Simulated annealing bij de reconstructie van positron-emissietomografie-beelden

Erik Sundermann

Promotor: Prof. dr. I. Lemahieu
Co-promotor: Prof. dr. ir. M. Vanwormhoudt

Proefschrift ingediend tot het behalen van de graad van
Doctor in de Toegepaste Wetenschappen

Vakgroep Elektronica en Informatiesystemen
Voorzitter: Prof. dr. ir. J. Van Campenhout
Faculteit Toegepaste Wetenschappen
Academiejaar 1997-1998

UNIVERSITEIT GENT

Dankwoord

Dit proefschrift is het resultaat van meer dan vier jaar doctoraatsonderzoek. Gedurende deze periode heb ik kunnen rekenen op de medewerking van een aantal personen. Het is dan ook gepast hen hiervoor expliciet te bedanken.

In de eerste plaats wens ik mijn erkentelijkheid te betuigen aan mijn promotoren prof. dr. Ignace Lemahieu en prof. dr. ir. Marc Vanwormhoudt. Hun voortdurende aanmoedigingen waren een belangrijke stimulans om mijn onderzoek voort te zetten, vooral op de momenten dat het allemaal wat minder vlot ging. Zij hebben mij steeds de vrijheid gelaten om zelfstandig de richting van mijn onderzoek te bepalen en hebben mij de mogelijkheid geboden om via internationale conferenties en cursussen in contact te treden met de buitenwereld.

Daarnaast dank ik ook prof. dr. ir. Jan Van Campenhout, voorzitter van de vakgroep ELIS. Ondanks zijn drukbezette agenda heeft hij steeds de nodige tijd vrijgemaakt om het onderzoek van de verschillende doctoraatsstudenten op de voet te volgen. Door zijn kritische en opbouwende tussenkomsten heb ik vaak de oplossing voor problemen gevonden op niet voor de hand liggende plaatsen.

Dr. ir. Pol Desmedt heeft mijn handje vastgehouden tijdens mijn eerste stapjes in de exotische wereld van de Bayesiaanse statistiek. Zowel tijdens de voorbereiding van mijn afstudeerwerk als tijdens de eerste jaren van mijn doctoraat was hij altijd bereid mijn vele vragen te beantwoorden. Naast zijn wetenschappelijke prijs ik ook zijn diplomatieke kwaliteiten.

Gedurende de voorbije jaren heb ik deel uitgemaakt van de steeds groeiende Medisip-onderzoeksgroep binnen de vakgroep ELIS. Op de eerste plaats zijn er de (ex-)PET-mensen lic. Yves Vander Haeghen en ir. Filip Jacobs. Met hen heb ik menig vruchtbaar gesprek kunnen voeren i. v. m. mijn onderzoek. De contacten met de overige leden van de onderzoeksgroep hebben in belangrijke mate bijgedragen tot de verruiming van mijn wetenschappelijke kennis. Verder waren ing. Erik Nolf en ing. Johan Keppens steeds bereid mij te helpen met computer- en andere technische probleempjes. Dokter Patrick Santens van het UZ Gent, dienst

Neurologie, heeft mij kennis laten maken met de klinische aspecten van PET-beeldvorming. Ik wens hen hiervoor allen te bedanken.

De (spijtig genoeg) occasionele contacten met de andere leden van de PET-gemeenschap in België heb ik steeds als zeer leerrijk ervaren. Ik bedank dan ook prof. dr. ir. Johan Nuyts, prof. dr. Michel Defrise en prof. dr. Christian Michel. Verder hebben mijn mondelinge en elektronische contacten met prof. dr. Harrison Barrett op enkele cruciale momenten mijn onderzoek in de juiste richting geduwd.

Prof. dr. Ignace Lemahieu, dr. ir. Rik Van de Walle en lic. Yves Vander Haeghen waren bereid binnen een (zeer) korte tijdspanne dit manuscript kritisch na te lezen. Hun suggesties hebben bijgedragen tot de leesbaarheid van dit proefschrift.

Dit onderzoek werd voor het grootste deel verricht binnen het kader van een specialisatiebeurs van het Vlaams Instituut voor de bevordering van het Wetenschappelijk-Technologisch Onderzoek in de Industrie. Zonder hun financiële steun zou dit werk waarschijnlijk nooit tot stand gekomen zijn.

Tenslotte ben ik ook dank verschuldigd aan mijn familieleden en vrienden. In de eerste plaats aan mijn ouders, die gedurende mijn volledige studies steeds achter mij hebben gestaan en mij zoveel mogelijk hebben geholpen. Mijn (aanstaande) echtgenote Miranda heeft tijdens de voorbije periode, en vooral tijdens het schrijven van dit proefschrift, het nodige moeten doorstaan. Zij krijgt hiervoor niet alleen mijn dank, maar bovendien een dikke kus. Als laatste moet ik nog Pingu bedanken, zonder wiens aanwezigheid dit proefschrift nooit "tijdig" voltooid geraakt zou zijn.

ir. Erik Sundermann,
24 april 1998.

UNIVERSITEIT GENT

Faculteit Toegepaste Wetenschappen **6 november1998**

DOCTORAAT

Titel proefschrift: Simulated annealing bij de reconstructie van positron-emmissietomografie-beelden

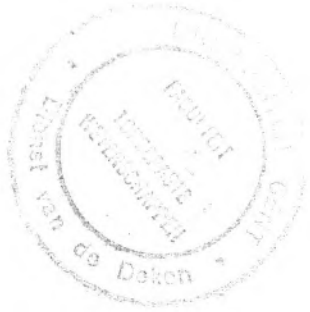
Aantal delen: 1

Ingediend door: **ir. Erik SUNDERMANN**

Promotor(en): Prof. I. LEMAHIEU

Datum openbare verdediging: 23 juni 1998

Behaalde graad: de grootste onderscheiding



Inhoudsopgave

Hoofdstuk 1

Inleiding

Tegenwoordig beschikt de medicus over een groot arsenaal aan medische onderzoekstechnieken die hem moeten helpen bij het stellen van een correcte diagnose. Oorspronkelijk gaven enkel uitwendige tekenen zoals temperatuur, huidskleur, bloeddruk e. d. een indicatie van de inwendige toestand van de patiënt. Invasieve technieken zoals het nemen van bloedstalen of weefselstalen, eventueel zelfs kijkoperaties, geven meer rechtstreekse informatie maar zijn ook meer belastend voor de patiënt. Recent zijn een aantal technieken ontwikkeld die toelaten op een niet-invasieve manier het inwendige van de patiënt te visualiseren. De grondslag voor deze methoden werd gelegd met de ontdekking van Röntgenstralen in 1895. Wilhelm Röntgen stelde vast dat hij door bestraling de inwendige structuur van zijn hand zichtbaar kon maken op een fotogevoelige plaat. Röntgenstralen worden geattenueerd door het weefsel, waardoor het mogelijk is densiteitsverschillen te visualiseren. Het nadeel van deze methode is dat een superpositiebeeld ontstaat, waarbij de driedimensionale anatomische structuur geprojecteerd wordt tot een tweedimensionaal beeld. De ontwikkeling van tomografische methoden bracht een oplossing voor dit probleem. Door deze methoden werd het mogelijk een beeld te reconstrueren van de weefseleigenschappen in een aantal vlakke doorsneden van de patiënt. Dit beeld is echter niet rechtstreeks beschikbaar maar moet m. b. v. computers berekend worden uit de meetgegevens. Een van deze tomografische technieken is positron-emissietomografie (PET). Deze techniek onderscheidt zich van een aantal andere beeldvormingstechnieken in het feit dat de gevisualiseerde informatie van metabolische aard is. Dit onderzoek is gewijd aan de ontwikkeling van een reconstructiealgoritme voor PET-beelden op basis van simulated annealing.

We bespreken de opbouw van dit proefschrift. Dit onderzoek is een combinatie van diverse onderzoeksdomeinen. Hoofdstukken 2, 3 en 4 vormen een inleiding en theoretische bespreking van deze diverse domeinen. De eigenlijke studie van het reconstructiealgoritme wordt besproken in hoofdstukken 5, 6 en 7. De onderzoeksresultaten worden tenslotte samengevat in hoofdstuk 8.

In hoofdstuk 2 wordt PET gesitueerd binnen de diverse medische beeldvormingstechnieken. De fysische principes van het meetproces worden kort besproken. Het reconstructieprobleem wordt toegelicht en een aantal bestaande reconstructietechnieken worden overlopen. Een belangrijk nadeel van PET is de slechte kwaliteit van de gereconstrueerde beelden wanneer men slechts over een klein aantal metingen beschikt.

Het Bayesiaanse formalisme vormt het algemeen theoretisch kader waarbinnen dit onderzoek zich situeert. Dit formalisme laat toe voorkennis (a priori-kennis) te gebruiken om de kwaliteit van de gereconstrueerde beelden te verbeteren. Een uitvoerige bespreking van de Bayesiaanse theorie voor beeldverwerkingsproblemen wordt gegeven in hoofdstuk 3. Hierbij worden beeldreconstructieproblemen besproken vanuit een algemeen oogpunt van parameterschatting. Dit geeft aanleiding tot beeldschatters die gebruik maken van de a posteriori-distributie. We voeren het Markov-Random-Veldmodel en de Gibbs-distributie in om een uitdrukking te kunnen opstellen voor deze a posteriori-distributie. Verder worden een aantal optimalisatiemethoden en bemonsteringsmethoden besproken die toelaten het beeld te schatten a. h. v. de a posteriori-distributie.

Een aantrekkelijke methode om de a priori-kennis te modelleren wordt gevormd door het gebruik van niet-convexe functies. Deze functies geven echter aanleiding tot lokale optima tijdens de beeldreconstructie. Vandaar de noodzaak om gebruik te maken van aangepaste optimalisatiemethoden. Simulated annealing is een optimalisatiemethode die in staat is eventuele lokale optima te verlaten en het globale optimum te vinden. De techniek simulated annealing en de diverse aspecten ervan worden uitvoerig besproken in hoofdstuk 4.

Gebruik makend van de theorie uit de voorgaande hoofdstukken wordt een reconstructiealgoritme voor PET-beelden op basis van simulated annealing ontwikkeld en bestudeerd. De belangrijkste aspecten ervan zijn het generatiemechanisme, het afkoelingsschema en de kostfunctie. Deze aspecten worden in detail besproken in hoofdstukken 5 en 6.

Hoofdstuk 7 tenslotte is gewijd aan een implementatie van het reconstructie-algoritme voor reële data. De overgang van gesimuleerde naar reële meetgegevens en de invloed ervan op de resultaten uit de voorgaande hoofdstukken wordt bestudeerd. De gereconstrueerde beelden worden vergeleken met beelden die afkomstig zijn van andere reconstructietechnieken. Tenslotte worden ook een aantal mogelijkheden nagegaan om het reconstructiealgoritme te versnellen.

Hoofdstuk 2

Positron-emissietomografie

2.1 Inleiding

In dit hoofdstuk wordt de techniek positron-emissietomografie (PET) geïntroduceerd. In een eerste paragraaf worden de meest bekende tomografische visualisatiemethoden overlopen. We bespreken vervolgens kort de fysische principes die aan de basis liggen van PET. Hierbij vermelden we ook enkele frequent gebruikte radio-isotopen en tracermoleculen. Daarna bespreken we de meetprocedure en vestigen de aandacht op enkele inherente problemen die zich voordoen tijdens het detectieproces. Tenslotte wordt het reconstructieprobleem toegelicht en een beknopt overzicht gegeven van de bestaande reconstructiemethoden.

2.2 Tomografische methoden

We herhalen dat tomografische methoden ons in staat stellen een beeld te vormen van de weefseleigenschappen in vlakke doorsneden van de patiënt. Afhankelijk van de gebruikte stralingsbron maken we het onderscheid tussen transmissietomografie (externe stralingsbron) en emissietomografie (door een radioactieve stof toe te dienen wordt het lichaam zelf de stralingsbron). De vier meest bekende technieken zijn CT, MR, SPECT en PET. We bespreken in deze paragraaf kort de basisprincipes van deze technieken.

De eerste techniek is computertomografie (CT). Hierbij wordt gebruik gemaakt van een uitwendige Röntgenstralingsbron en -detectoren. Elke detector meet de integraal van de absorptiecoëfficiënten van de verschillende weefseltypes over de strook tussen stralingsbron en detector. Door een groot aantal stroken te

bemeten is het mogelijk om een beeld te reconstrueren dat overeenstemt met de densiteitsverdeling over de onderzochte vlakke doorsnede van de patiënt. De bekomen informatie is bijgevolg uitsluitend van anatomische aard. Deze methode is dermate verfijnd dat beelden met een zeer hoge spatiale resolutie (tot 0.1 mm) bekomen kunnen worden. Op CT-beelden zijn de aanwezige botstructuren zeer duidelijk zichtbaar. Door de kleine densiteitsverschillen van de diverse types van zacht weefsel vertonen CT-beelden echter een relatief laag contrast voor zacht weefsel.

Magnetische-resonantiebeeldvorming (MR) is een tweede veel gebruikte visualisatiemethode. We bespreken kort het fysische principe van het detectieproces bij MR. Waterstofkernen bezitten een zgn. kernspin, waardoor zij zich gedragen als minuscule magneetnaaldjes. Bij aanleggen van een sterk extern magneetveld (grootteorde 1 T) zullen de kernspins in evenwicht een precessiebeweging uitvoeren rond de richting van het externe veld. De kernspins worden vervolgens uit evenwicht gebracht m. b. v. een radiofrequente puls (RF-puls). Na afleggen van de RF-puls zullen de kernspins relaxeren naar hun evenwichtstoestand en hierbij zal een RF-golf uitgestuurd worden. Het is deze golf die het gemeten MR-signaal vormt. Door een gradiënt aan te leggen bovenop het homogene magneetveld kan het signaal spatiaal gecodeerd worden. Het gemeten signaal staat o. a. in verhouding tot de aanwezige hoeveelheid waterstofkernen. Dit resulteert in een veel beter zacht-weefselcontrast in vergelijking met CT-beelden. Daar tegenover staan de lagere spatiale resolutie (typisch 0.3 mm) en de langere meettijden (typisch enkele minuten). Opnieuw is de bekomen informatie hoofdzakelijk van anatomische aard. Recent wordt echter ook onderzoek verricht naar de toepassing van MR voor functioneel onderzoek [VOos97].

Naast anatomische informatie is het voor de medicus dikwijls van belang om ook over metabolische informatie te beschikken. Technieken zoals SPECT (Single Photon Emission Computed Tomography) en PET laten toe het metabolisme van de patiënt te bestuderen. Beide technieken maken gebruik van het toedienen van een radioactieve stof aan de patiënt i. p. v. een externe stralingsbron. Bij SPECT worden radio-isotopen aangewend die bij verval een foton uitzenden. Dit foton wordt gedetecteerd m. b. v. uitwendige detectoren. De hoek waaronder het foton kan invallen op de detector wordt beperkt door het plaatsen van loden collimatoren voor de detector. Elke detector meet de integraal van de aanwezige activiteit over een kegel. Uit een voldoende aantal metingen kan een beeld gereconstrueerd worden van de activiteitsverdeling in de patiënt. Er zijn echter

een aantal belangrijke nadelen verbonden aan SPECT. Zo zijn bv. de gebruikte gamma-emittors meestal lichaamsvreemde stoffen (bv. ^{99}Tc) die slechts moeizaam door lever en nieren verwerkt kunnen worden. Tevens worden bij SPECT een laag aantal tellen gedetecteerd per detector, waardoor een aanzienlijke hoeveelheid ruis op de gereconstrueerde beelden aanwezig is. Bovendien is de resolutie van SPECT uiterst beperkt (typisch 12-15 mm [Webb88]) wegens de inherente onzekerheid betreffende de richting van waaruit het foton afkomstig is. Tenslotte zijn de gereconstrueerde beelden niet rechtstreeks bruikbaar voor kwantitatieve studies. Door attenuatie wordt nl. slechts een gedeelte van de ontstane fotonen gedetecteerd. Om hiervoor te corrigeren is het nodig de plaats te kennen waar het foton gevormd werd. Aangezien deze plaats onbekend is, is het onmogelijk een exacte attenuatiecorrectie uit te voeren. Er werden echter methoden voorgesteld voor een berekende attenuatiecorrectie [Nuyt91].

Een tweede methode om metabolische informatie te bekomen is PET [TerP80]. Deze methode zal in wat hierna volgt uitvoeriger beschreven worden. We vermelden echter nu reeds dat gebruik gemaakt wordt van positron-emittors. Door annihilatie ontstaan twee fotonen die gelijktijdig gedetecteerd worden. Hierdoor is er geen nood aan mechanische collimatie. Bovendien zijn de meest gebruikte radio-isotopen organische stoffen. De resolutie bij PET bedraagt typisch ongeveer 4 mm [Webb88]. Hoewel dit niet vergelijkbaar is met de resolutie van CT of MR, is het toch beduidend beter dan bij SPECT. Daarnaast is de sensitiviteit van PET zeer hoog. Zo kunnen concentraties tot picomolaire orde van grootte gevisualiseerd worden (dit is ongeveer 2 grootteorden kleiner dan bij SPECT, [Jone96b]). Bovendien kan op eenvoudige wijze een attenuatiecorrectie uitgevoerd worden, hetgeen een kwantitatieve analyse van de gegevens toelaat. Kwantitatieve analyse van PET-data vormt momenteel een belangrijk onderzoeksdomein. Zo worden bv. dynamische studies vaak gebruikt voor de berekening van tijdsactiviteitscurven over specifieke beeldregio's (ROI's, regions of interest) [Myer96, Patl83]. Een specifiek voorbeeld van kwantitatieve analyse voor neurologie is de bepaling van receptordichtheden in het centraal zenuwstelsel [Blom90]. Verder vermelden we nog de kwantificatie van de infarctgrootte van het myocard [VdHa95]. Tegenover de hoger vernoemde voordelen staan de aanzienlijk hogere kostprijs van een PET-scanner (in vergelijking met SPECT) en de behoefte aan een cyclotron in de nabijheid van het PET-centrum voor de productie van tracers. De rol van PET binnen het spectrum van medische beeldvorming wordt verder besproken in o. a. [Jone96b, Jone96a].

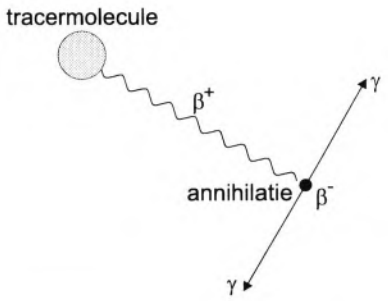
Tenslotte vermelden we nog de recente ontwikkeling van zgn. coïncidentie-SPECT [Jarr96]. Door gebruik te maken van een SPECT-camera met twee overstaande detectoren is het mogelijk PET-beeldvorming te verrichten. Wegens het uiterst beperkt aantal detectoren is voor een vergelijkbare toegediende dosis de sensitiviteit echter beduidend lager dan bij een PET-scanner.

2.3 Positron-emissietomografie

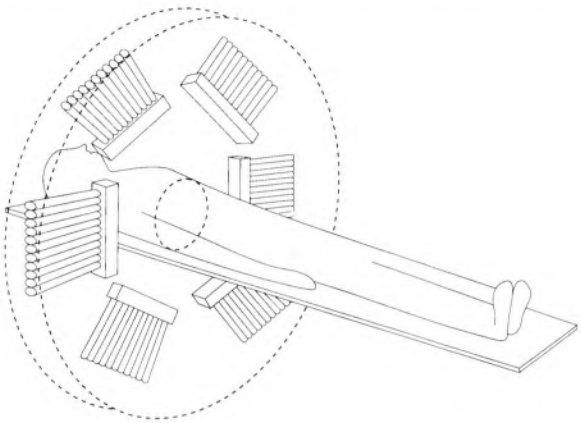
Zoals we reeds vermeld hebben wordt bij PET gebruik gemaakt van positron-emittors. Dit zijn radio-isotopen die bij verval een positron uitsturen. We komen hier in de volgende paragraaf op terug. Het ontstane positron annihileert vrijwel onmiddellijk met een elektron uit het weefsel. Bij deze annihilatiereactie ontstaan twee fotonen met energie 511 keV. Bij verwaarloosbare kinetische energie van het positron zullen de beide fotonen uitgezonden worden met een onderlinge hoek van 180°. Het principe van de positronannihilatie is schematisch weergegeven in figuur 2.1. Het lichaam van de patiënt wordt isotroop verondersteld, zodat elke richting waaronder de fotonen uitgezonden worden even waarschijnlijk is. De vrije weglengte van het positron (t. t. z. de gemiddelde afgelegde weg alvorens annihilatie optreedt) en de kinetische energie (waardoor de onderlinge hoek niet exact 180° bedraagt) zorgen samen voor een afwijking tussen de reële en de gedetecteerde plaats van annihilatie. Ten gevolge van deze afwijking bedraagt de fysische ondergrens voor de praktisch haalbare resolutie 1-3 mm, afhankelijk van het gebruikte radio-isotoop [Dere75]. De reële resolutie is verder o. m. te wijten aan de afmetingen van de detectoren.

Rond het lichaam van de patiënt wordt een ring van detectoren opgesteld (figuur 2.2). Wanneer twee detectoren gelijktijdig een foton detecteren treedt een zgn. coïncidentiemeting op. Uit de meting van een coïncidentie besluiten we dat zich in de strook tussen beide detectoren een positronannihilatie heeft voorgedaan en bijgevolg dat er zich in die strook een zekere hoeveelheid van de te onderzoeken stof bevindt (figuur 2.3). Een coïncidentiemeting verschaft echter geen informatie over de exacte locatie van de annihilatie binnen de projectiestrook.

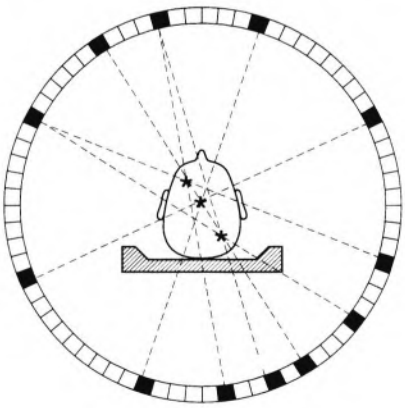
We introduceren een geïdealiseerd scannermodel. Voor dit model veronderstellen we een scanner die uit twee (lineaire) detectorbanken bestaat. Elke detectorbank is opgebouwd uit D elementaire detectoren. We veronderstellen dat deze elementaire detectoren niet met elkaar interageren (bv. van elkaar gescheiden door een oneindig dunne en perfect afschermende wand) en enkel fotonen detecteren



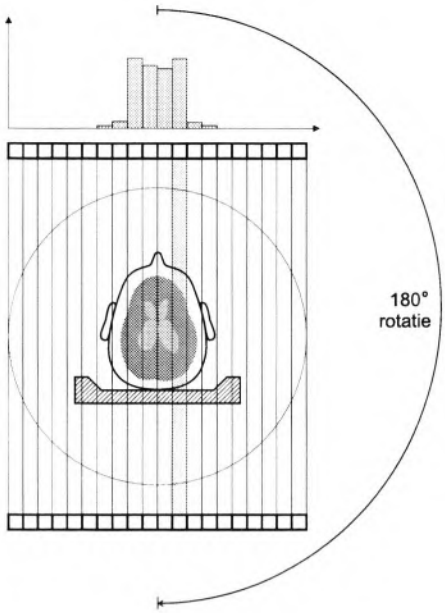
Figuur 2.1: Schematische voorstelling van positronannihilatie.



Figuur 2.2: Typische opstelling van een PET-scanner.

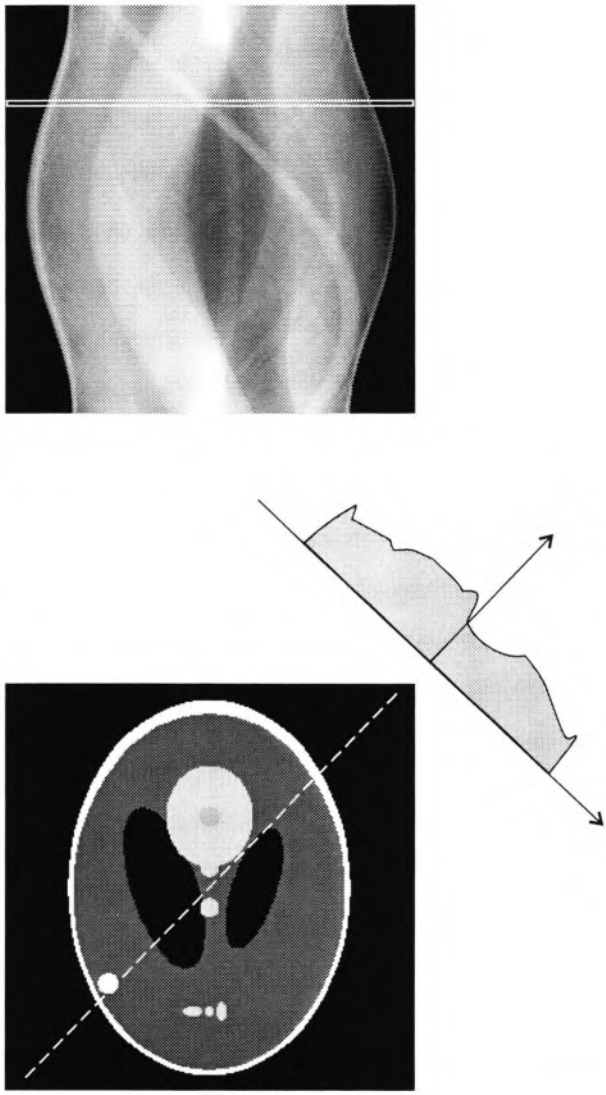


Figuur 2.3: Het principe van het detectieproces bij PET.



Figuur 2.4: Illustratie van het geïdealiseerde scannermodel: twee overstaande detectoren meten de activiteit die zich in de strook tussen beide detectoren bevindt.

die loodrecht op het detectoroppervlak invallen. We benadrukken dat het hier om hypothetische detectoren gaat. Twee overstaande detectoren meten bijgevolg de integraal van de activiteit die zich in de strook tussen beide detectoren bevindt (figuur 2.4). Op deze manier wordt gedurende een bepaalde tijd de tracerverdeling gemeten. Vervolgens worden de detectorbanken over een hoek $\Delta\alpha$ geroteerd, waarna de meting herhaald wordt. Wanneer we op deze manier M metingen uitvoeren ($\Delta\alpha = 180°/M$) hebben we de volledige omtrek van de patiënt gescand en beschikken we over $D \times M$ meetwaarden. Deze data worden opgeslagen in een matrix (M rijen, D kolommen) die we het sinogram noemen. Elke rij uit het sinogram correspondeert bijgevolg met de projectie van de activiteitsverdeling onder een welbepaalde hoek. Dit wordt geïllustreerd in figuur 2.5. We merken tenslotte op dat de metingen die afkomstig zijn van een reële (bv. cirkelvormige) scannerconfiguratie steeds zó herschikt kunnen worden dat zij overeenstemmen met metingen volgens het ideale scannermodel; dit wordt in de praktijk dan ook steeds uitgevoerd.



Figuur 2.5: Illustratie van de sinogrammatrix: elke rij correspondeert met de projectie van de activiteitsverdeling onder een welbepaalde hoek.

isotoop	$T_{1/2}$	$E_{\beta+}$
^{11}C	20.4 min	960 keV
^{13}N	10.0 min	1190 keV
^{15}O	2.05 min	1720 keV
^{18}F	109.6 min	635 keV

Tabel 2.1: Enkele veelgebruikte positronemittors met corresponderende halveringstijden en positronenergieën.

2.4 Positronemittors

Er bestaan een groot aantal positronemitterende radio-isotopen. Positronen komen o. a. vrij bij het β-verval van onstabiele protonrijke isotopen. De meeste PET-studies worden uitgevoerd met organische positronemittors (^{11}C, ^{13}N, ^{15}O, ^{18}F) [Stöc93]. Een overzicht van de meest gebruikte positronemittors wordt gegeven in tabel 2.1. Ten gevolge van de korte levensduur van deze isotopen moeten de gebruikte producten ter plaatse geproduceerd worden. Hierdoor is er noodzaak aan een cyclotron in de omgeving van het PET-centrum. Alleen tracers op basis van ^{18}F kunnen eventueel getransporteerd worden naar PET-centra zonder cyclotron.

De positronemitterende radio-isotopen moeten ingebouwd worden in moleculen (zgn. tracermoleculen) alvorens ze toegediend kunnen worden aan de patiënt. Een tracermolecule is een molecule waarvan een atoom gesubstitueerd is door een gelijkaardig radio-isotoop. Hierdoor wordt de metabolische werking van de molecule niet beïnvloed maar kan de verspreiding van de tracer in het lichaam bestudeerd worden. Het metabolisme van de patiënt zal er voor zorgen dat de tracer in de verschillende organen terechtkomt. De tracer moet uiteraard gekozen worden in functie van het metabolisme dat men wenst te bestuderen. PET wordt hoofdzakelijk toegepast bij onderzoek naar de werking van de hersenen (neurologie), de werking van het hart (cardiologie) en het opsporen van tumoren (oncologie). De meest gebruikte tracermoleculen zijn weergegeven in tabel 2.2, met vermelding van het gebruikte radio-isotoop. Zo wordt O_2 gebruikt om het zuurstofmetabolisme in de hersenen te onderzoeken. In het lichaam wordt CO_2 omgezet tot ^{15}O-gelabeld water, wat toelaat de bloedperfusie in organen en weefsels te bestuderen. Thymidine is een bouwsteen voor DNA en is bijgevolg een goede merker voor weefsel met een hoge celproliferatie (tumoren). Glucose en FDG (2-fluoro-2-deoxy-D-glucose) laten allebei toe het glucosemetabolisme te onderzoeken. Het

tracer	isotoop	metabolisme
O_2	^{15}O	zuurstofmetabolisme
CO_2	^{15}O	bloedperfusie
thymidine	^{11}C	DNA-synthese
glucose	^{11}C	glucosemetabolisme
FDG	^{18}F	glucosemetabolisme

Tabel 2.2: Een overzicht van de meest gebruikte tracermoleculen met het corresponderende radio-isotoop en het onderzochte metabolisme.

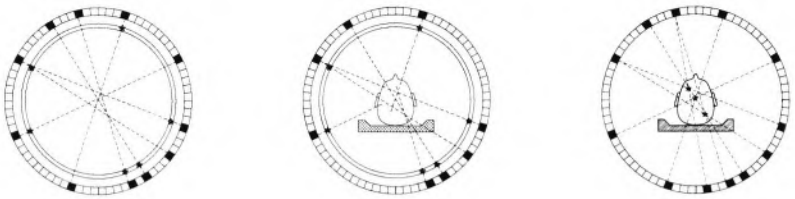
is nadelig dat bij gebruik van ^{11}C-gelabelde glucose ook het metabolisme van de afbraakprodukten zichtbaar wordt. Bij FDG daarentegen valt de afbraak na enkele stappen stil t. g. v. de aanwezigheid van ^{18}F. FDG is momenteel de meest universele tracermolecule en wordt zowel toegepast voor neurologisch, cardiologisch als oncologisch onderzoek [Stöc93].

2.5 Het meetproces

De PET-metingen dienen gecorrigeerd te worden voor de efficiëntie van de detectoren en de attenuatie die in de patiënt optreedt. Daarom bestaat de PET-meetprocedure uit 3 metingen: de blancoscan, de transmissiescan en de emissiescan (figuur 2.6). Bij de blancoscan wordt enkel een (uniforme) ringbron gemeten zonder patiënt. We verwachten dat het bijbehorende sinogram uniform is. Afwijkingen van een uniform patroon zijn een gevolg van de verschillen in efficiëntie tussen de verschillende detectoren. Bij de transmissiescan wordt dezelfde ringbron gemeten, ditmaal terwijl de patiënt zich in de scanner bevindt (zonder tracer toe te dienen). Door de aanwezigheid van de patiënt worden fotonen geattenueerd (verstrooid, geabsorbeerd) [Desm91]. Het verschil tussen de blancoscan en de transmissiescan laat toe de attenuatiefactoren te berekenen. De derde meting tenslotte wordt uitgevoerd na het verwijderen van de ringbron en het toedienen van de tracer aan de patiënt. Deze meting wordt gecorrigeerd voor detectorefficiëntie en attenuatie volgens

$$\text{gecorrigeerde meting} = \text{emissiescan} \times \frac{\text{blancoscan}}{\text{transmissiescan}}. \tag{2.1}$$

Tijdens het verdere verloop van dit betoog zullen we met "het sinogram" steeds het gecorrigeerde sinogram bedoelen.



(a) (b) (c)

Figuur 2.6: De drie metingen van een PET-meetprocedure: (a) de blancoscan, (b) de transmissiescan, (c) de emissiescan.

Daarnaast zijn er nog een aantal fenomenen die zich voordoen en aanleiding geven tot fouten. Zo kan één van beide ontstane fotonen door Compton-verstrooiing afgebogen worden en daardoor niet op een detector terechtkomen. Anderzijds zijn er fotonparen waarvan de baan zich niet in het vlak van de detectoren bevindt, maar waarvan toevallig toch één foton gedetecteerd wordt. Dit geeft aanleiding tot zgn. singles. Wanneer twee zulke singles per toeval gelijktijdig op een detector invallen ontstaat een zgn. random-coïncidentie. Fotonparen waarvan één foton door Compton-verstrooiing afgebogen wordt, maar die toch allebei gedetecteerd worden geven aanleiding tot zgn. scatter. Tenslotte zullen er coïncidenties verloren gaan t. g. v. dode tijd van de detectoren.

2.6 Beeldreconstructiemethoden

De twee belangrijkste aspecten van het reconstructieprobleem waarmee we bij PET geconfronteerd worden zijn enerzijds het feit dat de gemeten data eigenlijk projectiedata zijn en anderzijds dat de data ruis bevatten. Deze ruis kan in de meeste praktische gevallen gemodelleerd worden als Poisson-ruis [Rowe92]. Er zijn verschillende technieken ontwikkeld voor het beeldreconstructieprobleem. We kunnen deze technieken onderverdelen in drie groepen: analytische algoritmen, iteratieve algoritmen en statistische algoritmen. Voor een meer uitvoerige bespreking van deze reconstructiealgoritmen (zowel 2D als 3D) verwijzen we o. a. naar [Herm79, Defr90, Town93, Jaco96]. In deze paragraaf geven we een korte beschrijving.

Zoals we reeds aangehaald hebben in paragraaf 2.3 bevatten de metingen geen rechtstreekse informatie betreffende de ruimtelijke verdeling van de tracer in de patiënt, maar informatie over strookintegralen van de tracerconcentraties. Wan-

neer de metingen beschikbaar zijn voor alle hoeken tussen 0°en 180°(continu verloop) en voor alle laterale detectorposities (continu verloop), dan kan het beeld gereconstrueerd worden m. b. v. de inverse Radon-transformatie. Het eerste type reconstructiealgoritme, t. t. z. het analytische algoritme, is een aanpassing van de inverse Radon-transformatie in geval men slechts over een discreet aantal meetpunten beschikt. De methode bestaat uit een convolutiestap en een terugprojectiestap, vandaar de benaming gefilterde terugprojectie (filtered backprojection, FB) [Herm79, Kak88]. Deze techniek is, hoofdzakelijk omwille van zijn snelheid, de standaard reconstructiemethode voor PET. De kwaliteit van gefilterde-terugprojectiebeelden daalt echter sterk bij een laag aantal coïncidenties (zie ook verder). Aangepaste FB-algoritmen die elementen van een statistische reconstructietechniek bevatten worden voorgesteld door o. a. [Hebe92a, Fess93].

