore both LBSF and NOBSF configurations are investigated. The emitter of the simulated solar cells is homogeneous (75 Ω/square) with a diffusion depth of 0.4 μm. For devices featuring a local BSF, the aluminum diffusion is described by a Gaussian doping concentration profile with a junction depth of 5 μm and a peak doping of $2.5\text{x}10^{19}$ cm^{-3}.

6.2 Physical models

The physical models adopted for the numerical simulations include the high-field and doping dependent mobility model (also known as Philips Unified Models) as well the Schenk band-gap narrowing. Fermi statistics is adopted in order to correctly deal with high doping concentration regions in the emitter and in the BSF region. Minority carrier lifetime for Cz c-Si material and surface recombination velocities at passivated interfaces have been modeled in order to take into account the dependence of the cell efficiency on the material quality and process conditions. For bulk recombinations are used the Scharfetter conditions, well-calibrated in order to take into account a proper value of the minority carrier lifetime in the boron-doped base region. The surface recombination velocities at passivated interfaces, which play a key role in the device analysis, are accounted by using a doping dependent surface Schockley-Read-Hall model. *[61]*

In order to calculate more realistic values of the fill factor and of the efficiency, the parasitic contact and finger resistances have been included.

6.2.1 Optical simulation

Optical generation rate profiles are calculated on the basis of a simulation of plane-waves propagation in silicon assuming direct illumination with a standard AM1.5G spectrum (1000 W/m^2). In the following table the parameters of the simulated solar cells are summarized:

Parameter	Description	Value	
W_{sub}	Front contact pitch	2	mm
W	Wafer thickness	180	µm
W_M	Front finger width	100	µm
L_M	Front finger length	3	cm
H_M	Front finger thickness	20	µm
RSh_{em}	Emitter sheet resistance	75	Ω/sq
JD_e	n+ emitter diff. depth	0.4	µm
JD_b	Al-BSF diff. depth	5	µm
$N_{b,pk}$	Al-BSF peak doping concentration	$2.5 \cdot 10^{19}$	cm^{-3}
ρ_m	Front Metal resistivity	$6x10^{-6}$	Ω·cm

Table 6.1 – Parameters of the simulated solar cells.

The simulated device feature textured front surfaces. It is to be considered the shadowed caused by front fingers ideal. Since it has been performed a 1-D optical simulation, the internal bottom reflection coefficient R_{bi} is assumed uniform at the rear interface and it is weighted by the metallization fraction according to ***[11]***

$$R_{bi} = \left(R_{bi,p} - 0.25f\right) \quad (2)$$

where $R_{bi,p}$ is the internal bottom reflection coefficient of the silicon-dielectric interface, set to 0.90. The internal bottom reflection coefficient of the Al/p-Si interface is assumed equal to 0.65.

6.3 Results

6.3.1 Dependence of the output parameters on the metallization fraction

The dependence of the output cell parameters on the metallization fraction f has been calculated for the case p=500 µm with hole diameter ranging from 25 µm up to 400 µm. the short-circuit current density (J_{sc}), the open circuit voltage (V_{oc}), the fill factor (FF) and the efficiency η are reported in the following figures:

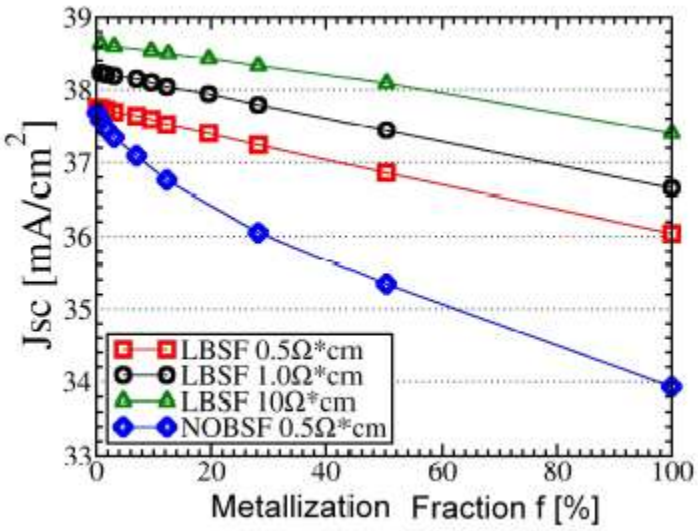


Figure 6.3 - Dependence of the short-circuit current density (J_{sc}) for LBSF and NOBSF cells on metallization fraction.

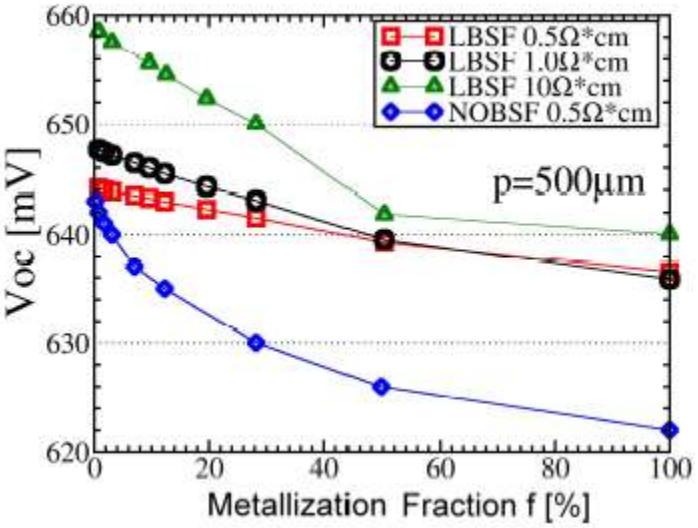


Figure 6.4 - Dependence of the open-circuit voltage (Voc) for LBSF and NOBSF cells on metallization fraction.

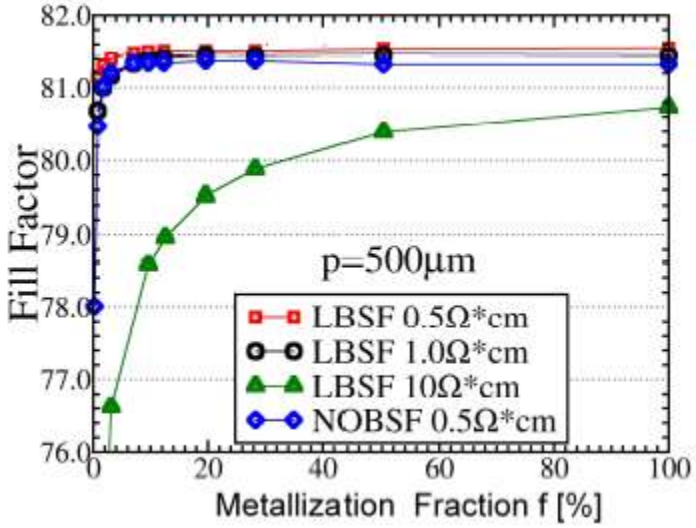


Figure 6.5 - Dependence of the fill factor (FF) for LBSF and NOBSF cells on metallization fraction.

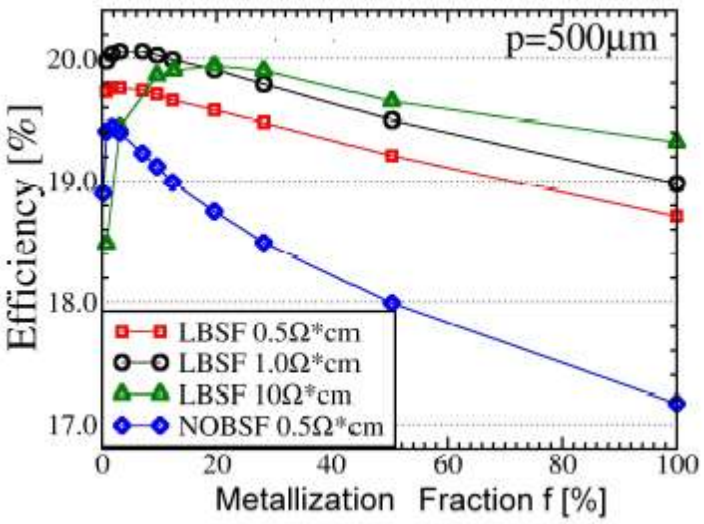


Figure 6.6 - Dependence of the efficiency for LBSF and NOBSF cells on metallization fraction

For the considered devices and substrate resistivity values, by decreasing the metallization fraction *f*, we observe and increase of both J_{sc} and V_{oc} due to the reduction of the effective back surface recombination velocity and to the increase of the effective internal bottom reflectivity. Moreover, the considered base resistivity values result in a different impact of the bulk recombination and of the rear surface recombination losses. In particular, the calculated bulk minority carrier lifetime τ_n is equal to 1.136 ms and 39.8 µs at N_{sub}=1.368x10^{15} cm^{-3} ($\rho_{SUB} = 10\ \Omega cm$) and N_{sub}=3.255x10^{16} cm^{-3} ($\rho_{SUB} = 0.5\ \Omega cm$), respectively. As a consequence, larger values of V_{oc} and J_{sc} are observed for $\rho_{SUB} = 10\ \Omega cm$. In addition, for the NOBSF cell, both J_{sc} and V_{oc} are smaller than those of LBSF cells because of the larger recombination rates in the rear side of the cell. It is also important to note that, for a given ρ_c, the presence of the LBSF leads to a weaker dependence of both V_{oc} and J_{sc} on *f* because of a reduced impact of the surface recombination losses at the rear surface.

By decreasing the metallization fraction, the base spreading and the contact series resistances $R_{S,CB}$ increase, leading to a degradation of the FF. It is worth noting that the NOBSF cell the back contact resistivity is $\rho_{CB} = 1.434\text{x}10^{-3}\ \Omega cm^2$ since N_{sub} = 3.255x10^{16} cm^{-3} leading to $R_{S,CB}$ = 0.73 Ω at f = 0.2%, but for the LBSF cell (independently of the substrate doping concentration), due to the presence of the Al-BSF diffusion, $\rho_{CB} = 3.015\text{x}10^{-7}\ \Omega cm^2$ therefore, $R_{S,CB}$ = 0.15 mΩ at f = 0.2%.

When comparing the three different substrate resistivity values, two consideration arise:

- a lower FF is observed for $\rho_{SUB} = 10\ \Omega cm$
- a stronger dependence of the FF on the metallization fraction is shown in the case of $\rho_{SUB} = 10\ \Omega cm$ because of the larger 3-D spreading effect

the efficiency trade-off due to the opposite trends of V_{oc} (J_{sc}) and FF as a function of f leads to an optimum value f_0 is within the range 1-3% for $\rho_{SUB} = 10\ \Omega cm$ and for $\rho_{SUB} = 1.0\ \Omega cm$, while for $\rho_{SUB} = 10\ \Omega cm$, f_0 is shifted to a larger value (around 20%) due to the stronger degradation of the FF at low metallization fractions.

Furthermore, the maximum efficiency is observed for $\rho_{SUB} = 1.0\ \Omega cm$ (η=20.06% at $f_0 = 3.14\%$), leading to an **efficiency improvement Δη equal to 1.08%** calculated with respect to the case of full-metalized rear side (*f*=100%), for which η=18.98%. As expected, for a given substrate resistivity, the LBSF cells feature a larger efficiency with respect to the NOBSF cell.

However, the gain in efficiency is reduced down to 0.31% at the optimum metallization fraction (the NOBSF cell features the maximum efficiency η=19.45% at *f*=1.77% while the LBSF cell with $\rho_{SUB} = 0.5\ \Omega cm$ reaches the maximum efficiency η=19.76% at *f*=3.14%.
The simulation results are summarized in the following table:

	s	f_o	J_{sc}	V_{oc}	FF	Eff
	μm	%	mA/cm	mV		%
LBSF 0.5	100	3.14	37.69	644	81.4	19.7
LBSF 0.5	-	100	36.04	637	81.5	18.7
LBSF 1.0	100	3.14	38.19	647	81.1	20.0
LBSF 1.0	-	100	36.65	636	81.4	18.9
LBSF 10	250	19.6	38.43	652	79.5	19.9
LBSF 10	-	100	37.40	640	80.7	19.3
NOBSF 0.5	75	1.77	37.46	641	80.9	19.4
NOBSF 0.5	-	100	33.93	622	81.3	17.1

Table 6.2 – Output cell parameters calculated at the optimum metallization fraction f_0 and at *f*=100% (fill-contacted rear side).

6.3.2 Collection efficiency of photo-generated carriers

It has been calculated the collection efficiency η_c of the photo-generated electron-hole pairs at the contacts as the ratio of the short circuit current to the photon current within the range 600-1100 nm for the LBSF cells. The situation is reported in the following figure. *[11]*
For long wavelengths, corresponding to photon absorption close to the back surface, the collection efficiency increases significantly due to the reduced rear recombination losses.
It is worth noting that for $\rho_{SUB} = 10\ \Omega cm$ the rear point contact geometry has a smaller impact on the collection efficiency with respect to the device featuring $\rho_{SUB} = 0.5\ \Omega cm$, for which the effects of the recombination losses strongly influences the contribution to short circuit current density by large wavelengths. In fact, for instance, at λ=1100 nm and for $\rho_{SUB} = 10\ \Omega cm$, $\Delta\eta_c$=0.01, while for $\rho_{SUB} = 0.5\ \Omega cm$, $\Delta\eta_c$=0.10.

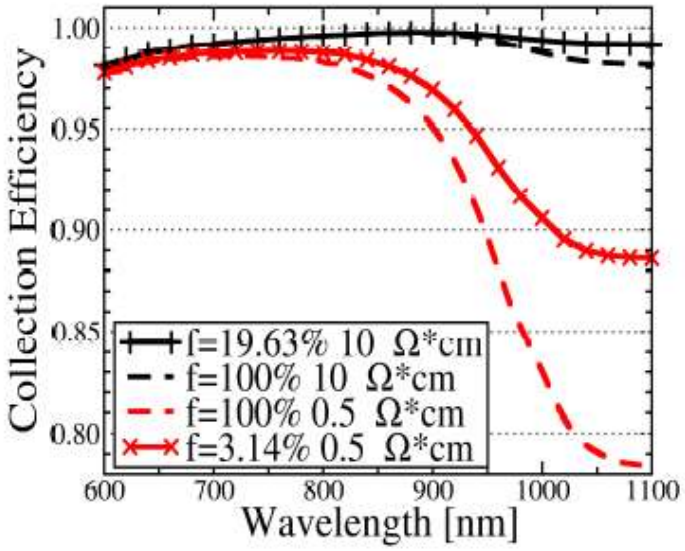


Figure 6.7 – Collection efficiency from 600 nm to 1100 nm for theLBSF cells with ρ_{sub}=0.50 Ωcm

6.4 Conclusions

The dependence of the electrical output parameters on the metallization fraction *f* for rear point contact solar cells (NOBSF and LBSF) have been investigated by means of 3-D numerical simulations. The performed analysis highlights a trade-off between the reduction of the rear surface recombination losses at passivated interfaces and the increase of the internal bottom reflectivity and of the series resistance. For a substrate resistivity of $\rho_{SUB} = 10\ \Omega cm$ a relatively large optimum value of metallization fraction f_0 (around 20%) has been calculated due to the strong effect of the spreading resistance on the fill factor, while for $\rho_{SUB} = 0.5\ \Omega cm$ and $\rho_{SUB} = 1.0\ \Omega cm$ f_0 is in the range 1-3% for both NOBSF and LBSF cells. The maximum efficiency (20.06%) is obtained for LBSF cell with $\rho_{SUB} = 1.0\ \Omega cm$, due to relatively smaller effective rear surface recombination velocities. However, for the lowest considered substrate resistivity ($\rho_{SUB} = 0.5\ \Omega cm$), the NOBSF cells shows the larger efficiency improvement ($\Delta\eta$=2.28%) with respect to the full-metalized rear side cell.

7 Analysis and optimization of a homogeneous emitter solar cell

7.1 The tool employed: TCAD Sentaurus

Sentaurus TCAD is a framework made up by different tools used to model and simulate different kinds of devices. Especially, the following tools have been employed: Sentaurus Structure Editor (SSE) and Sentaurus Device (SD). The first one, SSE, allows the scientist to create 2D and 3D geometrical structures and to model their physical properties (like the material and the doping profile), the second one allows to simulate the optical and electrical features of the device. Moreover, the first tool cooperates with the generator of the mesh on which the second tool will work on. Below, I will describe the input and output files, used by the different tools. *[9]*

Within SSE, **the input file** (*.Scm) is made up by a series of commands, whose syntax is based on the scripting language called Scheme. These instructions are used to create the geometrical structure, to define the contacts, to add to the model the doping profile (that can be a constant, analytical or externally generated), and, finally, to allow the tool to communicate with the mesh generator.

These last commands allow us to create the input file (boundary_fps.tdr and command_dvs.cmd) needed by the mesh generator. It is recalled (always by a command) by the SSE and it generates a file (grid_msh.tdr) containing the geometrical data and the doping profile related to the mesh nodes. This file is an input one for the simulator (SD). The other input to the simulator (command_dvs.cmd) is a command file that needs the scientist to select the physical model that better suits to the problem, the mathematics methods to solve it and the output to analyze. Finally, the simulator output (*_des.plt, *-des.tdr) described the electrical features of the device.