We gaan ervan uit dat, in afwezigheid van ruis, het verband tussen het beeld X en de metingen Y geschreven kan worden m. b. v. de transfertmatrix Ψ:

$$Y = \Psi X. \tag{2.2}$$

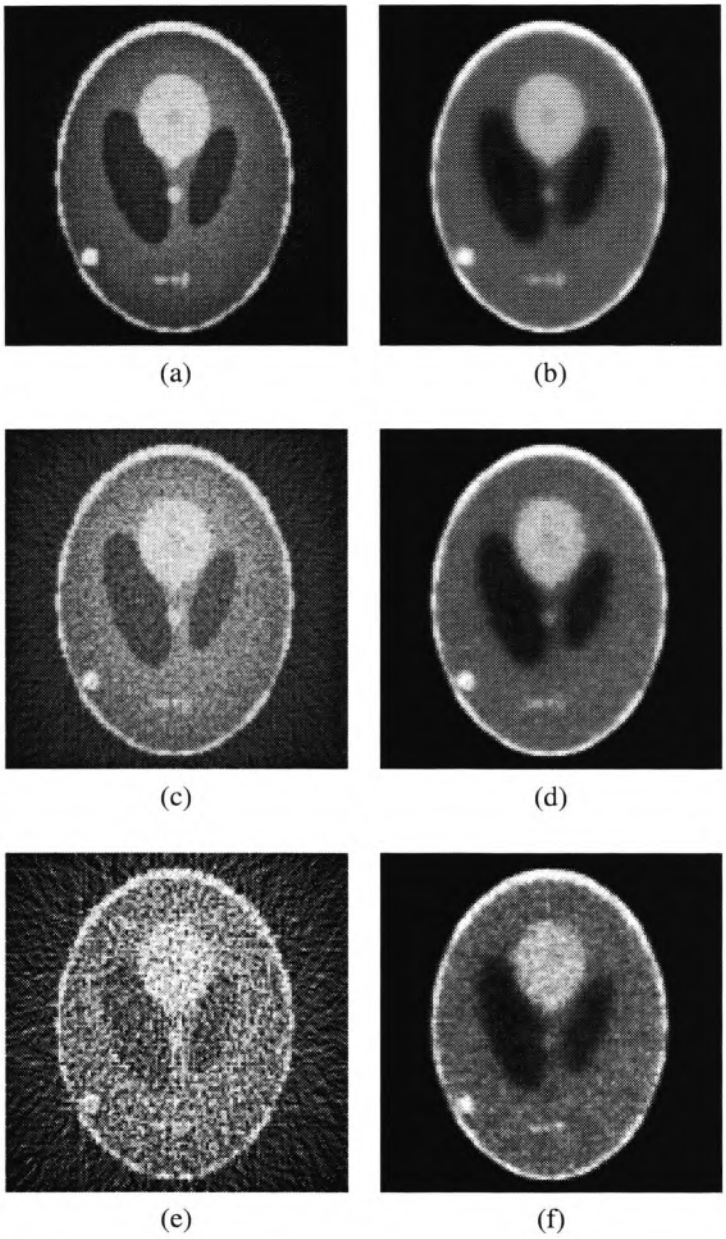
Deze transfertmatrix wordt ook de projectiematrix genoemd. We merken hierbij op dat X en Y kolommatrices zijn. De tweede groep reconstructietechnieken beschouwen het reconstructieprobleem als het inverteren van een stelsel lineaire vergelijkingen. Wegens de grote dimensie van de projectiematrix geeft dit aanleiding tot iteratieve methoden. De bekendste iteratieve reconstructiemethoden zijn ART (Algebraic Reconstruction Technique), MART (Multiplicative ART) en SMART (Simultaneous MART). Deze technieken worden o. a. besproken in [Stam88, VDij92, Byrn93].

Beide bovenstaande technieken (analytische en iteratieve methoden) nemen aan dat elk beeld aanleiding geeft tot één unieke set metingen en houden bijgevolg geen rekening met ruis op de meetdata. In het geval van PET is deze ruis echter inherent aanwezig. In de eerste plaats zijn de metingen het gevolg van het radioactief verval van isotopen. Bovendien hebben de in paragraaf 2.5 besproken fenomenen (Compton-verstrooiing, random coïncidenties, dode tijd van de detectoren) ook invloed op de metingen. Hierdoor is het onmogelijk exact te voorspellen hoeveel coïncidenties gedetecteerd zullen worden voor een gegeven activiteitsverdeling, zodat ook het inverse probleem sterk bemoeilijkt wordt. We merken op dat de PET-metingen gemodelleerd kunnen worden als statistisch onafhankelijke Poisson-grootheden [Desm95]. Het statistisch karakter van de metingen leidt tot streepvormige artefacten in het beeld bij de gefilterde-terugprojectietechniek.

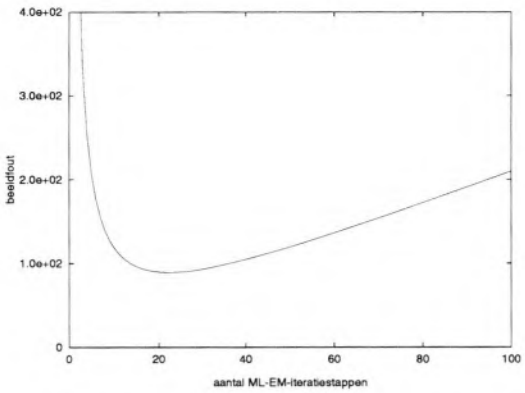
Deze artefacten bemoeilijken het stellen van een correcte diagnose door de medicus. Wegens het Poisson-karakter neemt de invloed van ruis op de beeldkwaliteit toe bij een afnemend aantal coïncidenties (figuur 2.7).

De derde groep bestaat uit de statistische algoritmen. Zij houden rekening met de ruis op de metingen door het detectieproces te modelleren als een stochastisch proces. Het gereconstrueerde beeld zal algemeen gezocht worden als zijnde het beeld dat de a posteriori-waarschijnlijkheidsdistributie maximaliseert. We zullen hierop in detail terugkomen tijdens het volgende hoofdstuk. We vermelden nu reeds dat ML-EM (Maximum Likelihood Expectation Maximization) de bekendste statistische reconstructiemethode is [Rock77, Shep82, Shep84]. Zoals in figuur 2.7 duidelijk te zien is zijn de ML-EM-gereconstrueerde beelden veel minder onderhevig aan artefacten t. g. v. van ruis. Een vergelijking tussen deterministische en statistische reconstuctiealgoritmen wordt o. a. gegeven in [Bouc96].

Bij het ML-EM-algoritme, wat een iteratief algoritme is, worden we echter met het probleem van ruisdeterioratie geconfronteerd [Snyd87]. Dit betekent dat gedurende de eerste iteratiestappen de kwaliteit van het beeld stijgt, maar dat naarmate het contrast toeneemt (d. w. z. dat hogere-frequentiecomponenten in het beeld worden geïntroduceerd) de oneffenheden in het beeld ook toenemen en de beeldkwaliteit opnieuw afneemt. We merken op dat momenteel een intuïtieve interpretatie van de begrippen "beeldkwaliteit" en "beeldfout" volstaan; deze grootheden zullen later gedefinieerd worden. In figuur 2.8 is te zien dat het verloop van de beeldfout i. f. v. het aantal ML-EM-iteratiestappen een duidelijk minimum vertoont. Om het probleem van ruisdeterioratie tegen te gaan worden een aantal technieken voorgesteld. Een eerste en populaire methode is het gebruik van een stopcriterium om het iteratieproces af te breken wanneer de minimale waarde van de beeldfout bereikt wordt [Vekl87, Llac88, Llac89, Hebe88, Hebe90, Tzan93, Kont94, Kont96]. Het is echter in de praktijk moeilijk om een goede uitdrukking op te stellen voor dit stopcriterium. Bovendien kan men zich de vraag stellen waarom men gebruik maakt van een maximumlikelihoodcriterium wanneer dit maximum vermeden wordt. Een tweede methode is het beperken van de hoge frequenties in het beeld door gebruik te maken van alternatieve basisfuncties die enkel lage-frequentiecomponenten bevatten (zoals "Sieves" [Snyd85, Mill86, Poli88] en "blobs" [Lewi92, Mate96b]). Het nadeel van deze methode is dat werkelijke hoge-frequentiecomponenten van de tracerverdeling hierdoor eveneens onderdrukt worden. Een derde methode is het gebruik van voorkennis over de klasse van te reconstrueren beelden. Het Bayesiaanse



Figuur 2.7: Simulatie van de invloed van ruis bij een afnemend aantal coïncidenties: (a), (c) en (e) gefilterde terugprojectie en (b), (d) en (f) ML-EM-reconstructie; (a) en (b) 4.10^7 tellen, (c) en (d) 4.10^6 tellen, (e) en (f) 4.10^5 tellen.



Figuur 2.8: Verloop van de beeldfout i. f. v. het aantal iteratiestappen voor het ML-EM-algoritme [Desm95].

formalisme laat toe op een elegante en theoretisch gefundeerde manier a priori informatie te incorporeren in het reconstructiealgoritme. We zullen deze methode in het volgende hoofdstuk nader bestuderen.

2.7 Conclusie

Tijdens dit hoofdstuk hebben we PET gesitueerd binnen het gamma van medische beeldvormingstechnieken. Technieken als PET en SPECT laten toe het metabolisme van de patiënt te visualiseren. De voordelen van PET t. o. v. SPECT zijn de hogere resolutie, de hogere sensitiviteit, de mogelijkheid tot kwantitatieve analyse en het gebruik van kortlevende organische radio-isotopen (wat minder belastend is voor de patiënt).

De belangrijkste aspecten van het reconstructieprobleem bij PET zijn de aard van de metingen (t. t. z. projectiemetingen) en de aanwezigheid van ruis. Deze ruis heeft storende streepvormige artefacten voor gevolg bij gefilterde terugprojectie. Statistische reconstructiealgoritmen pakken het probleem van ruis aan door het detectieproces als een stochastisch proces te modelleren. Bij ML-EM worden we echter geconfronteerd met het probleem van ruisdeterioratie. We zullen trachten dit probleem op te lossen door gebruik te maken van a priori informatie om de kwaliteit van de gereconstrueerde beelden te verbeteren.

Hoofdstuk 3

Het Bayesiaanse formalisme

3.1 Inleiding

Aangezien het Bayesiaanse formalisme voor beeldanalyse en -reconstructie het algemeen theoretisch kader vormt waarbinnen dit onderzoek zich situeert, geven we in wat volgt hiervan een overzicht. Binnen de Bayesiaanse context wordt een probleem van beeldverwerking algemeen beschouwd als een probleem van parameterschatting. Het gezochte beeld, dat we voorstellen door X, wordt geschat aan de hand van een stel waarnemingen Y. Kenmerkend voor de Bayesiaanse aanpak is dat hiervoor gebruik gemaakt wordt van voorkennis i. v. m. het te schatten beeld. Deze voorkennis wordt gecombineerd met de informatie uit de waarnemingen door toepassing van de regel van Bayes. Deze algemene basisformule uit de waarschijnlijkheidsrekening drukt de voorwaardelijke waarschijnlijkheid uit dat een gebeurtenis A optreedt, wanneer we met zekerheid weten dat een (andere) gebeurtenis B plaatsvindt:

$$\begin{aligned} p(A|B) &= \frac{p(A \cap B)}{p(B)} \\ &= \frac{p(B|A)\, p(A)}{p(B)}. \end{aligned} \tag{3.1}$$

We merken hierbij op dat aan de begrippen "waarschijnlijkheid" en "probabiliteit" de Bayesiaanse betekenis wordt toegekend, d. w. z. een betekenis van "aannemelijkheid" [Desm95]. We herschrijven deze uitdrukking door A en B te vervangen door het beeld X en de waarnemingen Y en geven een Bayesiaanse interpretatie

aan de verschillende termen:

$$p(X|Y) = \frac{p(Y|X)p(X)}{p(Y)}. \tag{3.2}$$

De waarschijnlijkheidsdistributie in het linker lid, $p(X|Y)$, wordt de a posteriori-waarschijnlijkheidsdistributie genoemd. Zij drukt de waarschijnlijkheid van de mogelijke beelden uit nadat de waarnemingen verricht zijn; vandaar de benaming "a posteriori". De kennis van $p(X|Y)$ verschaft de volledige statistische beschrijving van het schattingsprobleem. In het rechterlid vinden we enerzijds $p(Y|X)$. Deze term wordt de voorwaartse of directe waarschijnlijkheidsdistributie genoemd en drukt de waarschijnlijkheid uit dat een beeld X aanleiding zal geven tot waarnemingen Y. De karakteristieken van het meetproces bepalen $p(Y|X)$. De tweede term die we beschouwen is de a priori-waarschijnlijkheidsdistributie $p(X)$. Deze is onafhankelijk van de metingen Y en drukt de voorkennis uit omtrent het te schatten beeld. Wanneer we bv. weten dat het beeld moet bestaan uit rechte lijnen, dan zullen beelden die een rooster afbeelden een grotere a priori-waarschijnlijkheid hebben dan beelden die cirkels of ellipsen bevatten. Tenslotte vinden we nog de waarschijnlijkheidsdistributie $p(Y)$ in de noemer. Deze wordt de globale waarschijnlijkheidsdistributie genoemd en speelt de rol van normalisatieconstante.

Voor een meer gedetailleerde beschrijving van het Bayesiaanse waarschijnlijkheidsrekenen en de hypothesen die aanleiding geven tot de regel van Bayes (3.2) verwijzen we naar [Lore90, Bern94, Desm95]. Hoewel enkele onderzoekers (zoals Besag [Besa74, Besa86, Besa93] en Hunt [Hunt77]) vroeger reeds voorstelden om het Bayesiaanse formalisme toe te passen op problemen van beeldverwerking, heeft vooral de publicatie in 1984 van het artikel van Geman *et al.* [Gema84] de aanzet gegeven tot talrijke toepassingen. Een overzicht van de toepassingsmogelijkheden van deze theorie voor beeldverwerking wordt gegeven door o. a. [Demo89] en [Hart87]. Een uitgebreide beschrijving van de Bayesiaanse theorie, met inbegrip van diverse toepassingen voor beeldverwerking, wordt gegeven in het overzichtswerk van Winkler [Wink95]. Als recente toepassingen van het Bayesiaanse formalisme in een medische context vermelden we, naast de vele tomografische reconstructiemethoden die verder besproken zullen worden, het onderzoek van Malfait naar ruisonderdrukking met behulp van wavelets [Malf94, Malf95] en visualisatie van de neurale activiteit a. h. v. MEG-metingen (magnetoencefalogram) [Phil96].

We geven een kort overzicht van de opbouw van dit hoofdstuk. In een eerste inleidende paragraaf definiëren we het begrip "beeld" en introduceren een aantal grootheden en notaties. Gedurende het ganse hoofdstuk zal het reconstructieprobleem opgevat worden vanuit het oogpunt van parameterschatting. In een tweede paragraaf bespreken we daarom algemeen het probleem van parameterschatting. We introduceren drie verschillende schatters voor het beeld X. Deze schatters maken steeds gebruik van de a posteriori-distributie $p(X|Y)$; het is daarom van belang dat we beschikken over een uitdrukking voor de a posteriori-distributie. Hiervoor introduceren we achtereenvolgens de begrippen Markov-Random-Veldmodel en Gibbs-distributie. Deze laten toe een uitdrukking voor de a priori-distributie en tenslotte ook voor de a posteriori-distributie op te stellen. Nadat we over een uitdrukking voor de a posteriori-waarschijnlijkheidsdistributie beschikken herbekijken we de diverse schatters. Het reconstructieprobleem kan steeds geformuleerd worden als een optimalisatieprobleem. Daarom worden enkele frequent gebruikte optimalisatiemethoden besproken. Afhankelijk van de uitdrukking voor de schatter moet gebruik gemaakt worden van bemonsteringsmethoden, zodat ook een aantal bemonsteraars geïntroduceerd worden. In een laatste paragraaf tenslotte bestuderen we een bijzonder beeldmodel (het duale beeldmodel) voor de incorporatie van anatomische a priori-informatie.

3.2 Het beeldmodel

We zullen het beeldmodel o. a. gebruiken om onze voorkennis omtrent de te schatten beelden voor te stellen. Het beeldmodel dat gehanteerd zal worden is het random-veldmodel. Dit betekent dat een beeld X beschouwd wordt als een verzameling van N toevalsgrootheden $\{X_1, \ldots, X_N\}$. Voor de eenvoud zullen we ons in wat volgt beperken tot een bespreking van monochrome beelden die voorgesteld worden als een rooster van beeldpixels. Dit is van toepassing op het overgrote deel van de medische beeldverwerkingsapplicaties. De verzameling van roosterplaatsen (pixellocaties) stellen we voor door $S = \{1, \ldots, N\}$. In dit geval wordt elke toevalsgrootheid X_i geassocieerd met de intensiteit van de i^{de} beeldpixel. Een analoge bespreking kan echter gegeven worden voor kleurenbeelden, waarbij – afhankelijk van de kleurrepresentatie – verscheidene toevalsgrootheden overeenstemmen met elke beeldpixel (bv. 3 voor een RGB-representatie). Anderzijds beperkt deze theorie zich niet enkel tot beelden die d. m. v. pixels voorgesteld worden. Meer algemeen kunnen we stellen dat voor een beeldrepresentatie met behulp van N basisfuncties geldt dat X_i correspondeert met de coëfficiënt van de i^{de} basisfunctie.

We stellen de verzameling van mogelijke waarden van elke toevalsveranderlijke voor door Λ. Afhankelijk van de aard van Λ maken we het onderscheid tussen een discreet random veld en een continu random veld. We spreken van een discreet random veld wanneer Λ bestaat uit een eindig aantal discrete waarden, t. t. z. $\Lambda = \{I_1, \ldots, I_k\}$. Dit is bv. nuttig voor het modelleren van gediscretiseerde, gelabelde of gesegmenteerde beelden. Anderzijds spreken we van een continu random veld wanneer Λ een gesloten interval is, t. t. z. $\Lambda = [I_{min}, I_{max}]$. In een context van digitale beeldverwerking zullen we ons verder beperken tot een bespreking voor discrete random velden. De toestandsruimte voor beelden (de beeldruimte) zullen we voorstellen door Ω, zodat $\Omega = \Lambda^N$. We voeren ook de verkorte notatie X_J in, waarbij $J \subset S$:

$$X_J = \{X_i | i \in J\}. \tag{3.3}$$

Zo stelt bv. $X_{S/i}$ de verzameling van alle pixels behalve de i^{de} pixel voor.

We introduceren enkele waarschijnlijkheidsdistributies die van belang zullen zijn tijdens de verdere bespreking van het Bayesiaanse formalisme:

- $p(X)$: de waarschijnlijkheidsdistributie van het (volledige) beeld X; dit drukt de waarschijnlijkheid uit dat het beeld X voorkomt;
- $p(X_i|X_{S/i})$: de conditionele waarschijnlijkheidsdistributie van de i^{de} pixel (geconditioneerd op de kennis van alle overige pixelwaarden); dit drukt de waarschijnlijkheid uit dat de i^{de} pixel de waarde X_i aanneemt wanneer de overige pixelwaarden gekend zijn;
- $p(X_i)$: de marginale waarschijnlijkheidsdistributie van de i^{de} pixel; dit drukt de waarschijnlijkheid uit dat de i^{de} pixel de waarde X_i aanneemt, onafhankelijk van de waarden van de overige pixels.

We merken hierbij op dat er opzettelijk gekozen is voor een zekere ambiguïteit in de notaties, met de bedoeling deze niet nodeloos lang te maken. Zo zal de notatie $p(X)$ zowel de betekenis hebben van de waarschijnlijkheidsdistributie van de toevalsgrootheid X als de waarschijnlijkheid dat X een bepaalde waarde aanneemt. Uit de context zal echter steeds duidelijk zijn welke de correcte interpretatie is.

We herhalen dat we het beeldmodel zullen gebruiken om onze voorkennis omtrent de te schatten beelden voor te stellen. Hiervoor zijn verschillende beeldmodellen mogelijk. Een eerste eenvoudige mogelijkheid bestaat erin het beeld te

modelleren als N onafhankelijke toevalsgrootheden

$$p(X) = \prod_{i=1}^{N} p(X_i). \tag{3.4}$$

Dit model is echter ongeschikt omdat het niet toelaat spatiale correlaties te beschrijven. We merken op dat (3.4) overeenstemt met de maximum-entropieoplossing wanneer men enkel over de afzonderlijke $p(X_i)$ beschikt.

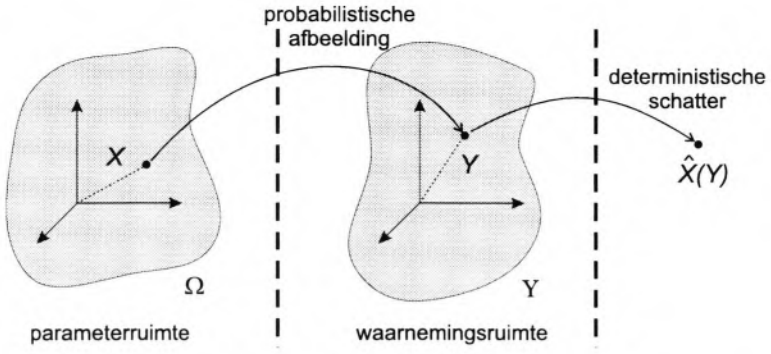
Een tweede mogelijkheid bestaat erin het beeld als een multidimensionale Gauss-distributie te modelleren. Hierbij rijst echter het probleem dat we hiervoor moeten beschikken over de gemiddelde pixelintensiteiten en de covariantiematrix. In praktijk is het zeer moeilijk hiervoor een gepaste keuze te maken.

Een derde mogelijkheid tenslotte bestaat erin het beeld te beschrijven aan de hand van conditionele waarschijnlijkheidsdistributies $p(X_i|X_J; J \subset S/i)$. Dit laat toe lokale beeldeigenschappen zoals vlakheid en randen te modelleren. Het is deze laatste aanpak die we verder zullen uitwerken. Het probleem dat zich hierbij voordoet is het volgende: kan en hoe kan de gezamenlijke waarschijnlijkheidsdistributie $p(X)$ gegenereerd worden uit alle conditionele distributies? De oplossing voor dit probleem zal geleverd worden door het verband tussen het Markov-Random-Veldmodel en de Gibbs-distributie (paragraaf 3.4).

3.3 Parameterschatting

We herhalen dat een probleem van beeldverwerking algemeen beschouwd wordt als een probleem van parameterschatting. Daarom geven we een kort overzicht van de problematiek van parameterschatting. Een meer uitgebreide beschrijving van deze klasse van problemen wordt o. a. gegeven door [VTre68].

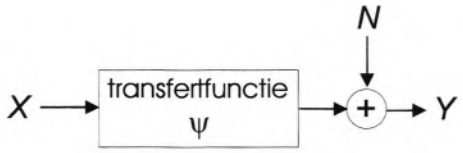
Algemeen kunnen we een schattingsprobleem voorstellen zoals in figuur 3.1. We beschouwen de parameter X als een toevalsveranderlijke met waarschijnlijkheidsdistributie $p(X)$. We merken op dat het begrip "parameter" in multidimensionale zin geïnterpreteerd moet worden. Een punt in de parameterruimte (de bron) geeft aanleiding tot waarnemingen Y. Deze waarnemingen worden voorgesteld als een punt in de waarnemingsruimte Υ. De afbeelding van de parameterruimte Ω naar de waarnemingsruimte Υ is probabilistisch van aard. Na waarneming van Y trachten we de bron X te schatten. We noteren de schatter als $\hat{X}(Y)$.



Figuur 3.1: Schematische voorstelling van het probleem van parameterschatting.

We concretiseren het abstracte model van bron en waarnemingen uit figuur 3.1. Vaak wordt voor problemen van beeldreconstructie en -analyse het schattingsmodel van figuur 3.2 gehanteerd. We merken op dat dit model perfect past binnen het model van figuur 3.1, dat evenwel meer algemeen is. In dit nieuwe model wordt de bron gevormd door het beeld X. De waarnemingen Y kunnen enerzijds een vervormde versie van het beeld X zijn of anderzijds metingen van fysische grootheden die een gevolg zijn van het beeld. Het verband tussen het beeld en de waarnemingen (de metingen) wordt gegeven door de transfertfunctie Ψ. Deze transfertfunctie zal in vele gevallen opgebouwd zijn uit een translatie-invariante puntspreidingsfunctie en een (eventueel niet-lineaire) transformatie. Verder onderstellen we nog dat er zich additieve ruis N op de metingen bevindt. De karakteristieken van het waarnemingssysteem worden volledig bepaald door de transfertfunctie Ψ en het model voor de ruis N.

We illustreren dat dit waarnemingsmodel zowel toegepast kan worden op beeldreconstructieproblemen als op beeldanalyseproblemen. Bovendien is de interpretatie van de begrippen bron en metingen afhankelijk van het probleem dat men wenst op te lossen. Bij de reconstructie van PET-beelden bv. interpreteren we het beeld van de distributie van de onderzochte tracer in de patiënt als de bron en het sinogram (t. t. z. de coïncidenties die door de detectoren gemeten worden) als de waarnemingen. Voor neurologisch onderzoek naar activatiecentra in de hersenen bv. kan het wenselijk zijn de verschillende geactiveerde centra in een PET-beeld van een label te voorzien. In dit geval interpreteren we het gereconstrueerde PET-beeld als de waarnemingen en beschouwen we de (gelabelde) activatiecentra als de bron.



Figuur 3.2: Schematische voorstelling van een waarnemingsmodel voor beelden.

Wanneer de grootheden X en Y vastgelegd zijn definiëren we de verliesfunctie (loss function) $L(X, \hat{X}(Y))$ als een functie die de kost (d. w. z. de penalisatie) van de schatting $\hat{X}(Y)$ uitdrukt. Dit betekent de penalisatie wanneer men er niet in slaagt de werkelijke bron X te bepalen en zich tevreden moet stellen met de schatting $\hat{X}(Y)$. In de meeste praktische gevallen zal deze verliesfunctie enkel afhangen van het verschil $L(X - \hat{X}(Y))$. Deze verliesfunctie dient zó gekozen te worden dat de waarde ervan overeenstemt met de tevredenheid van de waarnemer i. v. m. de schatting. In praktijk is het echter moeilijk een analytische vorm voorop te stellen die overeenstemt met de subjectieve grootheid "tevredenheid". Daarenboven moet de keuze van de verliesfunctie resulteren in een oplosbaar probleem! In vele gevallen blijkt echter dat eenzelfde schatting optimaal is voor een groot aantal verliesfuncties [VTre68].

We merken op dat de waarde van $L(X, \hat{X}(Y))$ niet berekend kan worden vermits we niet beschikken over de werkelijke bron X. Uitgaande van de verliesfunctie definiëren we daarom het risico $\mathcal{R}$ van de schatter als de verwachtingswaarde van de verliesfunctie, t. t. z.

$$\mathcal{R} = E[L(X, \hat{X}(Y))] = \int_{\Upsilon} dY \int_{\Omega} L(X, \hat{X}(Y))\, p(X, Y)\, dX. \tag{3.5}$$

We herschrijven de gezamenlijke waarschijnlijkheidsdistributie als

$$p(X, Y) = p(Y)\, p(X|Y) \tag{3.6}$$

en substitueren dit in (3.5).

$$\mathcal{R} = \int_{\Upsilon} p(Y)\, dY \int_{\Omega} L(X, \hat{X}(Y))\, p(X|Y)\, dX. \tag{3.7}$$

De Bayesiaanse schatter is de schatter die het risico minimaliseert. Zowel $p(Y)$ als de binnenste integraal zijn positief, zodat de minimalisatie van (3.7) correspondeert met de minimalisatie van de binnenste integraal. De schatting wordt dus

gegeven door

$$\hat{X}(Y) = \arg\min_{\hat{X}} \int_{\Omega} L(X, \hat{X}(Y))\, p(X|Y)\, dX. \tag{3.8}$$

Afhankelijk van de keuze van de verliesfunctie leidt dit tot verschillende criteria. We bespreken hieronder kort de meest gebruikte verliesfuncties.

3.3.1 De MAP-schatter

$$\begin{aligned} L_{MAP}(X, \hat{X}(Y)) &= 1 - \delta(X - \hat{X}(Y)), \\ \hat{X}_{MAP}(Y) &= \arg\max_{X} p(X|Y). \end{aligned} \tag{3.9}$$

De MAP-schatter (maximum a posteriori) vindt de parameterwaarde waarvoor de a posteriori-waarschijnlijkheid maximaal is (vandaar de benaming). Substitueren we de regel van Bayes in (3.9) en maken we gebruik van het feit dat de logaritme een monotone functie is, dan vinden we dat

$$\begin{aligned} \hat{X}_{MAP}(Y) &= \arg\max_{X} \frac{p(Y|X)\, p(X)}{p(Y)} \\ &= \arg\max_{X} [\ln p(Y|X) + \ln p(X)]. \end{aligned} \tag{3.10}$$

Aangezien zowel $p(Y|X)$ als $p(X)$ gekend zijn, kan de MAP-schatting geformuleerd worden als een optimalisatieprobleem, waarbij de uitdrukking $\ln p(Y|X) + \ln p(X)$ de rol van kostfunctie vervult. We zullen in paragraaf 3.8 een overzicht geven van de meest gebruikte optimalisatietechnieken.

Een bijzonder geval van de MAP-schatting ontstaat wanneer de a priori-distributie uniform gekozen wordt over de volledige toestandsruimte. We merken op dat de uniforme distributie correspondeert met de maximum-entropie-oplossing wanneer er geen voorkennis beschikbaar is [Jayn68]. In dit geval verdwijnt de tweede term in de bovenstaande uitdrukking en vinden we de bekende uitdrukking voor de Maximum-Likelihoodschatting (ML):

$$\hat{X}_{ML}(Y) = \arg\max_{X} [\ln p(Y|X)]. \tag{3.11}$$

Omgekeerd kunnen we stellen dat de MAP-schatting een uitbreiding is van de ML-schatting door toevoeging van een extra term $\ln p(X)$ [Lian89]. Vandaar dat

MAP ook vaak "penalized ML" of "weighted ML" genoemd wordt. Zoals gezegd drukt deze extra term onze voorkennis uit. Wanneer het inverse probleem goed gesteld is zal $\ln p(Y|X)$ een duidelijke piek vertonen rond de ML-oplossing. In dit geval zal de extra a priori-term geen invloed hebben op de ligging van het maximum. Wanneer echter de transfertfunctie tussen X en Y een grote nulruimte heeft, dan zullen een groot aantal oplossingen voldoen aan het maximum likelihood-criterium. In dit geval zal de toevoeging van een extra term $\ln p(X)$ een regulariserend effect hebben. Algemeen kunnen we stellen dat het Bayesiaanse formalisme een theoretisch gefundeerd model biedt dat toelaat aanvullende informatie te gebruiken om de beeldkwaliteit te verbeteren.

3.3.2 De MPM-schatter

De benaming MPM staat voor maximale marginale a posteriori's (Maximum of Posterior Marginals):

$$L_{MPM}(X, \hat{X}(Y)) = 1 - \sum_{i=1}^{N} \delta(X_i - \hat{X}_i(Y)),$$
$$\hat{X}_{MPM}(Y) = \{\arg\max_{X_i} p(X_i|Y), i = 1, \ldots, N\}. \tag{3.12}$$

De MPM-schatter vindt een oplossing door elke component afzonderlijk te schatten als de waarde die de marginale a posteriori-waarschijnlijkheidsdistributie maximaliseert. De MPM-schatter is vooral bruikbaar voor classificatieproblemen (bv. weefsellabeling of segmentatie) omdat hij tracht elke pixel afzonderlijk correct te identificeren (in tegenstelling tot de MAP-schatter, waarbij de foute identificatie van één pixel even zwaar bestraft wordt als de foute identificatie van alle pixels). Het probleem is echter dat de marginale distributie voor elke pixel afzonderlijk berekend moet worden. Dit kan enkel gebeuren door marginalisatie van de totale a posteriori-distributie $p(X|Y)$, zodat deze distributie gekend moet zijn (dit in tegenstelling tot de MAP-schatter, waar men niet over $p(Y|X)$ moet beschikken). De exacte berekening hiervan is echter onmogelijk, zodat een beroep gedaan zal moeten worden op bemonsteringsmethoden om de MPM-schatting te benaderen. Deze bemonsteringsmethoden worden verder besproken in paragraaf 3.9.

3.3.3 De MMSE-schatter

De benaming MMSE staat voor minimale gemiddelde kwadratische fout (Minimum Mean Square Error):

$$\begin{aligned} L_{MMSE}(X, \hat{X}(Y)) &= \|X - \hat{X}(Y)\|^2, \\ \hat{X}_{MMSE}(Y) &= E[X|Y] = \int_{\Omega} X\, p(X|Y)\, dX. \end{aligned} \tag{3.13}$$

De MMSE-schatter vindt als resultaat het a posteriori-gemiddelde. Deze methode heeft als voordeel dat foute oplossingen in toenemende mate gepenaliseerd worden naarmate zij meer afwijken van de correcte schatting (in tegenstelling tot beide vorige schatters, die enkel het onderscheid tussen een correcte en een foutieve schatting maken). Het probleem is echter dat een exacte berekening van de MMSE-schatting onmogelijk is, aangezien hiervoor de a posteriori-waarschijnlijkheidsdistributie volledig gekend moet zijn. Meestal wordt de MMSE-schatting benaderd door bemonsteringsmethoden te gebruiken. We verwijzen ook hier naar paragraaf 3.9 voor een bespreking van bemonsteringsmethoden.

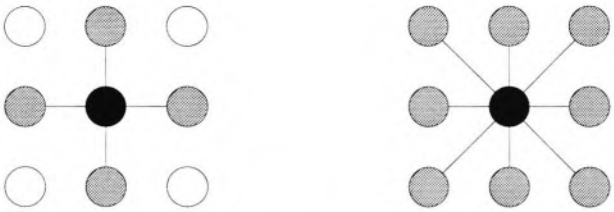
3.4 Het Markov-Random-Veldmodel

Op het einde van paragraaf 3.2 werd voorgesteld om spatiale correlaties in het beeld te modelleren aan de hand van conditionele waarschijnlijkheidsdistributies $p(X_i|X_J; J \subset S/i)$. We stuiten hierbij echter op de vraag hoe de gezamenlijke waarschijnlijkheidsdistributie $p(X)$ afgeleid kan worden uit de diverse conditionele termen. Hiervoor introduceren we een speciaal random-veldmodel, t. t. z. het Markov-Random-Veldmodel (MRV) [Azen87a, Azen87b, Wink95].

Het MRV is een tweedimensionale extensie van de Markov-keten. We bespreken daarom kort het model van de Markov-keten. We beschouwen een reeks toevalsgrootheden $\{Z_1, \ldots, Z_i, \ldots\}$, waarbij Z_i de i^{de} toestand van een systeem beschrijft (bv. de toestand van het systeem op tijdstip $t_0 + i\,\Delta t$). We noteren de "voorgeschiedenis" van de i^{de} toestand (d. w. z. de voorgaande $i-1$ toestanden) als $\mathcal{Z}_i = \{Z_1, \ldots, Z_{i-1}\}$. De reeks toevalsgrootheden vormt een Markov-keten wanneer geldt dat

$$p(Z_i|\mathcal{Z}_i) = p(Z_i|Z_{i-1}). \tag{3.14}$$

Dit betekent dat de huidige toestand van het systeem enkel afhankelijk is van de vorige toestand, maar niet van de voorgeschiedenis van het systeem (t. t. z. de



Figuur 3.3: Eerste- en tweede-orde omgevingsstructuur.

wijze waarop de vorige toestand tot stand is gekomen). Intuïtief kunnen we dit omschrijven als een "systeem met een beperkt geheugen".

Wanneer we het model van de Markov-keten willen uitbreiden tot beelden moeten we het begrip omgeving introduceren. We herhalen dat $S = \{1, \ldots, N\}$ de verzameling van alle pixellocaties voorstelt. We introduceren een omgeving van pixel s als een deelverzameling $\mathcal{S}_s \subset S$. Wanneer we voor elke pixel s een omgeving $\mathcal{S}_s$ definiëren, ontstaat een verzameling van N deelverzamelingen van S, die we voorstellen door $\mathcal{S} = \{\mathcal{S}_s | s \in S\}$. We noemen $\mathcal{S}$ een omgevingsstructuur op S wanneer geldt dat:

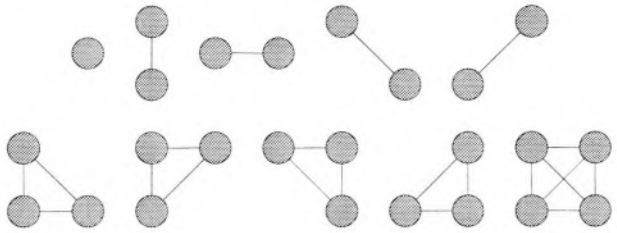
1. $s \notin \mathcal{S}_s$;
2. $s \in \mathcal{S}_r \Leftrightarrow r \in \mathcal{S}_s$.

Veel gebruikte omgevingsstructuren zijn de eerste-orde (4 naburige pixels) en de tweede-orde (8 naburige pixels) omgevingsstructuur (figuur 3.3). We noteren de combinatie van het rooster en de erop gedefinieerde omgevingsstructuur als $(S, \mathcal{S})$. Dit kan geïnterpreteerd worden als een graaf waarvan de knopen bepaald worden door S en de takken door $\mathcal{S}$ (de takken geven de nabuur-relatie aan).

We definiëren een Markov Random Veld als volgt: X is een MRV met betrekking tot $(S, \mathcal{S})$ als geldt dat:

1. $p(X) > 0, \forall X \in \Omega$;
2. $p(X_i | X_{S/i}) = p(X_i | X_{\mathcal{S}_i})$.

Met andere woorden: de conditionele waarschijnlijkheid van een pixelwaarde is enkel afhankelijk van de naburige pixelwaarden en niet van het volledige beeld. Deze eigenschap van beperkte afhankelijkheid noemt men de Markov-eigenschap.



Figuur 3.4: Verschillende types van klieken die corresponderen met een tweede-orde-omgevingsstructuur.

We herhalen dat we voorkennis wensen uit te drukken m. b. v. de a priori-distributie $p(X)$. De invoering van het MRV-model levert – tot nog toe – enkel een uitdrukking voor de conditionele distributies $p(X_i|X_J: J \subset S/i)$. Vooraleer we het verband kunnen leggen tussen beide distributies moeten we het begrip "kliek" (clique) invoeren. Een kliek $C \subset S$ is een verzameling pixelplaatsen die ofwel uit slechts één pixel bestaat ofwel uit pixels die tot elkaars omgeving behoren. De verzameling van alle klieken m. b. t. $(S, \mathcal{S})$ stellen we voor door $\mathcal{C}$. Figuur 3.4 stelt de verschillende types van klieken voor die corresponderen met een tweede-orde-omgevingsstructuur. De verzameling $\mathcal{C}$ ontstaat dan door alle kliektypes op alle mogelijke roosterposities te plaatsen. We vestigen er de aandacht op dat $\mathcal{C}$ rechtstreeks volgt uit de keuze voor de omgevingsstructuur $\mathcal{S}$. In het vervolg van deze uiteenzetting wordt gebruik gemaakt van klieken om een belangrijk verband te leggen tussen een MRV en een Gibbs-distributie. We introduceren eerst dit laatste begrip.