7.2 Homogeneous emitter solar cell

A 2D model of a homogeneous emitter solar cell has been implemented. It considers a p-n^+ junction. The structure dimensions are reported in the table below. So, the structure is made up, from the top to the bottom, by **a metal layer, followed by a p^+ layer, a p layer, a n^+ layer and then an antireflection coating and the contact**. *[9]*

Geometrical parameter	Value
Number of contacts	77
Wafer width	156000 μm
Substrate width	2026 μm (=156000/77)
Substrate thickness (p)	180 μm
Contact width	75 μm
Contact thickness	0.05 μm
Layer width AR	0.03 μm
Emitter width (n^+)	0.35 μm
BSF (p^+) width	10.3 μm

Table 7.1 – Geometrical parameters of the analyzed homogeneous emitter solar cell

While, the outputs are calculated from the characteristic curve I-V, measured thanks to the SD tool. They are:

- Maximum power current
- Maximum power voltage
- Short-circuit current
- Open circuit voltage
- Maximum power
- Fill Factor
- Efficiency

7.3 TCAD/Optimization algorithm interface

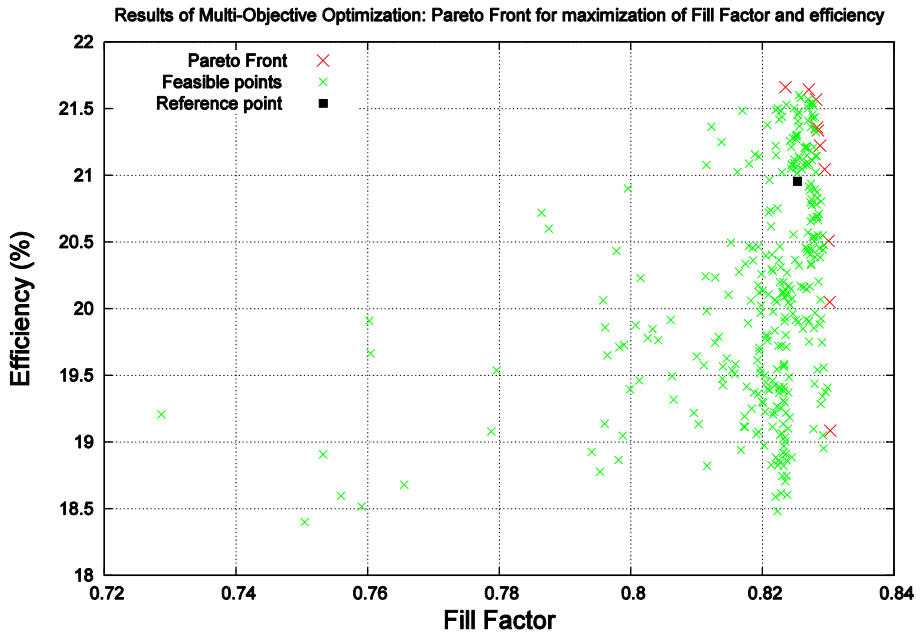
In order to optimize the cell performance, it is to operate on the geometric dimensions of the structure and on the physical properties (like the doping profile). Thus, the design variables (that are the decision variables) are properly managed by the optimization algorithm and included in the relevant file (*.Scm), that is the input file for the SSE tool. While, when the device is simulated, the characteristic curve I-V is extracted from the SD tool's output file (*_des.plt). From this curve, all the interesting parameters (or the objective functions) are extracted. Both scripts that insert and extract the information are written in Matlab®.

7.3.1 Optimization algorithm

It has been used a Genetic Algorithm, since it is a heuristic search and optimization method well suited to the cases when the objective function is not continuous, not derivable or strongly not linear. It enables the researcher to face problems of single (SOO) and multi-objective optimization (MOO) and also allows to adopt solution strategies that take into account constraints within the solution subset to solve constrained SOO and MOO problems.
More details over this topic can be found in "Clonal selection – An Immunological Algorithm for Global Optimization over Continuous Spaces" (Journal of Global Optimization, DOI 10.1007/s10898-011-9736-8)
My choice has been to use the Fill Factor and the Efficiency as my objective functions, while the geometrical parameters shown in the previous table have been chosen as decision variables.

7.4 Results

Simulating the solar cell and employing the numerical values shown in the previous table the result is **a Fill Factor equal to 0.82% and an Efficiency equal to 20.95%**. My results are further shown in the following figures.



The results in the previous figure have been gained running the simulation over the model of a HE solar cell, after that 300 different structures have been analyzed. These structures are compared among them as for Efficiency and Fill Factor. The points featured by the Pareto Optimality despite of the 300 remaining are underlined in red, in green it is possible to see all the other points. In black it has been underlined the point related to the reference structure (the one of the previous table).
The further step has been the consideration of a trade-off.

	Fill Factor	Efficienza(%)
Maximum Fill Factor	0,830297	19,083924
Maximum Efficiency	0,823503	21,659972
Best Trade off 1	0,827048	21,641942
Best Trade off 2	0,828174	21,566847

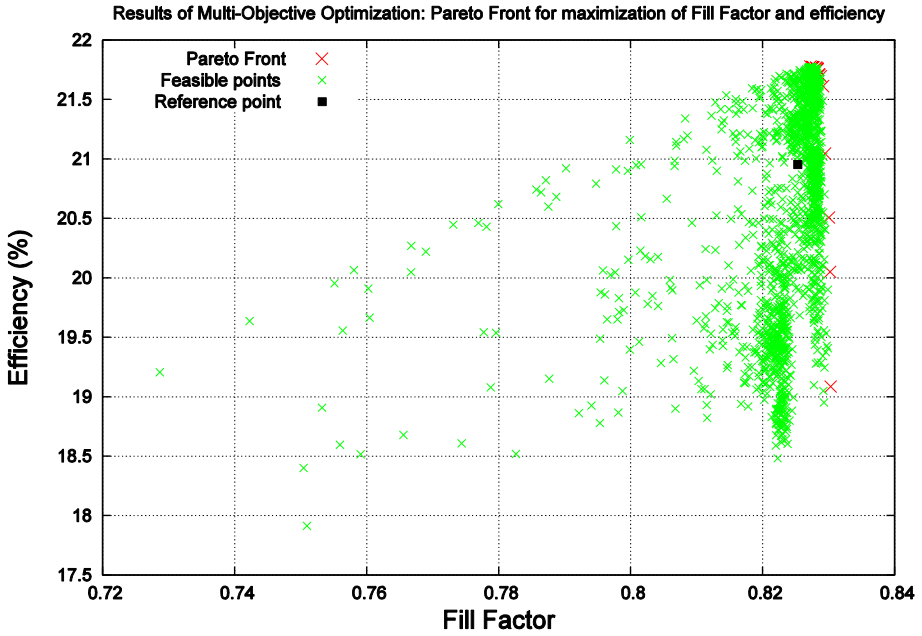
Table 7.2 – The points selected from the Pareto front in the previous figure.

In the first column of the previous table, the decision strategies have been shown. The first one and the second one are immediate, while the third and the fourth one have been gained through the **normalization of the two measures respect to the maximum value and then calculating the distance from the ideal point** (that is [1,1]).
Now, in the following table, I am going to show the percentage gain for both the interested parameters, related to the reference structure.

	Fill Factor Gain (%)	Efficiency Gain (%)
Maximum Fill Factor	+0,61	-8,93
Maximum Efficiency	-0,22	+3,37
Best Trade off 1	+0,21	+3,28
Best Trade off 2	+0,35	+2,92

Table 7.3 – Percentage gain (with respect to the reference structure)

In the following figure, I am going to show the results gained by the multi-objective optimization performed over the HE solar cell. These results have been extracted after that 1700 different structures have been analyzed. The structures are compared among them as for Efficiency and Fill Factor again. The points gaining the Pareto Optimality respect to all the other points are underlined in red, in green all the others. In black it is possible to look at the reference structure's performance (related to the parameters shown in the first table).



	Fill Factor	Efficiency(%)
Maximum Fill Factor	0,830297	19,083924
Maximum Efficiency	0,827228	21,777828
Best Trade off 1	0,828315	21,759444
Best Trade off 2	0,828365	21,753349
Best Trade off 3	0,828321	21,755498
Best Trade off 4	0,828116	21,768324
Best Trade off 5	0,827663	21,769921

Table 7.4 – The points selected from the Pareto front in the previous figure.

Again, the first two decision strategies are immediate, while the others are chosen by the normalization of the two measures respect to their maxima and then calculating the distance from the ideal point (again [1,1]).

	Fill Factor Gain (%)	Efficiency Gain (%)
Maximum Fill Factor	+0,61	-8,93
Maximum Efficiency	+0,23	+3,93
Best Trade off 1	+0,37	+3,84
Best Trade off 2	+0,37	+3,81
Best Trade off 3	+0,37	+3,82
Best Trade off 4	+0,34	+3,88
Best Trade off 5	+0,29	+3,89

Table 7.5 – Percentage gain (with respect to the reference structure)

Above, again, the fill factor and efficiency gains, related to the reference structure, are shown.

8 Thin-film solar cells

8.1 Introduction

The wafer thickness for sufficient absorption of the solar spectrum is >700 μm. This is quite a large thickness for a Si wafer and is not desirable for commercial production of solar cells for two reasons: the wafer cost can be very high and its effectiveness for collection of photogenerated carriers will be small because it is difficult to have a minority-carrier diffusion length (MCDL) comparable to such a large wafer thickness. Thus, for practical reasons, wafer thickness must be less than this value.

Clearly, when the thickness of a Si solar cell is reduced, some problems emerge. Just reducing the cell thickness will result in reduced absorption and thus, in a reduced photocurrent. To get a quantitative feeling of such a reduction in photocurrent, it is useful to look at the following figure, where the maximum achievable current density (MACD) generated by a planar solar cell, coated with an appropriate antireflection coating, is plotted against different cell thicknesses. *[1]*

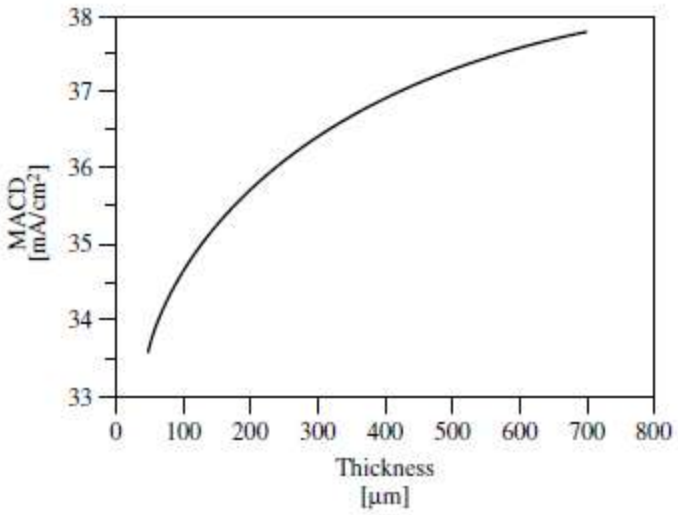


Figure 8.1 - Maximum achievable current density (MACD) from a planar, AR-coated Si solar cellas a function of cell thickness. These calculations assume an optimized AR coating and AM1.5 incident spectrum.

The previous figure shows that the photocurrent increases with an increase in thickness and saturates at a thickness of about 700 μm. At a thickness of about 300 μm, the current density is within 5% of the saturation value, which implies that a thickness of 300 μm is suitable for fabricating high-efficiency solar cells on planar substrates. This is fortunate because a similar demand on wafer thickness comes from requirements for maintaining a high yield in handling and processing other semiconductor devices.

Recently, however, there have been many advances in wafer handling and in the development of gentler processing methods to accommodate high throughput. These advances have sparked interest in using thinner substrates for two reasons:

- To reduce the amount of Si for each watt of PV energy generation. Because the PV industry has gone through periods of Si shortage, an efficient use of Si can minimize such hardships.
- To improve the efficiency of solar cells fabricated on low-cost substrates using improved cell designs.

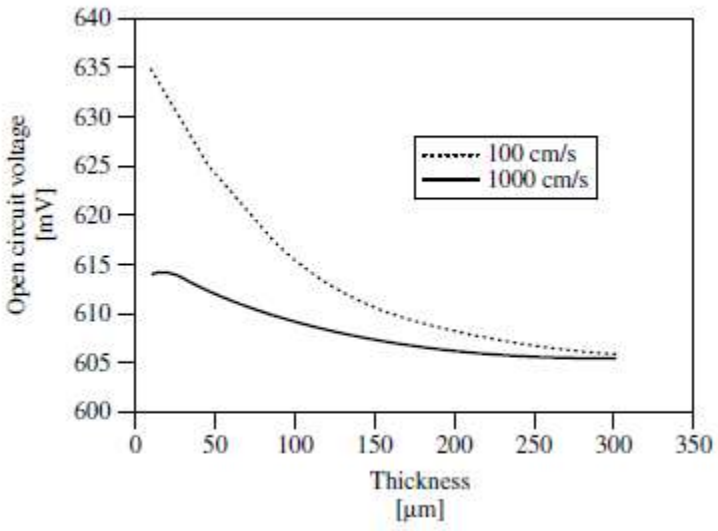


Figure 8.2 - V_{OC} of a Si solar cell as a function of thickness for high and low surface-recombination velocities.

As shown above, for a given material quality, both for high and lower recombination velocities, a reduction in the cell thickness can result in improving the open-circuit voltage (V_{oc}) and the fill factor. However, as the cell thickness is reduced, the surface recombination becomes an increasingly important component of the total recombination. In particular, surface recombination can severely degrade V_{oc}. Thus, thinner cells can yield higher voltages and higher fill factors if the surface recombination demand are met. However, they can suffer a loss in the photocurrent unless the optical losses associated with thickness reduction are compensated through superior light trapping design. If these conditions are met, thinner cells can be more efficient than their thicker counterparts, especially when considering a reduction in Silicon used to manufacture this kind of cells.

8.2 Optimization techniques

The thin-film devices, featured by a thickness from hundreds of nanometers up to a few μms, require efficiency optimization criteria different from the ones previously analyzed. In particular, techniques based on the Internal Quantic Efficiency (IQE) maximization, through the solar radiation confinement, have been experimented. Apart from the materials used (amorphous, crystalline or polycrystalline silicon) and from the model accuracy (in order to take into account, for instance, the Auger recombination effects), these efforts leaded to the definition of either a regular or an irregular texturing structure of the silicon surface, showing as a result in both cases an increase of the solar radiation contribution to the optical absorption of the cell. The photons hitting the cell surface are partly reflected away, partly absorbed by the material and, finally, partly transmitted through it. Anyway, only the absorbed photons give a contribution to the electrical conversion process. If we assume k as a refraction index for the material, the following parameter, also known as an **absorption coefficient** α gives us the measure of how much and at which wavelengths photons can penetrate the material before they are absorbed: *[2]*

$$\alpha = \frac{4\pi k}{\lambda} \quad [cm^{-1}]$$

as a result, photons featured by different wavelengths can penetrate up to different depths the material before they get absorbed. This depth is equal to the absorption coefficient's inverse.
So, the following strategies have been implemented to improve the cell efficiency:

- minimization of the metal contact on the cell surface, because photons hitting the contacts do not take part in the conversion process.
- Reduction of reflection losses, through the coating of the cell surface with an antireflection coating (AR).
- Reduction of reflection and transmission losses through the cell structure design, in order to increase the light optical path within the cell, to increase the probability to get the photons absorbed (IQE improvement).

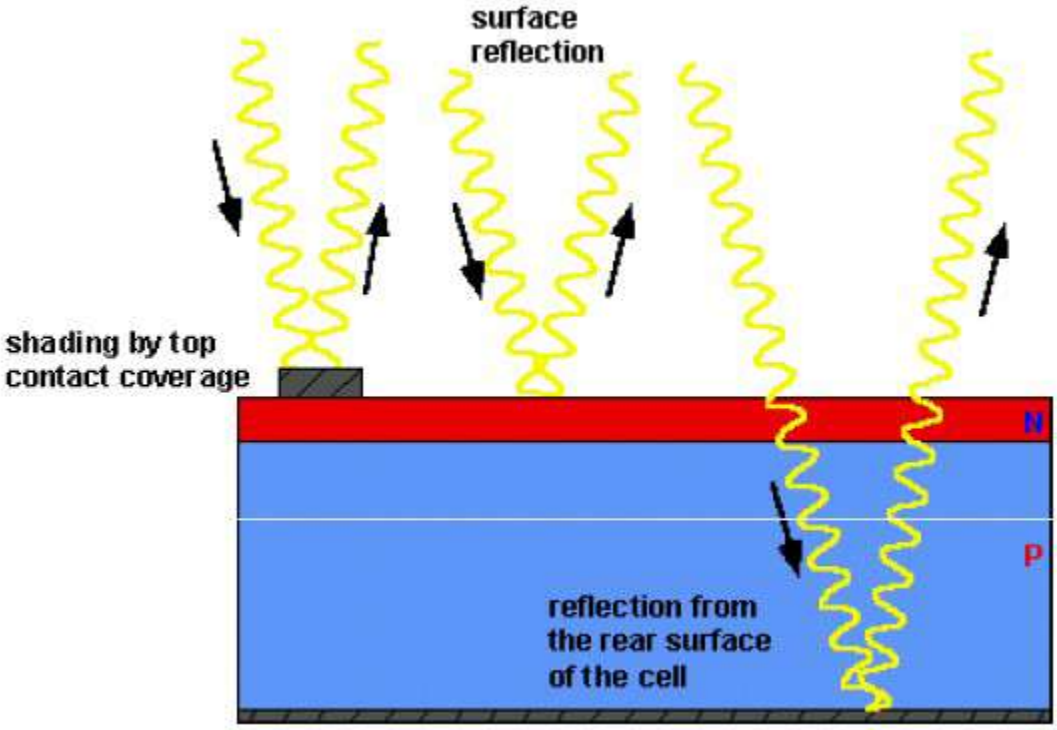


Figure 8.3 - The most important reflection losses for a photovoltaic cell

8.3 Anti Reflection Coating (ARC)

The ARC technology is based on the interference effect. Through the implementation of a dielectric layer on the cell surface, it is possible to minimize the interferences that are responsible for the light reflection away from the device. An adequate thickness dielectric layer is able to get a difference between the phases of the wave reflected from the cell coating and the wave reflected from the semiconductor substrate. This difference can eliminate the reflection phenomenon, avoiding the dispersion of reflected energy. The coating layer thickness is chosen taking into account the incident ray wavelength. Exactly, the dielectric material wavelength must be a quarter of the incident light wavelength. So, if we assume a material featured by a refraction index n_1 and a wavelength of the light equal to $\lambda=\lambda_0$, the dielectric thickness needed to coat the cell would be *[2]*

$$d_1 = \frac{\lambda_0}{4n_1}$$

If $\lambda \neq \lambda_0$, the reflection effect is still present, but it is still much weaker than in the no-antireflection coating case. If we consider the following antireflection layer ARC, with a thickness d_1, the presence of three different reflection indexes (n_0, n_1 and n_2), an optimal thickness dielectric layer on the left and no ARC on the right:

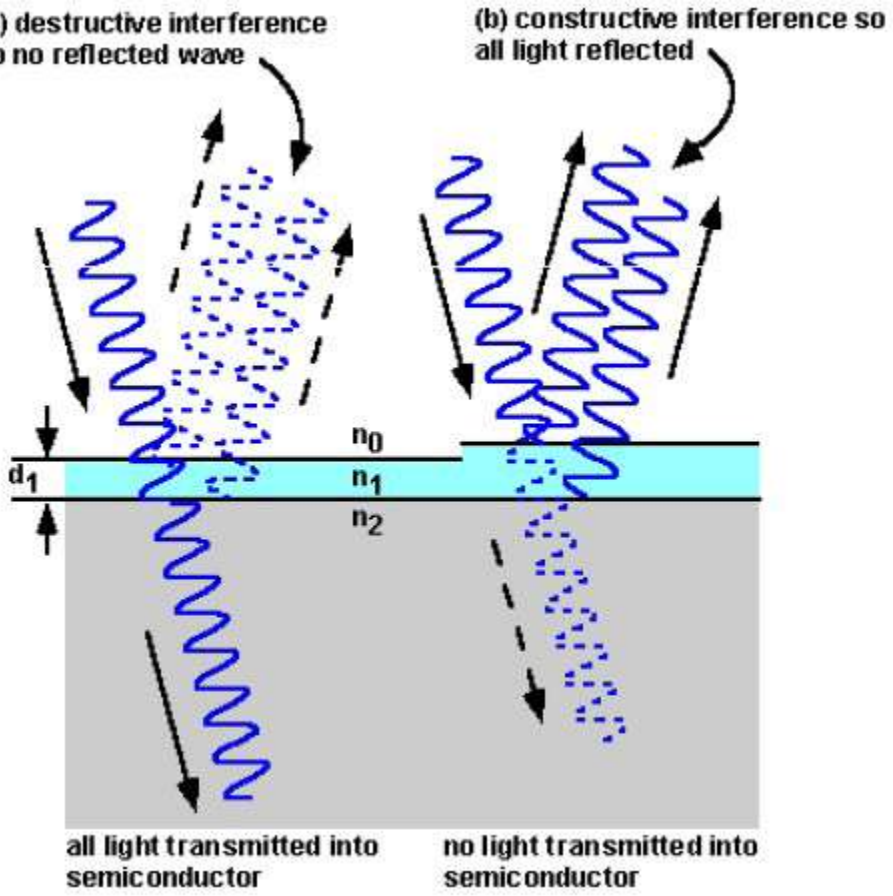


Figure 8.4 – Reflection and transmission phenomena in an ARC equipped device. This coating eliminates the reflection phenomenon through the interferences.

the reflection index would be *[2]*

$$R = \frac{r_1^2 + r_2^2 + 2r_1 r_2 \cos(2\beta)}{1 + r_1^2 r_2^2 + 2r_1 r_2 \cos(2\beta)}$$

$$with \quad r_1 = \frac{n_0 - n_1}{n_0 + n_1}, \quad r_2 = \frac{n_1 - n_2}{n_1 + n_2}, \quad \beta = \frac{2\pi n_1 d_1}{\lambda_0}$$

$$when \quad n_1 d_1 = \frac{\lambda_0}{4}$$

$$then\ R = R_{min} = \left(\frac{n_1^2 - n_0 n_2}{n_1^2 + n_0 n_2}\right)^2$$

So, R_{min}=0 if it is true $d_1 = \frac{\lambda_0}{4n_1}$.

By the use of multiple AR layers it is possible to further reduce the amount of light reflected away, through the reduction of the **reflectance**, anyway, because of its high cost, a double ARC is used only in high efficiency cells.

8.4 Texturing

Texturing is a process through which some microstructures are created within the silicon surface, using adequate corrosion techniques. Alkaline solutions, based on KOH and NAOH can corrode the silicon, creating pyramids with a squared base in random positions. Modern manufacturing processes can control the depth of these structures within the material, controlling the temperature and the time during which the corrosion reaction takes place. Light is reflected from one pyramid to another, and it results in absorption increase. Rough surfaces and an asymmetric structure can improve the results of this method. Texturized surfaces, firstly used to reduce the reflectance, are now considered as a valid technology to increase the optical path. *[63]*

Let us now consider the following parameter:

$$n = \frac{c}{v}$$

is called **refraction index** of a given medium with in respect to the vacuum. Where *c* is the speed of the light in the vacuum and *v* is the speed of the light in a given medium.

The geometrical path of a ray of light within a medium with a given refraction index is the following one:

$$d = tv = t\frac{c}{n}$$

while the **optical path** is:

$$\Delta = dn = ct$$

that is the space travelled by the light in the vacuum in the same time needed to travel the distance *d* within the given medium.

As a result, the texturing in a thin-film cell can lead to:

- higher light absorption
- lower recombination rate within the bulk (higher V_{oc})

8.5 Light trapping

In order to maximize the open circuit voltage, it is necessary to reduce as much as possible the recombination effects over the whole cell. It is generally reached through thinner layers, that, on the contrary, have the drawback to generate lower absorptions in respect to thicker layers. So, if we want to use thinner layers, without compromising the light absorption, it is needed to guarantee, within the thin layer, an optical path of the same length as in the thick layer case. That is why light trapping techniques are getting largely used in modern photovoltaic industry. The purpose is to trap the solar light within the cell to guarantee the highest absorption probability, increasing the reflection within the same cell. It is generally reached making the shadowed part of the cell act as a mirror. The process consists in the modification of the plane surfaces geometry, through, for instance, the texturing techniques. *[64]*

9 Thin-film cells optical model

In this paragraph an optical model for thin-film silicon solar cells will be introduced. Surfaces and interfaces will be assumed to have small roughness (about 20% of the thickness).

9.1 Multilayer thin-film structure

Our model is based on the cell's structure in figure below *[3]*

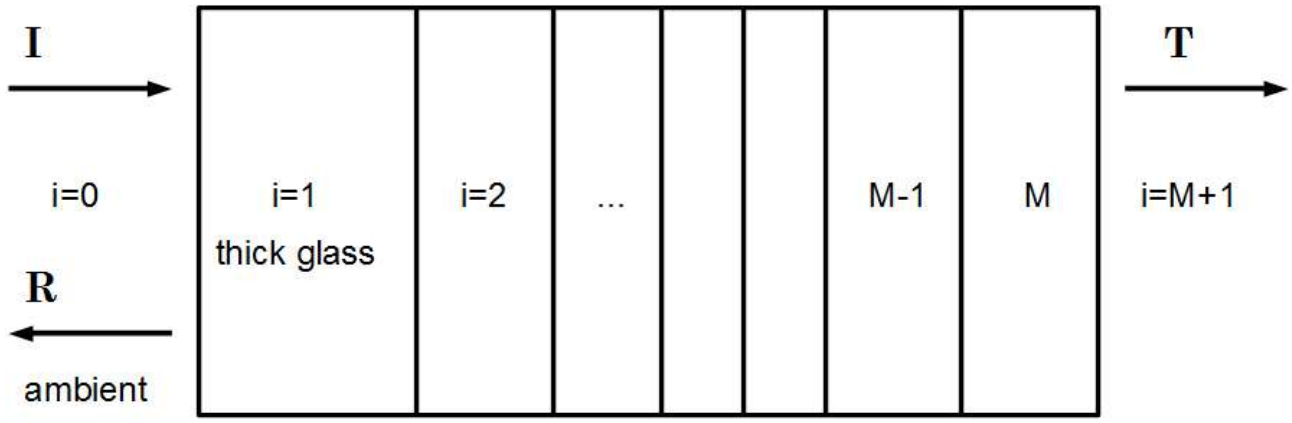


Figure 9.1 – A thin-film cell multilayer structure scheme.

It is a pile of M+1 layers with different optical properties in which the first film (labelled by i=1 and having a larger thickness) is generally made up by silicon dioxide (glass), while the last one (M+1) is a metal layer. The light horizontally hits the layer i=0, near to the glass layer and, generally, there is no transmission if electromagnetic radiation from the last metal layer to the ambient. The active part of the cell is made up by the M-2 layers, featured by a sub-micrometric thickness. Interfaces among adjacent layers are considered as diffusive regions so they can be used to adequately diffuse the light in order to increase the effective optical path through light trapping techniques. *[1]*

9.2 Parameters used in the simulation

A structure as in the following figure is featured by different electromagnetic field amplitudes, each one related to a different interface. *[3]*

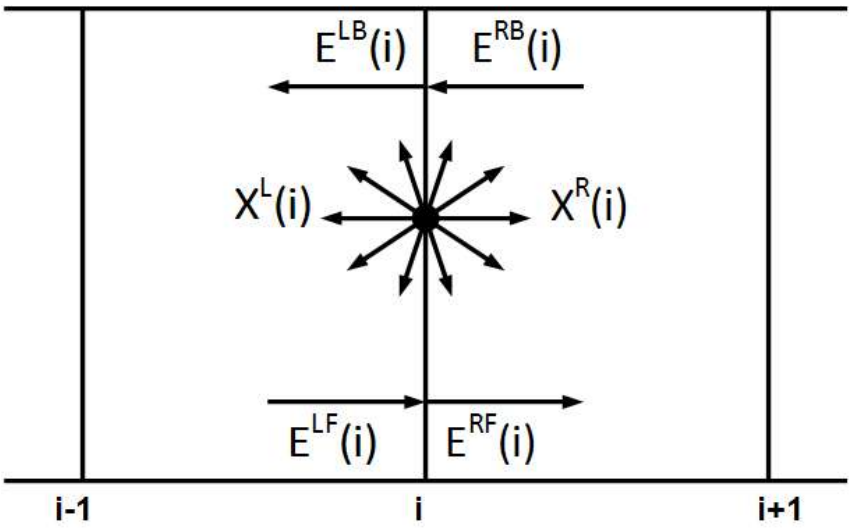


Figure 9.2 – The electromagnetic field amplitudes featuring each interface.

- E^{LF} is the left-forward amplitude
- E^{RB} is the right-backward one
- E^{LB} is the left-backward one
- E^{RF} is the right-forward one
- X^L and X^R are the random scattered amplitudes to the left and the right for the i^{th} layer considered

Each layer is characterized by some internal specific parameters, like the following ones:

- n(i) and K(i), that are the real and complex part of the refraction index
- d(i), the layer thickness
- σ(i) that is the roughness parameter, used to act on the intensity of the light's component diffused by the interface

As for the simulation code, it is to say that it is made up by two modules:

- The computation of the electromagnetic radiation propagation within the multilayer region through the **Matrix method**. For each layer of the cell, the code calculates the different components of the electromagnetic field: transmitted, diffused, absorbed and reflected.
- The statistical computation of the random part through the use of **Monte Carlo methods**. The diffused component can be either again absorbed by the structure or emitted into the ambient. This second module calculates the contribution given by the diffused component to the total absorption. Since the diffused radiation is made up by a stream of photons with different angles with respect to the interfaces, the Monte Carlo simulation follows statistically the path of a certain number of photons.

9.3 Computation of the coherent absorption through the Matrix Method

The Matrix Method allows us to exactly calculate the coherent electromagnetic radiation propagation in a structure like of the previous figure. The calculation is carried out:

- associating a matrix operator for each layer of the structure that realizes the computation of the radiation propagating within the same layer (reflected, diffused, absorbed radiation and this last one only if the extinction coefficient K(i) is not equal to zero).
- associating a matrix operator for any structure's interface responsible for the relation between an electric field's component and transmission and reflection coefficients related to the materials of the considered interface

The parameters featuring the ith layer are:

- the thickness d(i)
- the complex refraction index $N(i) = n(i) - ik(i)$ where $n(i)$ is the real refraction index and $k(i)$ is the extinction coefficient, typical of the material.

The ith interface is featured by:

- the ruggedness $\sigma(i)$
- the angular diffused scattering $P_{scatt}\ (i, [\theta])$ distribution function to calibrate
- the complex reflection and transmission coefficients (depending on the wavelength λ and the hitting radiation), given b the following equations

$$r(i) = \left[\frac{N(i-1) - N(i)}{N(i-1) + N(i)}\right], \quad t(i) = \left[\frac{2N(i-1)}{N(i-1) + N(i)}\right]$$

If the interface is rugged, the previous formulas are modified according to the "scalar scattering theory", from the layer i-1 to i:

$$r_F(i) = \left[\frac{N(i-1) - N(i)}{N(i-1) + N(i)}\right] S_r(i), \quad t_F(i) = \left[\frac{2N(i-1)}{N(i-1) + N(i)}\right] S_t(i)$$

and from the layer i to i-1:

$$r_B(i) = \left[\frac{N(i) - N(i-1)}{N(i-1) + N(i)}\right] S_r(i+1), \quad t_B(i) = \left[\frac{2N(i)}{N(i-1) + N(i)}\right] S_t^{-1}(i)$$

where S_t and S_r are the scattering coefficients defined by the following equations:

$$S_r(i) = e^{-\frac{1}{2}\left(\frac{2\pi n(i-1)\sigma(i)}{\lambda}\right)^2} \quad and \quad S_t(i) = e^{-\frac{1}{2}\left(\frac{2\pi [n(i-1) - n(i)]\sigma(i)}{\lambda}\right)^2}$$

The relations bounding the electromagnetic field amplitudes and the reflection and transmission coefficients, can be expressed as:

$$\begin{bmatrix} E^{LF}(i) \\ E^{LB}(i) \end{bmatrix} = \begin{bmatrix} \frac{1}{t_F(i)} & -\frac{r_B(i)}{t_F(i)} \\ \frac{r_F(i)}{t_F(i)} & \frac{t_F(i)t_B(i) - r_F(i)r_B(i)}{t_F(i)} \end{bmatrix} \begin{bmatrix} E^{RF}(i) \\ E^{RB}(i) \end{bmatrix} = \overleftrightarrow{I}(i) \begin{bmatrix} E^{RF}(i) \\ E^{RB}(i) \end{bmatrix}$$

Since the amplitudes $E^{RF}(i)$ and $E^{RB}(i)$ are related to the amplitudes $E^{LF}(i+1)$ and $E^{LB}(i+1)$ by the following:

$$E^{LF}(i+1) = E^{RF}(i)e^{i\beta(i)d(i)} \;\; and \;\; E^{RB}(i) = E^{LB}(i+1)e^{i\beta(i)d(i)}$$

it is possible to express them also as:

$$\begin{bmatrix} E^{RF}(i) \\ E^{RB}(i) \end{bmatrix} = \overleftrightarrow{L}(i) \begin{bmatrix} E^{LF}(i+1) \\ E^{LB}(i+1) \end{bmatrix}$$

$$\text{with } \overleftrightarrow{L}(i) = \begin{bmatrix} e^{i\beta(i)d(i)} & 0 \\ 0 & e^{-i\beta(i)d(i)} \end{bmatrix} = \begin{bmatrix} e^{\frac{2\pi}{\lambda}N(i)d(i)} & 0 \\ 0 & e^{-\frac{2\pi}{\lambda}N(i)d(i)} \end{bmatrix}$$

the effect given by a multilayer structure can be described as a composition of the effects coming from the single layers, in such a way that the relations between the amplitudes of the hitting radiation I, of the reflected one R and of the transmitted one T:

$$\begin{bmatrix} I \\ R \end{bmatrix} = \overleftrightarrow{I}(1)\overleftrightarrow{L}(1)\overleftrightarrow{I}(2)\overleftrightarrow{L}(2) \dots \overleftrightarrow{I}(M)\overleftrightarrow{L}(M)\overleftrightarrow{I}(M+1) \begin{bmatrix} E^{RF}(i+1) \\ E^{RB}(i+1) \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \begin{bmatrix} T \\ 0 \end{bmatrix}$$

and the related intensities:

$$R = Re(R(N(0)R)^*)I \;\; and \;\; T = Re(T(N(M+1)T)^*)I$$

where I is the hitting wave's intensity.
for the ith layer, by the use of the transmitted radiation T it is possible to calculate:

$$\begin{bmatrix} E^{RF}(i) \\ E^{RB}(i) \end{bmatrix} = \overleftrightarrow{L}(i)\overleftrightarrow{I}(i+1)\overleftrightarrow{L}(i+1) \dots \overleftrightarrow{I}(M)\overleftrightarrow{L}(M)\overleftrightarrow{I}(M+1) \begin{bmatrix} T \\ 0 \end{bmatrix}$$

where T has been calculated from the equation $T = \frac{1}{S_{11}}$

It is possible to calculate the absorption profile intensity within the ith layer, that contains:

- the resulting forward wave $E^f(i,x) = E^{RF}e^{-i\beta_i x}$
- the resulting back wave $E^B(i,x) = E^{RB}e^{i\beta_i x}$

The resulting wave has:

- <u>electric field amplitude</u> $E(i,x) = E^{RF}(i,x) + E^{RB}(i,x) = E^{RF}(i)e^{-i\beta_i x} + E^{RB}e^{-i\beta_i x}$

- magnetic field amplitude $H(i,x) = N(i)|E^{RF}(i,x) - E^{RB}(i,x)|$

Through the Poynting's vector $P(i,x) = Re[E(i,x)H(i,x)^*]I$
it is possible to calculate the local and the layer's absorption, respectively:

$$A(i,x) = -\frac{d\boldsymbol{P}(i,x)}{dx} \quad and \quad \boldsymbol{\Gamma}(\mathrm{i}) = \int_0^{\mathrm{d(i)}} \mathrm{A(i,x)dx}$$

then, from the law of conservation of energy we have

$$I = R + T + \sum_i \boldsymbol{\Gamma}(i) + \sum_i \boldsymbol{X}(i)$$

and the diffused light's intensity is calculated as:

$$X(i) = \alpha n(i)[|E^{RF}(i)|^2|r_F(i)|^2\{1 - S_r^2(i)\} + |E^{RB}(i)|^2|t_F(i)|^2\{1 - S_t^2(i)\} + |E^{RF}(i+1)|^2|r_B(i)|^2\{1 - S_r^2(i+1)\} + |E^{RB}(i+1)|^2|t_B(i)|^2\{1 - S_t^{-2}(i)\}]$$

where α is a normalization parameter.
The diffused component $\boldsymbol{X(i)}$ can be partially re-absorbed, partially reflected and partially transmitted by one of the structure's layer. The method discussed so far, anyway, do not allow us to carry out an accurate evaluation of the way the diffused light is redistributed. The following paragraph is going to introduce a method able to do that.