3.5 De Gibbs-distributie

De Gibbs-distributie is een speciale vorm van de Boltzmann-distributie, die gegeven wordt door

$$p(X) = \frac{1}{Z} \exp(-\beta H(X)). \tag{3.15}$$

Hierin stelt X een algemene toevalsgrootheid voor, $H(X)$ is de energiefunctie en β de regularisatieparameter. De partitiefunctie Z, die afhankelijk is van de waarde van β, vervult de rol van normalisatieconstante en wordt gegeven door

$$Z = \sum_{\Omega} \exp -\beta H(X). \tag{3.16}$$

De vorm van de energiefunctie $H(X)$ is willekeurig voor de Boltzmann-distributie. We spreken evenwel van een Gibbs-distributie wanneer $H(X)$ geschreven kan worden als de som van potentiaalfuncties

$$H(X) = \sum_{C \in \mathcal{C}} V_C(X). \tag{3.17}$$

De potentiaalfunctie $V_C(X)$ is enkel functie van de waarden van de pixels die tot de kliek C behoren. De waarde ervan stemt overeen met de bijdrage tot $H(X)$ van de pixels van de kliek C. We merken echter op dat een pixel tot meerdere klieken zal behoren en bijgevolg ook aanleiding zal geven tot meerdere potentiaalbijdragen. De potentiaalbijdrage van een kliek C moet dan ook geïnterpreteerd worden als de bijdrage die de pixels als groep, door hun onderlinge "interactie" (we herhalen dat alle pixels van een kliek elkaars buren zijn), leveren tot de totale energie van het systeem. De verzameling van alle klieken $\mathcal{C}$ kan dus ook beschouwd worden als de verzameling van alle groepen van pixels die onderling afhankelijk zijn.

Een belangrijke eigenschap is de equivalentie tussen een Gibbs-distributie en een MRV. Dit wordt uitgedrukt door het Hammersley-Clifford-theorema, dat stelt dat X een MRV is m. b. t. $(S, \mathcal{S})$ enkel en alleen indien geldt dat $p(X)$ een Gibbs-distributie is. Een bewijs voor dit theorema wordt o. a. gegeven door [Besa74] en [Wink95]. Het belang van dit theorema ligt in het verband dat hiermee gelegd wordt tussen de gezamenlijke waarschijnlijkheidsdistributie (Gibbs-distributie) en de conditionele waarschijnlijkheidsdistributies (MRV).

$$\begin{aligned} p(X_i|X_{S/i}) &= \frac{p(X_i, X_{S/i})}{p(X_{S/i})} \\ &= \frac{p(X)}{\sum_{X_i \in \Lambda} p(X)} \\ &= \frac{\exp\left(-\beta \sum_{C \in \mathcal{C}} V_C(X)\right)}{\sum_{X_i \in \Lambda} \exp\left(-\beta \sum_{C \in \mathcal{C}} V_C(X)\right)} \\ &= \frac{\exp\left(-\beta \sum_{C:i \in C} V_C(X)\right)}{\sum_{X_i \in \Lambda} \exp\left(-\beta \sum_{C:i \in C} V_C(X)\right)}. \end{aligned} \tag{3.18}$$

De tweede uitdrukking volgt door marginalisatie; de vierde uitdrukking wordt bekomen door de gemeenschappelijke factoren in teller en noemer weg te delen. De gemeenschappelijke factoren zijn de deelpotentialen die overeenstemmen met klieken waartoe X_i niet behoort; vandaar $C : i \in C$, d. w. z. enkel de klieken waartoe X_i behoort blijven behouden. Uit deze formule blijkt duidelijk de

equivalentie tussen de conditionele waarschijnlijkheidsdistributies en de potentiaalfuncties. Het opstellen van een beeldmodel aan de hand van conditionele waarschijnlijkheden is hiermee gereduceerd tot de keuze van een gepaste set potentiaalfuncties. In de meeste gevallen zullen daarenboven de potentiaalfuncties enkel per type kliek gekozen worden.

We illustreren het gebruik van het MRV-model. Veronderstel twee beelden $X^{(1)}$ en $X^{(2)}$ waarvan enkel een beperkt aantal pixelwaarden van elkaar verschillen. Een grootheid die veelvuldig gebruikt wordt is de verhouding van waarschijnlijkheden van beide beelden

$$\frac{p(X^{(1)})}{p(X^{(2)})}. \tag{3.19}$$

Dit leidt tot de berekening van het energieverschil

$$\Delta H = H(X^{(2)}) - H(X^{(1)}). \tag{3.20}$$

Aangezien slechts enkele pixelwaarden verschillen moeten de potentiaalbijdragen enkel berekend worden voor de klieken waarvan de verschillende pixels deel uitmaken. Voor lokale omgevingsstructuren geeft dit aanleiding tot een sterke vereenvoudiging van de berekeningen.

3.6 De a priori-waarschijnlijkheidsdistributie

De keuze van de a priori-waarschijnlijkheidsdistributie moet een weergave zijn van de voorkennis omtrent het te schatten beeld. De extra kennis die ingebracht wordt moet toelaten de beeldkwaliteit te verbeteren. Anderzijds mag deze voorkennis geen overheersend sturende rol spelen m. b. t. het beeld dat gereconstrueerd moet worden. Wanneer bv. de a priori-waarschijnlijkheidsdistributie oplegt dat de tracerverdeling binnen een patiënt steeds overeenstemt met de tracerverdeling binnen een "model"-individu, dan loopt men het risico dat anomalieën niet gedetecteerd kunnen worden. Het zijn echter vaak juist deze anomalieën die de medicus relevante informatie kunnen verschaffen.

De meeste medische reconstructiealgoritmen (bv. ML-EM) zijn onderhevig aan het probleem van ruisdeterioratie [Snyd87]. In dit geval zullen naburige pixels sterk verschillende intensiteiten aannemen. De meeste a priori-termen zullen daarom een vorm van vlakheid trachten op te leggen. In de meest algemene

vorm bestaat de a priori-waarschijnlijkheidsdistributie uit bijdragen van alle kliektypes die een gevolg zijn van de gekozen omgevingsstructuur. In de praktijk wordt echter uitsluitend rekening gehouden met potentiaalbijdragen t. g. v. twee-pixelinteracties. Dit komt er op neer dat de potentiaalbijdragen horend bij klieken die bestaan uit 1 pixel of meer dan 2 pixels op nul gesteld worden. Bovendien worden deze twee-pixelinteractietermen enkel functie gemaakt van het intensiteitsverschil tussen beide pixels:

$$H(X) = \sum_{i \in S} \sum_{j \in \mathcal{S}_i} b_{ij} \phi(X_i - X_j). \tag{3.21}$$

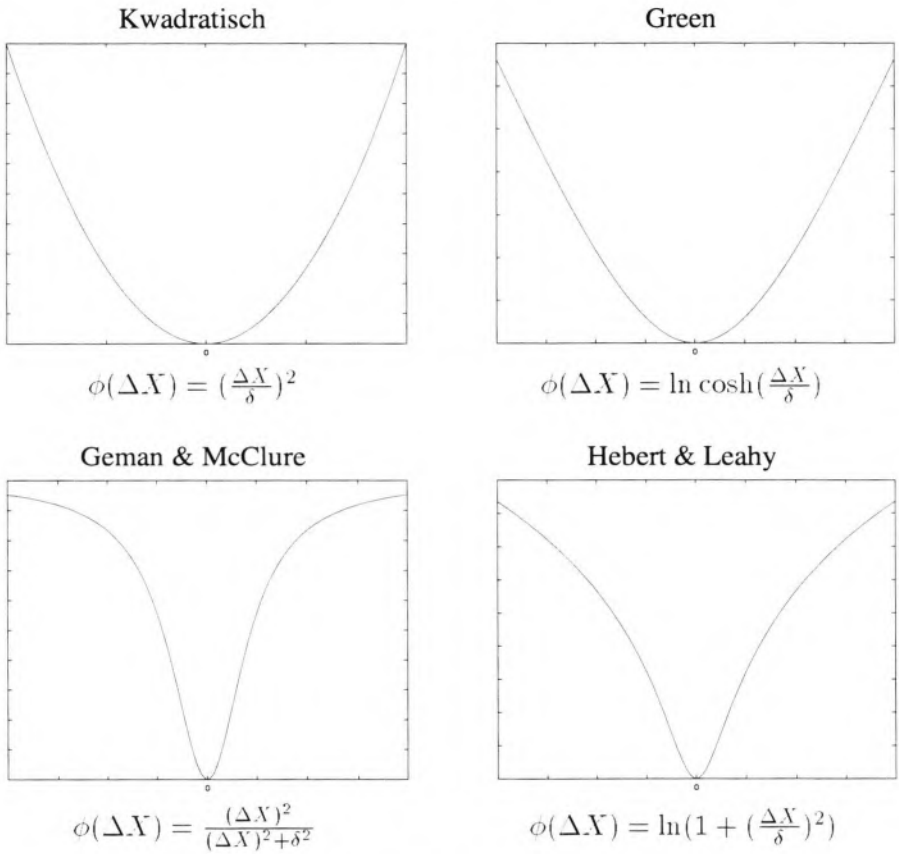
Een overzicht van enkele veel gebruikte potentiaalfuncties wordt gegeven in figuur 3.5. We verwijzen naar [Boum94] en [Gind93a] voor toepassingen waarbij gebruik gemaakt wordt van hiërarchische MRV-modellen om ook hogere-orde-interacties te modelleren. Jennison *et al.* gebruiken een verfijnd MRV-model waarbij randen door i. p. v. tussen pixels lopen [Jenn88].

Alvorens de bespreking van deze potentiaalfuncties voort te zetten, definiëren we eerst het begrip convexe functie. Volgens Rao is een functie $f(x)$ convex wanneer, voor $\alpha \in [0, 1]$, geldt dat [Rao65]

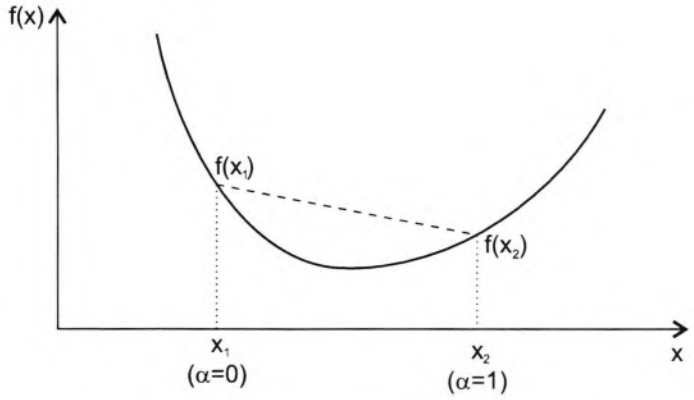
$$f((1-\alpha)x_1 + \alpha x_2) \leq (1-\alpha) f(x_1) + \alpha f(x_2). \tag{3.22}$$

Deze definitie wordt geïllustreerd a. h. v. figuur 3.6. Vervolgens kunnen de potentiaalfuncties ingedeeld worden in twee groepen, nl. convexe functies (figuur 3.5, kwadratisch en Green) en niet-convexe functies (figuur 3.5, Geman & McClure en Hebert & Leahy). Een veel gebruikte convexe potentiaalfunctie is bv. de kwadratische term. Convexe functies hebben als voordeel dat zij na sommatie opnieuw een convexe functie opleveren. Wanneer de totale energiefunctie $H(X)$ convex is zal het reconstructieprobleem geen lokale minima vertonen. Uit diverse toepassingen blijkt dat een convexe term uiterst geschikt is om kleine intensiteitsvariaties te onderdrukken, maar totaal ongeschikt is voor het modelleren van randen. De verklaring hiervoor is dat intensiteitsverschillen sterker dan proportioneel gepenaliseerd worden, waardoor (relatief grote) intensiteitsverschillen die voorkomen langs randen in het beeld uitgevlakt worden.

De twee belangrijkste lokale karakteristieken van functionele medische beelden (bv. PET en SPECT) zijn enerzijds lokale vlakheid (binnen homogene gebieden, bv. de opname van FDG in grijze hersenmassa) en anderzijds het bestaan van scherpe randen op de scheiding tussen lokaal vlakke gebieden (bv. het verschil in FDG-opname tussen witte en grijze massa). Een bruikbaar MRV-model



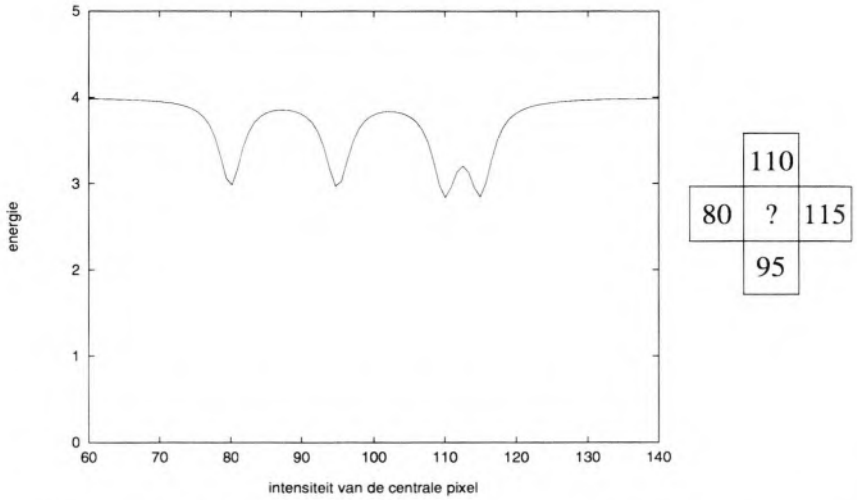
Figuur 3.5: Overzicht van de meest gebruikte a priori-potentiaalfuncties.



Figuur 3.6: Illustratie van de definitie van een convexe functie.

voor functionele beelden zal dan ook in staat moeten zijn beide eigenschappen te modelleren. Niet-convexe potentiaalfuncties sluiten nauwer aan bij dit model [Lian91]. De motivatie voor het gebruik van niet-convexe functies is in essentie de volgende: kleine intensiteitsverschillen worden verondersteld afkomstig te zijn van ruis op de waarnemingen en moeten bijgevolg onderdrukt worden; grote intensiteitsverschillen daarentegen worden verondersteld essentiële beeldkarakteristieken (randen) te zijn en moeten bijgevolg minder zwaar bestraft worden [Geig91, Gema92]. Wanneer het intensiteitsverschil een drempelwaarde δ overschrijdt wordt slechts een relatief kleine extra penalisatie toegekend voor een verdere toename van het intensiteitsverschil. De parameter δ bepaalt bijgevolg welke intensiteitsverschillen als ruis en welke als rand beschouwd worden [Han93]. Een bespreking van niet-convexe potentiaalfuncties voor computervisie wordt gegeven door Li *et al.* [Li95a, Li95b, Li95c, Li97].

Het belangrijkste nadeel van niet-convexe potentiaalfuncties is het ontstaan van een groot aantal lokale minima. We illustreren dit aan de hand van het voorbeeld in figuur 3.7. De som van 4 potentiaalbijdragen (t. g. v. een centrale pixel en zijn 4 naburen) wordt uitgezet i. f. v. de intensiteit van de centrale pixel. Hoewel in dit voorbeeld de δ-waarde onrealistisch laag gekozen wordt, illustreert het duidelijk het ontstaan van lokale minima. Algemeen zal de totale energie van een beeld bestaan uit een groot aantal potentiaaltermen. Deze potentiaalbijdragen zijn gelijk van vorm maar hebben een verschillende ligging van het minimum (omdat de pixelintensiteiten verschillen over het beeld). Bij sommatie kan dit, afhankelijk van de δ-waarde, aanleiding geven tot een groot aantal lokale minima. Vandaar



Figuur 3.7: Illustratie van het ontstaan van lokale minima bij niet-convexe potentiaalfuncties: som van 4 potentiaalbijdragen (Geman & McClure, $\delta = 2$) i. f. v. de intensiteit van de centrale pixel.

de noodzaak om aangepaste optimalisatiemethoden te gebruiken. Het zijn deze lokale minima die ertoe leiden dat simulated annealing als optimalisatiemethode gekozen wordt.

Behalve de hierboven besproken continue potentiaalfuncties kan ook gebruik gemaakt worden van een discreet MRV-model. Een eenvoudig voorbeeld wordt gegeven door:

$$\begin{aligned} H(X) &= \sum_{i \in S} \sum_{j \in \mathcal{S}_i} V(X_i, X_j), \\ V(X_i, X_j) &= \begin{cases} 0, & X_i = X_j \\ 1, & X_i \neq X_j. \end{cases} \end{aligned} \tag{3.23}$$

Dit soort modellen is vooral bruikbaar bij classificatie-toepassingen zoals bv. labeling van weefseltypes en segmentatie. We gaan hier echter niet dieper op in.

Een enigzins afwijkend a priori-model dat gebruik maakt van een mediaanfilter wordt beschreven door Alenius *et al.* [Alen94, Alen97]. Een klasse van a priori-distributies die buiten het MRV-model vallen zijn distributies die gebaseerd zijn op het maximum-entropieprincipe [Gull78]. Hierbij wordt de a priori-distributie rechtstreeks over het volledige beeld gedefinieerd (waardoor de behoefte aan het MRV-model wegvalt). Afhankelijk van de gekozen entropie-uitdruk-

king (bv. Shannon-entropie) geeft dit aanleiding tot verschillende uitdrukkingen voor de a priori-distributies. We verwijzen o. a. naar [Desm95] voor een gedetailleerde studie van deze vorm van a priori-distributies.

3.7 De a posteriori-waarschijnlijkheidsdistributie

Tot hier toe hebben we ons beperkt tot de a priori-waarschijnlijkheidsdistributie in de bespreking van het MRV-model en de Gibbs-distributie. De in paragraaf 3.3 besproken schatters steunen echter op de a posteriori-distributie (de MAP-schatter bv. zoekt het beeld dat de a posteriori-waarschijnlijkheidsdistributie maximaliseert). Het blijkt echter dat ook de a posteriori-distributie steeds geschreven kan worden als een Gibbs-distributie. Gebruik makend van de regel van Bayes vinden we dat

$$\begin{aligned} p(X|Y) &= \frac{p(Y|X)p(X)}{p(Y)} \\ &= \frac{1}{p(Y)} \exp\left(\ln p(Y|X) + \ln p(X)\right). \end{aligned} \tag{3.24}$$

We herschrijven dit als

$$p(X|Y) = \frac{1}{Z_{X|Y}} \exp\left(-H_{X|Y}(X,Y)\right). \tag{3.25}$$

waarbij $Z_{X|Y}$ de partitiefunctie voor de a posteriori-distributie voorstelt en de energiefunctie gegeven wordt door

$$\begin{aligned} H_{X|Y}(X,Y) &= -\ln p(Y|X) - \ln p(X)) \\ &= H_{Y|X}(X,Y) + \beta H_X(X). \end{aligned} \tag{3.26}$$

De energiefunctie bestaat uit twee termen: enerzijds een term die afhankelijk is van de directe waarschijnlijkheidsdistributie en die we de dataterm zullen noemen, en anderzijds een a priori-term. In hoofdstuk 6 zullen we een gedetailleerde analyse van beide termen maken. We voeren hier reeds de notaties in die gehanteerd zullen worden in hoofdstuk 6, t. t. z.

$$H(X,Y) = H_D(X,Y) + \beta H_P(X). \tag{3.27}$$

De ontstane a posteriori-energiefunctie $H_{X|Y}(X,Y)$ bestaat uit een sommatie van termen die we interpreteren als potentiaalfuncties $V_{X|Y}(X,Y)$. De pixels

die gemeenschappelijk voorkomen in een potentiaalterm $V_{X|Y}(X, Y)$ behoren tot een kliek en zijn bijgevolg elkaars naburen. Op deze manier ontstaat een omgevingsstructuur $\mathcal{S}_{X|Y}$ voor de a posteriori-distributie. Deze omgevingsstructuur is de unie van de omgevingsstructuren t. g. v. de directe en de a priori-waarschijnlijkheidsdistributies:

$$\mathcal{S}_{X|Y} = \mathcal{S}_{Y|X} \cup \mathcal{S}_X. \tag{3.28}$$

Hoewel we de a posteriori-waarschijnlijkheidsdistributie dus steeds in Gibbs-vorm kunnen schrijven, zal deze slechts in een beperkt aantal gevallen ook aanleiding geven tot een lokale omgevingsstructuur. Voor het geval van PET zal echter $(S, \mathcal{S}_{X|Y})$ een volledig geconnecteerd graaf vormen, d. w. z. dat de omgeving van de i^{de} pixel gegeven wordt door

$$\mathcal{S}_i = S/i. \tag{3.29}$$

Dit is een gevolg van het specifieke verband tussen X en Y (Y ontstaat door de berekening van strookintegralen over X voor verschillende richtingen). Toch is het ook in zulke gevallen nuttig om de a posteriori-distributie als een Gibbs-distributie te beschouwen.

Een interessante extensie van de theorie van Gibbs-distributies wordt gegeven door Hanson *et al.* [Hans95]. Wanneer de negatieve logaritme van de waarschijnlijkheidsdistributie geïnterpreteerd wordt als een potentiaal, dan kan de gradiënt ervan beschouwd worden als een kracht. De grootte van deze kracht in de omgeving van de oplossing kan dan beschouwd worden als een maat voor de "betrouwbaarheid" van deze oplossing.

Een belangrijke doch onbekende parameter van de a posteriori-waarschijnlijkheidsdistributie is de regularisatieparameter β. De waarde ervan bepaalt het relatieve belang van de directe waarschijnlijkheidsdistributie en de a priori-distributie. In de meeste gevallen wordt de waarde van β geschat uit een calibratie m. b. v. een set van referentiebeelden [Chin92, Chin93]. Voor PET-reconstructie beschikken we echter niet over zulke referentiebeelden. Daarom moet β rechtstreeks uit de waarnemingen geschat worden. Vanuit een Bayesiaans oogpunt beschouwen we deze parameter als een extra toevalsveranderlijke. We gaan er van uit dat deze distributie uniform is (elke β-waarde is a priori even waarschijnlijk), zodat we de schatting van β beschouwen als een ML-schatting.

De ML-schatting voor β wordt gegeven door de maximalisatie van

$$p(Y|\beta) = \int_{\Omega} p(Y|X)\, p(X|\beta)\, dX. \tag{3.30}$$

De a priori-distributie $p(X|\beta)$ is afhankelijk van β

$$p(X|\beta) = \frac{1}{Z(\beta)} \exp(-\beta H(X)), \tag{3.31}$$

waarin de β-afhankelijkheid van de partitiefunctie expliciet aangegeven wordt. Men kan aantonen dat

$$p(Y|\beta) = \frac{Z(Y, \beta)}{Z(\beta)}. \tag{3.32}$$

waarbij

$$Z(Y, \beta) = \int_{\Omega} \exp(\ln p(Y|X) - \beta H(X)) dX \tag{3.33}$$

de partitiefunctie voor $p(X|Y)$ voorstelt. Wanneer we de afgeleide van (3.32) aan nul gelijkstellen verkrijgen we een vergelijking voor de ML-schatting van β:

$$E[H(X)|Y, \beta] = E[H(X)|\beta]. \tag{3.34}$$

Een analytische oplossing van deze vergelijking is echter onmogelijk, vermits dit een integratie over de a posteriori-distributie en de a priori-distributie in respectievelijk het linker- en rechterlid impliceert. Bekijken we een EM-algoritme om β te bepalen (het EM-algoritme wordt meer in detail besproken in de volgende paragraaf),

$$\begin{aligned} &\text{E-stap} : H^k(X) = E[H(X)|Y, \beta^k] \\ &\text{M-stap} : \beta^{k+1} \text{ volgt uit} E[H(X)|\beta] = H^k(X) \end{aligned} \tag{3.35}$$

dan zien we dat elke stap opnieuw bestaat uit de integratie over één van beide distributies (a posteriori en a priori). We merken hierbij op dat een bovenindex k de waarde tijdens de k^{de} iteratiestap voorstelt. De bepaling van een gepaste waarde van β kan dus enkel opgelost worden door gebruik te maken van rekenintensieve benaderingsmethoden. Deze methode om β als een extra te schatten parameter tijdens het reconstructieproces te beschouwen, wordt o. a. toegepast door [John91, Lalu92, Fess96b].

Een veralgemeende MAP-methode om gelijktijdig het beeld X en de regularisatieparameter β te schatten werkt als volgt:

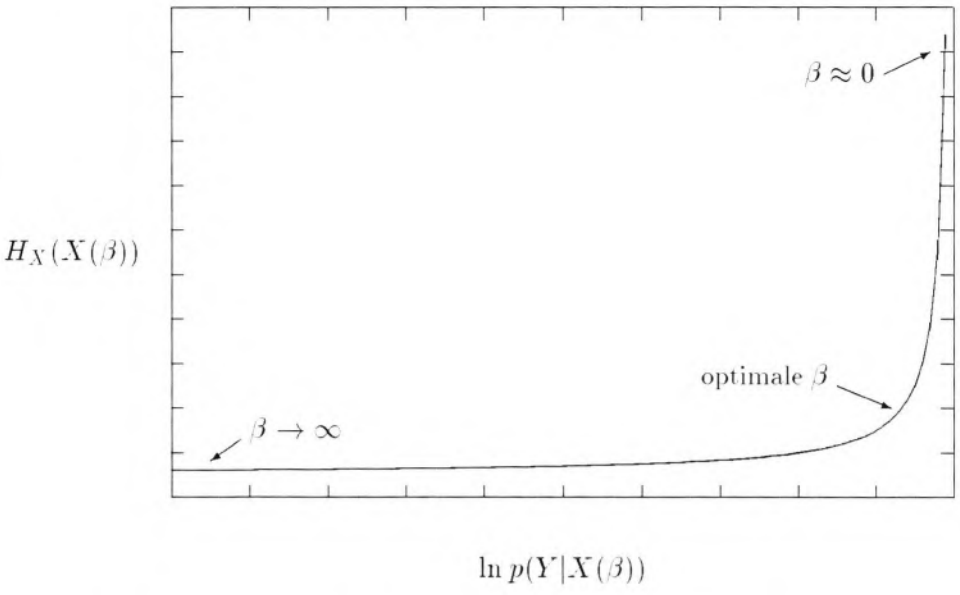
1. kies een initiële waarde β^0;
2. bepaal een nieuwe MAP-schatting X^{k+1}, uitgaande van β^k;
3. bepaal een nieuwe ML-schatting β^{k+1} op basis van X^{k+1};
4. herhaal tot wanneer zowel X als β geconvergeerd zijn.

Deze methoden zijn o. a. onderzocht door Manbeck [Manb90]. Uit zijn onderzoek blijkt dat deze methoden rekenintensief zijn en praktisch enkel bruikbaar voor optimalisatietechnieken die de MAP-oplossing voor X snel bepalen. Dit leidt ertoe dat deze methoden praktisch onbruikbaar zijn in combinatie met niet-convexe a priori-distributies. Recent onderzoek van Higdon *et al.* bepaalt de regularisatieparameter door de beide partitiefuncties te bemonsteren m. b. v. MKMC-methoden (zie paragraaf 3.9) [Higd97], wat echter een nog sterkere toename van de rekentijd veroorzaakt.

De meeste onderzoekers bepalen de waarde van de regularisatieparameter m. b. v. trial en error. Deze methoden zijn meestal gebaseerd op de zgn. L-curve [Hans92]. Voor een groot aantal β-waarden wordt het bijbehorende gereconstrueerde beeld $X(\beta)$ bepaald. Wanneer de a priori-Gibbs-energie $H_X(X(\beta))$ van het beeld $X(\beta)$ uitgezet wordt i. f. v. de logaritmische directe waarschijnlijkheid $\ln p(Y|X(\beta))$, dan ontstaat een L-vormige curve zoals weergegeven in figuur 3.8. Voor $\beta = 0$ vinden we de ML-schatting, zodat $p(Y|X)$ maximaal is; dit beeld zal typisch sterke intensiteitsverschillen bevatten zodat $H_X(X)$ een hoge waarde heeft. Anderzijds zal voor $\beta \to \infty$ de oplossing enkel bepaald worden door de voorkennis, zodat $H_X(X)$ minimaal wordt. De oplossing zal echter vrijwel onafhankelijk zijn van de waarnemingen, zodat $p(Y|X)$ klein is. De ideale β-waarde is deze waarvoor $H_X(X)$ klein en $p(Y|X)$ groot is. Dit punt kan echter enkel gevonden worden door de MAP-reconstructie uit te voeren voor een aantal β-instellingen en de L-curve uit te zetten.

3.8 Optimalisatiemethoden

In paragraaf 3.3 werd reeds opgemerkt dat voor de bepaling van de MMSE-schatting en de MPM-schatting een beroep gedaan moet worden op bemonsteringsmethoden om de benodigde distributies (a posteriori-distributie voor MMSE, marginale distributies voor MPM) te bepalen. In een volgende paragraaf zullen we dieper ingaan op enkele van de meest gebruikte bemonsteringstechnieken. Deze technieken zijn echter computationeel intensief. Daarom wordt voor het



Figuur 3.8: L-curve: verloop van de a priori-Gibbs-energie $H_X(X)$ vs. de logaritme van de directe waarschijnlijkheid $\ln p(Y|X)$.

merendeel van de Bayesiaanse beeldreconstructietoepassingen (of beeldanalysetoepassingen in het algemeen) de MAP-schatting berekend. We herhalen dat de MAP-schatter correspondeert met

$$\begin{aligned} \hat{X}_{MAP}(Y) &= \arg\max_X \frac{p(Y|X)\,p(X)}{p(Y)} \\ &= \arg\max_X [\ln p(Y|X) + \ln p(X)]. \end{aligned} \tag{3.36}$$

Dit betekent dat de berekening van de MAP-schatting overeenstemt met een optimalisatieprobleem met als kostfunctie $\ln p(Y|X) + \ln p(X)$. We bespreken kort enkele van de meeste toegepaste optimalisatietechnieken. Een eenvoudige implementatie van deze technieken kan gevonden worden in [Pres92].

3.8.1 Expectation Maximization-algoritmen

De populairste en meest bestudeerde Bayesiaanse reconstructiemethode voor tomografische toepassingen is zonder twijfel de EM-methode. Dit algoritme werd geïntroduceerd door Dempster [Demp77] en Shepp en Vardi [Shep82, Shep84]. In

de meeste gevallen wordt deze optimalisatiemethode gecombineerd met de ML-schatter tot het bekende ML-EM-algoritme. Het EM-algoritme bestaat uit twee stappen, nl.

1. de E-stap (Expectation step): berekent de conditionele verwachtingswaarde $E[\ln p(X|Y: \hat{X}^i)]$, waarbij $\hat{X}^i$ de i^{de} schatting voor X is;
2. de M-stap (Maximization step): de volgende schatting $\hat{X}^{i+1}$ is dan gelijk aan de intensiteiten die de conditionele verwachtingswaarde maximaliseren

$$\hat{X}^{i+1} = \arg\max_{X} E[\ln p(X|Y: \hat{X}^i)]. \tag{3.37}$$

Een belangrijk voordeel van dit algoritme is het gegarandeerd positief zijn van de oplossing (wat een vereiste is) en de relatief eenvoudige implementatie. Een nadeel van deze methode is de trage convergentie. Een uitbreiding van het ML-EM-algoritme die rekening houdt met meetfouten (random-coïncidenties en scatter) wordt besproken in [Olli94]. Recent onderzoek heeft verder geleid tot verbeterde algoritmen met een snellere convergentie en/of een hogere beeldkwaliteit. We vermelden o. a. OSEM (Ordered Subset EM) [Huds94], het gebruik van multi-roostermethoden [Rang88, Pan91, Doni94, Doni95], frequentie-afhankelijke versterking en attenuatie-compensatie [Nuyt93], het gebruik van een dempingsmatrix voor de pixelaanpassingen [Tana92] en ML/WLS (een hybride ML/gewogen-kleinste-kwadratenmethode) [Brow92]. Diverse parallelle implementaties van ML-EM wordt beschreven in [Bast93, Chen91a, Chen94, Raja94]. Tenslotte wordt een ML-EM-algoritme voor partiële metingen besproken in [Desm95, Brow94].

Voor het geval van ML-EM kan de convergentie naar de ML-oplossing op theoretische gronden bewezen worden [Shep82]. De uitbreiding van deze methode voor een algemene MAP-schatting met willekeurige a priori-distributie is echter niet triviaal. Verschillende onderzoekers hebben onafhankelijk van elkaar veralgemeningen van het EM-algoritme onderzocht. Herman *et al.* breiden het EM-algoritme uit voor Gaussiaanse a priori-distributies [Levi87, Herm89]. Green [Gree90a] bewijst dat een aangepaste implementatie van de M-stap resulteert in een "vertraagd" EM-algoritme (One Step Late-EM) dat convergeert naar het maximum van de a posteriori-distributie. De convergentie van OSL-EM wordt verder onderzocht in [Lang90]. De Pierro levert een theoretische studie van de convergentie-eigenschappen van een veralgemeend EM-algoritme voor een brede klasse van a priori distributies [DePi95]. Ook Hebert *et al.* ontwikkelen een veralgemeende EM-methode (GEM, Generalized EM) [Hebe89, Hebe92b, Hebe92c]. In [Lang95] wordt een vergelijking gemaakt van de verschillende EM-algoritmen.

3.8.2 Gradiënt-gebaseerde algoritmen

De meest bekende gradiënt-gebaseerde methoden zijn de maximale-gradiëntenmethode (steepest descent) en de toegevoegde-gradiëntenmethode (conjugate gradient). Algemeen kunnen we een gradiëntenmethode schrijven als

$$\hat{X}^{i+1} = \hat{X}^i - F\left(\nabla p(X|Y)|_{\hat{X}^i}\right). \tag{3.38}$$

Bij elke iteratiestap wordt het volledige beeld aangepast. De methoden verschillen van elkaar in de keuze voor $F(\nabla p(X|Y))$. Gradiëntenmethoden zijn potentieel zeer snel. Het positief karakter van de oplossing is echter niet inherent en moet expliciet opgelegd worden [Mumc94]. Het grootste probleem is echter de convergentie van deze methode. In zijn klassieke vorm convergeert een gradiëntenalgoritme steeds naar een lokaal optimum. We merken op dat een lokaal optimum (toevallig) kan samenvallen met het globaal optimum.

Zonder hierop dieper in te willen gaan vermelden we nog dat de convergentie verbeterd kan worden door gebruik te maken van zgn. preconditioners [Clin93]. Dit betekent dat bij elke iteratiestap de gradiëntenmatrix vermenigvuldigd wordt met een correctiematrix. Lange *et al.* [Lang87] bewijzen dat het EM-algoritme voor PET gereduceerd kan worden tot een aangepast maximale-gradiëntenalgoritme (preconditioned steepest gradient) waarbij de correctiematrix een diagonaalmatrix is met op de diagonaal een geschaalde versie van de pixelintensiteiten van de huidige schatting $\hat{X}^i$. Kaufman [Kauf87] komt tot een gelijkaardige versnelde versie van het EM-algoritme op basis van de toegevoegde-gradiëntenmethode.

3.8.3 Pixel-gebaseerde algoritmen

Deze methoden passen elke pixel $\hat{X}_k$ om beurt aan volgens de regel:

$$\hat{X}_k^{i+1} = \arg\max_{X_k} p(X_k|Y, \hat{X}^i_{S/k}). \tag{3.39}$$

Deze methoden zijn o. a. bekend als ICM (Iterated Conditional Mode, [Besa86, Besa93]) en ICD (Iterative Coordinate Descent, [Boum96]). Deze techniek is vergelijkbaar met de Gauss-Seidel methode voor het oplossen van stelsels differentiaalvergelijkingen. Het voordeel van deze methode is de zeer snelle convergentie en het eenvoudige inbouwen van de positiviteitsvoorwaarde. Bovendien biedt deze methode aantrekkelijke perspectieven voor implementatie op vectorprocessoren. De snelle convergentie is echter ook het belangrijkste nadeel van de

methode. Bij aanwezigheid van lokale optima zal het algoritme vrijwel onmiddellijk gevangen worden in een kwalitatief inferieur lokaal optimum. Deze methode werd in eerste instantie geïmplementeerd voor Gaussiaanse a priori-distributies [Saue91, Saue92, Saue93], maar later ook uitgebreid voor niet-Gaussiaanse convexe distributies [Boum96]. Recent onderzoek van Fessler *et al.* zet de veralgemeende EM-methode van De Pierro [DePi95] om tot een pixelgebaseerde optimalisatietechniek [Fess97].