9.4 Light's diffused component evaluation through Monte Carlo method

The diffused component of the electromagnetic radiation can be, as already stated, either re-absorbed by the structure's layer or emitted in the ambient. The evaluation of the contribution of the diffused radiation on the total absorption of the structure requires a calculation method able to follow the trajectory of a certain number of diffused photons (about 10^5). The method implemented in the simulation code is stochastic and it is based on a Monte Carlo simulation. The algorithm that follows the diffused photon's trajectory works in the following way:

- the radiation's diffused component is considered apart from the coherent part, in order to calculate it with the Monte Carlo method. This method allows us to probabilistically calculate the diffused photons' optical paths (Monte Carlo particles) as they are a series of events with a given probability: propagation within the layer (with a variable angle), absorption, specular reflection, coherent refraction, diffused reflection and diffused refraction.
- The hitting photon, comes through the first interface on the border of the layer and is propagated within it following the optical path. If Θ is the angle between the photon's propagation direction and the line vertical to the layers, the probability of the photon's absorption within the i^{th} layer is given by the Beer's law:

$$P^{abs}(i) = 1 - e^{-\frac{2\pi k(i)d(i)}{\lambda cos\theta}}$$

- if the photon is not absorbed, it can interact with one of the interfaces that are the borders of the layer and it depends on the Θ angle's value. The probability that the photon is reflected is:

$$R(a,b) = 0.5\left\{\left|\frac{\frac{N(a)}{cos\theta_a} - \frac{N(b)}{cos\theta_b}}{\frac{N(a)}{cos\theta_a} + \frac{N(b)}{cos\theta_b}}\right|^2 + \left|\frac{N(a)cos\theta_a - N(b)cos\theta_b}{N(a)cos\theta_a + N(b)cos\theta_b}\right|^2\right\}$$

where θ_a and θ_b are bounded by the Snell's law:

$$n(a)sin\theta_a = n(b)sin\theta_b$$

while the reflected or transmitted scattering probabilities are:

$$P^R_{scatt}(a,b) = 1 - e^{-\left(\frac{2\pi n(a)\sigma(a,b)cos\,\theta_a}{\lambda}\right)^2}$$

$$P^T_{scatt}(a,b) = 1 - e^{-\left(\frac{2\pi (n(a)-n(b))\sigma(a,b)cos\,\theta_a}{\lambda}\right)^2}$$

- four combinations of the events coming from the interaction between the hitting photon and the interface are possible, with an associated probability:

a) diffused reflection $R * P^R_{scatt}$
b) specular reflection $(1-R) * [1 - P^R_{scatt}]$
c) diffused refraction $[1-R] * P^T_{scatt}$
d) coherent refraction $(1-R) * [1 - P^T_{scatt}]$

- If the selected event is a scattering one, the new angle is evaluated by the angular distribution preset. The angular diffusion probability is to be considered as a function to be calibrated on the basis of the type of interface.
- The end of the optical path can be recognized either through the photon's absorption in a layer or the propagation towards the ambient. By the evaluation of a large number of photons' trajectories, it is possible to calculate both the lost and trapped diffused radiation's quantities.

9.5 Matlab simulation code

9.5.1 Input data and optical calibration

The *trapping_coer* script in the simulation code requires some input data:

- inpfile, a text file containing the geometrical parameters and the indexes of the materials of the cell for the M layers and the M+1 interfaces. They are, respectively, (thicknesses d(i), ruggednesses σ(i), material indexes I, containing the information about the optical materials' features, retrieved from an associated database
- datafile, a text file containing the refraction indexes n(λ) and the extinction coefficients based on the different wavelengths (from 300 to 1300 nm) of 19 different materials
- iglass, to be indicated before the code runs, that is the number of layers of the inpfile cell.

i-glass	1	2	3	4	5	6	7
d(i)	3500000	800	20	1000	20	20	150
σ(i)	0	160	2	100	2	2	15
i-type	19	13	7	5	6	18	14

Table 9.1 – Example of a text file (inpfile) for a simulated cell.

9.5.2 Computation of the absorbed radiance

For any wavelength, the simulation code calculates the coherent intensity component.

- Absorbed in the different layers $A = \sum_i A_i$
- Diffused X
- Reflected by the multilayer R and, eventually, the transmitted one T

After the matrix method calculation for rough surfaces, we will have:

$$I = R + T + A + X > T + R + A$$

The absorption in the different layers of the diffused absorbed $A_d(i)$, reflected R_d and transmitted T_d components is then calculated through the Monte Carlo method. A good statistical estimation is gained through the use of about 10^5 particles per layer. The calculated values are then summed to get the total values for the different quantities:

$$A_{tot}(i) = A(i) + A_d(i),\ \ R_{tot} = R + R_d,\ \ T_{tot} = T + T_d$$

The current version of the code only considers a Lambertian angular distribution and the outcome angle of the diffused photons is calculated stochastically through this distribution.
In the database we can find the following materials:

1	Intrinsic cristalline silicon
2	Intrinsic amorphus silicon
3	n-type amorphus silicon
4	p-type amorphus silicon
5	Intrinsic microcrystalline silicon
6	n-type microcrystalline silicon
7	p-type microcristalline silicon
8	ZnO sputt undoped
9	ZnO sputt lowdoped
10	ZnO sputt doped
11	ZnO sputt highdoped
12	ZnO LPCVD
13	SnO_2 APCVD
14	Ag
15	Al
16	ZnO+Ag (smooth)
17	ZnO+Ag (rugged)
18	ZnO+Al (smooth)
19	Glass

Table 9.2 – Materials included in the database.

9.5.3 Simulation and results

The simulation code has been tested when using different kinds of photovoltaic cells. The cells, manufactured using well-known materials, are case-studies through which it is possible to analyze the results obtained and, then, to evaluate the interesting parameters.

- Single-junction cells
  The cell in the figure below is in the characteristic *p-i-n* configuration. The chosen materials and the related geometrical features are the following ones:

Thickness [nm]	Material	Ruggedness [nm]
350	Glass	$\sigma = 0$
800	SnO_2	$\sigma = 160$
20	μc-Si:p	$\sigma = 2$
10^6	μc-Si:i	$\sigma = 100$
20	μc-Si:n	$\sigma = 2$
20	ZnO:Al	$\sigma = 2$
150	Ag	$\sigma = 15$

Table 9.3 – Chosen materials and their geometric features

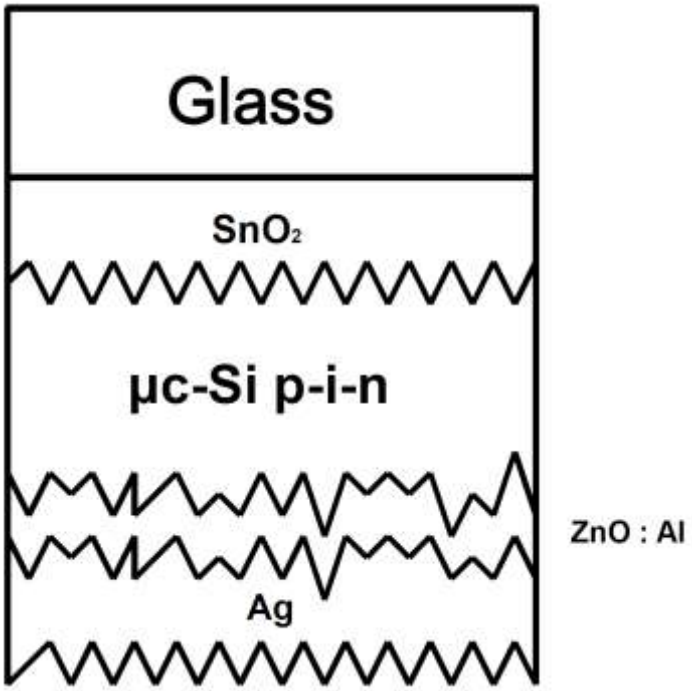


Figure 9.1 – A p-i-n solar cell

The roughness is about 10% of the thickness, apart for the glass ($\sigma = 0$) and for the TCO layer, that has a roughness equal to 20% of the thickness. The same simulation has been carried out changing the material contained in the TCO layer, but maintaining the same thickness and roughness values for each layer.

Thickness [nm]	Material	Ruggedness [nm]
350	Glass	$\sigma = 0$
800	ZnO_2 highdoped	$\sigma = 160$
20	μc-Si:p	$\sigma = 2$
10^6	μc-Si:i	$\sigma = 100$
20	μc-Si:n	$\sigma = 2$
20	ZnO:Al	$\sigma = 2$
150	Ag	$\sigma = 15$

Table 9.4 - The chosen materials and the related geometric features

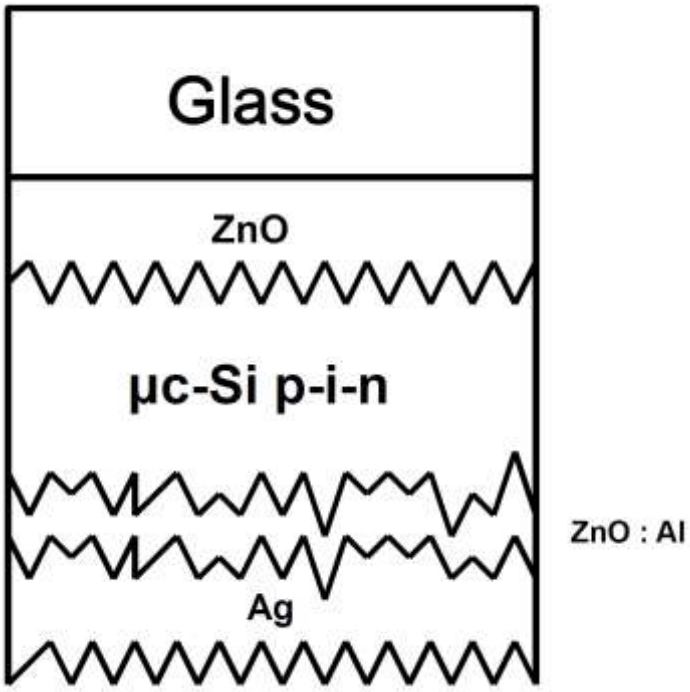


Figure 9.2 – The modified solar cell

The simulation results (using 10^5 Monte Carlo particles) for the respective cells are reported in the following figures. The charts represent the absorption for the single layers of the cell, when changing the wavelength.