3.8.4 Simulated annealing

Hoewel we in het volgende hoofdstuk uitvoerig zullen ingaan op deze methode, vermelden we hier toch de belangrijkste eigenschap. Zoals gezegd zal het gebruik van niet-convexe a priori-distributies aanleiding geven tot een groot aantal lokale optima. De drie hierboven besproken optimalisatiemethoden zullen met zekerheid in een lokaal optimum eindigen. De kwaliteit van deze oplossing (t. t. z. hoe dicht bevindt het bekomen lokaal optimum zich bij het globale optimum) is bovendien onbekend. Simulated annealing is een optimalisatiemethode waarvan de convergentie naar een globaal optimum aangetoond kan worden onder bepaalde voorwaarden. Het nadeel van deze methode is echter de sterke toename in benodigde rekentijd.

We vermelden nog dat op basis van het simulated annealing-algoritme voortzettingstechnieken (continuation techniques) opgesteld zijn [Gind93b, Rang92]. Meestal zijn dit heuristische technieken die een compromis zijn tussen de ICM-methode en simulated annealing, zowel wat snelheid als kwaliteit betreft. Typisch convergeren deze methoden naar een verbeterd lokaal optimum. Deze voortzettingstechnieken zijn hoofdzakelijk ontwikkeld door Bilbro *et al.* en worden gemiddeld-veldannealing genoemd (Mean Field Annealing) [Bilb89a, Bilb89b, Bilb91b, Bilb91a, Hiri89, Han92, Wang95, Wang96].

3.9 Markov Keten-Monte Carlo bemonstering

In tegenstelling tot de MAP-schatting moet voor de MPM-schatting en de MMSE-schatting de a posteriori-distributie gekend zijn. Aangezien deze niet exact berekend kan worden, zal men trachten deze te benaderen door gebruik te maken van Markov-Keten-Monte-Carlomethoden (MKMC). Enerzijds zijn deze methoden Monte-Carlomethoden, wat betekent dat zij gebruik maken van toevalsgetallen om monsters uit de toestandsruimte Ω te selecteren. Anderzijds wor-

den deze monsters op een sequentiële manier gegenereerd, zodanig dat elk nieuw monster op een specifieke manier van het vorige monster afhankelijk is. De opeenvolging van monsters vormt dan een Markov-keten. De benaming MKMC-bemonsteraars verwijst naar beide karakteristieken.

Algemeen is de bedoeling van bemonsteraars om grootheden van de volgende vorm te benaderen:

$$\overline{Q} = E[Q(X)] = \int_{\Omega} Q(X)\, p(X)\, dX. \tag{3.40}$$

Hierbij gaan we er van uit dat de waarschijnlijkheidsdistributie $p(X)$ een Gibbs-distributie is. Het is onmogelijk de grootheid $\overline{Q}$ exact te berekenen, omdat de toestandsruimte Ω meestal enorm groot zal zijn en omdat hiervoor de partitiefunctie gekend moet zijn. We zullen daarom een schatting $\hat{Q}$ bepalen aan de hand van een reeks monsters. De bemonsteraar dient zó gekozen te worden dat voor een voldoende groot aantal monsters M de reeks $\{X^i, i = 1, \ldots, M\}$ representatief is voor de waarschijnlijkheidsdistributie $p(X)$. Dit betekent dat

$$\hat{Q} = \frac{1}{M} \sum_{i=1}^{M} Q(X^i) \overset{M \to \infty}{\longrightarrow} \overline{Q} = E[Q(X)]. \tag{3.41}$$

Bij klassieke Monte-Carlomethoden zou men deze grootheid benaderen door op willekeurige wijze M monsters X^i te selecteren uit de toestandsruimte Ω, zodat de monsters uniform verdeeld zijn over de toestandsruimte Ω. Er geldt dan dat

$$\hat{Q} = \frac{\sum_{i=1}^{M} Q(X^i)\, p(X^i)}{\sum_{i=1}^{M} p(X^i)}. \tag{3.42}$$

Bij MKMC-bemonsteren worden de monsters X^i echter niet op uniforme wijze gekozen, maar in overeenstemming met hun waarschijnlijkheid. Zoals gezegd vormen de opeenvolgende monsters een Markov-keten. De opeenvolgende monsters zijn dus duidelijk afhankelijk, in tegenstelling tot de monsters bij het klassieke Monte-Carloproces. Onderstel dat de i^{de} toestand uit de Markov-keten als X^i genoteerd wordt. Uitgaande van deze toestand wordt een kandidaat $\tilde{X}^{i+1}$ voor de $(i+1)^{\text{de}}$ toestand gegenereerd. De waarschijnlijkheid dat deze toestand voorgesteld wordt (als X^i de vorige toestand is) definiëren we als de propositiewaarschijnlijkheid $p_p(\tilde{X}^{i+1}|X^i)$. De waarschijnlijkheid dat deze voorgestelde

toestand $\tilde{X}^{i+1}$ de nieuwe toestand X^{i+1} wordt, wordt gegeven door de aanvaardingswaarschijnlijkheid $p_a(\tilde{X}^{i+1}|X^i)$. Bij niet-aanvaarding wordt de vorige toestand behouden:

$$\begin{array}{lll} X^{i+1} = \tilde{X}^{i+1} & \text{met waarschijnlijkheid} & p_a(\tilde{X}^{i+1}|X^i), \\ X^{i+1} = X^i & \text{met waarschijnlijkheid} & 1 - p_a(\tilde{X}^{i+1}|X^i). \end{array} \tag{3.43}$$

De keuze van de propositiewaarschijnlijkheid en van de aanvaardingswaarschijnlijkheid bepalen de bemonsteraar. De bekendste bemonsteraars zijn de Gibbsbemonsteraar en de Metropolis-bemonsteraar.

3.9.1 De Metropolis-bemonsteraar

De Markov-keten wordt gestart vanuit de begintoestand X^0. Bij elke stap zal een nieuwe kandidaat-toestand $\tilde{X}^{i+1}$ gegenereerd worden door een willekeurige en kleine verandering van de vorige toestand X^i. Voor beelden betekent dit bv. dat één of enkele pixelwaarden veranderd worden. De mogelijke veranderingen bepalen een omgeving van de huidige toestand X^i. De voorgestelde nieuwe toestand wordt willekeurig gekozen uit de omgeving van de vorige toestand, wat betekent dat de propositiewaarschijnlijkheid uniform verdeeld is over de alle kandidaattoestanden. We berekenen de verhouding

$$\begin{aligned} r &= \frac{p(\tilde{X}^{i+1})}{p(X^i)} \\ &= \exp\left(-\beta(H(\tilde{X}^{i+1}) - H(X^i))\right). \end{aligned} \tag{3.44}$$

De aanvaardingswaarschijnlijkheid wordt dan gegeven door

$$p_a(\tilde{X}^{i+1}|X^i) = \min(r, 1). \tag{3.45}$$

Wanneer we stellen dat $\Delta H = H(\tilde{X}^{i+1}) - H(X^i)$, kunnen we deze uitdrukking herschrijven als

$$p_a(\tilde{X}^{i+1}|X^i) = \begin{cases} 1 & \text{als } \Delta H < 0 \\ \exp(-\beta \Delta H) & \text{als } \Delta H > 0. \end{cases} \tag{3.46}$$

Aangezien de voorgestelde veranderingen telkens klein zijn, is de keuze van de begintoestand X^0 van belang. Daarom worden meestal de eerste K overgangen verder niet in beschouwing genomen. De eerste K overgangen noemt men de

inloopperiode van de Metropolis-bemonsteraar. Na deze inloopperiode bevindt de bemonsteraar zich in evenwicht, wat betekent dat monsters gegenereerd worden die in overeenstemming zijn met de distributie $p(X)$ [Wink95]. De lengte van de inloopperiode K en het aantal overgangen M dat nodig is om een betrouwbare schatting $\hat{Q}$ te verkrijgen is uiteraard sterk afhankelijk van de toestandsruimte Ω en de waarschijnlijkheidsdistributie $p(X)$.

Het grote voordeel van de Metropolis-bemonsteraar is de eenvoudige implementatie. De convergentie kan echter traag verlopen wanneer de aanvaardingswaarschijnlijkheid laag ligt, waardoor een groter aantal monsters genomen moet worden.

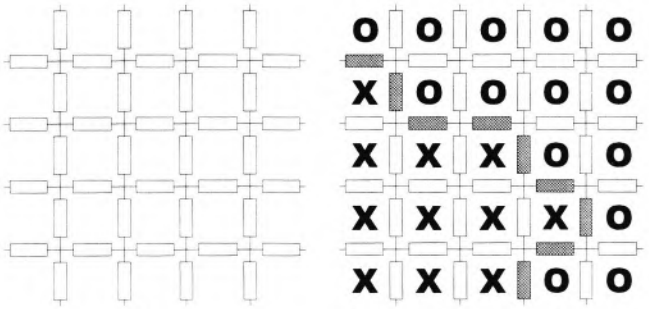
3.9.2 De Gibbs-bemonsteraar

Bij de Metropolis-bemonsteraar werd de overeenkomst met de gewenste distributie bewerkstelligd door de keuze van de aanvaardingswaarschijnlijkheid; bij de Gibbs-bemonsteraar daarentegen wordt deze overeenstemming bereikt door de keuze van de propositiewaarschijnlijkheid. De Gibbs-bemonsteraar maakt gebruik van de conditionele waarschijnlijkheid als propositiewaarschijnlijkheid. Onderstellen we bv. dat $\tilde{X}^{i+1}$ enkel van X^i verschilt in de k^{de} pixellocatie, dan is

$$p_p(\tilde{X}^{i+1}|X^i) = p(X_k^i = \tilde{X}_k^{i+1}|X_{\mathcal{S}_k}^i). \tag{3.47}$$

De aanvaardingswaarschijnlijkheid is steeds 1, wat betekent dat elke voorgestelde toestand steeds aanvaard wordt. Het voordeel van de Gibbs-bemonsteraar is de snellere convergentie, aangezien er geen voorgestelde overgangen verworpen worden. Het nadeel is echter dat het in praktijk vaak moeilijk is om monsters te genereren volgens de conditionele distributie.

Hoewel nog een aantal andere bemonsteraars bestaan (zoals o. a. de Hastingsbemonsteraar), worden in het merendeel van de toepassingen de beide bovenstaande bemonsteraars gebruikt. Voor beeldanalyse en -reconstructie bestaan deze toepassingen voornamelijk uit de bepaling van de a posteriori-distributie (bv. voor de MMSE-schatter), de marginale distributies (bv. voor de MPM-schatter), de partitiefunctie en de variantie (of ev. hogere orde momenten). We zullen echter in hoofdstuk 4 zien dat ook het simulated-annealingalgoritme beschouwd kan worden als een aangepaste versie van de Metropolis-bemonsteraar.



Figuur 3.9: Schematische voorstelling van het duale beeldmodel met pixellocaties en randlocaties.

3.10 Het duale beeldmodel

Als laatste onderdeel vermelden we nog een bijzonder type beeldmodel dat geïntroduceerd werd door Geman *et al.* [Gema84, Gema90]. Het betreft hier een duale voorstellingswijze, waarbij het beeld niet alleen gemodelleerd wordt door pixellocaties (t. t. z. intensiteiten) maar ook door lijnsegmenten op de pixelgrenzen (de randlocaties). Dit beeldmodel kunnen we schematisch voorstellen zoals in figuur 3.9 a. Het volledige beeld wordt voorgesteld door $\{X, L\}$. De N pixellocaties behouden dezelfde betekenis als in het voorgaande en worden nog steeds voorgesteld door X. De M randlocaties ($M \approx 2N$) worden voorgesteld door L. Elke randlocatie L_{ij} correspondeert met de aanwezigheid of afwezigheid van een rand tussen de pixels i en j. Randlocaties kunnen enkel de waarden 1 en 0 aannemen, waarbij de waarde 0 betekent dat een rand op deze plaats aanwezig is. Dit wordt geïllustreerd door het voorbeeld in figuur 3.9 b. De pixels voorgesteld door $\times$ stellen een eerste beeldregio voor, waarbinnen de intenstiteit vrijwel constant is. Analoog voor de tweede beeldregio, voorgesteld door $\circ$. De donkere randlocaties stemmen overeen met de aanwezigheid van een rand en bevatten bijgevolg de waarde 0; de overige randlocaties bevatten waarde 1.

Zoals eerder besproken in paragraaf 3.6 zijn convexe potentiaalfuncties voor de a priori-waarschijnlijkheidsdistributie zeer goed in staat tot het modelleren van lokale vlakheid, maar niet tot het modelleren van randen. Niet-convexe potentiaalfuncties daarentegen kunnen beide eigenschappen modelleren maar geven aanleiding tot lokale optima, waardoor de noodzaak ontstaat om trage optimalisatiemethoden (zoals simulated annealing) te gebruiken. Het duale beeldmodel tracht

dit probleem op te lossen door de tussenliggende randlocaties te gebruiken als "schakelaars" voor de potentiaalbijdrage van naburige pixels. Hierdoor ontstaat een energiefunctie van de vorm

$$H(X, L) = \sum_{i=1}^{N} \sum_{j \in \mathcal{S}_i} [L_{ij}\phi(X_i - X_j) + \gamma_{ij}(1 - L_{ij})] + \Xi(L). \tag{3.48}$$

De eerste term is de potentiaalbijdrage $\phi(X_i - X_j)$ t. g. v. het intensiteitsverschil tussen twee naburige pixels, vermenigvuldigd met de tussenliggende randlocatie. Pixels waartussen zich een rand bevindt zullen bijgevolg niet bijdragen tot $H(X, L)$. De tweede term $\gamma_{ij}(1 - L_{ij})$ beperkt het aantal randlocaties door een extra energiebijdrage voor elk aanwezig randsegment. De derde term $\Xi(L)$ tenslotte is functie van het volledige randproces en moet de vorming van onderling verbonden en gesloten randsegmenten bewerkstelligen.

Hoewel dit model op het eerste gezicht een zeer aantrekkelijk alternatief is voor niet-convexe potentiaalfuncties, moeten toch een aantal opmerkingen gemaakt worden [Li94]. In eerste instantie neemt het aantal te schatten parameters sterk toe ($N + M$ i. p. v. N). Daarnaast is het moeilijk een gepaste keuze te maken voor de penalisatie γ_{ij} van elk randsegment. Deze keuze berust voornamelijk op ervaring en is sterk beeldafhankelijk. Tenslotte is de functie $\Xi(L)$ ter bevordering van aaneengesloten randsegmenten meestal complex van vorm.

Ondanks deze nadelen wordt het duale beeldmodel met succes toegepast bij neurologische studies. Het laat toe anatomische informatie, verkregen uit hoofdzakelijk MR-beelden, te gebruiken als a priori-informatie voor PET-reconstructie. Hiervoor moeten de MR-beelden eerst geregistreerd worden met de bijbehorende PET-data en vervolgens gesegmenteerd. Met registratie bedoelen we het ev. verschuiven, schalen en roteren van de MR-beelden zodat de locatie van het hoofd exact overeenstemt in beide beelden. De randen die afkomstig zijn van het gesegmenteerde MR-beeld worden vervolgens gebruikt als initiële waarden voor de randlocaties van de PET-reconstructie. De onderliggende veronderstelling is dat de anatomische grenzen (van het MR-beeld) een goede eerste benadering zijn voor de randen die bij de reconstructie van het PET-beeld zullen gevormd worden. De grenzen van functionele gebieden zullen nl. vrijwel nooit de anatomische grenzen snijden. We vermelden een aantal publicaties waarin onderzoek naar het gebruik van anatomische a priori-informatie voor PET-reconstructie beschreven wordt [Leah91, Yan92, Ouya94, Arde93, Arde96, Lipi96, Lipi97]. Een meer algemene methode voor het gebruik van anatomische informatie m. b. v. een hiërarchisch MRV-model wordt beschreven in [John93, Chen91b].

3.11 Conclusie

In dit hoofdstuk hebben we een overzicht gegeven van het Bayesiaanse formalisme voor beeldreconstructie en -analyse. We trachten de kwaliteit van de gereconstrueerde beelden te verbeteren door gebruik te maken van a priori-informatie. Hiervoor wordt het beeld gemodelleerd als een MRV. Onze voorkennis wordt vertaald in de keuze van gepaste potentiaalfuncties voor de Gibbs-distributie. We onderscheiden twee verschillende modellen. Convexe potentiaalfuncties geven aanleiding tot optimalisatieproblemen die met eenvoudige technieken opgelost kunnen worden; zij laten echter niet toe randen te modelleren, waardoor de gereconstrueerde beelden vaak te vlak zijn. Niet-convexe potentiaalfuncties daarentegen maken het mogelijk zowel lokaal vlakke gebieden als randen te modelleren. Dit geeft echter aanleiding tot lokale optima tijdens de reconstructie. Klassieke (snelle) optimalisatiemethoden zullen steeds eindigen in een lokaal optimum. Simulated annealing stelt ons in staat het globale optimum te vinden, weliswaar met een sterke toename van de reconstructietijd tot gevolg.

Hoofdstuk 4

Simulated annealing

4.1 Inleiding

Een groot aantal problemen kunnen beschreven worden als het zoeken naar een goede of de beste toestand van een systeem, waarbij het aantal mogelijke systeemtoestanden zeer groot is. Problemen van deze klasse worden algemeen optimalisatieproblemen genoemd. Wanneer het aantal mogelijke syteemtoestanden eindig of aftelbaar oneindig is, wordt de term "combinatorische optimalisatieproblemen" gebruikt. Een aantal van de problemen die voorkomen in o. a. informatica en digitaal ontwerp kunnen geformuleerd worden als combinatorische optimalisatieproblemen. De kwaliteit van een systeemtoestand wordt geëvalueerd aan de hand van een kostfunctie. Deze kostfunctie associeert met elke systeemtoestand een reëel getal. De keuze van deze kostfunctie volgt in de meeste gevallen rechtstreeks uit de formulering van het probleem. Algemeen onderstellen we dat lagere kostfunctiewaarden overeenstemmen met betere systeemtoestanden. De optimale toestand correspondeert bijgevolg met de minimale waarde van de kostfunctie. Deze veronderstelling schaadt niet aan de algemeenheid, vermits een probleem dat op natuurlijke wijze geformuleerd kan worden als een maximalisatieprobleem steeds omgezet kan worden naar een minimalisatieprobleem mits een extra minteken in de kostfunctie.

We stellen een combinatorisch optimalisatieprobleem formeel voor door het paar (Γ, K). De toestandsruimte (of configuratieruimte) Γ stelt de verzameling van alle mogelijke systeemtoestanden voor. Een systeemtoestand (of configuratie) stellen we voor door $t \in \Gamma$. De kostfunctie $K(t)$ associeert een reëel getal met

elke systeemtoestand:

$$K : \Gamma \to \mathbb{R} : t \mapsto K(t). \tag{4.1}$$

De kostfunctiewaarde $K(t)$ van de toestand t zullen we verder kortweg de kost van t noemen. Het combinatorisch optimalisatieprobleem bestaat erin een configuratie t_{opt} te vinden die K minimaliseert, t. t. z. waarvoor geldt dat

$$K(t_{opt}) = K_{min} = \min_{t \in \Gamma} K(t). \tag{4.2}$$

We merken op dat – hoewel een aantal begrippen overeenstemmen met het vorige hoofdstuk (o. a. toestandsruimte) – we ervoor opteren om in dit hoofdstuk afzonderlijke notaties te gebruiken om eventuele verwarring te voorkomen.

We illustreren de klasse van combinatorische optimalisatieproblemen aan de hand van het gekende probleem van de handelsreiziger. Veronderstel een aantal steden die door een handelsreiziger bezocht moeten worden. De handelsreiziger wenst hiervoor een route uit te stippelen waarvoor de af te leggen weg minimaal is. Een route wordt beschouwd als een ordening van alle steden en de afstand tussen twee opeenvolgende steden wordt in vogelvlucht gemeten. Hierbij wordt de extra voorwaarde opgelegd dat de route eindigt in dezelfde stad waar ze begon. Het ligt voor de hand dat de kostfunctie in dit geval de afgelegde weg is. De complexiteit van het probleem blijkt uit het aantal mogelijke routes. Voor n steden zijn er nl. $(n-1)!$ mogelijkheden. Zelfs wanneer we er rekening mee houden dat een route en zijn inverse praktisch gesproken identiek zijn, leidt dit voor 100 steden tot 4.10^{157} mogelijke routes. Hoewel het handelsreizigerprobleem in eerste instantie artificieel kan lijken, kunnen een aantal problemen uit het digitaal ontwerp ertoe teruggebracht worden. Zo moeten bv. tijdens de fabricatie van een PCB (Printed Circuit Board) een aantal gaten geboord worden. Hierbij wenst men de totale afstand te minimaliseren die de boorkop moet afleggen. De benodigde boortijd zal nl. evenredig zijn met de af te leggen afstand. Daar waar een handelsreiziger-probleem met 100 steden onrealistisch lijkt, is een PCB-ontwerp met 100 boorgaten eerder courant [Otte89].

De meeste combinatorische optimalisatieproblemen behoren tot de klasse van de NP-complete problemen. Een NP-compleet probleem is een probleem waarvoor geen oplossingswijze gekend is die een exacte oplossing oplevert binnen een tijd die een polynomiale functie is van de complexiteit van het probleem. Dit betekent dat de rekentijd ten minste exponentieel toeneemt met de complexiteit. Bijgevolg is het praktisch onmogelijk een exacte oplossing te vinden voor

een NP-compleet probleem binnen een aanvaardbare rekentijd, zodat men een beroep moet doen op benaderingsmethoden. Deze benaderingsmethoden zijn in de meeste gevallen heuristisch van aard. De gebruikte heuristiek is echter vaak sterk probleemafhankelijk, zodat de ontwikkelde methode beperkt toepasbaar is. In de volgende paragraaf bespreken we kort enkele algemeen toepasbare combinatorische optimalisatietechnieken.

In het verdere verloop van dit hoofdstuk zullen we simulated annealing introduceren aan de hand van een meer eenvoudige techniek: iteratieve verbetering. We leiden het Metropolis-algoritme af uit de analogie met statistische mechanica en vestigen expliciet de aandacht op het verband met de Metropolis-bemonsteraar. We stellen een mathematisch model op om simulated annealing te beschrijven en bespreken aan de hand hiervan de (asymptotische) convergentie van het algoritme. Verder benutten we het verband met statistische mechanica om een aantal praktisch bruikbare grootheden af te leiden. Tenslotte bespreken we diverse technieken om de uitvoering van simulated annealing te versnellen.

4.2 Combinatorische optimalisatiemethoden

Recent onderzoek heeft een viertal nieuwe optimalisatietechnieken opgeleverd die in eerste instantie gebaseerd zijn op analogieën met natuurlijke systemen, maar toch ondersteund worden door een theoretisch kader. Deze technieken zijn simulated annealing, tabu search, genetische algoritmen en artificiële neurale netwerken. We lichten kort deze technieken toe; voor een meer gedetailleerde introductie en een beschrijving van de respectievelijke voor- en nadelen verwijzen we o. a. naar [Pirl92]. Een vergelijking tussen simulated annealing, tabu search en genetische algoritmen wordt gegeven door [Laur96], en tussen simulated annealing en neurale netwerken door [Jeff86]. Verder stellen Mahfoud *et al.* een hybried algoritme voor dat een combinatie is van simulated annealing en genetische algoritmen [Mahf93].

Simulated annealing berust op de analogie tussen het afkoelen van een vaste stof tot haar minimale energietoestand en de oplossing van een combinatorisch optimalisatieprobleem. De voordelen van deze methode zijn de eenvoudige implementeerbaarheid en de zeer algemene toepasbaarheid. Het grootste nadeel is de aanzienlijke rekentijd die nodig is om een oplossing te vinden. In dit hoofdstuk zullen we dieper ingaan op deze methode.

Tabu search is een zoekmethode waarbij m.b.v. een zgn. flexibel geheugen het zoekproces geoptimaliseerd wordt. Dit flexibele geheugen dient om structuren op te bouwen en uit te baten die voordeel trekken uit het recente verleden. De ontwikkeling van verfijnde methoden om dit geheugen efficiënt bij te houden en toe te passen in specifieke omstandigheden zijn cruciaal. De benaming van deze methode refereert naar de beperkingen ("taboes") die tijdens het verloop van het zoekproces opgelegd worden. Deze beperkingen uiten zich meestal als de uitsluiting van zoekalternatieven. De tabu-searchmethode wordt vooral toegepast voor optimalisatieproblemen in operationeel onderzoek.

Een genetisch algoritme is een adaptief algoritme dat gebaseerd is op de simulatie van genetische processen. De oplossing wordt gezocht door een populatie van mogelijke toestanden te laten evolueren volgens bepaalde voortplantingsregels. Een nieuwe generatie ontstaat door toestanden uit de vorige generatie te combineren tot nieuwe toestanden. Om de diversiteit van de populatie te garanderen worden ook mutaties toegelaten. De waarschijnlijkheid van reproductie van een toestand (d. w. z. de waarschijnlijkheid dat deze toestand combineert met een andere toestand om zo een nieuwe toestand aan te maken) is omgekeerd evenredig met de kost. Hierdoor zullen "goede" oplossingen reproduceren en "slechte" oplossingen afsterven (zgn. survival of the fittest) en zal de populatie evolueren naar een populatie met gemiddeld lagere kostfunctiewaarden. Cruciaal zijn de populatiegrootte, de voortplantingsregels en de keuze van de beginpopulatie.

De voorgaande methoden zijn sequentiële algoritmen, wat betekent dat een reeks van toestanden gevormd wordt die convergeert naar de gezochte oplossing. In de natuur herkennen we echter vaak ook parallelle algoritmen. Een aantal biologische perceptietaken (bv. visie) kunnen als optimalisatieproblemen beschouwd worden. Gezien de hoeveelheid data en de snelheid waarmee biologische informatiesystemen (bv. de menselijke hersenen) een oplossing vinden, is het duidelijk dat een hoge mate van parallellisatie toegepast wordt. Deze systemen worden gemodelleerd m. b. v. neurale netwerken, t. t. z. een groot aantal geconnecteerde neuronen. Deze netwerken moeten zorgvuldig getraind worden alvorens zij in staat zijn het optimalisatieprobleem op te lossen; na training zijn zij echter vrijwel ogenblikkelijk in staat de oplossing te berekenen. Bij deze techniek zijn de keuze van het netwerkmodel, het aantal neuronen en de keuze van de trainingsset van groot belang.

4.3 Simulated annealing

De techniek simulated annealing (gesimuleerd uitgloeien) werd geïntroduceerd door Metropolis *et al.* [Metr53]. Zij slaagden er in 1953 in om de evolutie van roostertoestanden van een metaal dat naar thermisch evenwicht streeft succesvol te modelleren m. b. v. Monte-Carlomethoden. Het duurde echter tot 1983 vooraleer Kirkpatrick *et al.* [Kirk83] de analogie met combinatorische optimalisatie benutten. Sindsdien is simulated annealing succesvol toegepast voor een brede waaier van optimalisatieproblemen. Een gedetailleerde beschrijving van simulated annealing wordt gegeven in het werk van Aarts en Van Laarhoven [Aart85, VLaa87] en Otten en Van Ginneken [Otte89]. Simulated annealing staat bekend onder een aantal andere benamingen, waarvan we vooral de benamingen "Metropolis-algoritme" en "stochastische relaxatie" (stochastic relaxation) vermelden. Deze laatste benaming wordt door Geman *et al.* gehanteerd in hun basisartikel over Bayesiaanse beeldverwerking [Gema84].

Simulated annealing wordt met succes toegepast voor een brede waaier van optimalisatieproblemen. Een zeer volledig overzicht van toepassingen is te vinden in de geannoteerde bibliografie van Collins *et al.* [Coll88]. De meeste toepassingen situeren zich op het gebied van digitaal ontwerp (plaatsing, routering e. d.). Een overzicht hiervan is te vinden in [Kirk83] en [VLaa87]. We vermelden nog het SAMURAI-algoritme van Catthoor *et al.* [Catt88a, Catt88b] en het werk van Van Marck [VMar93]. Daarnaast zijn er o. a. toepassingen voor het handelsreizigerprobleem [Bono84, Mosc90], partitioneringsmethoden voor grafen [Fu86, Bono86], matching [Lutt86], decompositie van probabilistische netwerken [Kjær91], een dipoolmodel voor biomagnetische data [Gers93, Seki92], erfelijkheid (ancestral inference) [Geye94] en modellering van distillatiekolommen [Port98].

Wat beeldverwerking betreft verwijzen we naar het overzichtswerk door Winkler [Wink95]. Daarnaast zijn er nog een aantal specifieke applicaties voor beeldsegmentatie [Tan91], ruisverwijdering [Carn85], labeling [Vand91, Vand93], beeldrestoratie [Robi98] en beeldmodellering m. b. v. een hiërarchisch MRV-model [Jeng90]. Nauw aansluitend bij het onderwerp van dit onderzoek vermelden we de resultaten van Han voor de reconstructie van tomografische beelden in het algemeen [Han93], van Barrett *et al.* [Smit83, Smit85a, Smit85b, Paxm85, Mage90, Gool90, Giro91, Seac93] en El Alaoui *et al.* [ElAl91] voor zgn. coded-aperture beeldvorming en van Webb [Webb89] en Kearfott [Kear90] voor de reconstructie van SPECT-beelden.

4.4 Iteratieve verbetering

Simulated annealing kan beschouwd worden als een veralgemening van een techniek die bekend staat als iteratieve verbetering (iterative improvement). We bespreken daarom eerst deze techniek. Iteratieve verbetering onderstelt, behalve de definitie van systeemconfiguraties en van een kostfunctie, de definitie van een generatiemechanisme. Het generatiemechanisme is een "voorschrift" om overgangen te maken van een configuratie naar een naburige configuratie d. m. v. een kleine wijziging. De definitie van een generatiemechanisme is bijgevolg equivalent met de definitie van een omgeving $\mathcal{N}_t$ van elke configuratie t: $\mathcal{N}_t$ bestaat uit alle configuraties die door één overgang bereikt kunnen worden vanuit t. We formuleren het algoritme nu als volgt. Uitgaande van een begintoestand t_0 wordt een opeenvolging van configuraties gegenereerd. Een kandidaat voor de nieuwe configuratie $\tilde{t}_k$ wordt willekeurig gekozen uit de omgeving $\mathcal{N}_{t_{k-1}}$ van de voorgaande toestand t_{k-1}. Deze kandidaattoestand wordt aanvaard als de nieuwe toestand t_k wanneer de kost daalt. Wanneer we stellen dat $\Delta K = K(\tilde{t}_k) - K(t_{k-1})$, dan kunnen we het aanvaardingscriterium schrijven als

$$\begin{array}{ll} t_k \leftarrow \tilde{t}_k & \text{als } \Delta K < 0, \\ t_k \leftarrow t_{k-1} & \text{als } \Delta K > 0. \end{array} \tag{4.3}$$

Het algoritme wordt beëindigd wanneer een toestand wordt bereikt die een lagere kost heeft dan alle naburige toestanden. Het groot voordeel van iteratieve verbetering is de zeer eenvoudige implementeerbaarheid. De nadelen zijn de volgende:

1. per definitie is de eindtoestand een lokaal minimum en er is geen informatie over de mate waarin dit lokaal minimum afwijkt van het globaal minimum;
2. het bereikte lokaal minimum is sterk afhankelijk van de gekozen begintoestand t_0 en het is zeer moeilijk om een goede keuze te maken voor deze begintoestand;
3. het is over het algemeen onmogelijk om een bovengrens voor de benodigde rekentijd te bepalen.

Op basis van deze nadelen kunnen we een aantal verbeteringen voor het algoritme formuleren:

1. het algoritme herhalen voor N willekeurig gekozen beginconfiguraties (we merken op dat met zekerheid een globaal minimum gevonden wordt wanneer $N \to \infty$, vermits het globaal minimum met waarschijnlijkheid 1 als beginconfiguratie zal voorkomen);

2. gebruik maken van informatie uit de voorgaande uitvoeringen om de keuze van de beginconfiguratie te verbeteren;

3. het generatiemechanisme uitbreiden (t. t. z. de omgeving $\mathcal{N}_t$ vergroten) om eventuele lokale minima te vermijden;

4. het beperkt aanvaarden van overgangen die corresponderen met een toename van de kost.

Het tweede en derde alternatief vereisen een grondige kennis van het probleem en zijn bijgevolg sterk probleemgebonden. Het eerste alternatief wordt succesvol toegepast wanneer het niet van essentieel belang is een oplossing te vinden die in de onmiddellijke omgeving van het globale minimum gelegen is. Wat de theoretische convergentie betreft is deze aanpak nl. niet beter dan het willekeurig bemonsteren van de toestandsruimte. Het vierde alternatief leidt tot de techniek van simulated annealing.

4.5 Het Metropolis-algoritme

Oorspronkelijk werd het Metropolis-algoritme ontwikkeld om het "annealen" (uitgloeien) van een vaste stof te simuleren. In de vaste-stoffysica wordt met annealen het proces bedoeld waarbij een vaste stof in een warmtebad eerst opgewarmd wordt tot de vloeibare fase (waarbij de atomen willekeurig bewegen, zodat de eventueel aanwezige roosterdefecten verdwenen zijn). Daarna wordt de vaste stof afgekoeld door het langzaam verlagen van de temperatuur van het warmtebad. Wanneer de afkoeling voldoende langzaam gebeurt zal uiteindelijk de grondtoestand van het rooster bereikt worden. Met de grondtoestand wordt de minimale energietoestand bedoeld, d. w. z. de toestand waarbij het aantal roosterdefecten – die met een toename van de energie corresponderen – minimaal is. Het afkoelingsproces kan als volgt beschreven worden. Vertrekkende van de maximale temperatuurwaarde wordt de temperatuur van het warmtebad verlaagd. Bij elke temperatuurwaarde wordt gewacht tot wanneer de vaste stof thermisch evenwicht bereikt heeft. Thermisch evenwicht wordt beschreven door de Boltzmanndistributie. Dit betekent dat de waarschijnlijkheid dat het systeem zich in een toestand met energie H bevindt, gegeven wordt door:

$$p(H) = \frac{1}{Z(T)} \exp(-\frac{H}{k_B T}). \tag{4.4}$$

Hierbij stelt T de temperatuur voor, $Z(T)$ de (temperatuurafhankelijke) partitiefunctie en k_B de constante van Boltzmann. De waarschijnlijkheid van lage-

energietoestanden stijgt naarmate de temperatuur afneemt. Wanneer de temperatuurwaarde 0 benadert zullen enkel de minimale-energietoestanden een van nul verschillende waarschijnlijkheid hebben.

Om de evolutie naar thermisch evenwicht te simuleren stellen Metropolis *et al.* een Monte-Carlomethode voor om een reeks van systeemtoestanden te genereren [Metr53]. Uitgaande van de huidige roostertoestand wordt een kandidaat voor de nieuwe toestand aangemaakt door een willekeurig gekozen atoom een kleine verplaatsing te geven. De corresponderende energieverandering ΔH wordt berekend. Wanneer de energie van de nieuwe roostertoestand lager is dan de voorgaande ($\Delta H < 0$) wordt het proces voortgezet met de nieuwe toestand. Wanneer de energie toeneemt wordt de nieuwe toestand aanvaard met waarschijnlijkheid

$$p(\Delta H) = \exp\left(-\frac{\Delta H}{k_B T}\right). \tag{4.5}$$

Deze aanvaardingsregel staat bekend als het Metropolis-criterium. Na een voldoende aantal overgangen bereikt de waarschijnlijkheidsdistributie van toestanden de Boltzmann-distributie van (4.4) [Metr53].

Het Metropolis-algoritme kan nu ook geformuleerd worden om een reeks toestanden van een combinatorisch optimalisatieprobleem te genereren. De beschrijving verloopt volledig analoog met de beschrijving van iteratieve verbetering uit de voorgaande paragraaf. Met behulp van een generatiemechanisme wordt tijdens de k^{de} iteratiestap een nieuwe toestand $\tilde{t}_k$ voorgesteld op basis van de voorgaande toestand t_{k-1}. Het corresponderend verschil in kost stellen we voor door $\Delta K = K(\tilde{t}_k) - K(t_{k-1})$. In de plaats van het aanvaardingscriterium (4.3) komt ditmaal het (probabilistische) Metropolis-criterium, waarbij H en T vervangen worden door de kostfunctie $K(t)$ en een pseudo-temperatuur τ:

$$\begin{array}{lll} p(t_k \leftarrow \tilde{t}_k) = 1 & \text{als} & \Delta K < 0, \\ p(t_k \leftarrow \tilde{t}_k) = \exp(-\frac{\Delta K}{\tau}) & \text{als} & \Delta K > 0. \\ p(t_k \leftarrow t_{k-1}) = 1 - \exp(-\frac{\Delta K}{\tau}) & \text{als} & \Delta K > 0. \end{array} \tag{4.6}$$

De parameter τ (die in feite $k_B T$ vervangt en dus de dimensie van een energie heeft) zullen we de temperatuur blijven noemen, hoewel er voor combinatorische optimalisatieproblemen geen voor de hand liggend analogon is voor de (fysische) temperatuur. De waarde van deze parameter zal verlaagd worden tijdens de uitvoering van het algoritme, zodat $\tau(k)$ functie is van k. Het verloop van $\tau(k)$ tijdens het iteratieproces noemen we het afkoelingsschema.

We hebben het Metropolis-algoritme ingevoerd als een uitbreiding van iteratieve verbetering. Anderzijds kunnen we iteratieve verbetering beschouwen als een speciaal geval van simulated annealing met temperatuurwaarde nul. In dit geval worden nl. enkel overgangen toegelaten die de energie verlagen. Daarom wordt iteratieve verbetering ook "quenched annealing" (afschrikken) genoemd.

4.6 Verband met de Metropolis-bemonsteraar

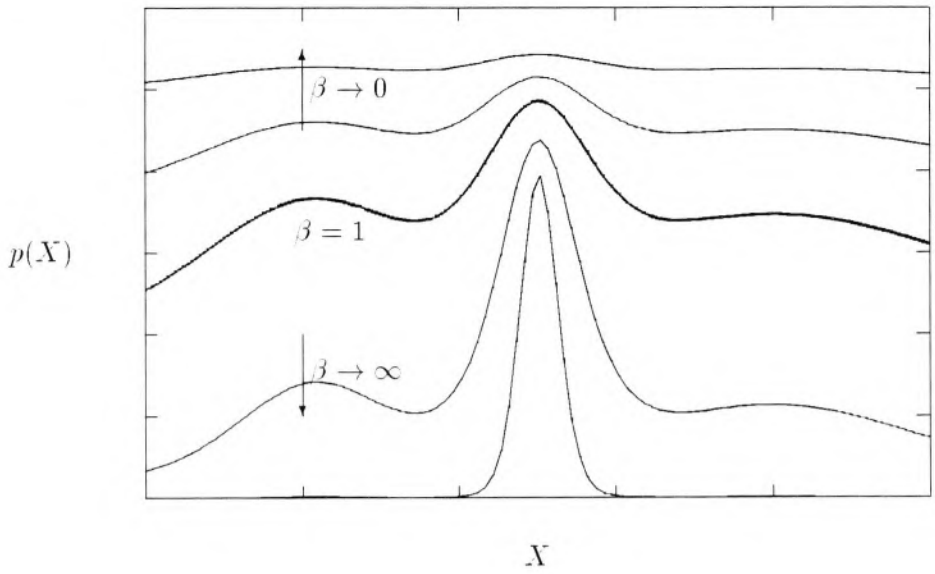
We leggen expliciet het verband tussen simulated annealing en de Metropolis-bemonsteraar uit paragraaf 3.9. We bemerken een duidelijke overeenkomst tussen de bovenstaande formule (4.5) en uitdrukking (3.46) voor de aanvaardingswaarschijnlijkheid van de Metropolis-bemonsteraar. We herhalen dat de Metropolis-bemonsteraar monsters genereert die representatief zijn voor de distributie

$$p(X) = \frac{1}{Z(\beta)} \exp(-\beta H(X)). \tag{4.7}$$

We interpreteren de parameter β als een inverse temperatuur

$$\beta = \frac{1}{\tau}. \tag{4.8}$$

Door de waarde van β te wijzigen blijft de ligging van maxima van de waarschijnlijkheidsdistributie ongewijzigd, maar wordt enkel de "gepiektheid" van $p(X)$ beïnvloed. We illustreren dit aan de hand van figuur 4.1. Hierin zetten we voor een willekeurig gekozen energiefunctie $H(X)$ de waarschijnlijkheidsdistributie $p_\beta(X)$ uit voor een aantal waarden van β. Het dient opgemerkt te worden dat de diverse curven op verschillende wijze genormeerd zijn, opdat het verschil in vorm duidelijk zichtbaar zou zijn. Voor $\beta = 1$ vinden we de originele distributie. Voor $\beta \to 0$ (d. w. z. temperatuur $\tau \to \infty$) zien we enerzijds dat $p_\beta(X)$ evolueert naar een uniforme distributie. Dit correspondeert met de vloeibare fase, waar de atomen willekeurig bewegen en elke toestand vrijwel dezelfde waarschijnlijkheid heeft. Anderzijds evolueert $p_\beta(X)$ naar een δ-functie rond het globale maximum van $p(X)$ (t. t. z. de minimale energietoestand) voor $\beta \to \infty$ (d. w. z. $\tau \to 0$). Hierin herkennen we opnieuw het Metropolis-algoritme. Door namelijk τ langzaam stapsgewijze te variëren van ∞ naar 0 en bij elke τ-waarde $p_\tau(X)$ te bemonsteren m. b. v. de Metropolis-bemonsteraar verkrijgen we een reeks monsters die evolueert van een uniforme verdeling naar monsters die enkel in de directe omgeving liggen van de toestand met maximale waarschijnlijkheid. Deze toestand correspondeert met de minimale waarde van de energie (de kostfunctie).



Figuur 4.1: Illustratie van het verband tussen simulated annealing en de Metropolis-bemonsteraar: verloop van de waarschijnlijkheidsdistributie voor verschillende temperatuurwaarden (bij een gelijke Gibbs-energiefunctie).

4.7 Mathematisch model

We voeren in deze paragraaf een mathematisch model in dat ons toelaat het simulated-annealingalgoritme te beschrijven. Uitgaande van een gekozen omgevingsstructuur kan simulated annealing beschouwd worden als een algoritme dat bij elke iteratiestap de huidige toestand tracht om te zetten naar één van de naburige toestanden. Dit mechanisme kan mathematisch het beste beschreven worden door een Markov-keten (zie ook paragraaf 3.4), t. t. z. een opeenvolging van toestanden waarbij de waarschijnlijkheid om in een nieuwe toestand terecht te komen enkel afhankelijk is van de voorgaande toestand [Azen88].

We herhalen dat tijdens de k^{de} iteratiestap, op basis van de voorgaande toestand t_{k-1}, een kandidaat $\tilde{t}_k$ voor de k^{de} toestand t_k gegenereerd wordt. We beschrijven de Markov-keten m. b. v. een transitiematrix $\mathbf{T}(k)$, waarbij $T_{ij}(k)$ de voorwaardelijke (transitie)waarschijnlijkheid voorstelt dat tijdens de k^{de} iteratiestap

het systeem overgaat van toestand i naar toestand j, t. t. z.

$$T_{ij}(k) = p(t_k = j | t_{k-1} = i). \tag{4.9}$$

Aangezien het aantal mogelijke systeemtoestanden eindig of aftelbaar oneindig is, kunnen we iedere toestand door zijn rangnummer karakteriseren. We noemen de toestand die overeenstemt met rangnummer i dan ook kortweg de toestand i. De transitiematrix $\mathbf{T}(k)$ is een $|\Gamma| \times |\Gamma|$-matrix (waarbij $|\Gamma|$ het aantal mogelijke toestanden voorstelt) die in zijn algemeenste vorm afhankelijk is van de iteratiestap k. Wanneer de transitiematrix onafhankelijk is van k noemen we de Markov-keten homogeen; in het andere geval spreken we van een inhomogene Markov-keten.

We introduceren daarnaast de toestandsdistributie $\pi(k)$, waarbij $\pi_i(k)$ de waarschijnlijkheid voorstelt dat i de k^{de} toestand van het systeem is. We kunnen dan $\pi_i(k)$ bepalen uit de recursiebetrekking

$$\pi_i(k) = \sum_{j=1}^{|\Gamma|} \pi_j(k-1)\, T_{ji}(k). \tag{4.10}$$

Uitgaande van de begintoestand t_0 (d. w. z. $\pi_i(0) = 0$ als $i \neq t_0$ en $\pi_i(0) = 1$ als $i = t_0$) vinden we voor een homogene Markov-keten (d. w. z. $\mathbf{T}$ onafhankelijk van k) dat

$$\pi(k) = \mathbf{T}^k\, \pi(0). \tag{4.11}$$

We ontbinden verder de transitiematrix $\mathbf{T}(k)$ in een generatiematrix $\mathbf{G}(k)$ en een aanvaardingsmatrix $\mathbf{A}(k)$. Hierbij stelt $G_{ij}(k)$ de waarschijnlijkheid voor dat tijdens de k^{de} iteratiestap een overgang van toestand i naar toestand j voorgesteld wordt en $A_{ij}(k)$ de waarschijnlijkheid om deze overgang te aanvaarden. Dit betekent dat

$$\begin{aligned} G_{ij}(k) &= p(\tilde{t}_k = j | t_{k-1} = i), \\ A_{ij}(k) &= p(t_k = j | \tilde{t}_k = j, t_{k-1} = i). \end{aligned} \tag{4.12}$$

De transitiewaarschijnlijkheid wordt dus gegeven door

$$T_{ij}(k) = \begin{cases} G_{ij}(k)\, A_{ij}(k). & \forall j \neq i \\ 1 - \sum_{l=1, l\neq i}^{|\Gamma|} G_{il}(k)\, A_{il}(k), & j = i. \end{cases} \tag{4.13}$$

We merken op dat tot nu toe geen specifieke voorwaarden opgelegd worden aan de generatie- en aanvaardingswaarschijnlijkheden. We noemen een algoritme dat

overeenstemt met (4.13) een veralgemeend simulated annealing algoritme. Deze klasse van algoritmen (probabilistic hill-climbing algorithms) wordt uitvoerig beschreven door Connors *et al.* [Conn89]. We bepalen nu concrete uitdrukkingen voor het generatiemechanisme $\mathbf{G}(k)$ en het aanvaardingscriterium $\mathbf{A}(k)$ in het specifieke geval van het Metropolis-algoritme. Voor wat de generatiewaarschijnlijkheid betreft wordt de voorgestelde toestand $\tilde{t}_k$ willekeurig gekozen uit de omgeving $\mathcal{N}_{t_{k-1}}$ van de voorgaande toestand t_{k-1}, zodat

$$G_{ij}(k) = \begin{cases} \frac{1}{|\mathcal{N}_i|}, & j \in \mathcal{N}_i \\ 0, & j \notin \mathcal{N}_i. \end{cases} \tag{4.14}$$

Hierbij stelt $|\mathcal{N}_i|$ het aantal naburen van i voor. Eventueel kan het generatiemechanisme wijzigen tijdens het iteratieproces, hetgeen betekent dat de omgevingsstructuur functie is van k. We gaan er in wat volgt echter van uit dat dit niet het geval is. De aanvaardingswaarschijnlijkheid wordt gegeven door het hoger vermelde Metropolis-criterium, t. t. z.

$$A_{ij}(k) = \begin{cases} 1, & K(j) \leq K(i) \\ \exp\left(-\frac{K(j)-K(i)}{\tau(k)}\right), & K(j) \geq K(i). \end{cases} \tag{4.15}$$

De afhankelijkheid van k uit zich dus enkel via de waarde van de temperatuurparameter $\tau(k)$. We onderscheiden bijgevolg twee formuleringen van het algoritme:

1. het homogene algoritme bestaat uit een sequentie van homogene Markov-ketens, waarbij de temperatuurwaarde constant gehouden wordt tijdens elke Markov-keten en verlaagd wordt tussen twee opeenvolgende Markov-ketens;

2. het inhomogene algoritme bestaat uit één enkele inhomogene Markov-keten waarbij de temperatuurwaarde continu verlaagd wordt.

Voor vrijwel alle toepassingen wordt gebruik gemaakt van het homogene algoritme. We merken op dat het model voor de afkoeling van een vaste stof (t. t. z. de temperatuur van het warmtebad stapsgewijze verlagen en bij elke temperatuurwaarde wachten tot wanneer zich thermisch evenwicht instelt) correspondeert met het homogene algoritme. In dit geval wordt het afkoelingsschema praktisch bepaald door 4 parameters: de beginwaarde van de temperatuur, de eindwaarde, het aantal iteraties per homogene Markov-keten en de temperatuurdaling tussen opeenvolgende Markov-ketens.

We merken op dat het begrip "Markov-keten" eigenlijk de benaming is van een bepaald type stochastisch proces. Deze benaming wordt echter door vrijwel

alle auteurs ook gebruikt om een praktische realisatie van dit proces aan te duiden. Zo zullen we een reeks overgangen die bij een constante temperatuurwaarde gegenereerd worden een homogene Markov-keten noemen.

4.8 Convergentie

Het simulated-annealingalgoritme vindt een optimale toestand (t. t. z. een toestand met minimale kost) na een voldoende aantal iteraties K wanneer geldt dat

$$p(t_K \in \Gamma_{opt}) = 1. \tag{4.16}$$

Hierbij stelt Γ_{opt} de verzameling van globale optima voor:

$$\Gamma_{opt} = \{t \in \Gamma | K(t) = K_{min}\}. \tag{4.17}$$

Het is duidelijk dat aan (4.16) enkel asymptotisch voldaan kan worden, t. t. z.

$$\lim_{k \to \infty} p(t_k \in \Gamma_{opt}) = 1. \tag{4.18}$$

We stellen een aantal voorwaarden op waaronder deze asymptotische convergentie geldt. Deze voorwaarden zijn verschillend voor het homogene en het inhomogene algoritme en worden dadelijk besproken. We benadrukken dat de convergentie van het simulated annealing algoritme dus steeds geïnterpreteerd dient te worden als een asymptotische convergentie.

We herhalen dat het homogene algoritme bestaat uit een opeenvolging van homogene Markov-ketens. We maken bij conventie een onderscheid tussen de indices k en l. Zo zal k staan voor de waarde tijdens de k^{de} iteratiestap, terwijl l staat voor de waarde tijdens de l^{de} homogene Markov-keten. Voor het homogene algoritme, waarbij de temperatuur τ_l enkel functie van l is, geldt (4.18) wanneer

1. de generatiematrix $\mathbf{G}(\tau_l)$ en de aanvaardingsmatrix $\mathbf{A}(\tau_l)$ aan bepaalde voorwaarden voldoen;

2. het temperatuurverloop voldoet aan

$$\lim_{l \to \infty} \tau_l = 0; \tag{4.19}$$

3. elke individuele Markov-keten oneindig lang is.

Deze voorwaarden zijn opgesteld door Romeo *et al.* [Rome85]. De eerste voorwaarde is enkel noodzakelijk voor een veralgemeend simulated-annealingalgoritme. De beide matrices moeten irreduceerbaar, aperiodisch en door een evenwichtsvergelijking met elkaar verbonden zijn. Voor het "klassieke" simulated-annealingalgoritme (t. t. z. het Metropolis-algoritme) is aan deze voorwaarden echter voldaan, zodat we hier niet dieper op ingaan. Het homogene algoritme convergeert dus asymptotisch wanneer de temperatuurwaarden van de opeenvolgende Markov-ketens naar 0 convergeren en elke individuele Markov-keten oneindig lang is. Hoewel dit resultaat van theoretisch belang is, is het onbruikbaar voor praktische implementaties van simulated annealing.

Nodige en voldoende voorwaarden voor de convergentie van het inhomogene algoritme worden o. a. geformuleerd door Geman *et al.* [Gema84], Hajek [Haje88], Gidas [Gida85, Gida89] en Mitra [Mitr86]. Bij het inhomogene algoritme varieert de temperatuurwaarde τ_k bij elke iteratiestap. Asymptotische convergentie geldt wanneer:

1. de generatiematrix $\mathbf{G}(\tau_l)$ en de aanvaardingsmatrix $\mathbf{A}(\tau_l)$ aan bepaalde voorwaarden voldoen;

2. het temperatuurverloop voldoet aan

$$\begin{aligned} &\lim_{k\to\infty} \tau_k = 0; \\ &\tau_k \geq \tau_{k+1}, \forall k; \end{aligned} \tag{4.20}$$

3. het afkoelingsschema, t. t. z. de sequentie $\{\tau_k\}$, niet sneller convergeert dan $\mathcal{O}(|\log k|^{-1})$.

De bespreking van de eerste voorwaarde is identiek aan deze voor het homogene algoritme. Bemerk dat door de tweede voorwaarde niet uitgesloten wordt dat de temperatuur constant gehouden wordt gedurende een aantal iteraties. De convergentie-eigenschappen van het inhomogene algoritme zijn bijgevolg ook bruikbaar voor het homogene algoritme. Wat de derde voorwaarde betreft vinden de hoger vermelde auteurs telkens een uitdrukking van de vorm

$$\tau_k \geq \frac{c}{\log k}, \tag{4.21}$$

waarbij c een constante is die afhankelijk is van de complexiteit van het probleem en de keuze van de kostfunctie. De concrete uitdrukkingen voor c verschillen echter per auteur. Het bewijs voor deze uitdrukking is gebaseerd op de zwakke

en sterke ergodiciteit van inhomogene Markov-ketens. We verwijzen hiervoor o. a. naar [VLaa87]. Men is het er echter algemeen over eens dat de theoretische voorwaarde van een logaritmisch afkoelingsschema in praktijk tot te langzame afkoeling leidt [Gema84, Aart85, VLaa87, Otte89, Azen92]. Daarom wordt in vrijwel alle praktische implementaties van simulated annealing gebruik gemaakt van een exponentieel afkoelingsschema van de vorm

$$\tau_{k+1} = \alpha \, \tau_k, \tag{4.22}$$

waarbij $\alpha < 1$. Goede waarden zijn gelegen tussen 0.8 en 0.98. Ook voor dit onderzoek zal gebruik gemaakt worden van een exponentieel afkoelingsschema. We merken op dat in dit geval asymptotische convergentie niet langer gegarandeerd is.

Hoffman *et al.* geven de theoretische afleiding van een optimaal afkoelingsschema [Hoff90]. Dit soort afleidingen zijn echter enkel mogelijk voor zeer eenvoudige optimalisatieproblemen (met triviale kostfuncties). De invloed van een variabele kostfunctie (t. t. z. een iteratie-afhankelijke kostfunctie) op de convergentie wordt onderzocht door Frigerio *et al.* [Frig93].

4.9 Verband met de statistische mechanica

We hebben het Metropolis-algoritme in paragraaf 4.5 afgeleid door gebruik te maken van de analogie tussen het traag afkoelen van een vaste stof en het oplossen van een combinatorisch optimalisatieprobleem. Het afkoelingsproces wordt in de vaste-stoffysica gemodelleerd m. b. v. de statistische mechanica. We zullen in deze paragraaf de analogie met de statistische mechanica verder benutten om een aantal praktisch bruikbare grootheden voor optimalisatieproblemen te berekenen. Deze analogie wordt o. a. beschreven in [Kirk83, Khac86, VLaa87, Swen87, Wang90].

De statistische mechanica geeft ons een inzicht in de collectieve (macroscopische) eigenschappen van een systeem met een groot aantal vrijheidsgraden, zoals bv. de atomen in een vaste stof. Aangezien dit aantal atomen typisch van de grootteorde $10^{23}/\text{cm}^3$ is, wordt experimenteel enkel het gedrag waargenomen dat statistisch het meest waarschijnlijk is. Dit gedrag kan bijgevolg gekarakteriseerd worden door een verwachtingswaarde, genomen over een Gibbs-ensemble van identieke systemen, en kleine fluctuaties rond deze verwachtingswaarde. We herhalen dat in dit ensemble de waarschijnlijkheid van een toestand X met energie

$H(X)$ gegeven wordt door de Boltzmann-distributie

$$p(X) = \frac{1}{Z(T)} \exp\left(-\frac{H(X)}{k_B T}\right). \tag{4.23}$$

waarbij k_B de constante van Boltzmann en T de temperatuur voorstelt. Ensemblegemiddelden kunnen berekend worden m. b. v. de partitiefunctie

$$Z(T) = \sum_{X \in \Gamma} \exp\left(-\frac{H(X)}{k_B T}\right). \tag{4.24}$$

De Helmholtz vrije energie $F(T)$, die evenredig is met de logaritme van $Z(T)$, levert informatie over de gemiddelde energie $\langle H(T) \rangle$ en de entropie $S(T)$:

$$F(T) = -k_B T \ln Z(T) = \langle H(T) \rangle - T\, S(T). \tag{4.25}$$

Boltzmann-gewogen ensemblegemiddelden kunnen steeds uitgedrukt worden i. f. v. afgeleiden van F. Zo kunnen we bv. de gemiddelde energie schrijven als:

$$\langle H(T) \rangle = -\frac{d \ln Z}{d(\frac{1}{k_B T})}. \tag{4.26}$$

De snelheid waarmee deze gemiddelde energie varieert i. f. v. de temperatuur wordt gekarakteriseerd door de soortelijke warmte $c(T)$

$$\begin{aligned} c(T) &= \frac{d\langle H(T) \rangle}{dT} \\ &= \frac{\langle H(T)^2 \rangle - \langle H(T) \rangle^2}{k_B T^2}. \end{aligned} \tag{4.27}$$

Een grote waarde van $c(T)$ correspondeert met een faseovergang van het systeem. In een context van optimalisatie geeft dit aan dat de temperatuur zeer langzaam verlaagd moet worden. We kunnen van de soortelijke warmte verder ook gebruik maken om de entropie $S(T)$ te bepalen uit de thermodynamische betrekking

$$\frac{dS(T)}{dT} = \frac{c(T)}{T}. \tag{4.28}$$

Wanneer $S(T_0)$ gekend is voor een (hoge) temperatuurwaarde T_0 vinden we na integratie dat

$$S(T) = S(T_0) - \int_T^{T_0} \frac{c(T)\, dT}{T}. \tag{4.29}$$

Wat de praktische bruikbaarheid van deze statistische grootheden betreft merken we op dat deze afgeleid moeten worden uit de partitiefunctie. In vele gevallen kan $Z(T)$ enkel via bemonsteringsmethoden bepaald worden. Toch maken Rose *et al.* succesvol gebruik van $Z(T)$ om een goede schatting van de begintemperatuur te maken [Rose90]. Deze methode wordt verder verfijnd door Varanelli *et al.* [Vara93]. Fu *et al.* gebruiken de vrije energie om een schatting te maken van de minimale kostfunctiewaarde $\langle K_{min} \rangle$ [Fu86]. Ettelaie *et al.* berekenen de residuele entropie als maat voor de kwaliteit van de gevonden oplossing [Ette85].

Men kan zich bedenkingen maken bij de analogie tussen de afkoeling van een vaste stof en combinatorische optimalisatie. Zo bestaat een vaste stof bv. uit identieke atomen en bijgevolg zal de grondtoestand een regelmatig kristalrooster zijn. Een typisch optimalisatieprobleem bestaat daarentegen uit een groot aantal niet-identieke en onderling niet-uitwisselbare elementen, waardoor een regelmatige oplossing hoogst onwaarschijnlijk is. Toch blijkt uit onderzoek in de vaste-stoffysica naar specifieke magnetische stoffen (zgn. Ising spin glasses) dat grootheden uit de statistische mechanica bruikbaar zijn bij de beschrijving van macroscopische eigenschappen van systemen die uit ongelijke elementen opgebouwd zijn [Kirk83].

4.10 Versnelling en parallellisatie

Zoals reeds aangehaald werd is het belangrijkste nadeel van simulated annealing het grote aantal iteraties (en bijgevolg de aanzienlijke rekentijd) dat nodig is om tot een oplossing te komen. Daarom zijn een aantal alternatieven ontwikkeld die tot doel hebben het algoritme te versnellen. We maken een onderscheid tussen aanpassingen van het "klassieke" sequentiële algoritme en parallelle algoritmen.

We vermelden eerst twee aangepaste sequentiële algoritmen. Het algoritme van Szu [Szu87a, Szu87b, Mats89] maakt gebruik van een Cauchy-distributie voor het generatiemechanisme. Hierdoor hebben "verre sprongen" (t. t. z. sprongen die ev. een lokaal minimum kunnen verlaten) een hogere waarschijnlijkheid, zodat sneller afgekoeld kan worden. Een parallelle implementatie van het algoritme van Szu wordt beschreven in [Witt91]. Een ander alternatief is het algoritme zonder verworpen overgangen van Greene *et al.* [Gree86]. Hierbij worden overgangen voorgesteld volgens hun aanvaardingswaarschijnlijkheid, zodat elke voorgestelde overgang aanvaard kan worden. Deze methode vereist het bijhouden van een tabel met de invloed van elke mogelijke overgang op de kostfunctie. Deze tabel dient

bij elke iteratiestap aangepast te worden. Dit kan echter slechts voor een klein aantal toepassingen efficiënt uitgevoerd worden.

De meest populaire versnellingsmethode voor simulated annealing is parallellisatie. We verwijzen eerst naar het meer algemene overzicht van parallelle algoritmen voor combinatorische optimalisatieproblemen door Pardalos *et al.* [Pard96]. Een uitstekend overzicht van de diverse parallellisatietechnieken voor simulated annealing wordt gegeven door Azencott [Azen92]. Daarnaast worden nog enkele probleemonafhankelijke parallelle implementaties voorgesteld in [Gree90b, Rous90, Lee95]. Volgens Azencott kan het onderscheid gemaakt worden tussen vier verschillende technieken. We geven een korte beschrijving van deze methoden; voor een meer gedetailleerde mathematische behandeling en eventuele toepassingen verwijzen we naar [Azen92]. Een vijfde en triviale parallellisatiemethode wordt niet vermeld, nl. een uitvoering van het sequentiële algoritme waarbij de benodigde berekeningen (zoals de kostfunctiewaarde) parallel uitgevoerd worden. Aangezien de parallellisatieaspecten in dit geval sterk probleemafhankelijk zijn gaan we hier niet dieper op in. In de onderstaande bespreking veronderstellen we dat we beschikken over N identieke processoren, elk met voldoende geheugen om het sequentiële algoritme te kunnen uitvoeren.

Een eerste methode bestaat uit N simultane uitvoeringen van hetzelfde sequentiële algoritme (d. w. z. zelfde generatiemechanisme, zelfde afkoelingsschema, zelfde aantal iteratiestappen). Dit betekent dat er geen communicatie is tussen de processoren tot wanneer het algoritme beëindigd wordt. Achteraf wordt uit de N oplossingen deze met de kleinste kost gekozen als uiteindelijke oplossing. Aangezien deze oplossing het resultaat is van een uitvoering van het sequentiële algoritme, blijven alle voorgaande (convergentie-)eigenschappen behouden. Vermits het hier om algoritmen met een eindig aantal iteratiestappen gaat, kunnen we de convergentie enkel uitdrukken in termen van een eindafwijking ϵ. Azencott bewijst dat, voor een vooropgestelde eindafwijking ϵ van de oplossing, het parallelle algoritme gemiddeld een versnelling N bereikt t. o. v. het sequentiële algoritme. Deze parallellisatiemethode is vooral bruikbaar voor architecturen waar communicatie tussen processoren tijdrovend is. We merken nog op dat dit algoritme ook beschouwd kan worden als een versnelling van het sequentiële algoritme. Veronderstel dat we over voldoende rekentijd beschikken om I sequentiële iteratiestappen uit te voeren. In [Dodd90] onderzoekt Dodd het gebruik van meerdere (onderstel M) korte uitvoeringen, elk met I/M iteratiestappen, versus één trage uitvoering met I iteratiestappen.

Een tweede methode bestaat uit een simultane uitvoering met periodieke interactie. We veronderstellen dat de N processoren met elkaar interageren op de tijdstippen $s, 2s, 3s, \ldots$. Na elke interactiestap vertrekken alle processoren van dezelfde toestand. Gedurende het tijdsinterval s voeren zij het sequentiële algoritme uit, waarbij elke processor dezelfde temperatuurwaarde hanteert. Na dit tijdsinterval wordt de eindtoestand met minimale kost gekozen en verdeeld over de processoren als nieuwe begintoestand. Dit mechanisme wordt herhaald voor n tijdsintervallen. De temperatuurwaarde kan enkel gewijzigd worden tussen opeenvolgende tijdsintervallen. De eindoplossing is de toestand met minimale kost op tijdstip ns. Deze toestand is opnieuw het gevolg van een uitvoering van het sequentiële algoritme, zodat ook hier de convergentie-eigenschappen behouden blijven.

Een derde methode maakt gebruik van interactie na elke iteratiestap. Opnieuw veronderstellen we dat elke processor uitgaat van dezelfde toestand. Na de berekening van één iteratiestap worden de resultaten van de N processoren achtereenvolgens geraadpleegd. Wanneer een processor een aanvaarde overgang retourneert, wordt deze nieuwe toestand over de processoren verdeeld en worden de resultaten van de overige processoren genegeerd. Enkel wanneer een processor een verworpen overgang retourneert wordt de volgende processor geraadpleegd, en dit tot wanneer een aanvaarde overgang gevonden wordt. Dit resulteert in een uitvoering van het sequentiële algoritme, waarbij voor N voldoende groot vrijwel elke iteratiestap tot een aanvaarde overgang zal leiden. Opnieuw blijven de convergentie-eigenschappen behouden. Deze methode is vooral succesvol in de latere fase van het algoritme (d. w. z. bij lage temperatuurwaarden), waar het merendeel van de voorgestelde overgangen verworpen wordt. Wegens het grote aantal communicatiestappen (na elke iteratiestap) is deze methode enkel geschikt voor computerarchitecturen waar de communicatie tussen processoren voldoende snel verloopt. In dit geval leidt deze methode vaak tot een effectieve versnellingsfactor N.

De vierde en laatste methode is de enige die fundamenteel afwijkt van het sequentiële algoritme. Deze methode gelijkt sterk op de derde methode. Opnieuw onderstellen we interactie na elke iteratiestap. Ditmaal worden echter de resultaten van de overige processoren niet verworpen wanneer een aanvaarde overgang gevonden wordt. Ook de overige aanvaarde overgangen worden uitgevoerd. Strikt genomen zijn deze overige beslissingen niet langer geldig nadat een (eerste) over-

gang aanvaard wordt. We veronderstellen echter dat de invloed van de kostfunctieverandering van één enkele overgang op het beslissingscriterium van eventuele volgende overgangen verwaarloosbaar klein is. Het is duidelijk dat voor deze parallellisatiemethode de convergentie-eigenschappen niet langer behouden blijven. Het aantal iteratiestappen dat in parallel geëvalueerd kan worden is uiteraard sterk probleemafhankelijk. In hoofdzaak is dit functie van de grootte van de toestandsruimte en de relatieve grootte van de kostfunctieverandering t. g. v. één overgang t. o. v. de kostfunctiewaarde [VLaa87]. Een grondige studie van deze parallellisatietechniek voor plaatsingsproblemen wordt gegeven door Greening [Gree95].

We vermelden enkele praktische toepassingen van de parallellisatie van simulated annealing: het gebruik van een "hypercube" multiprocessor voor plaatsingsproblemen [Bane90], het gebruik van een shared-memory multiprocessor voor plaatsing in VLSI [Krav87] en gedistribueerde implementaties voor het handelsreizigerprobleem [Allw89] en SAT-problemen [Sohn94, Sohn95].

Tenslotte wijzen we nog op een laatste en zeer veelbelovende versnellingsmethode, nl. de ontwikkeling van specifieke hardware. Voor een aantal toepassingen met eenvoudige kostfuncties (zoals handelsreizigerprobleem en graafpartitionering) vindt Abramson een enorme snelheidswinst [Abra92]. Zo blijkt dezelfde (sequentiële) implementatie op een PC-AT met specifiek hiervoor ontwikkelde hardware tot 80 maal sneller te zijn dan op een Cray Y/MP.

4.11 Conclusie

We hebben in dit hoofdstuk de combinatorische optimalisatiemethode simulated annealing in detail besproken. Deze methode onderscheidt zich van andere methoden in het feit dat onder bepaalde voorwaarden de oplossing convergeert naar het globale minimum van de kostfunctie, zelfs wanneer de toestandsruimte een groot aantal lokale minima bevat. Uit de voorgaande bespreking leiden we vier "ingrediënten" af die nodig zijn voor de implementatie van simulated annealing:

1. de keuze van configuraties (toestanden);
2. de keuze van een kostfunctie;
3. de keuze van een generatiemechanisme;
4. de keuze van een afkoelingsschema.

Het afkoelingsschema bepaalt het verloop van de temperatuurparameter τ tijdens het iteratieproces. We zullen in het verdere verloop van dit onderzoek gebruik maken van het homogene algoritme. Dit betekent dat de temperatuurwaarde constant gehouden wordt gedurende een aantal iteraties (homogene Markov-keten), waarna de temperatuur sprongsgewijze verlaagd wordt tussen opeenvolgende Markov-ketens. Dit betekent dat het afkoelingsschema volledig bepaald wordt door de volgende vier parameters:

1. de begintemperatuur;
2. de eindtemperatuur (stopcriterium);
3. de lengte van de Markov-ketens;
4. de grootte van de temperatuursprongen.

Elk van deze aspecten zal in de volgende hoofdstukken nauwkeurig onderzocht worden voor het specifieke geval van een simulated-annealingalgoritme voor de reconstructie van PET-beelden.

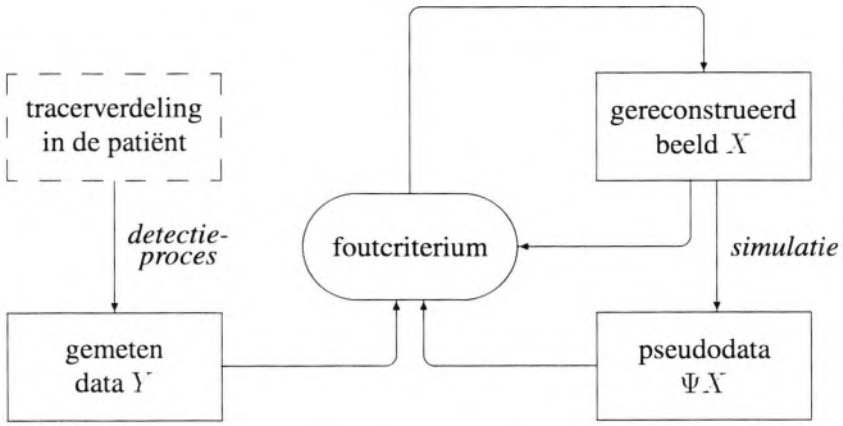
Hoofdstuk 5

Een reconstructiealgoritme op basis van simulated annealing

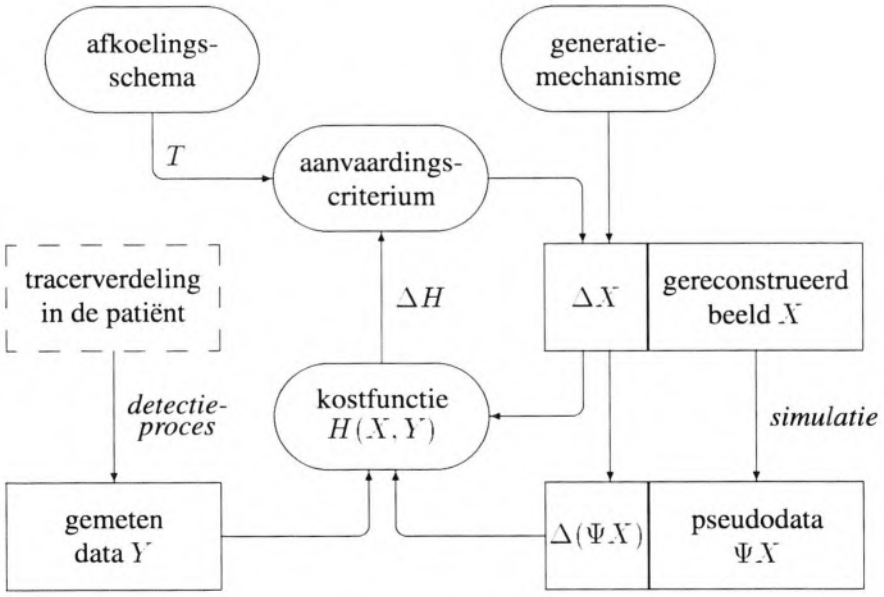
5.1 Inleiding

Algemeen kan een iteratief reconstructiealgoritme voor PET schematisch voorgesteld worden zoals in figuur 5.1. We herhalen dat we het beeld van de tracerverdeling over het inwendige van de patiënt trachten te reconstrueren. Hiertoe beschikken we over gemeten data Y die van de scanner afkomstig zijn. We trachten op iteratieve wijze het bijbehorende beeld X te reconstrueren. Door simulatie van het meetproces worden de corresponderende pseudodata ΨX berekend (de zgn. voorwaartse projectiestap). Met behulp van een foutcriterium evalueren we de "kwaliteit" van het beeld X. In het geval van maximum-likelihoodreconstructie bv. wordt de waarschijnlijkheid $p(Y|X)$ gemaximaliseerd. Dit foutcriterium geeft aanleiding tot een aanpassing van het beeld ΔX. Meestal zal deze aanpassing het gevolg zijn van wijzigingen in de pseudodata $\Delta(\Psi X)$ (de zgn. terugprojectiestap).

Het simulated-annealingalgoritme dat hier onderzocht wordt onderscheidt zich van de vorige methoden door de afwezigheid van een terugprojectiestap. De beeldverandering ΔX wordt op willekeurige wijze gegenereerd en het foutcriterium beslist enkel over aanvaarding of verwerping van de voorgestelde verandering. Er is echter geen rechtstreekse terugkoppeling tussen het foutcriterium en de voorgestelde wijziging ΔX. De schematische voorstelling van een simulated-annealingalgoritme voor de reconstructie van PET-beelden wordt weergegeven in figuur 5.2.



Figuur 5.1: Schematische voorstelling van een iteratief reconstructiealgoritme voor PET-beelden.



Figuur 5.2: Schematische voorstelling van een iteratief reconstructiealgoritme voor PET-beelden op basis van simulated annealing.