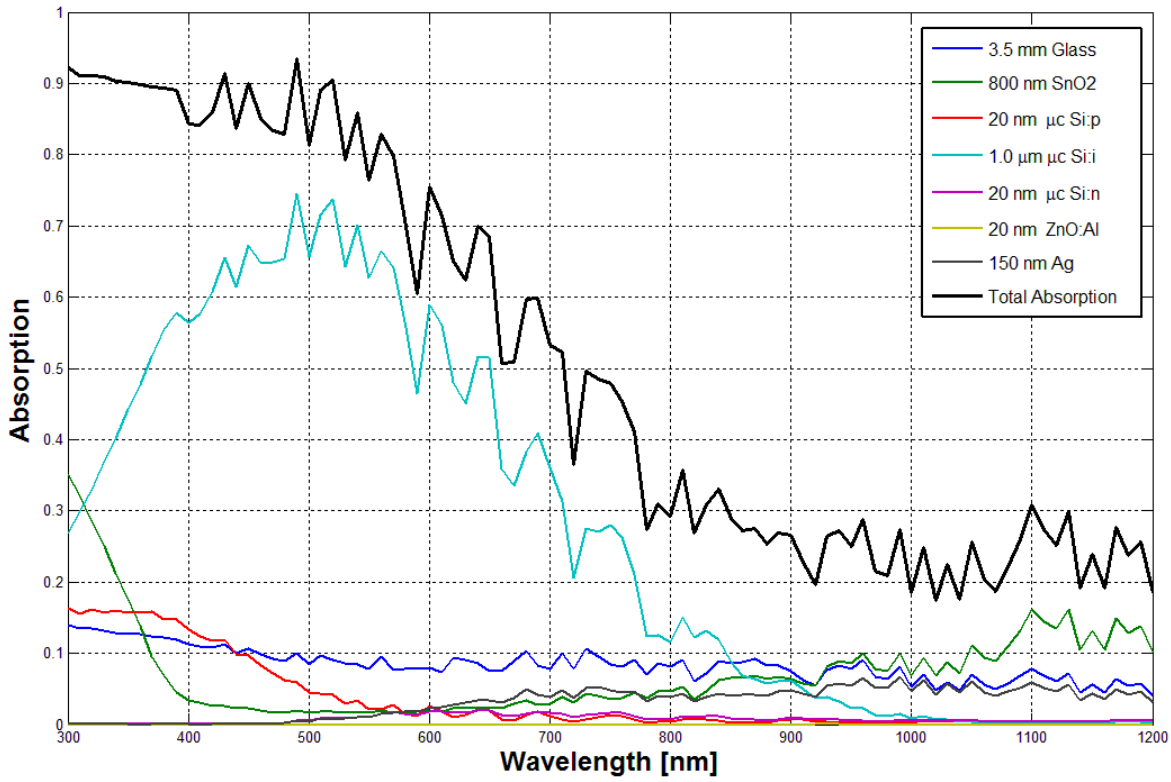


Figure 9.3 – Absorption versus wavelength for each layer for the first cell

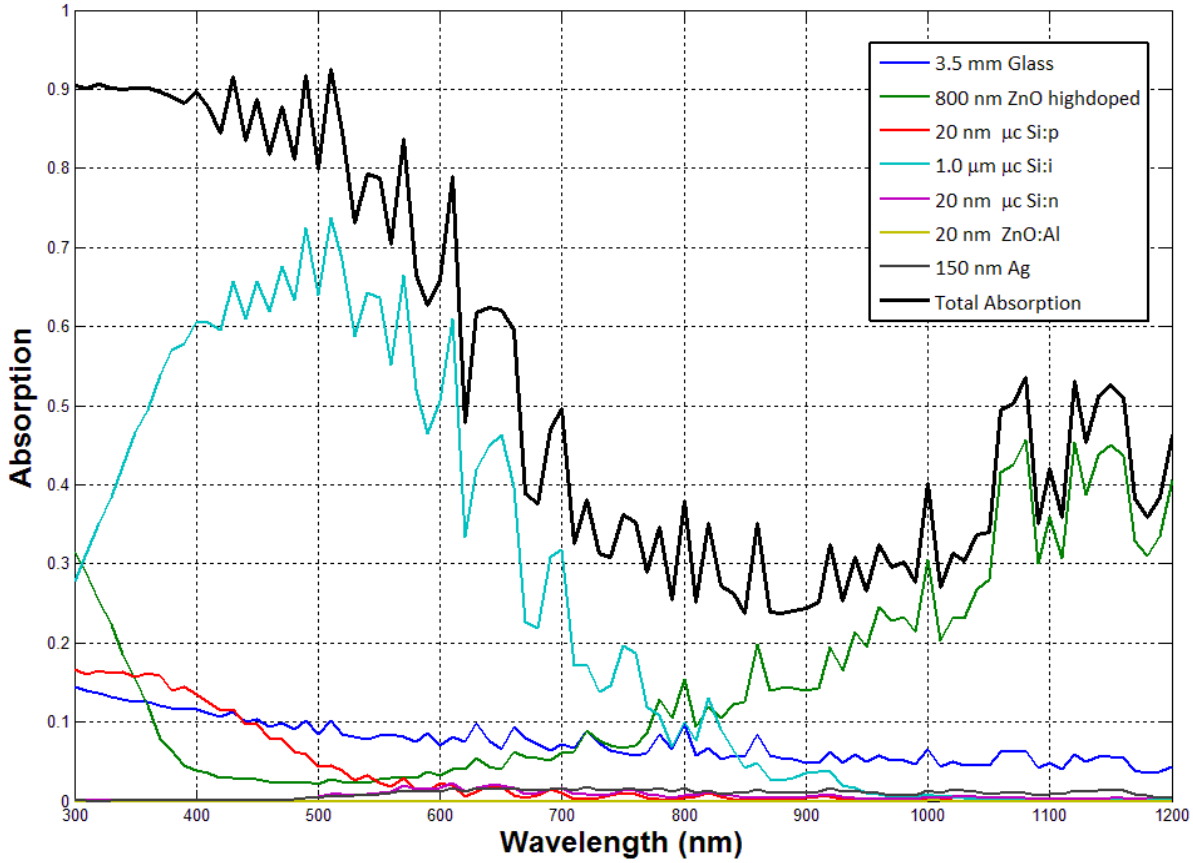


Figure 9.4 – Absorption versus wavelength for each layer for the second cell

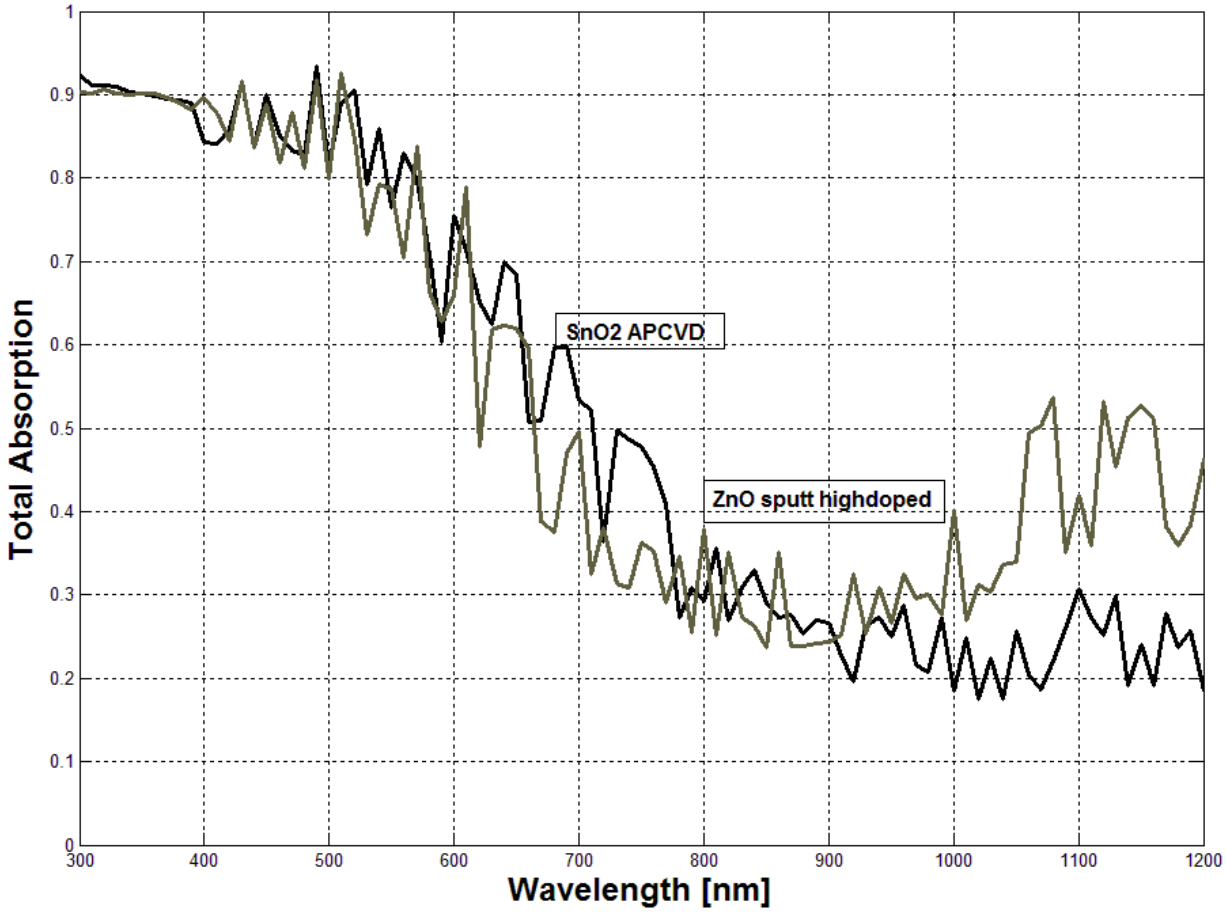


Figure 9.5 - Total absorption for the investigated cells

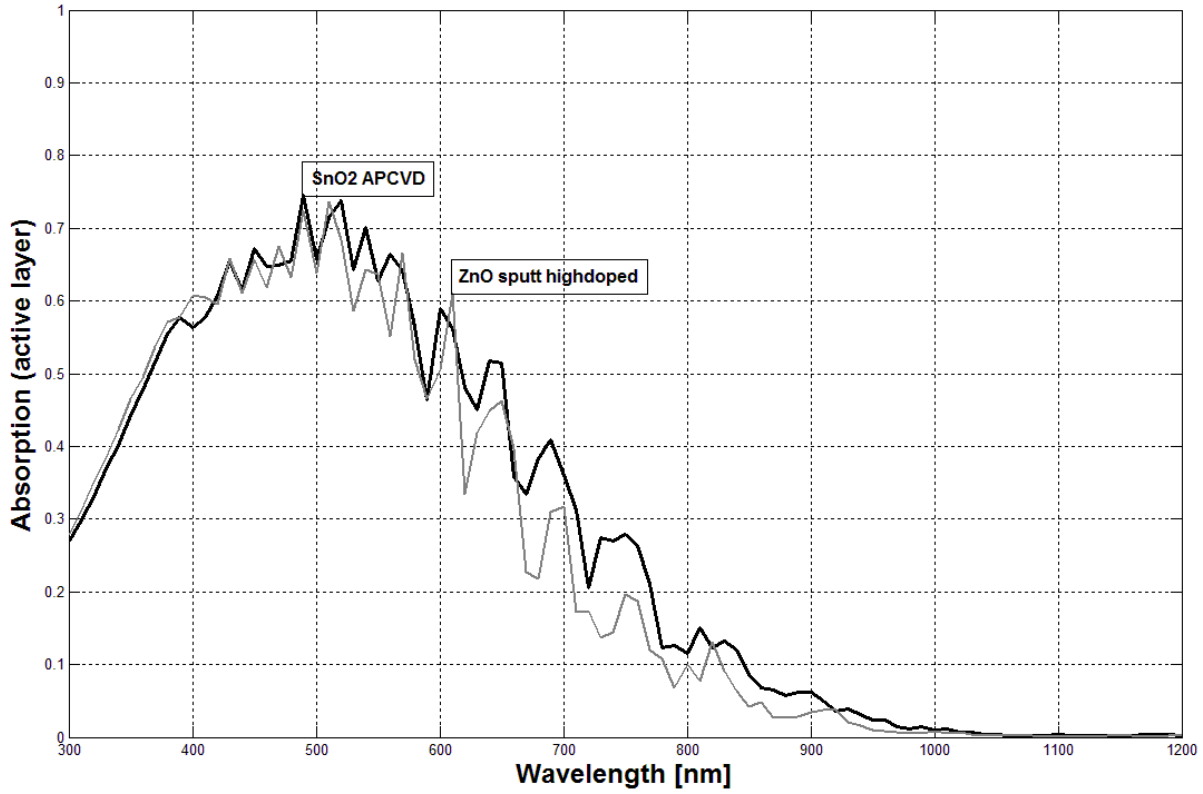


Figure 9.6 - Total absorption for the investigated cells within the active layer

9.5.4 Tandem cells

The materials used for the tandem cell are reported in the following table:

Thickness [nm]	Material	Ruggedness [nm]
350	Glass	$\sigma = 0$
600	SnO_2	$\sigma = 120$
20	a-Si:p	$\sigma = 2$
300	a-Si:i	$\sigma = 30$
20	a-Si:n	$\sigma = 2$
20	SnO_2	$\sigma = 2$
20	μc-Si:p	$\sigma = 2$
$1.7*10^6$	μc-Si:i	$\sigma = 170$
20	μc-Si:n	$\sigma = 2$
20	ZnO:Al	$\sigma = 2$
150	Ag	$\sigma = 15$

Table 9.5 - The chosen materials and the related geometric features for the tandem cell

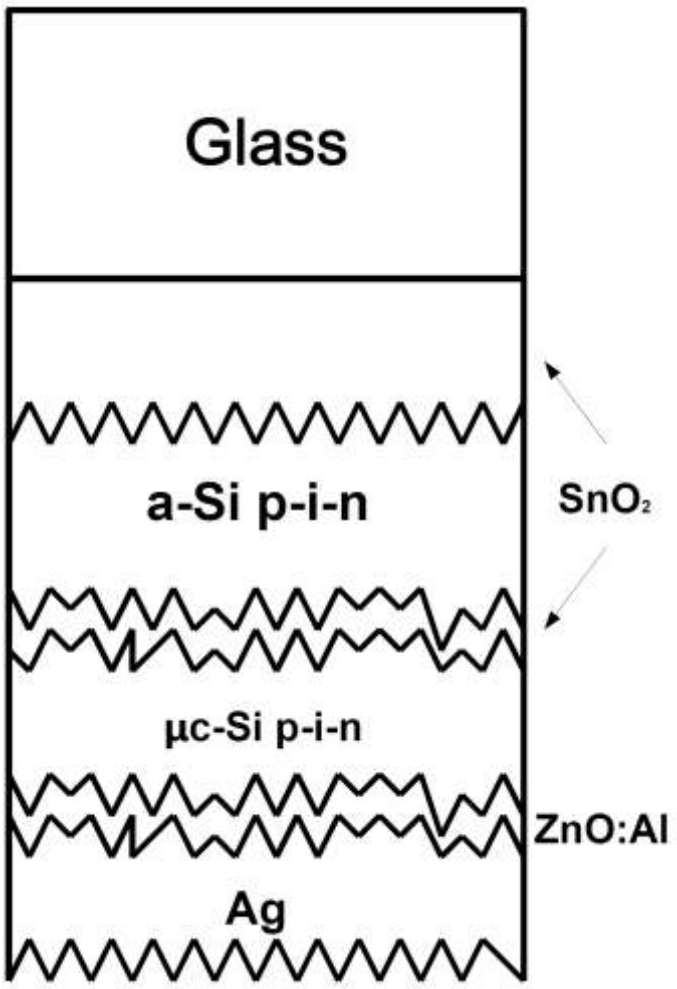


Figure 9.7 – The analyzed tandem cell structure.

When changing some materials and maintaining the same features, we tried to evaluate the following configuration:

Thickness [nm]	Material	Ruggedness [nm]
$3.5*10^6$	Glass	$\sigma = 0$
600	ZnO	$\sigma = 120$
20	a-Si:p	$\sigma = 2$
300	a-Si:i	$\sigma = 30$
20	a-Si:n	$\sigma = 2$
20	ZnO	$\sigma = 2$
20	μc-Si:p	$\sigma = 2$
$1.7*10^3$	μc-Si:i	$\sigma = 170$
20	μc-Si:n	$\sigma = 2$
20	ZnO:Ag	$\sigma = 2$
150	Ag	$\sigma = 15$

Table 9.6 - The chosen materials and the related geometric features for the modified tandem cell

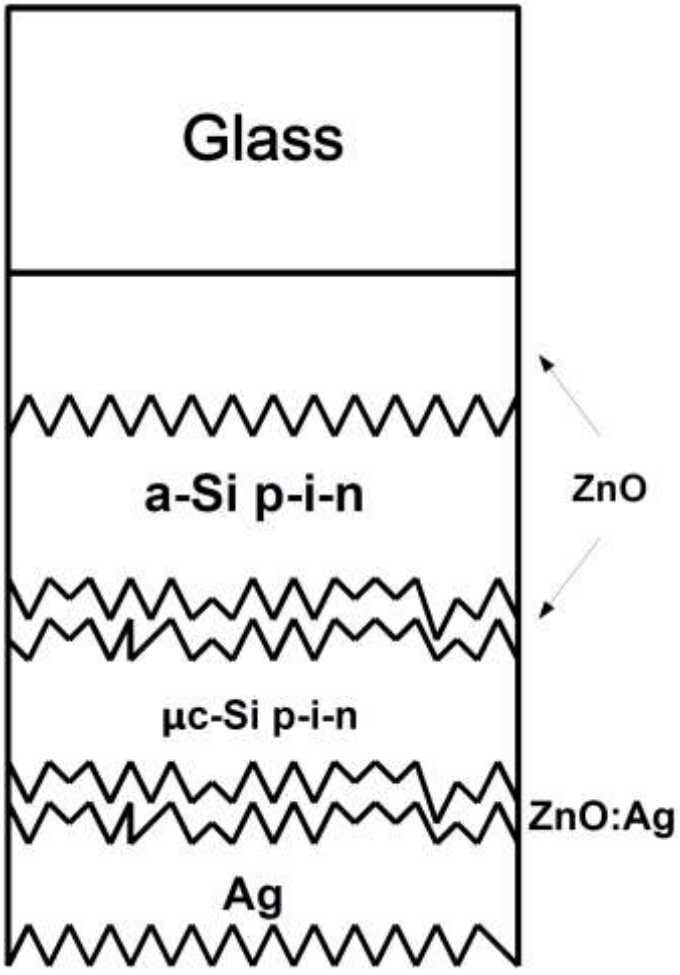


Figure 9.8 – The modified tandem cell

Simulation results (10^5 Monte Carlo particles) are reported in the following figures. The charts represent the absorption of the single layers of the cell, when changing the wavelength.

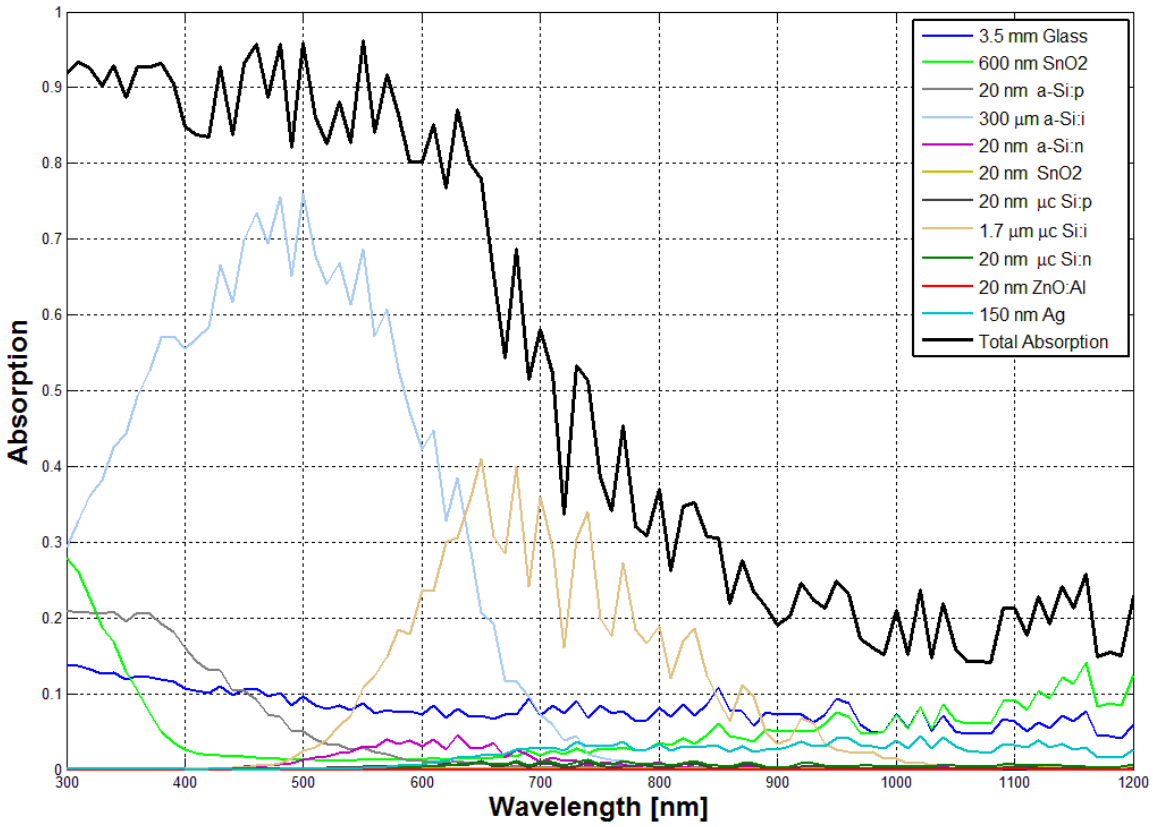


Figure 9. 9 - Absorption for each layer in the first tandem cell evaluated

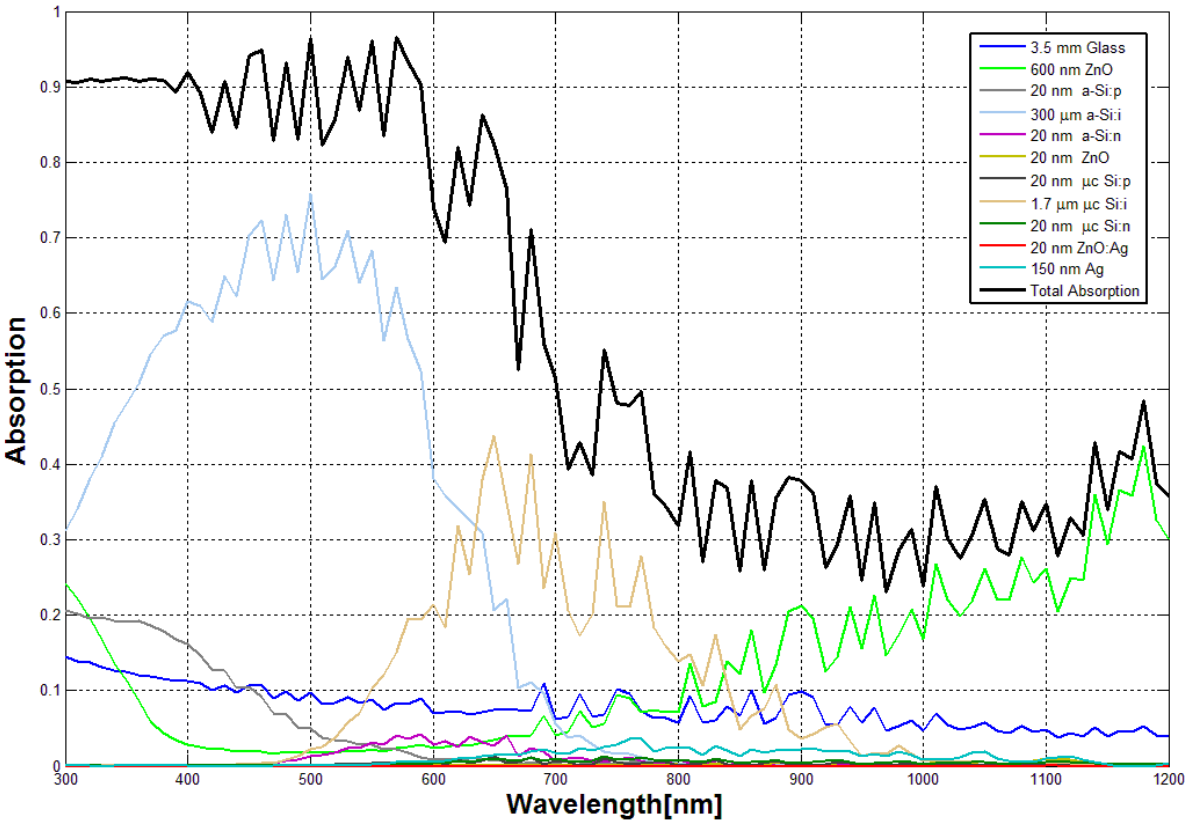


Figure 9.10 - Absorption for each layer in the second tandem cell evaluated

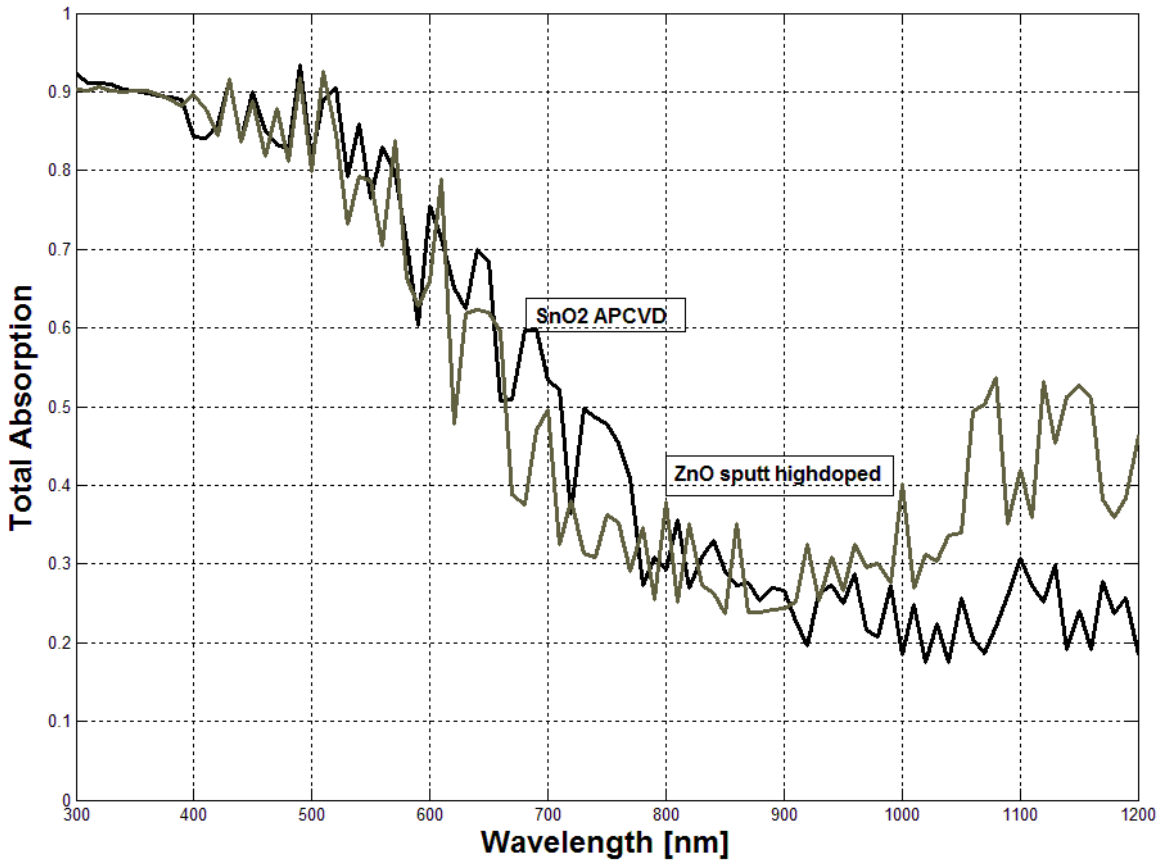


Figure 9.11 - Total absorption for the tandem cells evaluated

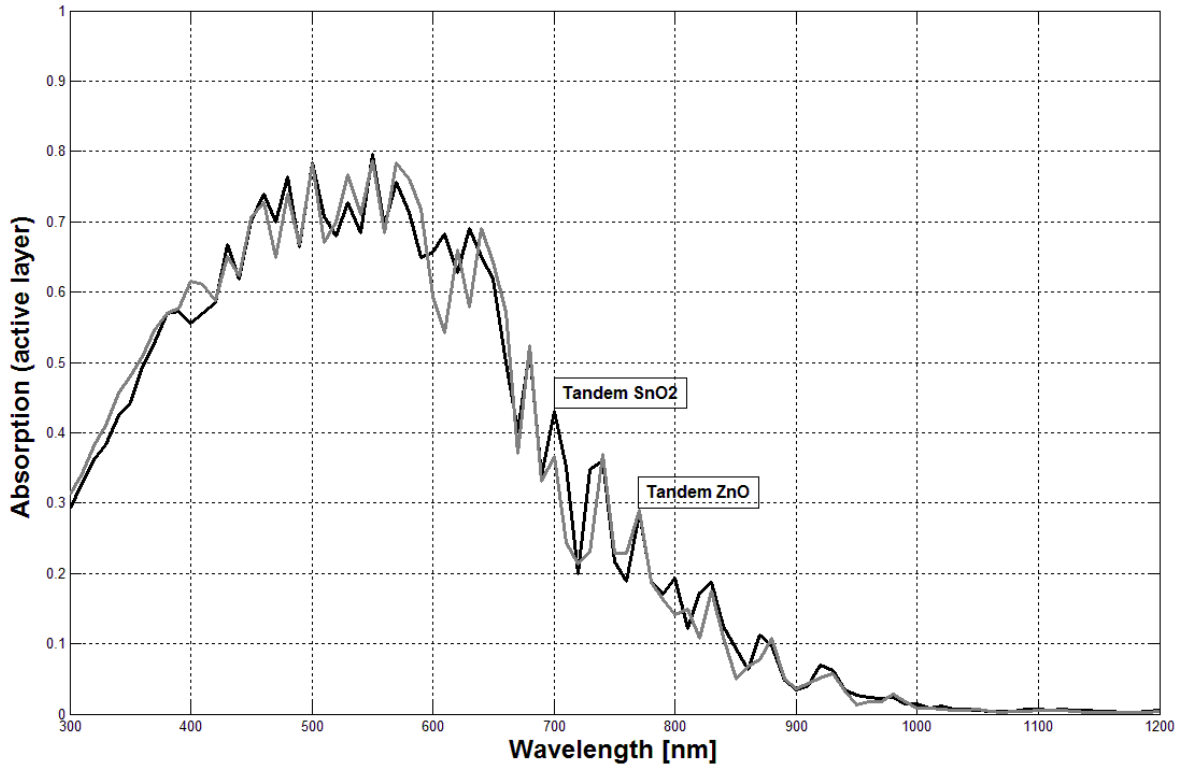


Figure 9.12 - Absorption within the active layer (a-Si:i+μc-Si:i) for the tandem cells evaluated

9.6 Results analysis

The results obtained from the simulations underline similar behaviors in both single junction and tandem cells. In particular, when comparing the total absorption, for both kinds of cells, it is possible to understand that the use of ZnO in the TCO layer, instead of SnO_2, can strongly increase the total absorption, for wavelengths between 900 and 1200 nm. Moreover, the total absorption in tandem cells maintain absorption values between 0.9 and 1.0 for wavelengths between 300 and 600 nm, thanks to the presence of a double active layer embracing amorphous intrinsic (a-Si:i) and microcrystalline intrinsic (μc-Si:i). It is clear that this values are much higher than the total absorption values of the single junction cells.

The results demonstrate that the tandem cell with the use of ZnO appears to be, among all the simulated ones, the one with the best conversion efficiency. That is why it has been chosen as a prototype cell to be implemented in an optimization routine. This optimization model, based on a Genetic Algorithm, is going to be discussed in the following paragraph.

10 Thin-film silicon solar cell optimization through a Genetic Algorithm

10.1 Introduction to Genetic Algorithms

A Genetic Algorithm is a global and stochastic search method based on the emulation of a biological evolution process. In biology, the term "evolution" is related to progressive and continuous modifications that generate, over a long enough period of time, substantial changes in living beings. This process can take place thanks to two events, that are the selective reproduction of new variants and the constant addition of these new variants to the original genetic data set. It is a process using the genetic transmission channel from an individual to his/her children also including the eventual random mutations that interfere with the genetic heritage. Thus, the biological evolution depends on a genetic selection process (random transmission and mutation), that is part of a global and larger natural selection process embracing also all the factors related to the environment adaptation and cooperation with the other organisms. *[51]*

The evolution idea, starting from the biological analogy, indicates, in a more general way, a search method working on a huge number of possible solutions, that are all the organisms able to reproduce themselves and adapt to the environment. Thus, the adaptation idea is the main principle leading the researcher to the best solution, that is the optimal one.

In a similar way, the Genetic Algorithm operate over a large set of potential solutions, and use the "survival" of the best adapted individuals in the current generation, as it is in nature, recombining in a adequate way these solutions in such a way that they can evolve towards the optimum, that is the nearest solution to the real solution of the problem.

Genetic Algorithms are applicable to the resolution of a large number of optimization problems, that cannot be solved through the classical algorithms, including those problems whose objective function is strongly discontinuous, not derivable, stochastic or strongly not linear; in general, they are algorithms suitable to the realization of parallel computations and to the research of a strategy that can choose the further sequences to optimize. These problems require, typically, the research of an optimum among a large set of solutions; research influenced as well by a large number of variables interfering with the solutions themselves (for instance, the Artificial Intelligence research is an example of the application of Genetic Algorithms).

10.2. GA theory in brief

Let us suppose we have a geometrical or physical parameters modifiable set

$$\lambda = \{\lambda_1, \lambda_2 \dots \lambda_k\}$$

whose values can vary within the parameters' space

$$\lambda_i^{min} \leq \lambda_i \leq \lambda_i^{max} \quad for\ i = 1,2, \dots, k$$

let us suppose, moreover, we have an observable measures set

$$g = g\{g_1, g_2 \dots g_M\}$$

where ***g*** depends on the k variables that are contained in the vector

$$g = g(\lambda_1, \lambda_2 \dots \lambda_k)$$

The general problem is to determine the value of the structure's parameters set, subjected to the specified constraints, in such a way that the following objective function would be maximum or minimal:

$$F[g(\lambda)]$$

the optimization algorithms are global search methods according to which, when fixed a set of initial values for the chosen parameters, the algorithm goes on according to a stochastic search towards the finding of the values of the parameters near to the global objective function maximum or minimum (this phenomenon is called convergence). *[7]*
Such algorithms are featured by some important features:

- they are not very dependent on the initial solution
- they do not require the objective function to have particular properties
- they produce a set of sub-optimal solutions (apart from the optimal one)
- they allow us to study problems featured by a large number of parameters
- their convergence is slow
- they require the calibration (based on one's experience) of some simulation parameters

An adequate combination of the optimizing parameters set produce an **individual**, that is a possible solution for the optimization problem.
A solution can be biuniquely codified in a binary code thanks to the J. Holland's (GA inventor) intuition.
The specific sequence (string) made up by 0 and 1 that constitute the individual (candidate solution) is called **chromosome** (codified candidate solution using a string of bits).
It is called **gene** encoding (binary for instance) of the λ parameter:

$$B_r = \left[\lambda_r^{min}, \lambda_r^{max}\right] \rightarrow \left\{\left(\alpha_1, \alpha_{n_r}\right): \alpha_j \in [0,1], \qquad j = 1,2, \dots n_r\right\}$$

$$with \sum_{r=1}^{k} n_r = n$$

The chromosome will be the sequence of the single individual's genes

$$p_n = \{B(\lambda_1), B(\lambda_2), \dots B(\lambda_k)\}$$

A set of individuals would be a **population**:

$$P = \{p_1, p_2, \dots, p_N\}$$

Considering that the algorithm leads to a temporal evolution of the population, we define **generation**, the population at a given time.
In nature, individuals reproduce themselves and mix their genes transmitting to the newborn individuals a new genetic heritage, that is a combination of the mother and father's one. The natural selection, that is the choice and reuse of either the best or the best adapted solutions, makes the strongest individuals survive and reproduce themselves, generating this way again the best adapted individuals or the one featured by the highest fitness to the environment (the solutions nearest to the optimum). Over the time, the average population fitness, going on along this selection criterion, will tend to increase generation after generation, determining the population's evolution.
We define **evolution** the iteration of the optimization process that allows us to modify the population's individual genes through a sequence of operators

$$P^i \rightarrow P^{i+1}$$

The parameter that allows us to evaluate the goodness of the found solution is defined **fitness**. Usually, the total fitness is defined as the weighted average of the fitness functions associated to the different observable measures:

$$F_p = F(p_n) = \sum_{j=1}^{J} w_j F_j [g_j(\lambda_n)]$$

where w_j are the weights, whose sum is

$$\sum_{j=1}^{J} w_j = 1$$

The theoretical structure for a GA is the following one: *[51]*

- Initial generation definition.
  Definition, even if randomly, of a first set of possible solutions for the relevant problem
- Evaluation of each solution and selection of the best one.
  Evaluation of each solution, associating at each of them a quality indicator (or fitness), in such a way that it is possible to sort them
- Definition of a new generation.
  Definition of a new group of solutions, through the adequate modification of the solutions with the highest quality, in order to make them evolve instead of the worst ones
- Conclusion of the elaboration.
  If either the defined number of iterations has been reached, or the best available solution quality is acceptable, it is possible to stop the algorithm, otherwise, it is needed to go again to the second step in order to define a new group of solutions. In the first case, the best individuals are defined "parents" of the following generation.

Starting from these individuals, an even number of individuals belonging to the next generation is generated. This process can take place through two genetic operators, whose task is to combine

genes (elements making up the chromosomes) of the different solutions, in order to explore the new ones:

- Mutation.
  It introduces in an aleatory way new genes in some chromosomes, creating individuals with completely new features. Thus, some chromosomes are randomly selected (according to the mutation probability) and in these chromosomes one or ore bits are changed randomly.
- Crossover.
  Emulating the reproduction, it realizes the both parents' genes combination.
  Let us look at the following example:

Elimination	ABCDE - FGH → ABCE - FGH	
Duplication	ABCDE - FGH → ABCBCDE - FGH	
Inversion	ABCDE - FGH → ADCBE - FGH	
Mutual translocation	ABCDE - FGH MNOCDE - FGH	

	MNOPQ - R ABPQ - R	
Crossover at one point		
(parent number 1)	011010 - 10100	(son number 1) 011010 - 01010
(parent number 2)	011101 - 01010	(son number 2) 011101 - 10100

The algorithm goes on until we do not get the convergence of the features, that is when the parents extracted from a population cannot be further improved. Generally, when starting from a starting population, a genetic algorithm produces new generations usually containing solutions better than the previous ones. If it happens, it is gained a general evolution of the generations towards the fitness function global optimum.
Thus, the generation process is iterated until a stopping condition is met. The most common ones are:

- a solution satisfying the minimum criteria is found
- the maximum number of generations declared at the beginning is reached
- the economic/time limit defined at the beginning is reached
- a manual control on the algorithm stop is created
- the best result already gained do not evolve until a given number of generations elapses
- within a given number of generations no individual within the solution space is found (the problem cannot be solved).

The following model could be considered for a GA:

Choose an initial population of chromosomes;
while termination condition not satisfied ***do***
repeat
if crossover condition satisfied ***then***
{select parent chromosomes;

choose crossover parameters;
perform crossover}
if mutation condition satisfied ***then***
{choose mutation points;
perform mutation};
evaluate fitness of offspring
until sufficient offspring created;
select new population:
endwhile

10.3 The optimization problem

According to the results gained in the previous section, the tandem cell is the best one as for the total absorption, among all the simulated cells. It is possible to further improve these starting values, through the modifications of thickness and roughness values of the different tandem cell layers. By increasing the thickness in the different layers and, especially, in the active layers, (in this case a-Si:I and μc-Si:i), it is possible to notice a proportional increase of the absorbed radiance. *[23]* Thus, the optimization and increase of the radiance could seem a banally solved problem. Actually, this thickness increase cannot be carried out without taking into account the economic cost of the manufacturing process to get materials like the crystalline silicon. That is why, the thickness increase must be accurately investigated. This last statement gives us some real production limits influencing the cell performance.
The optimization problem in our case has the target to find a good balance between the layer thickness and the production cost.
In order to get a realistic cost, since we do not know the real production cost, the following calculations will be based on these assumptions:

- Photovoltaic equipment power = 3 kW
- Income coming from the equipment sale = 10 k€
- Number of photovoltaic panels = 20
- Income coming from the single panel sale (Y) = 500 €
- Production cost for the single panel (C) = 250 €
- Net profit R_u = 250 €

10.4 Mathematical formalization

The optimization problem, formulated as above, requires the search of an optimum that should be a trade-off between the thickness variation and the layers materials production costs. So, the function to optimize is in this case a multiobjective function:

$$\max \ \{R_u = Y - C\}$$

while we want $\max Y \ \ and \ \ \min C$
with $Y = \alpha Q_e; \ \ C = T_{proc} R_c$
where α is a proportional factor, Q_e is the quantic efficiency and T_{proc} is the time needed to produce a film:

$$T_{proc} = T_{fixed} + T_{var} = T_{fixed} + \frac{X}{v}$$

where

- X is the thickness of the intrinsic layer in nm
- v is the velocity of the process = 0.2 nm/sec
- R_c is a "cost rate" expressed in [€/ton]

Moreover, let us assume that the ratio between T_{fixed} and T_{var} is 1:2.

Now, let us have a look at the following figure, representing the tandem cell to be optimized together with the variables chosen for the problem.

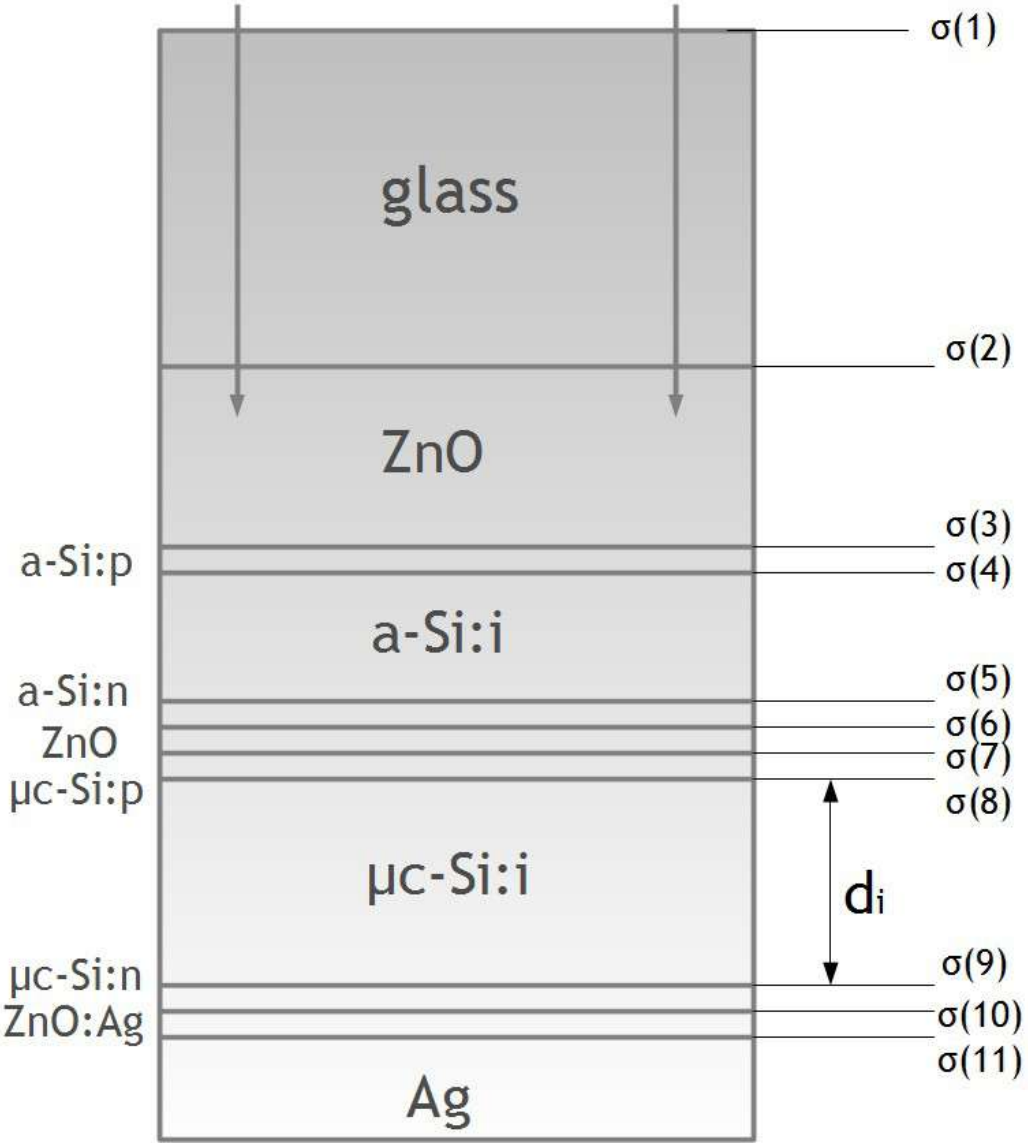


Figure 10.1 – Scheme for a tandem cell to be optimized together with the chosen variables.