Aan het begin van elke iteratiestap beschikken we over de huidige toestand van het gereconstrueerde beeld X, de overeenkomstige pseudodata ΨX en de bijbehorende kost $H(X, Y)$. Het generatiemechanisme genereert een willekeurige beeldverandering ΔX en de corresponderende aanpassingen van de pseudodata $\Delta(\Psi X)$ en van de kost ΔH worden berekend. Het aanvaardingscriterium beslist op basis van deze kostverandering of de voorgestelde ΔX al dan niet aanvaard wordt. Als aanvaardingscriterium wordt het Metropolis-criterium gebruikt (zie paragraaf 4.5). Wanneer de voorgestelde verandering de kost doet dalen ($\Delta H < 0$) wordt zij steeds aanvaard; wanneer de kost toeneemt ($\Delta H > 0$) wordt de verandering aanvaard met waarschijnlijkheid

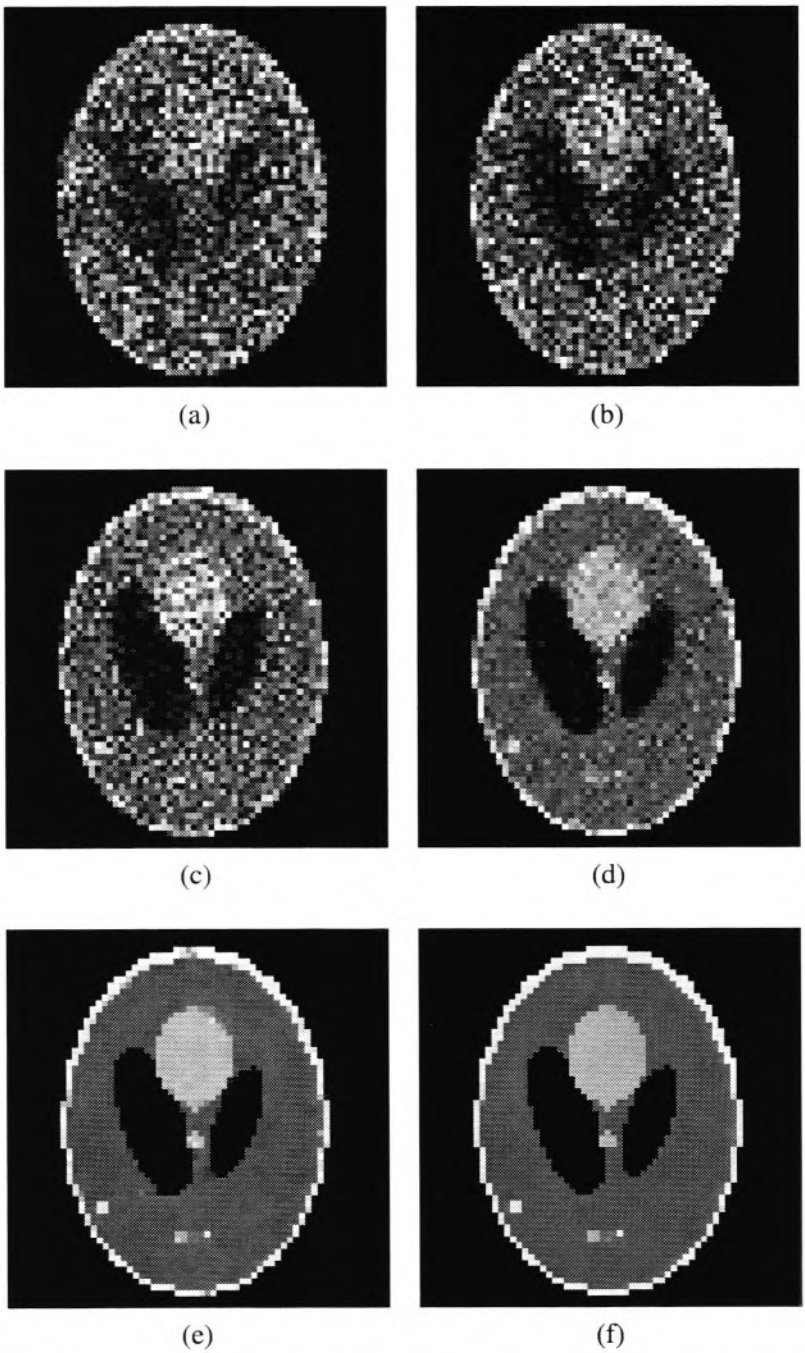
$$p(\Delta X) \sim \exp\left(-\frac{\Delta H}{\tau}\right). \tag{5.1}$$

Deze waarschijnlijkheid is o. a. afhankelijk van de waarde van de temperatuurparameter τ. We herhalen dat het afkoelingsschema het verloop van τ tijdens de reconstructie bepaalt.

Dit reconstructiealgoritme wordt geïllustreerd door figuur 5.3. Hierin wordt het gereconstrueerde beeld voorgesteld tijdens verschillende stadia van het reconstructieproces. Er is duidelijk te zien hoe het beeld als het ware "kristalliseert" naarmate het aantal iteraties toeneemt, d. w. z. naarmate de temperatuurwaarde afneemt.

In het schema van figuur 5.2 herkennen we opnieuw de drie belangrijkste aspecten van het simulated-annealingalgoritme: de kostfunctie, het generatiemechanisme en het afkoelingsschema. De keuze van de kostfunctie en de instellingen van de diverse parameters van het generatiemechanisme en het afkoelingsschema bepalen samen de kwaliteit van de gereconstrueerde beelden. De analyse van de gebruikte kostfunctie komt in een volgend hoofdstuk aan bod. In dit hoofdstuk bespreken we enkele specifieke aspecten van het generatiemechanisme en het afkoelingsschema meer in detail. Daarvoor bespreken we nog de evaluatiemethode voor de kwaliteit van de gereconstrueerde beelden, de gebruikte fantoombeelden en de modellering van het detectieproces.

In het verdere verloop van dit proefschrift zullen op diverse plaatsen rekentijden vermeld worden. Deze tijden zijn afkomstig van simulaties op een IBM RS6000 43P/140 met een PowerPC 604e 166MHz processor en 128MB geheugen.



Figuur 5.3: Het gereconstrueerde beeld tijdens verschillende stadia van de reconstructie: (a) 250.000 iteraties, (b) 500.000 iteraties, (c) 750.000 iteraties, (d) 1 miljoen iteraties, (e) 1.5 miljoen iteraties en (f) 2 miljoen iteraties.

5.2 Evaluatiemethode voor de beeldkwaliteit

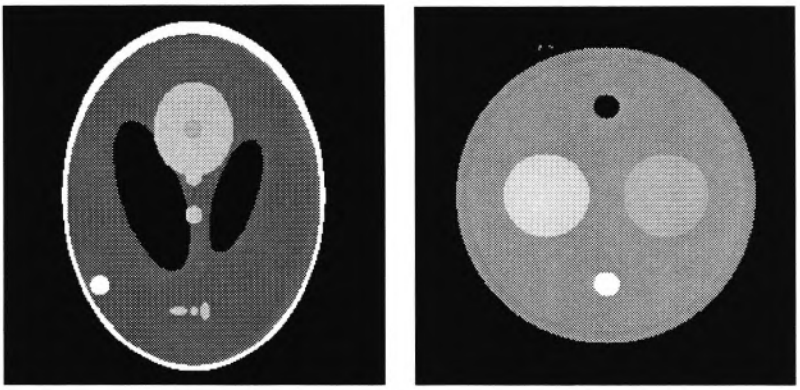
In recent onderzoek van o. a. Barrett *et al.* worden objectieve maten voor beeldkwaliteit geïntroduceerd. We geven hieronder een kort overzicht van deze theorie; voor een volledige bespreking verwijzen we o. a. naar [Barr90, Barr91, Barr92, Barr93, Barr95, Gool92, Herm91, Mate94, Mate96a, Eski96]. De onderzoekers gaan uit van het standpunt dat de kwaliteit van een beeld gedefinieerd moet worden in termen van "*hoe goed een bepaalde observator een bepaalde taak* (bv. het stellen van een diagnose) *kan vervullen aan de hand van dat beeld*". Elk objectief criterium voor beeldkwaliteit moet daarom uitgaan van de specificatie van de te vervullen taak en de keuze van de observator. Wat de mogelijke taken betreft wordt het onderscheid gemaakt tussen classificatietaken en estimatietaken; wat de observator betreft gebruikt men menselijke observatoren en mathematische observatormodellen (zoals bv. de Hotelling-observator). In het eenvoudige geval van binaire classificatietaken (bv. tumor aanwezig of afwezig) kan met behulp van een aantal menselijke observatoren een ROC-curve (Receiver Operating Characteristic) opgesteld worden. De oppervlakte onder de ROC-curve, die de relatieve waarschijnlijkheid van een valse positieve en een valse negatieve diagnose uitdrukt, is in dit geval een goede maat voor de beeldkwaliteit. Het uitvoeren van zulk een ROC-analyse met de medewerking van een aantal menselijke observatoren (bv. medische specialisten) is echter een zeer tijdrovende studie die in praktijk moeilijk uitvoerbaar blijkt te zijn. Het opstellen van mathematische observatormodellen die de performantie van de menselijke observator goed benaderen biedt in dit geval een uitkomst. Een rigoureuze evaluatie van de simulated-annealingreconstructiemethode aan de hand van observatormodellen valt echter buiten het bestek van dit onderzoek.

Een meer eenvoudige en veel gebruikte kwaliteitsmaat is de gemiddelde kwadratische afwijking t. o. v. een referentiebeeld. Er zijn echter zowel verschillende soorten kwadratische afwijkingen als verschillende soorten gemiddelden mogelijk. Enerzijds is er het feit dat we een continu object (de intensiteitsverdeling) reconstrueren als een gediscretiseerd beeld. We kunnen de kwadratische fout dus zowel continu (door over te gaan op een continue voorstelling van het beeld) als discreet (door over te gaan op een discrete voorstelling van het object) definiëren. Anderzijds kunnen we het gemiddelde beschouwen als een spatiaal gemiddelde, als een gemiddelde over een ensemble van ruisrealisaties van hetzelfde object of als een gemiddelde over een ensemble van objecten. Wat het spatiaal gemiddelde betreft kan dit zowel over het volledige beeld als over een specifieke beeldregio (ROI, Region Of Interest) geëvalueerd worden. Bij het gebruik van een gemid-

delde kwadratische afwijking als maat voor de beeldkwaliteit is het daarom van belang nauwkeurig te specifiëren welke gemiddelde kwadratische afwijking bedoeld wordt.

Daarnaast merken we op dat er geen direct verband bestaat tussen de gemiddelde kwadratische afwijking en de manier waarop het verschil tussen beelden visueel ervaren wordt. Wanneer de beoogde taak enkel bestaat uit visuele evaluatie van de beelden, zijn kwaliteitsmaten die bv. aangepast zijn aan de karakteristieken van het menselijk oog meer aangewezen [Wysz82]. Het gebruik van de gemiddelde kwadratische afwijking is echter te rechtvaardigen wanneer een kwantitatieve analyse van de gereconstrueerde beelden tot taak gesteld wordt [Herm89]. Dit is o. a. het geval bij de analyse van PET-beelden (zie paragraaf 2.2). Vandaar ook de keuze voor een gemiddelde kwadratische afwijking als kwaliteitsmaat voor de evaluatie van dit onderzoek.

Voor de optimalisatie van de verschillende aspecten van het reconstructiealgoritme werden hoofdzakelijk simulaties uitgevoerd die gebruik maken van softwarefantomen (hiermee worden softwarematig aangemaakte beelden bedoeld). De twee softwarefantomen waarvan gebruik wordt gemaakt voor de hierna besproken simulaties worden weergegeven in figuur 5.4. Het Shepp-Logan hoofdfantoom (SL-fantoom) is een softwarefantoom dat oorspronkelijk ontworpen is voor CT-studies. Het bestaat uit een aantal elliptische structuren en is een aanvaardbare benadering voor een transversaal densiteitsbeeld doorheen de hersenen van een mens [Kak88]. Het tweede fantoom (HC-fantoom) bestaat uit enkele cirkelvormige structuren ("warme" en "koude" regio's, Hot- en Cold-spots, [Kear90]). Beide fantomen zijn in overeenstemming met de algemeen gangbare onderstelling dat de klasse van te reconstrueren beelden in het geval van PET bestaat uit beelden die lokaal vlak zijn in regio's begrensd door scherpe grenzen. Strikt genomen stemt het SL-fantoom niet overeen met het model voor een transversale snede uit een PET-hersenscan. Zo heeft bv. de schedelrand in het SL-fantoom een hoge intensiteit, wegens de hoge densiteit van de schedel, terwijl we bij PET-studies (vrijwel) geen activiteit in de schedel zullen waarnemen. Toch zullen we gebruik maken van het SL-fantoom, aangezien het een veelgebruikt fantoom is in de medische beeldverwerking. We merken tenslotte nog op dat voor ons onderzoek de meeste studies uitgevoerd werden met een discretisatie van deze fantomen op een 64 x 64 rooster.



(a) (b)

Figuur 5.4: Gebruikte fantoombeelden: (a) SL-fantoom (Shepp-Loganfantoom) en (b) HC-fantoom.

Stellen we het fantoombeeld voor door $\tilde{X}$, dan berekenen we de corresponderende artificiële data Y als

$$Y = \Psi \tilde{X}. \tag{5.2}$$

Aangezien we gebruik maken van gediscretiseerde softwarefantoombeelden ligt het voor de hand om de gediscretiseerde kwadratische afwijking te gebruiken. Voorts beperken we ons tot een spatiaal gemiddelde over de pixels van een beeld. Gezien het aantal te optimaliseren onafhankelijke parameters is het wat rekentijd betreft praktisch onuitvoerbaar om deze simulaties te herhalen voor een ensemble van gelijkaardige ruisrealisaties van hetzelfde object. Dit betekent dat geen rekening werd gehouden met de invloed van een specifieke realisatie van een bepaald ruisniveau. Wel werd de invloed nagegaan van verschillende ruisniveaus op de parameterinstellingen. Om dezelfde reden (rekentijd) werd ook geen systematisch gemiddelde gemaakt over een ensemble van objecten. Wel werden een aantal optimalisaties uitgevoerd voor verschillende fantomen, in de hoop de beeldonafhankelijkheid van deze optimalisaties te mogen veralgemenen tot alle uitgevoerde optimalisaties. De keuze voor een spatiaal gemiddelde discrete kwadratische afwijking betekent dat we de beeldafwijking $E(X, \tilde{X})$ definiëren als

$$E(X, \tilde{X}) = \sum_{i=1}^{N} (X_i - \tilde{X}_i)^2, \tag{5.3}$$

waarbij N het aantal beeldpixels voorstelt. In wat volgt zal naar de bovenstaande uitdrukking gerefereerd worden als kortweg "de beeldfout".

5.3 Modellering van het detectieproces

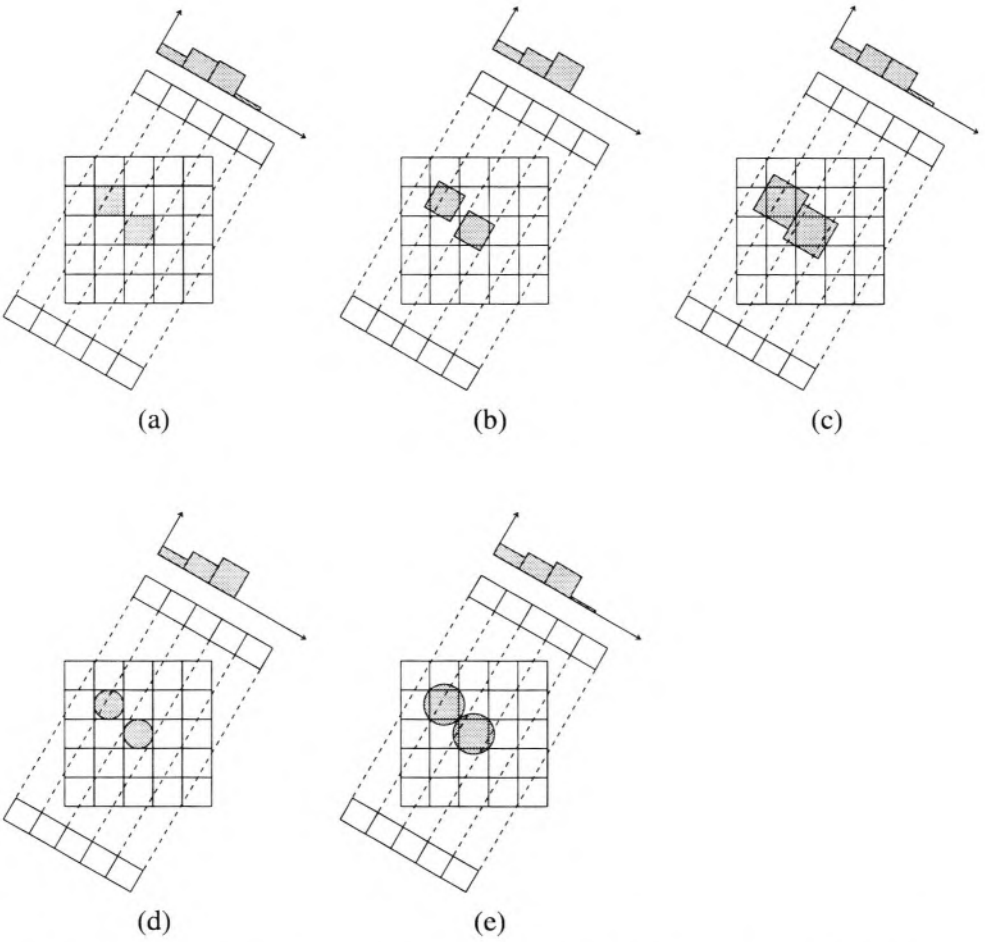
Een belangrijk aspect van het reconstructiealgoritme is de modellering van het detectieproces. Meer in het bijzonder besteden we aandacht aan de voorwaartse projectiestap. Bij het meetproces wordt een continue intensiteitsverdeling getransformeerd naar gediscretiseerde projecties. Tijdens het reconstructieproces echter wordt de gereconstrueerde intensiteitsverdeling discreet voorgesteld. Voor elke iteratiestap moeten de pseudodata berekend worden die overeenstemmen met de actuele toestand van het beeld. Dit betekent dat we een model moeten kiezen om de voorwaartse projectie van een gediscretiseerde intensiteitsverdeling te kunnen berekenen.

We gaan uit van het feit dat beeldpixels rechthoekige gebieden met constante intensiteit voorstellen. De meest voor de hand liggende keuze voor het projectiemodel is het geometrisch exacte model (zie figuur 5.5 a). Hierbij is de intensiteit die door een detectorpaar gemeten wordt evenredig met de oppervlakte van het deel van de pixel dat in de strook tussen beide detectoren gelegen is. Dit geometrisch exacte model is de beste keuze die we kunnen maken voor het projectiemodel, maar is echter ook relatief rekenintensief.

Aangezien deze voorwaartse projectie bij elke iteratiestap opnieuw berekend moet worden, gaan we op zoek naar andere projectiemodellen die het geometrisch exacte model voldoende benaderen en minder rekenintensief zijn. We hebben vier verschillende modellen onderzocht (zie figuur 5.5 b t. e. m. e). Hierbij worden de pixels onder elke projectiehoek voorgesteld als:

1. vierkanten met constante zijde, gealigneerd volgens de projectierichting;
2. vierkanten met variabele (hoekafhankelijke) zijde, gealigneerd volgens de projectierichting;
3. cirkels met constante diameter;
4. cirkels met variabele (hoekafhankelijke) diameter.

Bij de eerste twee modellen worden de pixels voorgesteld als vierkanten die steeds gealigneerd zijn met de projectierichting. Dit heeft als voordeel dat de intensiteitsverdeling zich herleidt van een 2-dimensionale distributie tot een lijndistributie. Anderzijds betekent dit dat bij elke projectiehoek telkens een ander object (t. t. z. telkens een andere representatie van hetzelfde object) geprojecteerd wordt. Door gebruik te maken van cirkels (de laatste twee modellen) beschikken we weliswaar



Figuur 5.5: Een schematische voorstelling van de verschillende onderzochte projectiemodellen: (a) het geometrisch exacte model, (b) een model met vierkanten met constante zijde, (c) een model met vierkanten met variabele zijde, (d) een model met cirkels met constante diameter en (e) een model met cirkels met variabele diameter.

projectiemodel	gem. fout	max. fout	rekentijd
constante vierkanten	0.32%	5.89%	66.05%
variabele vierkanten	0.37%	0.68%	68.58%
constante cirkels	0.32%	2.70%	93.10%
variabele cirkels	0.29%	1.94%	105.07%

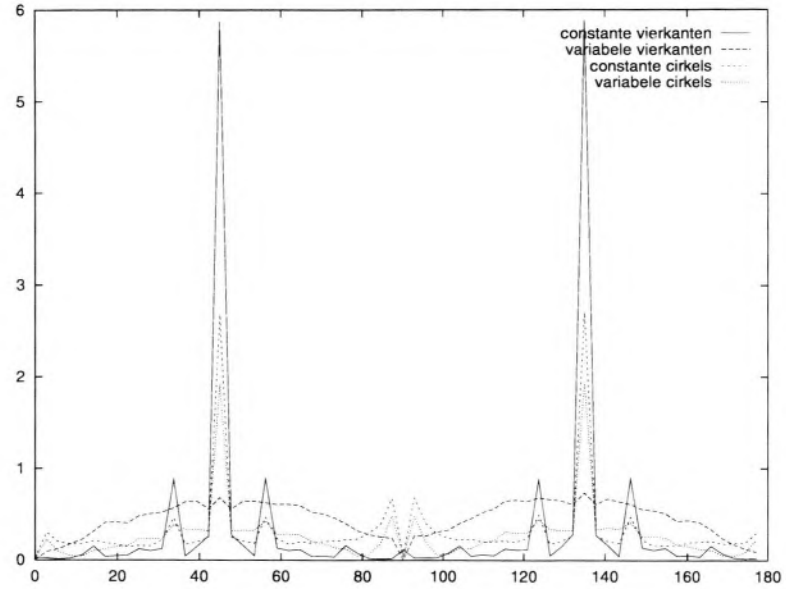
Tabel 5.1: Vergelijking van de vier onderzochte projectiemodellen: gemiddelde procentuele fout, maximale procentuele fout en benodigde rekentijd (geometrisch exacte model = 100%).

over een rotatie-invariante representatie van de pixels, maar blijft het noodzakelijk een 2-dimensionale intensiteitsdistributie te berekenen.

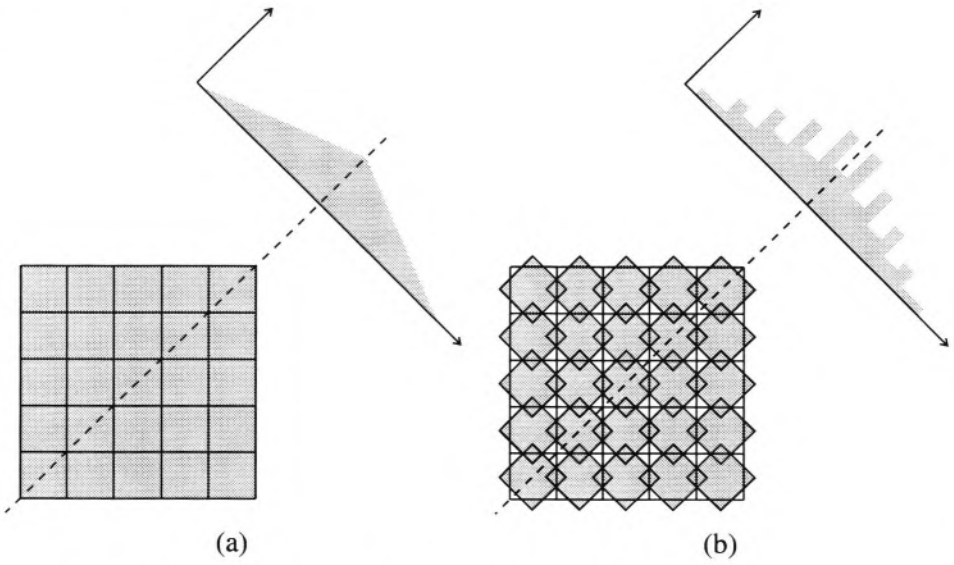
In het geval van modellen met constante breedte (zijde of diameter) is deze breedte gelijk aan de breedte van de pixel d; in het geval van modellen met variabele breedte wordt de breedte b onder elke projectiehoekhoek θ zó gekozen dat naburige vierkanten of cirkels steeds aansluitend zijn.

$$b = d(\sin\theta + \cos\theta) \tag{5.4}$$

Figuur 5.6 toont de procentuele afwijkingen voor elk model in vergelijking met het geometrisch exacte model, en dit in functie van de projectierichting. In tabel 5.1 zien we voor elk model de gemiddelde procentuele afwijking, de maximale procentuele afwijking en de benodigde rekentijd (waarbij de rekentijd voor het geometrisch exacte model gelijkgesteld werd aan 100%). We merken eerst en vooral op dat de gemiddelde procentuele afwijking voor elk onderzocht model kleiner is dan 0.5%, zodat we mogen stellen dat alle onderzochte modellen het geometrisch exacte model zeer goed benaderen. We zien echter dat beide modellen met constante breedte aanleiding geven tot grotere maximale afwijkingen. Uit figuur 5.6 blijkt dat deze maximale afwijkingen zich voordoen bij projectiehoeken van 45°en 135°. We kunnen dit als volgt verklaren: bij deze hoeken zijn de "lege ruimten" tussen naburige elementen gealigneerd en dit geeft aanleiding tot zichtbare artefacten. Dit wordt geïllustreerd in figuur 5.7. De beide modellen met variabele breedte zijn hieraan niet onderhevig, omdat de breedte steeds zo gekozen wordt dat naburige elementen aansluitend zijn. We merken ook op dat de beide cirkelmodellen qua rekentijd vergelijkbaar of zelfs trager zijn dan het geometrisch exacte model. Dit is te verklaren door het optreden van een boogcosinus in de berekening van de oppervlakte van een cirkelkoorde. Hoewel het model met vierkanten met variabele breedte de grootste gemiddelde afwijking veroorzaakt,



Figuur 5.6: Vergelijking van de procentuele fout van de projectiemodellen volgens de projectiehoek.



Figuur 5.7: Illustratie van de artefacten die ontstaan onder 45°en 135°bij modellen met constante breedte: projectie van een vlak beeld onder 45° m. b. v. (a) het geometrisch exacte model en (b) het model met vierkanten met constante breedte.

besluiten we toch dat dit model het beste het geometrisch exacte model benadert: het geeft geen aanleiding tot zichtbare artefacten (kleinste maximale afwijking) en kan ongeveer 30% sneller berekend worden dan het geometrisch exacte model.

5.4 Het generatiemechanisme

Zoals in hoofdstuk 4 reeds besproken werd kan het generatiemechanisme beschouwd worden als een "recept" dat het reconstructiealgoritme in staat stelt om d. m. v. kleine beeldwijzigingen over te gaan naar een naburig beeld. De keuze van een generatiemechanisme is equivalent met de definitie van een omgevingsstructuur. We benadrukken dat er een onderscheid gemaakt dient te worden tussen de omgevingsstructuur die verbonden is met het generatiemechanisme en de omgevingsstructuur die via het MRV-model aan de basis ligt van de a priori-waarschijnlijkheidsdistributie (zie hoofdstuk 3).

Het hier besproken generatiemechanisme verandert de intensiteit van één of meer willekeurig gekozen pixels tijdens elke iteratiestap. Dit betekent dat o. a. een keuze gemaakt moet worden i. v. m. het aantal gelijktijdig gewijzigde pixels (en het verband tussen de intensiteitsveranderingen), de pixels die voor verandering in aanmerking komen en de hoeveelheid intensiteit die gewijzigd wordt. Het veranderen van de pixelintensiteit gebeurt door toevoeging van een kleine (positieve of negatieve) hoeveelheid intensiteit, die we verder een intensiteitskorrel zullen noemen. In wat volgt worden de drie aspecten van het generatiemechanisme meer in detail besproken. We merken verder nog op dat deze keuze voor het generatiemechanisme niet de enige mogelijkheid is. In paragraaf 5.4.4 worden enkele alternatieve generatiemechanismen kort besproken.

5.4.1 De pixelkeuze

Onder de pixelkeuze verstaan we de keuze van de beeldpixels die voor aanpassing in aanmerking komen. In het meest algemene geval zijn dit alle mogelijke pixels. Wanneer echter de contour van het gescande object bij het begin van de reconstructie bepaald kan worden, kunnen we de efficiëntie van het generatiemechanisme sterk verbeteren door ons te beperken tot het aanpassen van de intensiteit van enkel de pixels die binnen de contour gelegen zijn.

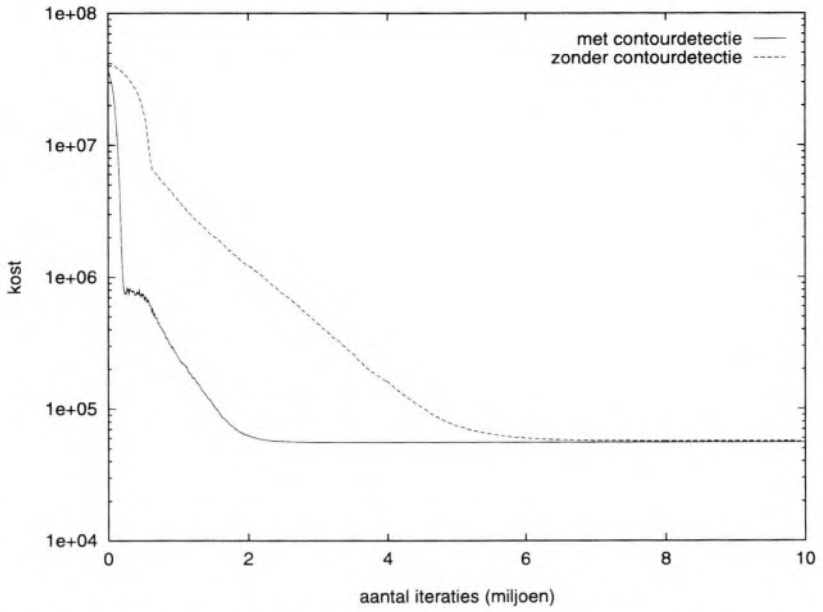
De gemeten data van een PET-scanner moeten steeds gecorrigeerd worden voor de attenuatie die optreedt in het gescande object. In de praktijk wordt deze

attenuatiecorrectie meestal afgeleid uit een extra transmissiescan van de patiënt (zie paragraaf 2.5). Een aantal auteurs beschrijven een techniek waarbij men deze gemeten attenuatiecorrectie kan benaderen door een berekende attenuatiecorrectie, waardoor er geen behoefte meer is aan de transmissiescan [Huan81, Berg82, Tomi87, Mich89, Sieg92]. De berekende attenuatiecorrectie wordt bepaald aan de hand van nulelementen in het sinogram. Elk nulelement impliceert nl. een strook nulpixels in het beeld. We hebben van deze techniek gebruik gemaakt om de contour van het gescande object te bepalen alvorens met het reconstructieproces te beginnen. Zoals in figuren 5.8 en 5.9 te zien is verhoogt een voorafgaande contourdetectie de efficiëntie van het generatiemechanisme tijdens het ganse verloop van de reconstructie. Het gebruik van contourdetectie heeft echter geen invloed op de eindkwaliteit van het gereconstrueerde beeld.

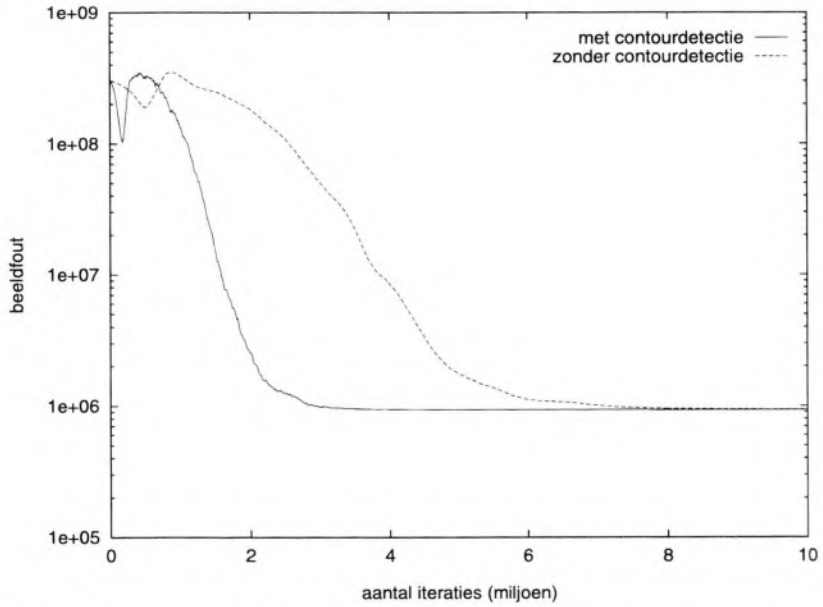
Het gebruik van contourdetectie is vooral zinvol bij fantoomstudies, omdat we met zekerheid weten dat nulelementen in het sinogram enkel en alleen mogelijk zijn indien alle bijbehorende beeldpixels intensiteit nul hebben. Bij de reconstructie van reële data moeten we dit echter nuanceren. In paragraaf 2.5 werd aangegeven dat zich tijdens het detectieproces een aantal fouten voordoen (t. g. v. Compton-verstrooiing, randomcoïncidenties en dode tijd van de detectoren). Deze fouten kunnen er enerzijds aanleiding toe geven dat een detectorpaar coïncidenties detecteert, hoewel er zich geen activiteit tussen de detectoren bevindt. Hiermee zou rekening gehouden kunnen worden door tijdens de contourdetectie een goed gekozen drempelwaarde te hanteren i. p. v. de waarde nul. Hierdoor neemt echter ook de kans op fouten toe. Anderzijds kan het voorkomen, vooral bij metingen met een laag aantal coïncidenties, dat een bepaald detectorpaar geen coïncidenties detecteert hoewel er toch activiteit aanwezig is. Het foutief nul stellen van alle bijbehorende pixels geeft aanleiding tot onaanvaardbare artefacten in het beeld. Omwille van deze problemen, samen met de vaststelling dat contourdetectie enkel de snelheid van het algoritme verbetert en geen invloed heeft op de beeldkwaliteit, hebben we voorafgaande contourdetectie enkel toegepast bij fantoomstudies.

5.4.2 De aanpassingsmethode

Onder de aanpassingsmethode verstaan we de keuze voor het aantal gelijktijdig aangepaste pixels en de relatie tussen de intensiteitsveranderingen van de verschillende pixels. Gebaseerd op [Frie74] en [Ustu91] onderscheiden we twee verschillende methoden. Een eerste methode bestaat erin bij elke iteratiestap een willekeurige beeldpixel te kiezen en hieraan een intensiteitskorrel toe te kennen. We noemen deze methode Grain Allocation Method of afgekort GAM. Ander-



Figuur 5.8: Verloop van de totale kost i. f. v. het aantal iteraties met en zonder voorafgaande contourdetectie (HC-fantoom, 6.10^6 tellen).



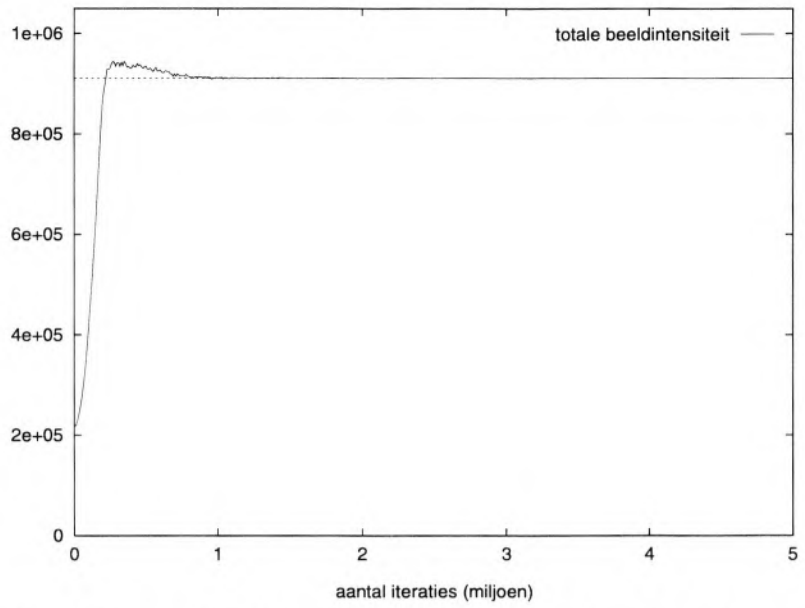
Figuur 5.9: Verloop van de beeldfout i. f. v. het aantal iteraties met en zonder voorafgaande contourdetectie (HC-fantoom, 6.10^6 tellen).

zijds kunnen we bij elke iteratiestap een willekeurig pixelkoppel kiezen en de overdracht van een intensiteitskorrel tussen beide pixels beschouwen. In dit geval spreken we van Grain Transfer Method of afgekort GTM.

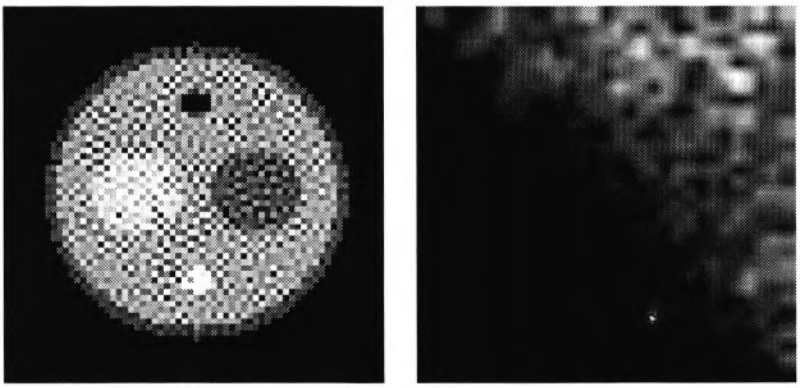
Het is duidelijk dat een algoritme dat uitsluitend gebruik maakt van GTM moet uitgaan van een beginbeeld met correcte totale intensiteit. Aangezien bij GTM-stappen enkel intensiteitsherverdeling plaatsvindt, wijzigt de totale beeldintensiteit niet. Deze totale intensiteit is echter niet bij voorbaat gekend; daarom zullen we de GTM-methode steeds moeten combineren met een aantal GAM-stappen om mogelijk te maken dat de totale intensiteit varieert.

Het is eenvoudig in te zien dat de dataterm van de kostfunctie meer gevoelig is aan de totale hoeveelheid intensiteit dan aan de plaats ervan. Fout gelokaliseerde intensiteit zal slechts onder welbepaalde projectiehoeken duidelijk zichtbaar zijn en zal dus slechts in bepaalde sinogramlijnen aanleiding geven tot een significante toename van de kost. Afwijkingen t. o. v. de correcte totale intensiteit zijn echter zichtbaar in elke sinogramlijn en zullen dus ook een sterkere stijging van de kost veroorzaken. Daarom zal de kostfunctie er voor zorgen dat tijdens de eerste iteratiestappen van de reconstructie zo snel mogelijk de correcte totale intensiteit in het beeld gevormd wordt. Dit is ook duidelijk te zien in figuur 5.10. Aangezien enkel GAM-overgangen de totale intensiteit kunnen wijzigen en de kostfunctie in deze fase nog niet plaatsgevoelig is, is het duidelijk dat intensiteit in eerste instantie op vrijwel willekeurige plaatsen terecht zal komen. Tijdens het verdere verloop van de reconstructie moet de aanwezige intensiteit herverdeeld worden. Het lijkt voor de hand liggend dat GTM-overgangen hiervoor efficiënter zullen zijn dan GAM-overgangen.

Omdat de toevoeging van een a priori-term de kostfunctie meer plaatsgevoelig maakt, hebben we in eerste instantie de invloed van de aanpassingsmethode onderzocht door enkel gebruik te maken van een dataterm als kostfunctie. Uit simulaties blijkt dat beide methoden (GAM en GTM) kwalitatief gelijkwaardige beelden opleveren, die echter sterk afwijken van het ideale beeld. De gereconstrueerde beelden bevatten storende "stervormige" artefacten, zoals duidelijk te zien is in figuur 5.11 a. Een Fourier-analyse van het foutbeeld (figuur 5.11 b) toont aan dat het foutbeeld vrijwel uitsluitend hoge-frequentiecomponenten bevat, wat betekent dat de afwijkingen in het beeld op zeer korte afstand (orde van pixelbreedte) gecompenseerd worden. Dit is in overeenstemming met de stervormige fouten die we waarnemen en heeft ertoe geleid om het GTM-mechanisme enig-



Figuur 5.10: Verloop van de totale beeldintensiteit i. f. v. het aantal iteraties (HC-fantoom, 6.10^6 tellen).



Figuur 5.11: Het gereconstrueerde beeld (a) wanneer enkel gebruik wordt gemaakt van een dataterm als kostfunctie en (b) 2D-FT van het foutbeeld (toenemende frequenties van links naar rechts en van onder naar boven).

zins aan te passen. We gaan ervan uit dat door GAM-overgangen de intensiteit in de omgeving van de correcte plaats terechtkomt. We laten bij GTM-overgangen daarom enkel nog intensiteitsoverdracht tussen naburige pixels toe. Uit simulaties blijkt echter dat deze nieuwe GTM-methode geen verbetering van de beeldkwaliteit met zich meebrengt. Om dit te verklaren bespreken we eerst kort het begrip lokaal minimum.

We definiëren de begrippen globaal minimum en lokaal minimum van een functie op een manier die het verband met het generatiemechanisme illustreert. We maken hierbij gebruik van de omgeving Ω van een punt $x_0 = (x_{0,1}, \ldots, x_{0,N})$ in de N-dimensionale ruimte; deze omgeving wordt gedefinieerd a. h. v. een verplaatsing $\delta = (\delta_1, \ldots, \delta_N)$ t. o. v. het centrale punt x_0:

$$\Omega = \{x \mid x = x_0 + \delta, \delta \in \mathcal{D}\}. \tag{5.5}$$

Zonder aan de algemeenheid te schaden stellen we dat $\mathcal{D}$ het produkt is van N intervallen $\mathcal{D}_i = [\delta_{i,min}, \delta_{i,max}]$, waarbij $\delta_{i,min} < 0$ en $\delta_{i,max} < 0$:

$$\begin{aligned} \mathcal{D} &= \mathcal{D}_1 \times \ldots \times \mathcal{D}_N \\ &= [\delta_{1,min}, \delta_{1,max}] \times \ldots \times [\delta_{N,min}, \delta_{N,max}]. \end{aligned} \tag{5.6}$$

Op een analoge manier introduceren we een gereduceerde omgeving Ω_0 van x_0 d. m. v. $\mathcal{D}_0 = \mathcal{D} \backslash \{(0, \ldots, 0)\}$. Het punt x_0 is een globaal minimum van de functie $f(x_1, \ldots, x_N)$ als

$$\begin{aligned} &\forall \delta \in \mathbb{R}^N \backslash \{(0, \ldots, 0)\} : \\ &\quad f(x_{0,1} + \delta_1, \ldots, x_{0,N} + \delta_N) \geq f(x_{0,1}, \ldots, x_{0,N}). \end{aligned} \tag{5.7}$$

Analoog stellen we dat x_0 een lokaal minimum is als

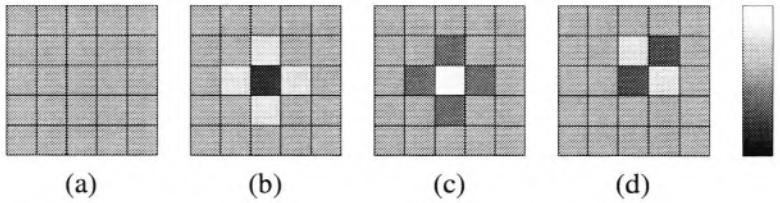
$$\begin{aligned} &\exists \mathcal{D}_0, \forall \delta \in \mathcal{D}_0 : \\ &\quad f(x_{0,1} + \delta_1, \ldots, x_{0,N} + \delta_N) \geq f(x_{0,1}, \ldots, x_{0,N}). \end{aligned} \tag{5.8}$$

We maken het onderscheid tussen mathematische lokale minima, die een gevolg zijn van het functieverloop, en lokale minima wegens de beperkingen van het generatiemechanisme. Een mathematisch lokaal minimum is – zoals hierboven gedefinieerd – een punt uit de N-dimensionale toestandsruimte waarvoor N δ-intervallen gevonden kunnen worden zodat elke wijziging $(\delta_1, \ldots, \delta_N)$ een toename in de functiewaarde veroorzaakt. Een lokaal minimum t. g. v. het generatiemechanisme daarentegen is een punt uit de toestandsruimte dat een lagere

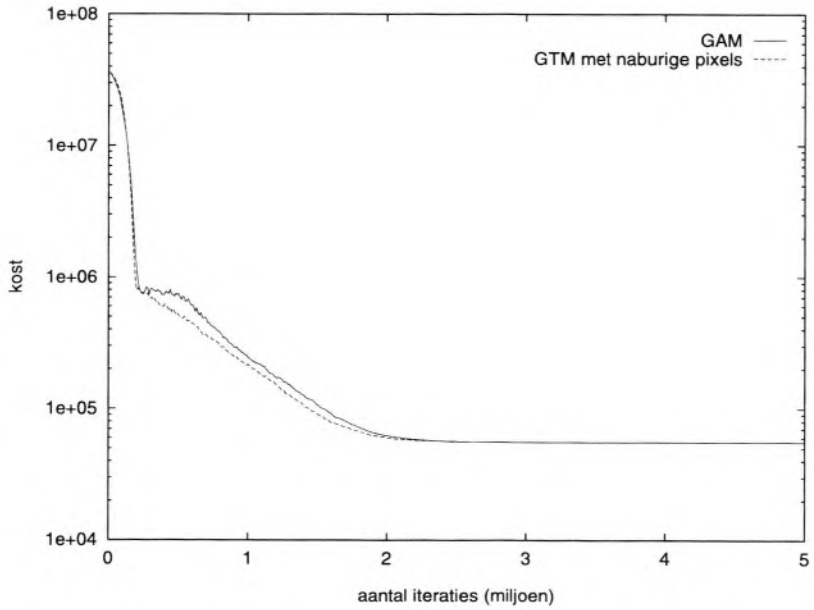
kostfunctiewaarde heeft dan alle punten die bereikt kunnen worden d. m. v. één iteratiestap. Dit betekent dat een ideaal generatiemechanisme (ideaal in de zin dat de lokale minima ervan overeenkomen met de mathematische lokale minima) moet toelaten om tijdens een iteratiestap ev. de waarde van elke veranderlijke aan te passen. Elk generatiemechanisme dat slechts enkele veranderlijken wijzigt per iteratiestap, introduceert een aantal lokale minima. Deze minima zijn het gevolg van het verschil tussen het mathematische begrip omgeving en de omgevingsstructuur die volgt uit het generatiemechanisme. We merken op dat een generatiemechanisme dat in overeenstemming is met de mathematische omgeving in praktijk onbruikbaar is. Wanneer toegelaten wordt dat tijdens één iteratiestap alle veranderlijken aangepast worden, wordt de efficiëntie van een iteratiestap verwaarloosbaar klein en bijgevolg wordt het benodigd aantal iteratiestappen onaanvaardbaar groot.

De beperkingen van het generatiemechanisme (en de aanpassingsmethode in het bijzonder) geven aanleiding tot een aantal artefacten die een lokaal minimum vormen. Enkele van deze artefacten worden schematisch weergegeven in figuur 5.12. Elke intensiteitswijziging of intensiteitsoverdracht tussen naburige pixels zal in eerste instantie de kostfunctie doen toenemen. Daarom werd gebruik gemaakt van aanpassingsmethoden die specifiek gericht zijn op de beeldfouten die we waarnemen. Zo werd bv. gebruik gemaakt van de zgn. "Mexicaanse hoed"-functie, waarbij intensiteitsoverdracht van een centrale pixel naar de vier naburige pixels beschouwd wordt. Toch kan zelfs het gebruik van specifieke aanpassingsmethoden de optredende artefacten niet volledig onderdrukken. We besluiten daarom dat de onderzochte methode weinig bruikbaar is wanneer de kostfunctie uitsluitend uit een dataterm bestaat.

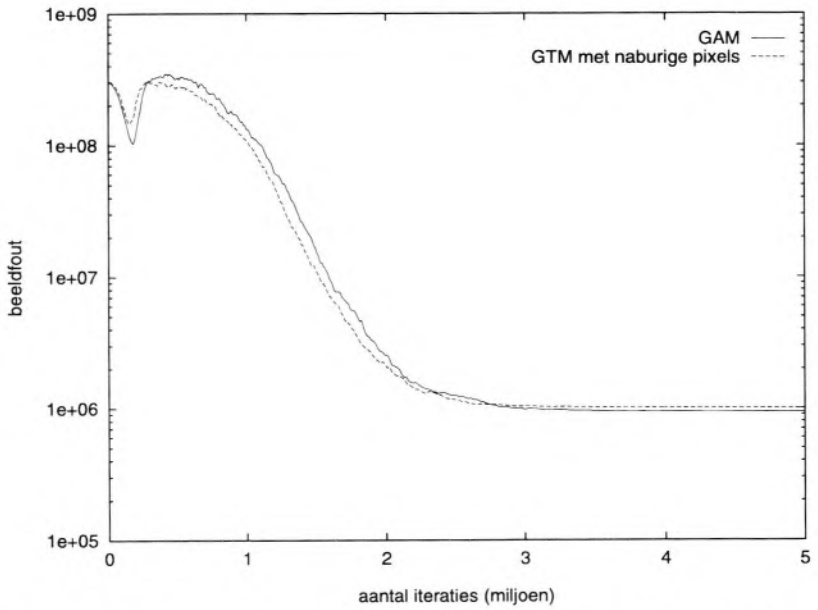
We herhalen dat de introductie van een a priori-term in de kostfunctie tot doel heeft de ruisdeterioratie van de gereconstrueerde beelden tegen te gaan. De artefacten die ontstaan door ruisdeterioratie zijn vergelijkbaar met de hierboven besproken artefacten. Het blijkt dat de toevoeging van een gepaste a priori-term aan de kostfunctie ook de artefacten t. g. v. de aanpassingsmethode onderdrukt. Uit figuren 5.13 en 5.14 blijkt dat bij gebruik van een a priori-term de invloed van de aanpassingsmethode op de beeldkwaliteit te verwaarlozen is. De benodigde rekentijd voor een GAM-overgang is echter beduidend kleiner dan voor GTM-overgangen, zoals blijkt uit figuur 5.15. We besluiten daarom enkel gebruik te maken van GAM-overgangen als aanpassingsmethode.



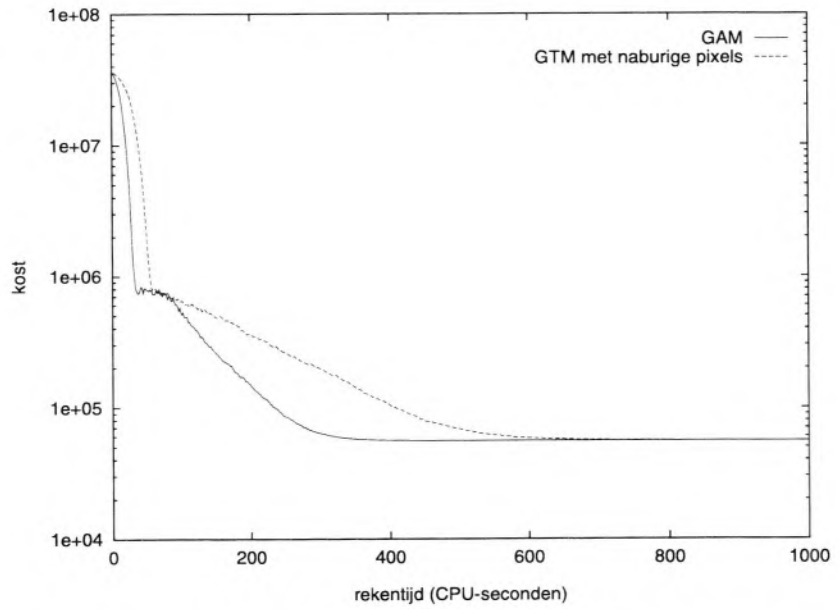
Figuur 5.12: Een schematische voorstelling van enkele van de optredende artefacten t. g. v. lokale minima van het generatiemechanisme; (a) stelt het ideale beeld voor.



Figuur 5.13: Verloop van de totale kost i. f. v. het aantal iteraties voor verschillende aanpassingsmethoden (HC-fantoom, 6.10^6 tellen).



Figuur 5.14: Verloop van de beeldfout i. f. v. het aantal iteraties voor verschillende aanpassingsmethoden (HC-fantoom, 6.10^6 tellen).



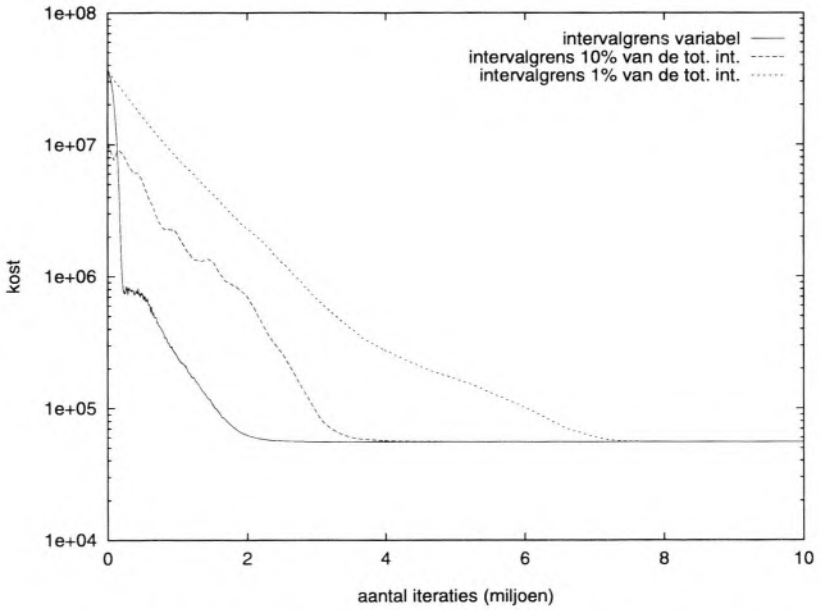
Figuur 5.15: Verloop van de totale kost i. f. v. de rekentijd voor verschillende aanpassingsmethoden (HC-fantoom, 6.10^6 tellen).

5.4.3 De korrelgrootte

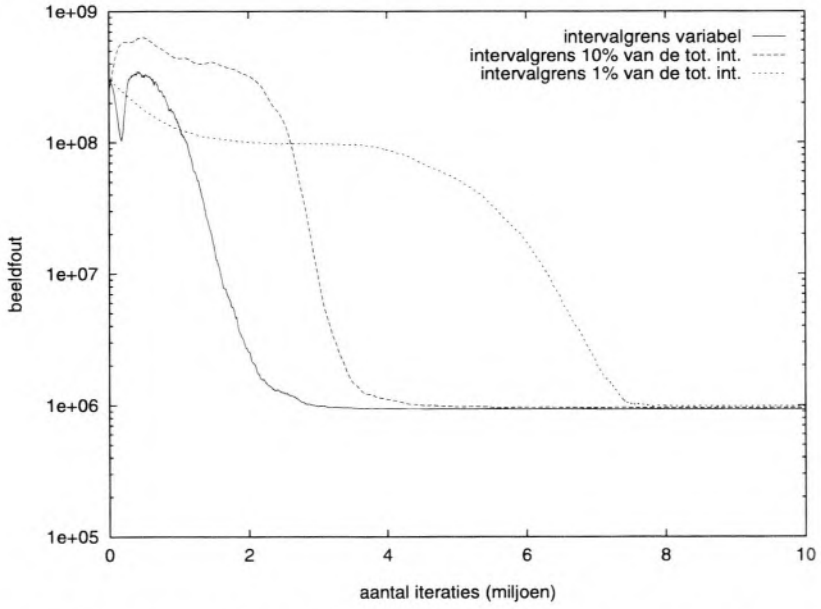
De intensiteit van de pixels wordt gewijzigd door toevoeging van een kleine hoeveelheid intensiteit, intensiteitskorrel genaamd. Het algoritme moet zowel positieve als negatieve intensiteitskorrels toelaten, maar er moet op toegezien worden dat de pixelintensiteit zelf nooit negatief wordt. De grootte van deze intensiteitskorrels moet klein zijn t. o. v. de gemiddelde pixelintensiteit, en dit vooral naar het einde van het reconstructieproces toe [VLaa87]. We merken hierbij op dat het Metropolis-criterium een zelfregulerende rol vervult. Enerzijds zal het gereconstrueerde beeld tijdens de finale iteratiestappen al van voldoende kwaliteit zijn, zodat het weinig waarschijnlijk is dat er grote intensiteitswijzigingen voorgesteld worden die een daling van de kost veroorzaken. Anderzijds zal door de lage temperatuurwaarde de aanvaardingswaarschijnlijkheid van grote intensiteitswijzigingen in positieve zin erg klein worden.

Zoals door de meeste auteurs voorgesteld wordt, hebben we de intensiteitskorrels gekozen volgens een uniforme distributie over een symmetrisch interval rond 0 $[-b, b]$ [Webb89, Kear90]. In eerste instantie hebben we de intervalgrensparameter b constant gehouden tijdens de reconstructie. We zien in figuur 5.16 dat voor een grote b-waarde de kost snel daalt tijdens de eerste iteratiestappen, maar dat verdere daling langzaam gebeurt. De voorgestelde intensiteitswijzigingen zijn gemiddeld (te) groot en worden dus meestal verworpen. Dit resultaat kunnen we ook afleiden uit figuur 5.18, waar we zien dat al snel meer dan 95% van de voorgestelde overgangen verworpen worden. Anderzijds blijkt uit figuur 5.16 dat voor kleine b-waarden de kost eveneens langzaam daalt. De efficiëntie van de voorgestelde intensiteitswijzigingen ligt echter veel hoger dan in het voorgaande geval (figuur 5.19). Tijdens een groot deel van de reconstructie maken de aanvaarde overgangen (positieve en negatieve) samen 90% uit van de voorgestelde overgangen. De intensiteitswijzigingen zijn echter te klein om de kostfunctie efficiënt te doen dalen. We vestigen er de aandacht op dat er weinig onderscheid is tussen aanvaarde en verworpen overgangen wat de rekentijd betreft. Het merendeel van de bewerkingen moet nl. uitgevoerd worden vooraleer de beslissing over aanvaarding genomen kan worden. Het is daarom aangewezen om de voorgestelde overgangen zó te kiezen dat de kans op aanvaarding groot is.

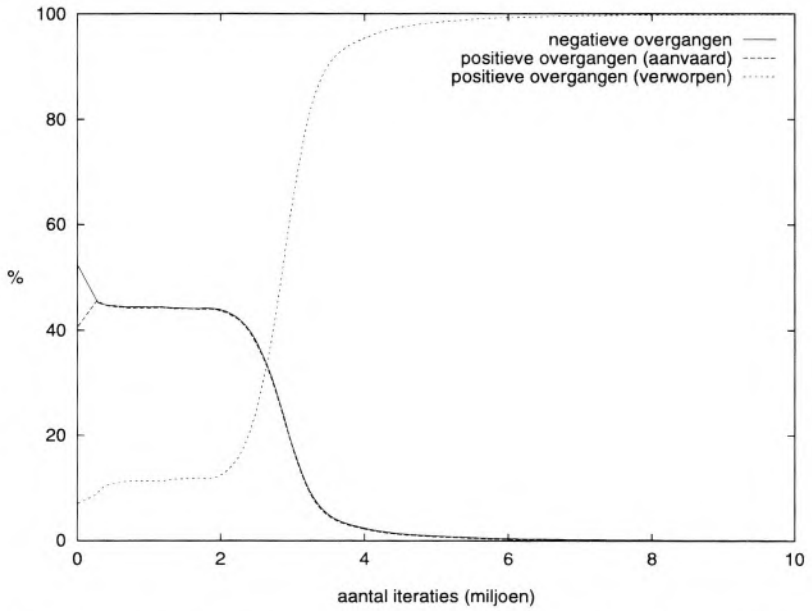
We hebben zelf een criterium ontwikkeld dat de waarde van de intervalgrensparameter b aanpast tijdens de reconstructie. Dit gebeurt door het verloop van de gemiddelde grootte van de aanvaarde intensiteitskorrels γ bij te houden. We merken hierbij op dat we met grootte van de intensiteitskorrel de absolute waarde



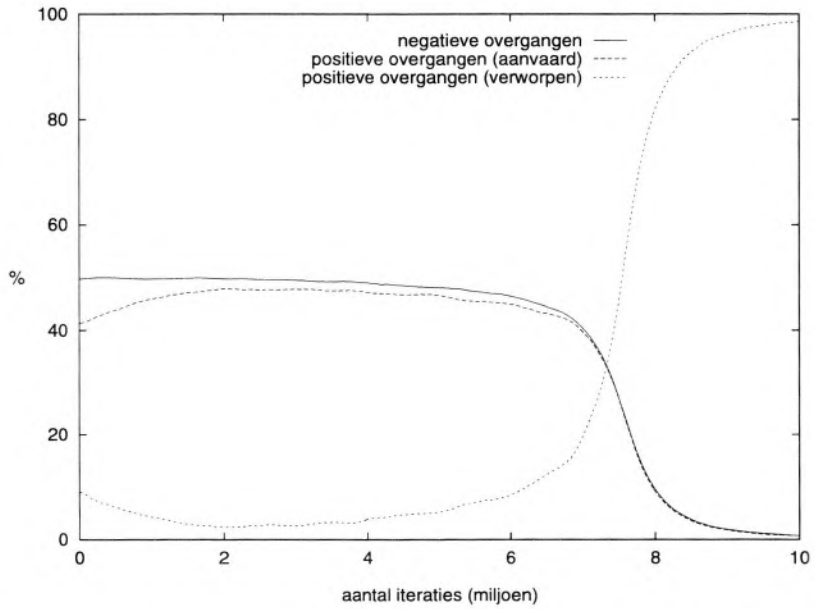
Figuur 5.16: Verloop van de totale kost i. f. v. het aantal iteraties voor verschillende intervalgrenzen voor de bepaling van de korrelgrootte (HC-fantoom, 6.10^6 tellen).



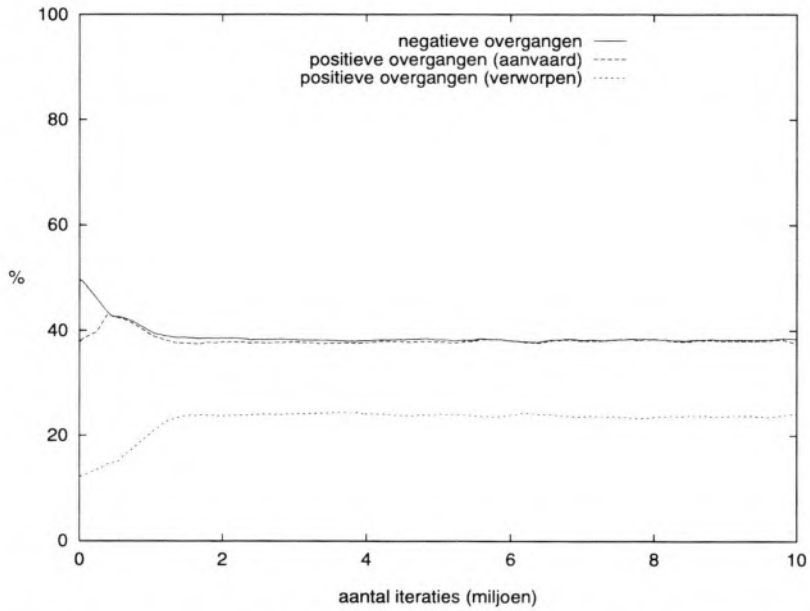
Figuur 5.17: Verloop van de beeldfout i. f. v. het aantal iteraties voor verschillende intervalgrenzen voor de bepaling van de korrelgrootte (HC-fantoom, 6.10^6 tellen).



Figuur 5.18: Procentuele verdeling van de verschillende overgangen i. f. v. het aantal iteraties voor een constante intervalgrens (10% van de maximale intensiteit) voor de bepaling van de korrelgrootte (HC-fantoom, 6.10^6 tellen).



Figuur 5.19: Procentuele verdeling van de verschillende overgangen i. f. v. het aantal iteraties voor een constante intervalgrens (1% van de maximale intensiteit) voor de bepaling van de korrelgrootte (HC-fantoom, 6.10^6 tellen).



Figuur 5.20: Procentuele verdeling van de verschillende overgangen i. f. v. het aantal iteraties voor een variabele waarde van de intervalgrens voor de bepaling van de korrelgrootte (HC-fantoom, 6.10^6 tellen).

van de intensiteitskorrel bedoelen. Door telkens

$$\beta = 2\gamma \tag{5.9}$$

te stellen, zorgen we ervoor dat de gemiddelde voorgestelde korrelgrootte overeenstemt met de waarde die het meest waarschijnlijk aanvaard wordt. In praktijk gebruiken we echter een b-waarde die iets groter is (factor 2.5 i. p. v. 2) om toe te laten dat de korrelgrootte indien nodig kan toenemen.

We zien in figuren 5.16 en 5.17 dat dit mechanisme met variabele intervalgrenswaarde tot beter resultaten leidt. De korrelgrootte is steeds aangepast aan de temperatuurwaarde, zodat de kost indien mogelijk snel kan dalen. De efficiëntie van dit mechanisme blijkt ook duidelijk uit figuur 5.20, waar we zien dat tijdens het ganse verloop van de reconstructie ongeveer 80% van de voorgestelde overgangen aanvaard worden. Tijdens de finale fase van de reconstructie zijn deze overgangen echter verwaarloosbaar klein geworden, zodat er geen merkbare toename van de beeldkwaliteit is (figuur 5.17).

We hebben ook onderzocht wat de invloed is van de distributie van de intensiteitskorrels. Zo hebben we gebruik gemaakt van een Gaussiaanse verdeling met variabele standaardafwijking σ i. p. v. een uniforme distributie met variabele intervalgrens b. Intuïtief zou men vermoeden dat een Gaussiaanse verdeling aanleiding geeft tot betere resultaten. Enerzijds worden meer kleine overgangen voorgesteld, hetgeen in overeenstemming is met [VLaa87]. Anderzijds blijven overgangen met grote intensiteitswijzigingen theoretisch mogelijk, wat nuttig kan zijn om ev. een lokaal minimum te verlaten.

We stellen eerst een uitdrukking op voor de optimale waarde van de standaardafwijking.

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \tag{5.10}$$

$$\begin{aligned} E[|x|] &= \int_0^{+\infty} x\, p(x)\, dx - \int_{-\infty}^{0} x\, p(x)\, dx \\ &= \frac{2\sigma}{\sqrt{2\pi}} \end{aligned} \tag{5.11}$$

Wanneer σ gekozen wordt i. f. v. de gemiddelde grootte van de aanvaarde intensiteitskorrels γ volgens

$$\sigma = \sqrt{\frac{\pi}{2}}\gamma \tag{5.12}$$

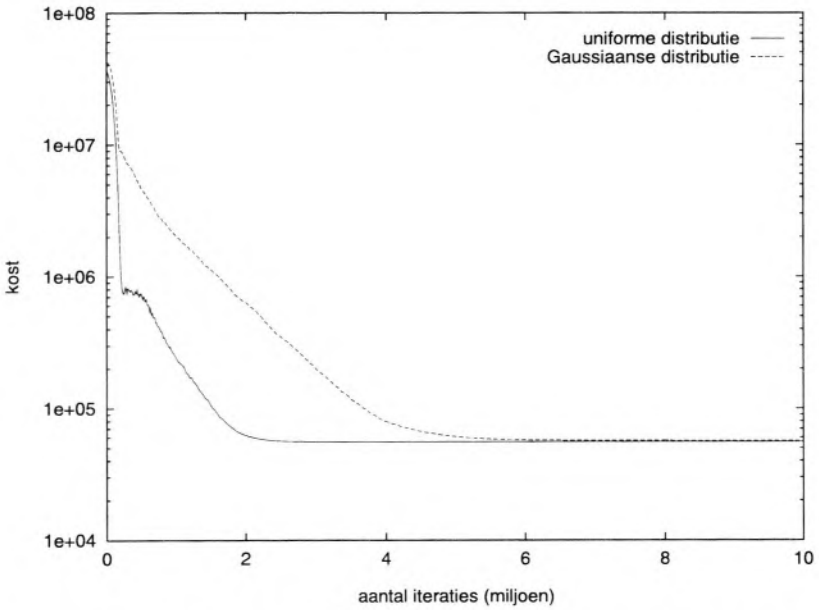
zullen de voorgestelde intensiteitskorrels opnieuw de optimale aanvaardingswaarschijnlijkheid hebben. Ook hier zal in praktijk een grotere waarde voor σ gebruikt worden om ev. een stijging van de korrelgrootte toe te laten.

Uit de simulatieresultaten in figuren 5.21 en 5.22 blijkt echter dat Gaussiaans verdeelde korrelgroottes minder efficiënt zijn. Dit is in tegenspraak met wat men intuïtief zou verwachten. Een mogelijke verklaring hiervoor is het relatief grotere aantal overgangen met kleine intensiteitswijzigingen in vergelijking met een uniforme distributie, waardoor ook de efficiëntie lager ligt. Dit is te zien in figuur 5.23, waar de optimale uniforme en Gaussiaanse distributie voor eenzelfde waarde van γ weergegeven zijn. We bemerken ook dat er gelijkenis bestaat tussen het verloop van de kost voor een Gaussiaanse distributie (figuur 5.21) en het verloop bij een uniforme distributie met constante en te kleine intervalgrens (figuur 5.16). Dit doet opnieuw vermoeden dat de problemen te wijten zijn aan overgangen met te kleine intensiteitswijzigingen. Daarnaast is er het feit dat de benodigde rekentijd voor het genereren van een willekeurige overgang uit een Gaussiaanse distributie groter is dan deze voor een uniforme distributie. Dit leidt ertoe te besluiten dat het gebruik van een uniforme distributie met variabele intervalgrens de meest aangewezen methode is voor de keuze van de korrelgrootte.

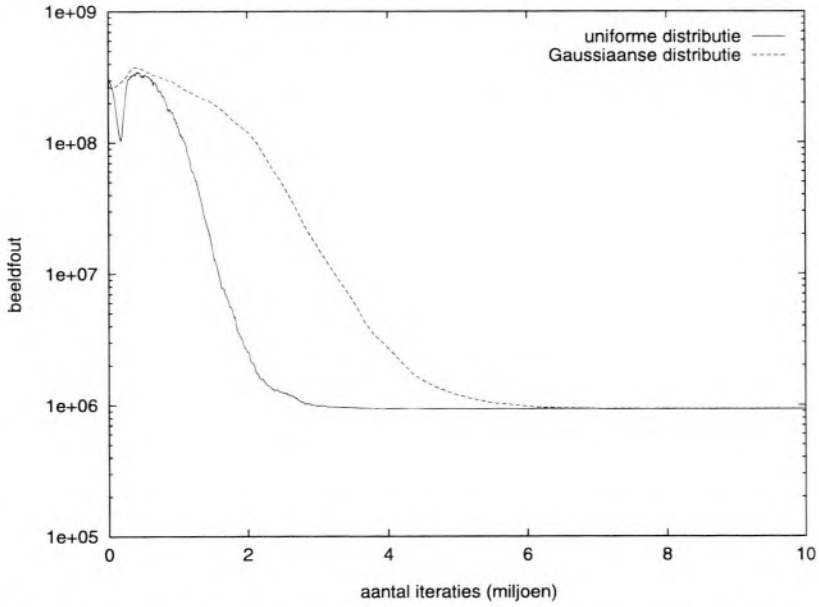
5.4.4 Alternatieve generatiemechanismen

We herhalen dat het besproken generatiemechanisme de intensiteit van één pixel of enkele willekeurig gekozen pixels wijzigt. We bespreken hieronder kort twee alternatieve generatiemechanismen.

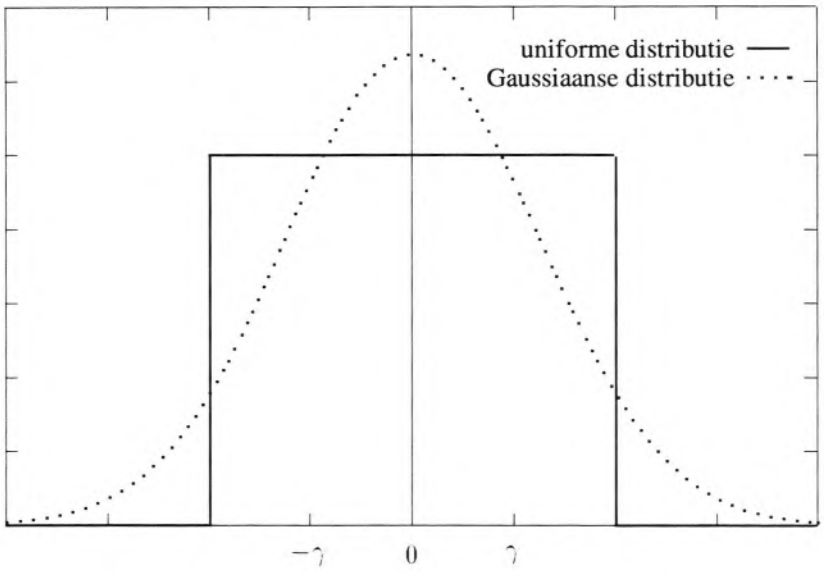
Als eerste alternatief zouden de voorgestelde overgangen gegenereerd kunnen worden in de dataruimte i. p. v. in de beeldruimte. De intensiteitsverandering van een punt uit de datamatrix (het sinogram) correspondeert met de intensiteitsverandering van de beeldpixels die gelegen zijn op de strook tussen een detectorpaar. De coördinaten van het datapunt bepalen de positie van het detectorpaar en de hoek van de projectiestrook. Het generatiemechanisme zou dan alle of enkele willekeurig gekozen pixels op deze projectiestrook kunnen aanpassen. De onderliggende motivatie voor dit alternatief is dat het nu gebruikte generatiemecha-



Figuur 5.21: Verloop van de totale kost i. f. v. het aantal iteraties voor verschillende distributies van de korrelgrootte (HC-fantoom, 6.10^6 tellen).



Figuur 5.22: Verloop van de beeldfout i. f. v. het aantal iteraties voor verschillende distributies van de korrelgrootte (HC-fantoom, 6.10^6 tellen).