Thus, we have chosen as our variables the thickness of the μc-Si:i layer and the roughness of the tandem cell layers, for a total amount of 12 variables.
Taking into account the material cost when choosing the intrinsic layer thickness and the fact that it is not possible to use on the interfaces a too high roughness value with respect the thickness itself, the following **constraints** have been imposed:

Interface	Roughness [nm]
1	$\sigma(1) = 0$
2	$d^*(2) \leq \sigma(2) \leq d^*(2)+d^*(2)*0.3$
I=3:11	$d^*(1) \leq \sigma(i) \leq d^*(i)+ d^*(i)*0.2$

Table 10.1 – Constraints imposed for the problem.

For the layer μc-Si:i thickness (d_i) is assumed the following: $(d_i - d_i * 0.3) \leq d_i \leq (d_i + d_i * 0.3)$, where d_i is the initial thickness.

10.5 Simulation and results

The optimization problem proposed in the previous paragraph has been performed using a Matlab simulation code, with the algorithm used in the section dedicated to the GAs on the multiobjective functions. *[12]*
The code calibration has been carried out by using:

- A population made up by 20 individuals
- 50 Generations as a maximum

The obtained results are the following ones:

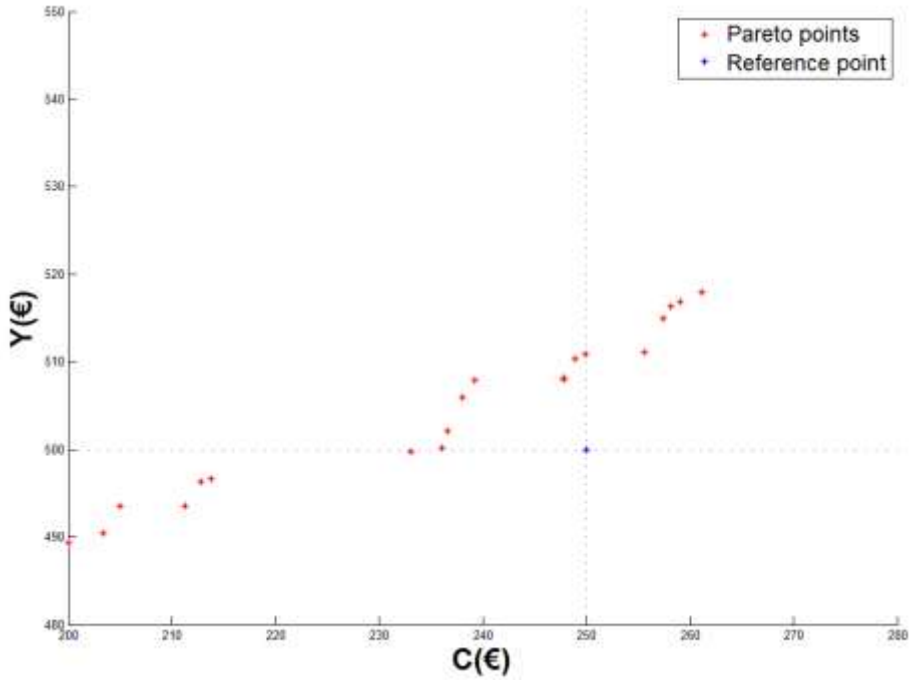


Figure 10.2 – Results obtained at the last generation, with respect to the initial reference point.

It is possible to notice that the reference point is (250, 500). In the chart we can find the results obtained from the last generation, while, in the following figure, it is possible to look at the results extracted from the last generation and at some selected points (in yellow):

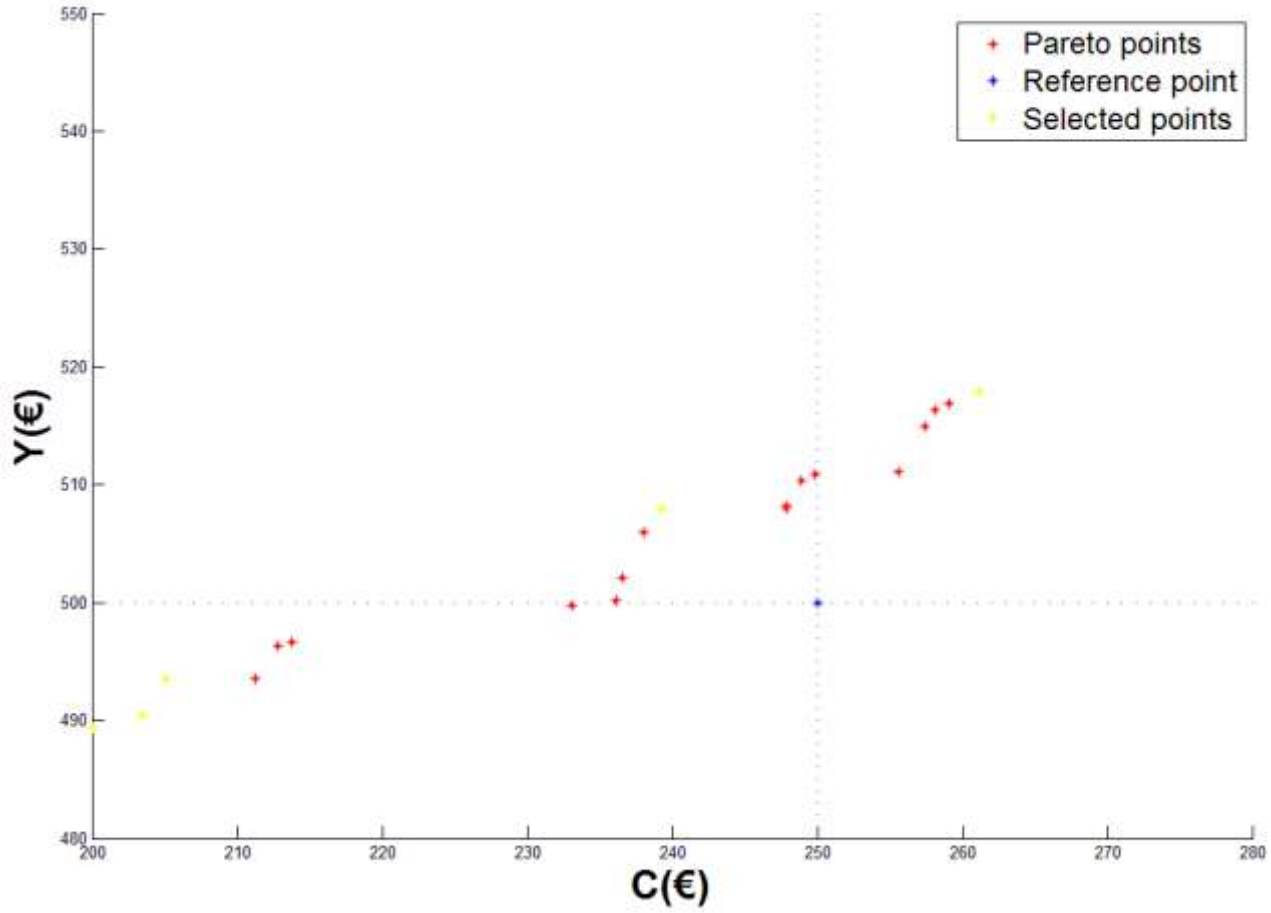


Figure 10.3 – Results obtained after the last generation, with special attention to the selected points (in yellow).

10.6 Results analysis

To better understand the Pareto charts shown before, it is necessary to define what we mean when referring to the optimal solution for a multiobjective programming problem.

If the solver has already found a Pareto optimum and he wants to further decrease the value of one or more objective functions, we have to be willing to accept a consequent increase in one or more of the other problem functions. Thus, we can state that, within the objective space, the Pareto optima are "*equilibrium points*" on the boundary of the image Z of the feasible region within the objective space ($Z = f(F)$). In the chart above, we can notice that, respect to the reference point (250, 500), the chart is divided in four parts. The point that improves, simultaneously, both Y and C is the one with (239, 508) as coordinates. This point, in yellow, is in the left-upper part of the chart.

Both the variables, Y and C, actually increase simultaneously. While, in the right-upper part of the chart, the Y values improve and the C values get worse and in the left-upper part the C values improve and the Y values worsen. This is the analytical demonstration of what we just stated about the Pareto optima.

11 Multiobjective optimization for effective solar cell design

Introduction

The method we are going to introduce is focused on the analysis and optimization of a solar cell. It is composed, essentially, by four algorithms. *[12]*

- Morris algorithm. Focused on sensitivity analysis, to analyze the parameters that are influential on the output and to select the sensitive parameters for the optimization.
- Multiobjective optimization. To optimize simultaneously the key objectives: fill factor and efficiency.
- Decision making. To select the best solutions of the optimization.
- Robustness analysis. To assess the robustness of the candidate solutions.

This methodology has been integrated with Synopsys TCAD Sentaurus and it has been tested and a 2D model for an homogeneous emitter (HE) solar cell.

11.1 The optimization algorithm

The new Electronic Design methodology chosen is based on Pareto Optimality, and it is called PAREDA. The proposed optimization algorithm is a stochastic *black-box* optimization algorithm inspired on the clonal selection principle derived from the immune system. A problem is an **antigen** and a candidate solution is a **B-cell**. The affinity between an antigen and a B-cell is given by the objective function(s) of the optimization problem. Each B-cell is thus a vector of n real values, where n is the dimension of the problem. Each candidate solution has associated an age τ that indicates the number of iterations since the last successful mutation and, initially, it is set to zero. An initial population $P^{(0)}$ of dimension d is randomly generated, with each variable constrained in the bounds. However, it could be useful to use and ad-hoc population to start the optimization process (for instance, performing a *local optimization*). PAREDA can take in input a *starting point* p_{st} and it uses this point to initialize one candidate solution of the population and the remaining *d-1* with perturbation of p_{st}. The algorithm is iterative, each iteration is made of a cloning, a mutation and a selection phase.

The algorithm stops when the maximum number of objective function evaluations is reached. The cloning phase is responsible for the production of copies of the candidate solutions. Each member of the population is cloned *dup* times producing a population P_{clo} of size *d x dup*, where each cloned candidate solutions takes the same age of its parent, simultaneously, the age of the parent is increased by one.

After the P_{clo} population is created, it undergoes to the mutation phase in order to find better solutions; in this phase, the hyper-mutation and hyper-macromutation are applied to each candidate solution. Firstly, Firstly, the hyper-mutation operator mutates a randomly chosen variable x_i of a given candidate solution using a *self-adaptive Gaussian mutation* computed as

$$x_i^{new} = x_i + \sigma N(0,1), \;\; where \;\; \sigma_i^{'} = \sigma_i e^{\left[(\tau N(0,1))+(\tau^{'} N_i(0,1))\right]}$$

Successively, the hyper-macromutation applies a convex perturbation to a given solution by setting

$$x_i^{new} = (1-\gamma)x_i + \gamma x_k$$

where x_i is a variable randomly chosen such that $x_i \neq x_k$ with $\gamma \in [0,1]$ a uniformly distributed random variable. Since variables x_i and x_k typically have different ranges, the valτue x_k is normalized within the range of x_i using the following equation:

$$x_k^{norm} = L_i + \frac{(x_k - L_k)}{(U_k - L_k)}(L_i - U_i) \qquad (1)$$

where L_i and U_i are the lower and upper bounds of x_i and L_k and U_k are lower and upper bounds of x_k. The value used to mutate the x_i variable is x_k^{norm}. These mutation operators are controlled by the specific mutation rate α; for the hyper-mutation we define $\alpha = e^{-\rho f}$, instead for the hyper macromutation we adopted $\alpha = \frac{1}{\beta} e^{-f}$ where f is the objective function value normalized in [0,1].
These operators are applied sequentially; the hyper-mutation operator acts on the P_{clo} producing a new population P_{hyp}. The hyper-macromutation mutates P_{hyp} generating the P_{macro} population. After mutations, the population P_{macro} is evaluated; if a candidate solution achieves a better objective function value, its age is set to zero otherwise it is increased by one. The *aging* operator is applied on $P^{(t)}$ and P_{macro}; it erases candidate solutions with an age greater than τ_b+1, where τ_b is a parameter of the algorithm. The deleted candidate solutions are stored into the archive BC_{arch}; since the archive contains at most s_a solutions, if there is enough space, the candidate solution is out into the first available locations, otherwise it is put in a random location. Finally, the selection is performed and the new population $P^{(t+1)}$ is created by picking the best individuals form the parents and the mutated candidate solutions; however, if $\left|P^{(t+1)}\right| < d$, $d - \left|P^{(t+1)}\right|$ candidate solutions are randomly picked from the archive and added to the new population. In many real applications, it is common to deal with constraints, which could be imposed on input and output value. In general, a constraint is a function *g(x)* that certificates if a solution for a given optimization problem is feasible or not. We consider constraints defined as $g(x): R^n \rightarrow R \;\; if \;\; g(x) \leq \theta$ where θ is a feasibility threshold. The algorithm considers the constraint values during the selection procedure. Given two individuals p_1 and p_2, if both are feasible the one with the lowest objective function value is picked; if p_1 is feasible and p_2 is unfeasible, p_1 is chosen, otherwise if p_1 and p_2 are unfeasible, the one with the lowest constraints violation is selected. In order to face and solve the circuit design problems of the next paragraphs, the parameters have been set to:

- $d = 20$
- $dup = 2$
- $\tau_b = 50$
- $\rho = 1$
- $\beta = 7$
- $s_a = 160$

to select the more robust and effective design, the optimization algorithm has been integrated with the following pre and post-processing methods:

- *sensitivity*
- *epsilon-dominance*
- *robustness analysis*

We will not give details on this methods, since here we just want to underline their utility in this work. In order to show how the methods have been integrated, we will show an example of circuit sizing.

11.2 Sensitivity analysis

It allows determining if there are some values of the parameter vector that do not affect the performances. In particular this analysis gives a value of mean and standard deviation for each parameter variations, where a high mean indicates a parameter with an important overall influence on the output and a high standard deviation indicates that either the factor is interacting with the others factors or the factor has nonlinear effects on the output. The output considered are the two objective functions. The following figure shows the results of the Sensitivity Analysis (SA) with respect to the first objective function (A), to the second one (B) and to the normalized sum of both of them (C). In this case all the considered parameters influence the two objective functions, but in some cases if a parameter has the values of μ and σ near to zero, they can be neglected in the next step, that is the optimization phase. So, when this step begins, the algorithm searches for the best designs that respect the specifications, therefore, it produces as output several feasible points, i.e. points that satisfies the constraints on the performance.

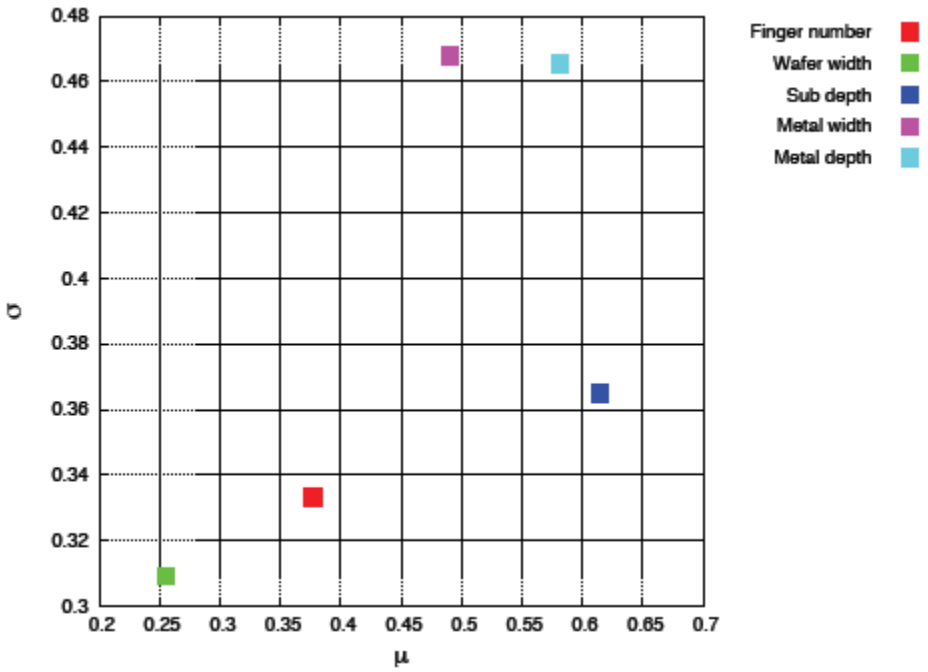


Figure 11.1 – Sensititvity analysis with respect to the relevant objective functions.

11.3 Robustness analysis

In order to add information and to select robust solutions, a further step is executed; the robustness analysis. For each solution is assigned an index of robustness called *Global Robustness* (GR). It is calculated by perturbing N times all the variables of a given candidate solution with Gaussian noise with zero mean and a standard deviation equal to 1% of parameters. The samples whose performance do not deviate more than 1% from the original values are considered robust and the value of GR is calculated as the ratio between robust samples and the total ones. At this point we can choose the most efficient and robust design. In particular, the following solutions have been selected: the solution minimizing the power consumption, robustness/power consumptions, gain/robustness and, finally, gain/power consumption/robustness. The first three ones are immediate, while the other ones are calculated by normalizing the objective functions and the robustness values, calculating the distance from the ideal point (*GR=1, PowerConsumption = 0, Gain =1)* and choosing the solutions at the minimum distance.

The last three selected points are optimal and robust circuits. It is important to underline that the best trade-off between the minimum power consumption, the maximum gain and robustness is an epsilon-dominance point. This shows how epsilon-dominance analysis combined with the robustness analysis, help to select an optimal and robust design from the feasible points found by the optimization algorithm.

Finally, it is important to note that the Robustness analysis algorithm has the same computational effort of the optimization algorithm for the nominal design problems; for both methods, the stopping criterion is the same: the maximum number of simulations performed.

11.4 Identifiability analysis

In statistics, identifiability is a property which a model must satisfy in order for precise inference to be possible. We say that the model is identifiable if it is theoretically possible to learn the true value of this model's underlying parameter after obtaining an infinite number of observations from it. Mathematically, this is equivalent to saying that different values of the parameter must generate different probability distributions of the observable variables. Usually the model is identifiable only under certain technical restrictions, in which case the set of these requirements is called the identification conditions.

A model that fails to be identifiable is said to be non-identifiable or unidentifiable. In some cases, even though a model is non-identifiable, it is still possible to learn the true values of a certain subset of the model parameters. In this case we say that the model is partially identifiable. In other cases it may be possible to learn the location of the true parameter up to a certain finite region of the parameter space, in which case the model is set identifiable.

Let $\wp = \{P_\theta : \theta \in \Theta\}$ be a statistical model where the parameter space Θ is either finite- or infinite-dimensional. We say that $\wp$ is **identifiable** if the mapping $\theta \mapsto P_\theta$ is one-to-one:

$$P_{\theta_1} = P_{\theta_2} \Longrightarrow \theta_1 = \theta_2 \ \ for\ all\ \theta_1, \theta_2 \in \Theta$$

This definition means that distinct values of θ should correspond to distinct probability distributions: if $\theta_1 \neq \theta_2$, then also $P_{\theta 1} \neq P_{\theta 2}$. If the distributions are defined in terms of the probability density functions, then two pdfs should be considered distinct only if they differ on a set of non-

zero measure (for example two functions $f_1(x)=\mathbf{1}_{0\leq x<1}$ and $f_2(x)=\mathbf{1}_{0\leq x\leq 1}$ differ only at a single point $x=1$ — a set of measure zero — and thus cannot be considered as distinct pdfs).
Identifiability of the model in the sense of invertibility of the map $\theta \mapsto P_\theta$ is equivalent to being able to learn the model's true parameter if the model can be observed indefinitely long.
Thus with an infinite number of observations we will be able to find the true probability distribution P_0 in the model, and since the identifiability condition above requires that the map $\theta \mapsto P_\theta$ be invertible, we will also be able to find the true value of the parameter which generated given distribution P_0.
So, we have performed the identifiability analysis to the Pareto optimal designs of the solar cell obtained during the optimization process as follows:

Parameter	p_i	r^2	cv	♯	Parameters Groups
No. of fingers	p_1	0.944	0.161	5	p1,p2,p3,p4,p5**
Wafer width	p_2	0.939	0.053	5	p1,p2,p3,p4,p5
Substrate depth	p_3	0.885	0.101	5	p1,p2,p3,p4,p5
Finger width	p_4	0.919	0.498	5	p1,p2,p3,p4,p5**
Finger depth	p_5	0.890	0.508	5	p1,p2,p3,p4,p5

Table 11.1 – identifiability analysis applied to the Pareto optimal design (non-dominated points) obtained during the optimization process.

and then, the same analysis have been applied as follows to the Pareto optimal and the ε-non dominated ($\varepsilon = 10^{-5}$) designs of the solar cell model obtained during the optimization process:

Parameter	p_i	r^2	cv	♯	Parameters Groups
No. of fingers	p_1	0.950	0.149	5	p1,p2,p3,p4,p5**
Wafer width	p_2	0.956	0.053	5	p1,p2,p3,p4,p5
Substrate depth	p_3	0.908	0.099	5	p1,p2,p3,p4,p5
Finger width	p_4	0.943	0.525	5	p1,p2,p3,p4,p5**
Finger depth	p_5	0.890	0.524	5	p1,p2,p3,p4,p5

Table 11.2 – Identifiability analysis applied to the Pareto optimal (non-dominated points) and the ϵ-non-dominated ($\epsilon = 10^{-5}$) designs of the solar cell model obtained during the optimization process.

where:

- p_i is the index of the functional relation according to the index of the response parameter
- r^2 indicates how much variance of the response can be explained by the predictor
- $cv = \frac{std\,(p)}{mean\,(p)}$ helps us to practically distinguish identifiable from non-identifiable parameters
- # indicates how often has this special tuple been found
- The parameters group indicated the tuples assigned to have a functional relation

the function relations are ranked according to the following criteria:

- the more often a functional relation has been found (#), the better
- the more variance of the response can be explained by the predictors (r^2), the larger the effect of the fixation of parameters on the standard deviations of the fitted parameters. Values of $r^2 > 0.9$ are very good and $cv > 0.1$ are given one. If, additionally, the functional relation has been found more than once, another * is assigned.

11.5 Experimental results

We change the parameters one at a time (OAT one-at-a-time design) several times and measure the mean and variance of the effect produced on the output. The parameters used are number of fingers, wafer width, substrate depth, metal width and metal depth. All the parameters result influential on the output. High mean indicates linear influence on the output. high standard deviation shows a non-linear influence on the output or a relation with another parameter.

A derivative-free, evolutionary, immunological algorithm has been used. It can handle discrete, integer and real variables and can face Constrained Single and Multi-Objective Optimization problems. The main features of the algorithm are:

- the cloning operator, which explores the neighborhood of a given solution
- the inversely proportional hypermutation operator, which perturbates each candidate solution as a function of its objective function value
- the aging operator, that eliminates the oldest candidate solutions from the current population in order to introduce diversity and thus avoiding local minima during the search process

The parameters selected are the same of the sensitivity analysis, their ranges are:

Parameter	Range	Step
Number of finger	25-125	1
Width wafer	136000-186000 μm	100 μm
Depth substrate	90-270 μm	0.01μm
Width metal	10-150 μm	0.01μm
Depth metal	0.01-0.10 μm	0.0001 μm

The objective functions are:

- Maximize Fill Factor
- Maximize Efficiency

the Pareto optimal points are shown in the figure below, together with the points that maximize both the objective functions, the best trade-off (by calculating the distance from the ideal point and choosing the points at minimum distance) and the other sampled and tested.

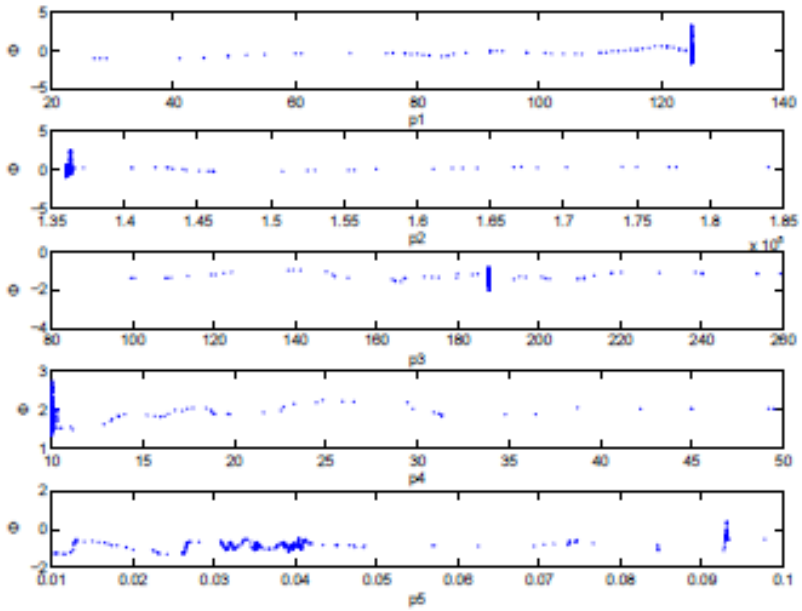


Figure 11.3 – Optimal transformations Θ (y axis) found for the five parameters p_i, i=1,…,5 (x axis) using the non-dominated points obtained during the optimization process. This plot shows the relation among these five parameters.

Conclusions

The development of this thesis has led to some innovative results.
The definition of a geometrical structure for a homogeneous emitter solar cell is able to improve, from a modeling standpoint, the efficiency performances of the analyzed cell.
The considered structure, taking into account a p-n^+ junction, starts from some geometrical parameters that are the initial values for the design variables. Then, through the use of a genetic algorithm combined with the TCAD Sentaurus tool for the mathematical modelization, the optimization starts and it has been left going on until it led to the solution of the multiobjective problem. So, after 300 generations, some results improving the literature values for the considered cell have been found. We have been able to push the cell until an efficiency of 20.65% and a fill factor of 0.83, while the literature reference structure barely get a 17% efficiency.
This is also a design issue because the algorithm involves the contacts design in order to maximize the current at collected at the contacts without increasing the natural shadowing effect.
So, Efficiency and Fill Factor have been chosen as objective functions while the geometrical parameters have been chosen as decision variables.
The second important result for the cell design has been obtained on the thin-film structures.
Under the pressure for cost reduction in photovoltaic industry, due to the large development of large scale panels manufacturing, the thin-film technology arises, as a way to reduce the wafer thickness to save material. These kind of cells are made up by many layers, once with a given thickness and roughness. In these cells, as it is clear, an efficient design is more important than ever.
The simulation in this case, consisted in a two steps module. The first one calculates the electromagnetic radiation through the matrix method, the second one uses the Monte Carlo method to calculate the light's diffused component. After that, again, a genetic algorithm has been used to optimize the cell behavior. In this step, the absorption of the cell for the different wavelength, coming from the previous optical calibration, led to a profit maximization problem, bound to the minimization of production cost for the panels against the layer thickness.
The results are related to the optical model for a thin-film cell. An open issue for future development of this work could be the inclusion of an electrical module along with the optical model gained from the calibration already performed in order to also evaluate the electrical effects.

References

1. Jeffery L. Gray, Blushan Sopori - *Handbook of photovoltaic science* – John Wiley & Sons Ltd, Chichester, West Sussex, England – 2003.

2. Hartmut Haug, Stephan W. Koch – *Quantum theory of the optical and electronic properties of semiconductors, 4th edition* – World Scientific Publishing Co. Pte. Ltd., Singapore – 2004.

3. Sheng S. Li – *Semiconductor Physical Electronics 2nd edition* – Springer Science+Business Media, LLC – 2006.

4. M.J.D. Powell – *Direct search algorithm for optimization calculations* – Dept. of Applied Mathematics and Theoretical Physics, University of Cambridge, England – March 1998.

5. Stefano Lucidi – *Corso di ottimizzazione* – Dipartimento di Informatica e Sistemistica, Università "La Sapienza", Roma – February 1999.

6. Tamara G. Kolda, Robert Michael Lewis, Virginia Torczon – *Optimization by direct search: new perspectives on some classical and modern methods* – Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, USA – 2003.

7. M.J.D. Powell – *A view of algorithm for optimization without derivatives* - Dept. of Applied Mathematics and Theoretical Physics, University of Cambridge, England – April 2007.

8. G. Liuzzi – *Programmazione multiobiettivo* – Dipartimento di Ingegneria Informatica, Automatica e Gestionale, Università "La Sapienza", Roma – 1999

9. G. Carapezza, A. Greco – *TCAD Sentaurus, modello cella solare a emettitore omogeneo* – Dipartimento di Matematica e Informatica, Università degli Studi di Catania – 2013

10. M. Zanuccoli, P.F.Bresciani, M. Frei, H.W. Guo, H. Fang, M. Agrawai, C. Fiegna, E. Sangiorgi – *Numerical simulation and modeling of monocrystalline selective emitter solar cells* - Bologna University, IUNET (Cesena), Applied Materials Inc., Santa Clara, USA – 2010

11. M. Zanuccoli, M. Frei, H.W. Guo, M. Agrawai, C. Fiegna, E. Sangiorgi – *Numerical simulation and modeling of rear point contact solar cells* - Bologna University, IUNET (Cesena), Applied Materials Inc., Santa Clara, USA – 2010

12. G. Carapezza, A. Greco, A. La Magna, G. Nicosia, V. Romano – *Optimization of thin-film solar cells* – (Preprint) – Dipartimento di Matematica ed Informatica, Università degli Studi di Catania – 2013

13. M. Echart – *Financing solar energy in the U.S.* – Scoping Paper – 1999

14. K. Jechoutec et alii – *The solar initiative* – World Bank – 1995

15. P. Sickenberger – *KfW Support Schemes for PV Financing in Germany* – 17th European PV Conference – Munich 2001

16. *Rural Energy and Development: improving energy supplies to two billion people* – World Bank 1996

17. *Solar Generation* – GreenPeace, for the European PV Industry Association – 2001

18. J. Wiles – *PV Power systems and the National electric code: suggested practices* – Sandia Nat Labs, Albuquerque, USA – 2001

19. D. Osborn – *Sustainable Orderly Development: the SMUD Experience* – UPEX – 2001

20. J. Gregory – *Financing Renewable Energy Projects: a guide for development workers* – Stockholm Environmental Institute – 1997

21. M. Green – *Solar cells: operating principles, technology and system applications* – Prentice Hall, Englewood Cliffs, New Jersey, USA – 1982

22. R. Pierret – *Modular series on solid state devices, volume VI: advanced semiconductior fundamentals* – Addison-Wesley, Reading, Massachusetts, USA – 1987

23. S. Sze – *Physics of Seminconductor devices, 2nd edition* – John Wiley & Sons, New York, NY – 1981

24. K. Böer – *Survey of Semiconductor Physics: electrons and other particles in bulk semiconductors* – Van Nostrand Reinhold, New York – 1990

25. J. Singh – *Physics of Semiconductor and their heterostructures* – McGraw-Hill, New York – 1993

26. J. Slotboom, H. De Graff – *Solid-state Electron* – 1976

27. J. Pankove, *Optical processes in semiconductors* – Dover Publications, New York – 1971

28. F. Sanii, F. Giles, R. Schwartz, J. Gray – *Solid state electron* – 1992

29. J. Gray – *Two dimensional modeling of silicon solar cells* – Ph.D. thesis, Purdue University, West Lafayette, Indiana, USA – 1982

30. M. Lundstrom – *Numerical analysis of silicon solar cells* – Ph.D. thesis, Purdue University, West Lafayette, Indiana, USA – 1980

31. L. Landau, E. Lifchitz – *Physique Statistique* – Mir, Moscow – 1967

32. R. Badescu – *Equilibrium and nonequilibrium statistical mechanics* – Wiley, New York – 1975

33. G. Araujo – *Limits to efficiency of single and multiple bandgap solar cells* – Adam Hilger, Bristol – 1990

34. J. Parrot – *Physical limitations to photovoltaic energy conversion* – Adam Hilger, Bristol – 1990

35. A. Brown, M. Green – *Limiting efficiency for a multi-solar cell containing three and four bands* – International Workshop on photovoltaics in nanostructures, Dresden, Germany – 2001

36. C. Ballif, F. Huljic, A. Hessler-Wysler – *Nature of the Ag-Si interface in screen printed contacts: a detailed transmission electron microscopy study of cross-sectional structures* – Proceedings of the 29th IEEE Photovoltaic Specialist Conference, New Orleans, USA – 2002

37. M. Laumanns, L. Thiele, K. Deb, E. Zitzler – *Combining convergence and diversity in evolutionary multiobjective optimization* – Evol. Computing – 2002

38. M. Morris – *Factorials sampling plans for preliminary computational experiments* – Technometrics – 1991

39. P. Nuzzo, X. Sun, C. Wu, F. D. Bernardinis, A. Sangiovanni-Vincentelli – *A platform-based methodology for system-level mixed-signal design* – EURASIP J. Embedded Syst. – 2010

40. M. Pavone, G. Narzisi, G.Nicosia – *Clonal selection, an immunological algorithm for global optimization over continuous spaces* – Journal of Global Optimization – 2012

41. R. Horst, H. Tuy – *Global Optimization: deterministic approaches* – Springer-Verlag – 1990

42. J. D. Pinter – *Global Optimization in action. Continuous and Lipschitz Optimization: algorithms, implementations and applications* – Kluwer Academic Publishers – 1996

43. R. Horts, P. Pardalos – *Handbook of global optimizations* – Kluwer Academic Publishers – 1995

44. R. P. Brent – *Algorithms for minimization without derivatives* – Prentice-Hall, Englewood Cliffs New Jersey, USA – 1973

45. K. W. Brodie – *A new direction set method for unconstrained minimization without evaluating derivatives* – J. Inst. Maths Applics, 1975

46. A. R. Conn, K. Scheinberg – *On the convergence of derivative-free methods for unconstrained optimization* – Cambridge University Press, Cambridge – 1997

47. C. Elster, A. Neumaier – *A grid algorithm for bound unconstrained optimization of noisy functions* – IMA Journal of Numerical Analysis – 1995

48. R. Fletcher – *Practical methods of optimization* – John Wiley & Sons, Chichester – 1987

49. P.E. Gill, W. Murray, M.H. Wright – *Practical optimization and machine learning* – Addison Wesley, Reading, Massachusetts, USA – 1981

50. L. Grippo, F. Lampariello, S. Lucidi – *Global convergence and stabilization of unconstrained minimization methods without derivatives* – Journal of Optim. Theory Appl. – 1988

51. D. E. Goldberg – *Genetic Algorithms in search, optimization and machine learning* – Addison-Wesley, Reading, Massachusetts, USA – 1989

52. R. Hooke, T.A. Jeeves – *Direct search solution of numerical and statistical problems* – Journal of Association of Comput. Mach. – 1961

53. P.J.M. Van Laarhoven, E.H.L. Aarts – *Simulated annealing: theory and applications* – Reidel Publishing Co. Dordrecht, The Netherlands – 1987

54. M.J.D. Powell – *A direct search optimization method that models the objective and constraint functions by linear interpolation* – Advances in optimization and numerical analysis, Kluwer Academic, Dordrecht, The Netherlands – 1994

55. M.J.D. Powell – *UOBYQA: unconstrained optimization by quadratic approximation* – Math. Programming B. – 2002

56. L. Armijo – *Minimization of functions having Lipschitz continuous first partial derivatives* – Pacific Journal of Mathematics, 1966

57. V. Pareto – *Course d'economie politique* – Rouge, Lausanne, Switzerland – 1896

58. M.S. Bazaraa, H.D. Sherali, C.M. Shetty – *Nonlinear programming: theory and algorithms* – Wiley, New York – 1979

59. K. Miettinen – *Nonlinear multiobjective optimization* – Kluwer Academic Publishers, Boston, Massachusetts, USA – 1999

60. M. Cid, N. Stem – *Phosphorus emitter and metal-grid optimization for homogeneous (n^+p) and double-diffused ($n^{++}n^+p$) emitter solar cells* – Material research – 2009

61. J. Nijs at al. – *Recent improvements in the screenprinting technology and comparison with the buried contact technology by 2-D simulation* – Solar energy materials and solar cells -1996

62. N. Stem, M. Cid – *Studies of phosphorus Gaussian profile emitter silicon solar cells* – Material research Vol. 4 – 2001

63. A. Goetzberger, J.Knobloch, B.Voss – *Crystalline silicon solar cells* – John Wiley & Sons – 1998

64. K. Varahramyan, J. Verret – *Solid-state electronics* – 1996

65. D.C. Langreth – *Linear and nonlinear transports in solids* – Plenum, New York – 1976

66. H. Haug, A.P. Jauho – *Quantum kinetics in transport and optics of semiconductors* – Springer, Berlin – 1996

67. W. Schäfer, M Wegener – *Semiconductor optics and transport phenomena* – Springer, Berlin - 2002