Figuur 5.23: Optimale vorm van een uniforme en Gaussiaanse distributie van de korrelgrootte, bij gelijke gemiddelde grootte van de aanvaarde intensiteitskorrels γ.

nisme enkele beeldpixels wijzigt, maar dat de evaluatie van de voorgestelde overgang hoofdzakelijk gebeurt in de dataruimte (m. b. v. de dataterm). Het is daarom onwaarschijnlijk dat een lokale afwijking in de dataruimte (bv. 2 naburige sinogramelementen die verschillen) effectief verholpen zal worden door willekeurig gekozen beeldpixels te wijzigen. Deze wijziging heeft nl. gevolgen in het volledige sinogram, waardoor de wijziging lokaal een verbetering kan betekenen, maar over het totale sinogram toch een stijging van de dataterm kan veroorzaken. Het potentiële nadeel van deze alternatieve aanpak is dat bij elke iteratiestap zowel een voorwaartse projectie als een terugprojectie berekend moet worden, waardoor de rekentijd sterk toeneemt. Daarnaast zijn een aantal positieve eigenschappen van de gebruikte methode (zoals het hogere contrast van de gereconstrueerde beelden) mede een gevolg van het gebruikte generatiemechanisme. Bij toepassing van een terugprojectie-generatiemechanisme bestaat de kans dat een aantal voordelen van de onderzochte methode verdwijnen en dat de gereconstrueerde beelden meer gelijkenis vertonen met bv. ML-EM reconstructies. Een meer grondige studie van dit alternatief werd echter niet uitgevoerd.

Een tweede alternatieve aanpak zou erin kunnen bestaan om overgangen te genereren aan de hand van een andere set basisfuncties dan de klassieke pixels. Zo zou bv. gebruik gemaakt kunnen worden van wavelets. De coëfficiënten van wavelet-basisfuncties hebben de eigenschap dat zij zowel spatiale als frequentiële informatie bevatten. Voor een bespreking van het begrip wavelets verwijzen we o. a. naar [Mall89, Riou91, Temm91, Bult95] ; een overzicht van de biomedische toepassingen van wavelets wordt gegeven in [Akay95, Aldr96, Unse96]. We suggereren enkel kort een aantal potentiële voordelen van het gebruik van wavelets in deze context. Wanneer tijdens de eerste fase van het reconstructieproces enkel laagfrequente wavelets toegelaten worden, ontstaan in het beeld enkel de algemene kenmerken. Pas in een latere fase kan toegelaten worden dat ook hoge frequenties in het beeld geïntroduceerd worden. Dit heeft als voordeel dat de intensiteit zich onmiddellijk op de juiste plaats bevindt in het beeld, en dat een herverdeling van de intensiteit enkel nodig is op plaatsen waar randen gevormd moeten worden. Bovendien kunnen we door een gelijkmatig toevoegen van hoge frequenties in het beeld het eventuele probleem van ruisdeterioratie beter controleren. Het toepassingsgebied van wavelet-basisfuncties reikt echter verder dan het generatiemechanisme.

Momenteel beperkt de toepassing van wavelets voor PET-reconstructie zich hoofdzakelijk tot aanpassingen van een iteratieve inversiemethode voor de Radontransformatie (d. w. z. gefilterde terugprojectie m. b. v. wavelets) [Sahi93, Sahi96b, Sahi96a, Bhat96, Dela95, Dela98, Hajj96, Kola94]. Wavelet-basisfuncties kunnen echter ook gebruikt worden om het beeld te reconstrueren als een aantal waveletcoëfficiënten. Deze aanpak hoeft zich niet te beperken tot het gebruik van simulated annealing, maar kan ook voor andere iteratieve reconstructiemethoden onderzocht worden [Wu93]. De onderliggende motivatie is de ruisonderdrukkende eigenschap van wavelets voor medische beelden [Malf94, Malf95] (enkel hoogfrequente ruis). We stellen vast dat gereconstrueerde beelden bij PET veel ruis bevatten en dat waveletcompressie van deze beelden (met bv. compressiefactor 10) vaak aanleiding geeft tot visueel betere beelden. Dit suggereert dat men zich tijdens de reconstructie met wavelet-basisfuncties zou kunnen beperken tot een kleine deelverzameling van de complete set basisfuncties zonder kwaliteitsverlies. Hierdoor kan de dimensionaliteit van het reconstructieprobleem sterk gereduceerd worden. Het probleem van deze aanpak is enerzijds dat niet bij voorbaat gekend is welke deelverzameling aangewezen is uit de complete set basisfuncties, zodat dit tijdens de reconstructie bepaald moet worden. Een tweede probleem is de berekening van de voorwaartse projectie en de terugprojectie. Bij klassieke

(pixelgebaseerde) reconstructiemethoden kan men gebruik maken van een geoptimaliseerde berekeningswijze van het radiologisch pad, wat resulteert in snelle projectiemethoden [Pete81, Jose82, Jaco98]. We beschikken echter niet over een efficiënt algoritme voor beide projecties tussen een representatie m. b. v. wavelets en de sinogramruimte zonder gebruik te moeten maken van een tussenrepresentatie als pixels. Een diepgaande studie van de verhouding tussen het voordeel van de reductie van de dimensionaliteit en het nadeel van de sterke toename van rekentijd voor beide projecties lijkt aangewezen.

5.5 Het afkoelingsschema

We herinneren er nogmaals aan dat het afkoelingsschema het verloop van de temperatuurparameter tijdens de reconstructie bepaalt. Bij het homogene simulated-annealingalgoritme daalt de temperatuur sprongsgewijze en vormen de iteraties bij constante temperatuurwaarde een Markov-keten. In dit geval wordt het afkoelingsschema volledig bepaald door vier parameters, nl. de begintemperatuur, de eindtemperatuur (t. t. z. een stopcriterium), de lengte van de Markov-ketens en de grootte van de tussenliggende temperatuursprongen. We bespreken elk van deze parameters meer in detail.

5.5.1 De begintemperatuur

We onderscheiden twee verschillende gevallen voor de bepaling van een beginwaarde voor de temperatuur τ. Enerzijds veronderstellen we dat de begintoestand (het beginbeeld) geen relevante informatie bevat en opgevat mag worden als een willekeurig beginbeeld. In dit geval moet de temperatuur voldoende hoog gekozen worden opdat bijna alle voorgestelde overgangen aanvaard worden. Anderzijds kan uitgegaan worden van een beginbeeld dat reeds in zekere mate een benadering is van het te reconstrueren beeld. In dit tweede geval is de keuze van de begintemperatuur van cruciaal belang. Een te hoge waarde voor τ zal (te) veel positieve overgangen toelaten, waardoor een deel van de al aanwezige informatie in het beeld vernietigd wordt. Een te lage waarde voor τ zal er echter voor zorgen dat het ev. lokale minimum van de begintoestand niet meer verlaten kan worden zodat geen verdere beeldverbetering mogelijk is. We bespreken beide gevallen meer in detail.

Willekeurig beginbeeld

In de literatuur over simulated annealing worden een aantal methoden voorgesteld voor de bepaling van een gepaste beginwaarde voor τ. Deze methoden zijn steeds geïnspireerd op de analogie tussen de statistische mechanica en simulated annealing en trachten de begintemperatuur zó te kiezen dat de waarschijnlijkheidsdistributie van toestanden de uniforme distributie over de toestandsruimte benadert. Dit betekent dat elke voorgestelde overgang – en dus meer in het bijzonder elke positieve overgang, vermits negatieve overgangen steeds aanvaard zullen worden – aanvaard moet worden. Hierbij worden met positieve en negatieve overgangen respectievelijk overgangen bedoeld die de kostfunctiewaarde doen toenemen of afnemen. We herhalen dat de aanvaardingswaarschijnlijkheid voor een positieve overgang met kosttoename ΔH gegeven wordt door

$$p(\Delta H) = \exp\left(-\frac{\Delta H}{\tau}\right). \tag{5.13}$$

Dit zou echter betekenen dat τ initieel oneindig groot gekozen moet worden. Daarom wordt gebruik gemaakt van het begrip aanvaardingsratio χ voor positieve overgangen. Deze aanvaardingsratio wordt gedefinieerd als de verhouding van het aantal aanvaarde positieve overgangen tot het aantal voorgestelde positieve overgangen. We merken op dat een te grote beginwaarde voor de aanvaardingsratio (en dus een te hoge τ-beginwaarde) de rekentijd nodeloos doet toenemen zonder de kwaliteit van de eindoplossing te beïnvloeden. Een te lage aanvaardingsratio daarentegen verhoogt het risico dat het algoritme vroegtijdig gevangen wordt in een lokaal minimum. De meeste auteurs zijn het erover eens dat $\chi_0 = 0.8$ een goede beginwaarde voor de aanvaardingsratio is (o. a. [VLaa87]).

Kirkpatrick *et al.* [Kirk83] stellen de volgende empirische regel voor: *Kies een voldoende grote waarde voor τ en voer een aantal willekeurige transities uit. Wanneer de aanvaardingsratio χ kleiner is dan een goed gekozen waarde χ_0, verdubbel dan de waarde van τ. Herhaal deze procedure tot wanneer de aanvaardingsratio groter wordt dan χ_0.* Dit betekent echter dat voor elke τ-waarde een testreeks van overgangen gegenereerd moet worden.

Een meer gesofisticeerde versie van deze methode wordt o. a. voorgesteld door Otten *et al.* [Otte89] en wordt door de meeste auteurs gehanteerd. Onderstel een testreeks van M willekeurige overgangen met kostverandering ΔH_i, waarvan N ($N \leq M$) positieve overgangen met kosttoename $\Delta H_i^{(+)}$. Gezocht is de waarde

van τ waarvoor

$$\frac{1}{N}\sum_{i=1}^{N}\exp\left(-\frac{\Delta H_i^{(+)}}{\tau}\right) = \chi_0. \tag{5.14}$$

De gemiddelde kosttoename t. g. v. positieve overgangen is

$$\overline{\Delta H}^{(+)} = \frac{1}{N}\sum_{i=1}^{N}\Delta H_i^{(+)}. \tag{5.15}$$

Benaderen we dan (5.14) door

$$\exp\left(-\frac{\overline{\Delta H}^{(+)}}{\tau}\right) = \chi_0. \tag{5.16}$$

dan vinden we dat

$$\tau = \frac{\overline{\Delta H}^{(+)}}{\ln\frac{1}{\chi_0}}. \tag{5.17}$$

Het voordeel van deze methode t. o. v. de eerste methode is dat de temperatuurwaarde berekend kan worden uit slechts één testreeks van overgangen. Voor $\chi_0 = 0.8$ leidt dit tot

$$\tau = 4.48\,\overline{\Delta H}^{(+)}. \tag{5.18}$$

Uit simulaties blijkt dat deze waarde aanleiding geeft tot een aanvaardingsratio van ongeveer 0.7, hetgeen afwijkt van de vooropgestelde waarde. Bovendien wordt nergens een argumentatie gegeven voor de overgang van (5.14) naar (5.16).

We hebben zelf een alternatieve methode ontwikkeld voor de bepaling van τ. We gaan hierbij opnieuw uit van een reeks van M willekeurige overgangen, waarvan N positieve overgangen, en zoeken de τ-waarde waarvoor (5.14) geldt. Vermits we τ zó zullen bepalen dat het gemiddelde van de exponentiële termen dicht bij de waarde 1 ligt ($\chi_0 = 0.8$), mogen we aannemen dat de afzonderlijke exponentiële termen ook voldoende in de nabijheid van 1 zullen liggen (m. a. w. $\tau \gg \Delta H$). We ontbinden daarom de exponentiële termen van (5.14) in een Taylorreeks en verwaarlozen de termen van tweede orde en hoger:

$$\begin{aligned}\exp\left(-\frac{\Delta H}{\tau}\right) &= 1 - \frac{\Delta H}{\tau} + \mathcal{O}\left(\frac{\Delta H}{\tau}\right)^2 \\ &\approx 1 - \frac{\Delta H}{\tau}.\end{aligned} \tag{5.19}$$

Bijgevolg moet τ voldoen aan

$$1 - \frac{\overline{\Delta H}^{(+)}}{\tau} = \chi_0, \tag{5.20}$$

of uiteindelijk

$$\tau = \frac{\overline{\Delta H}^{(+)}}{1 - \chi_0}. \tag{5.21}$$

We merken op dat deze benadering een overschatting van de gezochte waarde van τ zal opleveren ten gevolge van de verwaarlozing van de hogere-ordetermen in de Taylorontwikkeling. Zoals reeds gezegd heeft een overschatting van τ enkel een toename van de rekentijd voor gevolg, terwijl een onderschatting het risico op een vroegtijdig lokaal minimum verhoogt. Voor $\chi_0 = 0.8$ vinden we ditmaal dat

$$\tau = 5\,\overline{\Delta H}^{(+)}. \tag{5.22}$$

Uit simulaties blijkt dat deze nieuwe methode aanleiding geeft tot een aanvaardingsratio van ongeveer 0.82, wat de vooropgestelde waarde beter benadert. We concluderen daarom dat de nieuwe methode te prefereren is boven de eerder vermelde methoden.

Zinvol beginbeeld

De situatie van een zinvol beginbeeld doet zich bv. voor wanneer het reconstructiealgoritme vroegtijdig wordt afgebroken. Later kan het wenselijk zijn de reconstructie verder te zetten zonder dat men de beschikking heeft over de temperatuurwaarde bij stopzetting. Een andere mogelijkheid is het gebruik van simulated annealing als nabewerking van beelden die afkomstig zijn van bv. snellere reconstructietechnieken. In beide gevallen wenst men de al aanwezige informatie in het beeld optimaal te benutten. Bij een overschatting van de temperatuurwaarde zal de aanvaardingsratio te hoog zijn, zodat door positieve overgangen de beeldinhoud gedeeltelijk verloren gaat. Een onderschatting van de temperatuurwaarde verhoogt echter opnieuw het risico op een vroegtijdig lokaal minimum. Er zijn in de literatuur slechts sporadische verwijzingen te vinden naar dit probleem van temperatuurkeuze. Grover *et al.* [Grov87] spreken in deze context van "simulated sintering". Het probleem wordt echter enkel aangehaald zonder een oplossing te geven. De meeste auteurs maken gebruik van een zelfgekozen startwaarde gebaseerd op ervaring. Deze aanpak heeft echter geen theoretische basis en is zeer

sterk probleemafhankelijk.

Uitgaande van het artikel van Rose *et al.* [Rose90] hebben we zelf twee methoden ontwikkeld voor de bepaling van de begintemperatuur: een dynamische en een statische methode. Beide methoden zijn gebaseerd op het concept van thermisch evenwicht. Om tot een praktisch bruikbare mathematische formulering van het begrip "evenwicht" te komen herhalen we een aantal begrippen en notaties uit hoofdstuk 4 en voeren een aantal nieuwe begrippen in.

De reeks overgangen die door het simulated-annealingalgoritme bij constante temperatuurwaarde gegenereerd worden (het inhomogene algoritme) kunnen gemodelleerd worden als een Markov-keten. We stellen dat evenwicht bereikt is wanneer een stationaire waarschijnlijkheidsdistributie van toestanden bereikt is, d. w. z. wanneer de waarschijnlijkheid om in een bepaalde toestand te verkeren niet meer wijzigt bij verdere iteraties. We definiëren nu de evenwichtstemperatuur van een bepaalde toestand als de temperatuur waarvoor de Markov-keten die deze toestand als begintoestand heeft in evenwicht is. Dit betekent o. a. dat de kostfunctiewaarde van de opeenvolgende toestanden niet verandert, of nog dat de verwachtingswaarde voor de kostverandering nul is:

$$E[\Delta H] = \int_{-\infty}^{+\infty} \Delta H \, p(\Delta H) \, p_a(\Delta H) \, d(\Delta H) = 0. \tag{5.23}$$

Hierin is $p(\Delta H)$ de waarschijnlijkheid dat het simulated-annealingproces in evenwicht een overgang genereert die aanleiding geeft tot een kostverandering ΔH; $p_a(\Delta H)$ is de aanvaardingswaarschijnlijkheid van deze overgang. Zoals reeds gezegd is deze aanvaardingswaarschijnlijkheid 1 als $\Delta H < 0$ en $\exp(-\frac{\Delta H}{\tau})$ als $\Delta H > 0$. We kunnen daarom schrijven dat bij evenwicht moet gelden dat

$$\Delta H^{(-)} + \Delta H^{(+)} = 0, \tag{5.24}$$

waarbij

$$\begin{aligned} \Delta H^{(-)} &= \int_{-\infty}^{0} \Delta H \, p(\Delta H) \, d(\Delta H), \\ \Delta H^{(+)} &= \int_{0}^{+\infty} \Delta H \, p(\Delta H) \, \exp\left(-\frac{\Delta H}{\tau}\right) \, d(\Delta H). \end{aligned} \tag{5.25}$$

Van cruciaal belang in deze formule is de waarschijnlijkheidsdistributie van kostveranderingen $p(\Delta H)$, die we daarom meer in detail bestuderen.

We maken hiervoor gebruik van een aantal begrippen die geïntroduceerd werden in paragraaf 4.7. We herhalen dat het generatiemechanisme correspondeert met een (temperatuuronafhankelijke) generatiematrix $\mathbf{G}$ waarin elk element G_{ij} de waarschijnlijkheid voorstelt dat een overgang van toestand i naar toestand j gegenereerd wordt. De waarschijnlijkheidsdistributie van toestanden π drukt de waarschijnlijkheid π_i uit dat het systeem zich in toestand i bevindt; bij evenwicht is deze toestandsdistributie stationair. Verder noteren we nog de kostverandering t. g. v. de overgang van toestand i naar toestand j als

$$\Delta H_{ij} = H(\text{toestand } j) - H(\text{toestand } i). \tag{5.26}$$

We kunnen dan de waarschijnlijkheidsdistributie van kostveranderingen $p(\Delta H)$ schrijven als

$$p(\Delta H) = \sum_{i=1}^{N} \sum_{j=1}^{N} \pi_i \times \begin{cases} G_{ij}, & \Delta H_{ij} = \Delta H \\ 0, & \Delta H_{ij} \neq \Delta H. \end{cases} \tag{5.27}$$

We merken op dat deze distributie temperatuurafhankelijk is wegens de temperatuurafhankelijkheid van π.

Wanneer we de evenwichtstemperatuur wensen te bepalen die correspondeert met een bepaalde begintoestand, moeten we beschikken over $p(\Delta H)$. De bepaling van $p(\Delta H)$ is echter zelf temperatuurafhankelijk. Het is duidelijk dat het bemonsteren van $p(\Delta H)$ bij een foutieve temperatuurwaarde aanleiding zal geven tot een foutieve schatting voor de evenwichtstemperatuur. We kunnen echter het volgende iteratieve schema formuleren voor de temperatuurbepaling:

1. kies een arbitraire beginwaarde voor de temperatuur τ_0;
2. bemonster $p(\Delta H)$ bij de huidige temperatuurwaarde τ_n, uitgaande van de begintoestand;
3. bepaal $\Delta H_n^{(-)}$ en $\Delta H_n^{(+)}$ volgens (5.25);
4. bepaal de correctiefactor r_n volgens

$$r_n = -\frac{\Delta H_n^{(-)}}{\Delta H_n^{(+)}}; \tag{5.28}$$

5. als $r_n = 1$ dan is voldaan aan (5.24) en is τ_n de gezochte evenwichtstemperatuur; als $r_n \neq 1$ stel dan $\tau_{n+1} = r_n \tau_n$ en ga naar stap 2.

Het is duidelijk dat $r_n < 1$ correspondeert met een te groot aandeel van de positieve overgangen, wat betekent dat de aanvaardingswaarschijnlijkheid voor positieve overgangen en dus ook de temperatuurwaarde te groot is. Analoog betekent $r_n > 1$ dat het aandeel van de negatieve overgangen te groot is, zodat de aanvaardingswaarschijnlijkheid voor positieve overgangen en bijgevolg de temperatuurwaarde te klein is. Hiermee is evenwel geen bewijs geleverd voor de convergentie van het iteratieve schema. Uit simulaties is gebleken dat deze methode in alle onderzochte gevallen convergeert. De correctiefactoren r_n nemen echter snel waarden aan in de omgeving van 1, waardoor de convergentie zeer langzaam verloopt. Er moeten bijgevolg een relatief groot aantal iteraties (grootteorde 100) uitgevoerd worden alvorens de temperatuurwaarde een goede benadering van de evenwichtstemperatuur is. Daarnaast is er het belangrijke probleem dat bij elke temperatuurwaarde τ_n $p(\Delta H)$ opnieuw bemonsterd moet worden, wat betekent dat telkens een Markov-keten met een voldoende aantal overgangen berekend moet worden. We noemen deze methode van temperatuurschatting daarom de dynamische methode. De trage convergentie en de noodzaak om $p(\Delta H)$ telkens opnieuw te bemonsteren resulteren in een aanzienlijke rekentijd.

We trachten dit op te lossen door een statische benadering voor de waarschijnlijkheidsdistributie $p_s(\Delta H)$ te bepalen. Hiermee bedoelen we dat de statistiek van overgangen met ΔH bepaald wordt door overgangen te genereren, zonder evenwel deze overgangen te accepteren. Elke overgang zal dus steeds uitgaan van dezelfde begintoestand i. Hierdoor verdwijnt in (5.27) de stationaire toestandsdistributie π en dus ook de temperatuurafhankelijkheid, zodat

$$p_s^i(\Delta H) = \sum_{j=1}^{N} \begin{cases} G_{ij}, & \Delta H_{ij} = \Delta H \\ 0, & \Delta H_{ij} \neq \Delta H. \end{cases} \tag{5.29}$$

Het aandeel $\Delta H^{(-)}$ van de negatieve overgangen wordt dan een constante en enkel het aandeel van de positieve overgangen $\Delta H_n^{(+)}$ is functie van de temperatuur. De evenwichtstemperatuur kan dan uit (5.24) bepaald worden door gebruik te maken van bv. de bisectiemethode of de methode van Newton-Raphson. Enerzijds moet bij de statische methode slechts één keer een reeks overgangen gegenereerd worden. Anderzijds is het aantal benodigde iteratiestappen van zowel de bisectiemethode als de methode van Newton-Raphson voor het zoeken van een nulpunt van (5.24) veel kleiner dan het aantal iteratiestappen van de dynamische methode om de temperatuurwaarde te benaderen. Het is bijgevolg duidelijk dat de rekentijd voor de statische methode beduidend lager is. Uit simulaties blijkt dat de statische methode ongeveer 20 tot 30 keer sneller is dan de dynamische methode.

De mate waarin $p_s^i(\Delta H)$ een goede benadering is van $p(\Delta H)$ zal uiteraard afhankelijk zijn van de begintoestand i. Het is daarom van belang dat de begintoestand geen lokaal minimum vormt voor de kostfunctie. Rose *et al.* [Rose90] merken op dat de begintoestand daarom niet tot stand gekomen mag zijn d. m. v. een "gulzig" optimalisatiealgoritme (greedy algorithm). Met deze benaming worden optimalisatiealgoritmen bedoeld die steeds de kostfunctie zover mogelijk zullen doen afnemen en bijgevolg steeds een lokaal minimum als eindtoestand zullen opleveren (zoals bv. de toegevoegde-gradiëntenmethode, paragraaf 3.8). Eventueel kan toch gebruik gemaakt worden van een gulzige optimalisatiemethode, maar dan moet voor deze eerste reconstructie een andere kostfunctie gehanteerd worden. Wanneer de begintoestand geen lokaal minimum vormt voor de kostfunctie stellen Rose *et al.* vast dat voor plaatsingsproblemen in chipontwerp er een vrijwel perfecte overeenkomst bestaat tussen $p(\Delta H)$ en $p_s^i(\Delta H)$.

Er dient een belangrijke opmerking gemaakt te worden die zowel van toepassing is op de statische als op de dynamische methode. Wanneer nl.

$$\left| \int_{-\infty}^{0} \Delta H \, p(\Delta H) \, d(\Delta H) \right| > \left| \int_{0}^{+\infty} \Delta H \, p(\Delta H) \, d(\Delta H) \right|, \tag{5.30}$$

dan zal nooit voldaan kunnen worden aan (5.24). Dit is het geval wanneer, uitgaande van de gegeven begintoestand, meer en/of grotere negatieve dan positieve overgangen gegenereerd worden. Deze toestand kan zich voordoen als het beginbeeld van slechte kwaliteit is (hoge kostfunctiewaarde). Dit betekent dat in de praktijk de evenwichtstemperatuur enkel met succes bepaald kan worden voor beelden die reeds in voldoende mate een benadering zijn van het uiteindelijk te reconstrueren beeld.

We hebben de statische methode voor temperatuurbepaling getest door het reconstructieproces (uitgaande van een willekeurig beginbeeld) na een verschillend aantal iteraties af te breken. Van de aldus bekomen beelden werd de evenwichtstemperatuurwaarde bepaald en vergeleken met de temperatuurwaarde bij stopzetting. We kunnen vaststellen (tabel 5.2) dat er een zeer goede overeenkomst is tussen beide temperatuurwaarden.

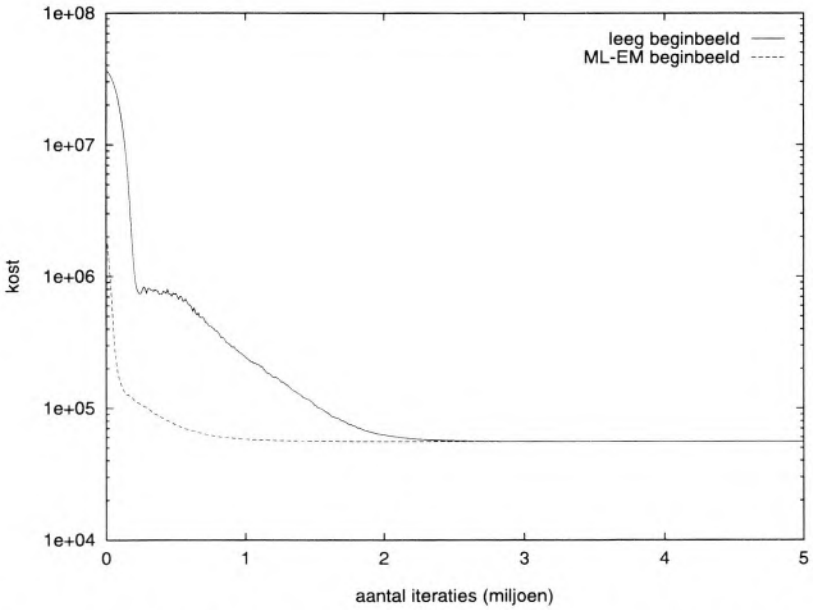
We hebben de methode ook getest door de evenwichtstemperatuur te schatten van beelden die bekomen zijn door ML-EM reconstructie. We kunnen in dit geval echter niet vergelijken met een gekende "correcte" temperatuurwaarde. Daarom

beginbeeld	temperatuurwaarde bij stopzetting	geschatte evenwichts-temperatuurwaarde
6.10^6 tellen, 1 miljoen iteraties	125.3	127.8
6.10^6 tellen, 1.5 miljoen iteraties	25.79	27.12
6.10^6 tellen, 2 miljoen iteraties	4.778	4.474
6.10^5 tellen, 1 miljoen iteraties	151.9	155.4
6.10^5 tellen, 1.5 miljoen iteraties	28.66	28.40
6.10^5 tellen, 2 miljoen iteraties	4.577	4.381

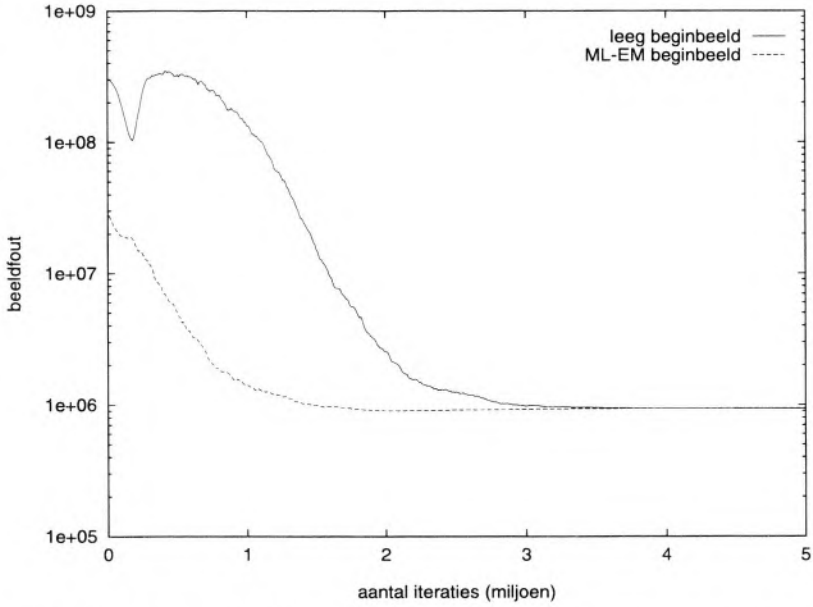
Tabel 5.2: Vergelijking van de geschatte evenwichtstemperatuur met de temperatuur bij stopzetting na een verschillend aantal iteraties (HC-fantoom).

valideren we de methode door de kwaliteit van de gereconstrueerde beelden te vergelijken met die van de beelden bekomen door reconstructie uitgaande van een willekeurig beginbeeld. In figuren 5.24 tot 5.27 zien we dat het gebruik van een ML-EM beginbeeld geen negatieve invloed heeft op de kwaliteit van het gereconstrueerde beeld. Dit bevestigt dat de geschatte beginwaarde voor de temperatuur voldoende hoog is, zodat de reconstructie niet vroegtijdig gevangen wordt in een lokaal minimum. In tabel 5.3 worden het aantal iteraties weergegeven die nodig zijn om 110% van de minimale waarde van de kost en van de beeldfout te bereiken, uitgaande van zowel een blanco als een ML-EM beginbeeld. We bemerken dat de reconstructie in het geval van een ML-EM beginbeeld beduidend sneller verloopt. Ongeveer de helft van het aantal iteraties nodig in het geval van een blanco beginbeeld volstaan om een beeld van gelijkaardige kwaliteit te reconstrueren. Dit bevestigt dat de begintemperatuur voldoende laag geschat wordt, zodat nuttig gebruik gemaakt wordt van de aanwezige informatie in het beginbeeld. We concluderen dat het gebruik van een ML-EM beginbeeld de benodigde rekentijd met ongeveer 50% kan reduceren zonder nadelige invloed op de eindkwaliteit. Hierbij wordt geen rekening gehouden met de rekentijd nodig om het beginbeeld te bekomen. Deze rekentijd is voor de meeste methoden (bv. ML-EM) echter minimaal een grootteorde kleiner dan de reconstructietijd voor het simulated-annealingalgoritme.

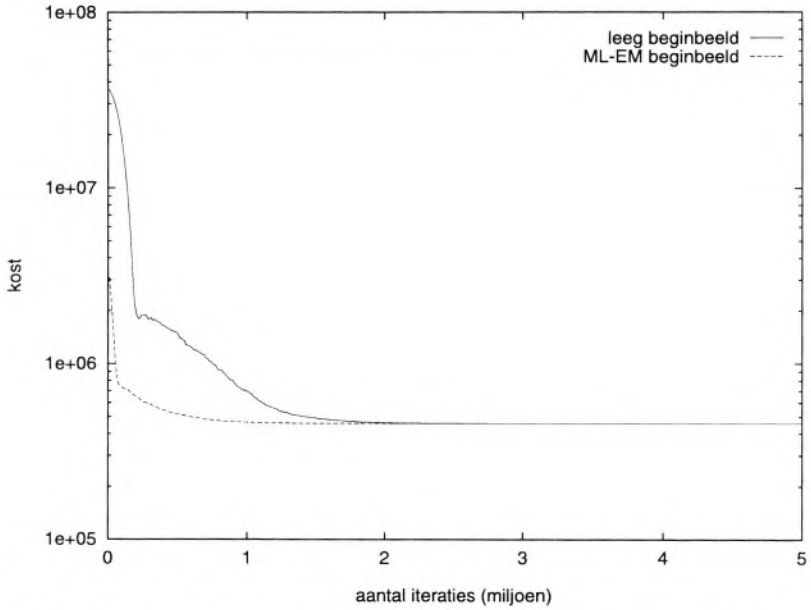
Tenslotte merken we nog op dat deze methode enkel bruikbaar is voor beelden die afkomstig zijn van reconstructiemethoden die vergelijkbaar zijn met het simulated-annealingalgoritme. Met "vergelijkbaar" wordt hier bedoeld dat de gereconstrueerde beelden van vergelijkbare kwaliteit zijn en gelijkaardige eigenschappen vertonen. Dit is o. a. het geval voor ML-EM reconstructie. Deze beel-



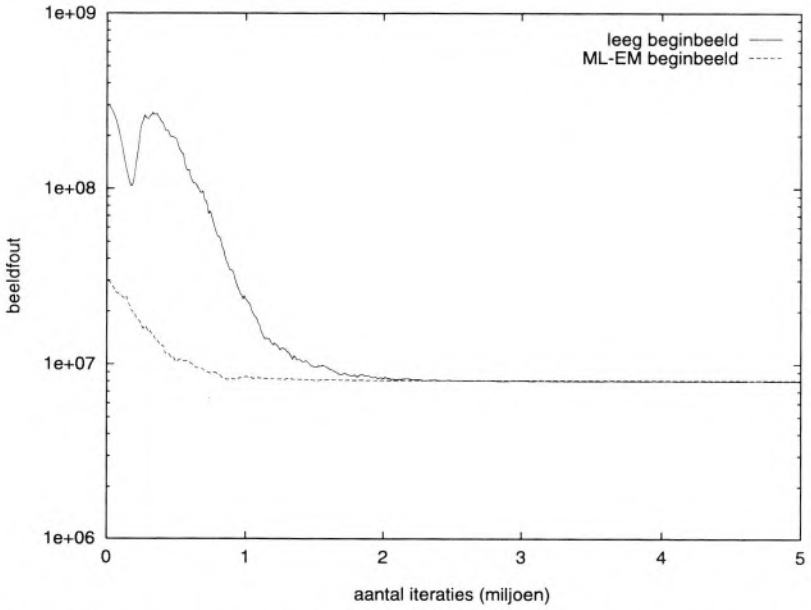
Figuur 5.24: Verloop van de totale kost i. f. v. het aantal iteraties voor reconstructies uitgaande van een leeg beeld en een ML-EM reconstructie als beginbeeld (HC-fantoom, 6.10^6 tellen).



Figuur 5.25: Verloop van de beeldfout i. f. v. het aantal iteraties voor reconstructies uitgaande van een leeg beeld en een ML-EM reconstructie als beginbeeld (HC-fantoom, 6.10^6 tellen).



Figuur 5.26: Verloop van de totale kost i. f. v. het aantal iteraties voor reconstructies uitgaande van een leeg beeld en een ML-EM reconstructie als beginbeeld (HC-fantoom, 6.10^5 tellen).



Figuur 5.27: Verloop van de beeldfout i. f. v. het aantal iteraties voor reconstructies uitgaande van een leeg beeld en een ML-EM reconstructie als beginbeeld (HC-fantoom, 6.10^5 tellen).

beginbeeld	kost	beeldfout
blanco, 6.10^6 tellen	2.1 miljoen	2.4 miljoen
ML-EM, 6.10^6 tellen	800.000	1.4 miljoen
blanco, 6.10^5 tellen	1.4 miljoen	1.7 miljoen
ML-EM, 6.10^5 tellen	600.000	800.000

Tabel 5.3: Maat voor de convergentie van de reconstructie uitgaande van een blanco beginbeeld en een ML-EM beginbeeld: aantal iteraties nodig om 110% van de minimale waarde van de kost en van de beeldfout te bereiken.

den vertonen nl. gelijkenissen met de beelden tijdens een bepaald stadium van de simulated-annealingreconstructie. Het is dan ook mogelijk een temperatuur te vinden waarvoor het ML-EM beginbeeld deel uitmaakt van de evenwichtsdistributie. Beelden gereconstrueerd met de gefilterde-terugprojectietechniek zijn hier niet bruikbaar. Deze beelden vertonen nl. streepvormige artefacten. Meer algemeen is een deel van de intensiteit gelegen "buiten" het object. Tijdens de reconstructie met simulated annealing zien we echter dat reeds in een vroeg stadium de intensiteit vrijwel volledig binnen het gescande object gelegen is. Dit betekent dat tijdens de reconstructie met simulated annealing er geen beelden gevormd worden die een gelijkaardig uitzicht hebben als FB-beelden. Bijgevolg zal aan het FB-beginbeeld een hoge evenwichtstemperatuur geassocieerd worden die correspondeert met een vroeg stadium in de reconstructie. In dit geval is geen noemenswaardige reductie van de reconstructietijd mogelijk.

5.5.2 De lengte van de Markov-ketens

Alvorens het gebruikte criterium voor de lengte van de Markov-ketens te bespreken, vestigen we er de aandacht op dat er een duidelijk verband bestaat tussen de lengte van de Markov-ketens en de methode voor temperatuurverlaging (besproken in de volgende paragraaf). Het is nl. de bedoeling dat de Markov-keten beëindigd wordt wanneer een evenwichtstoestand bereikt is. Wanneer bv. de temperatuurdaling tussen twee opeenvolgende Markov-ketens klein is, zal een klein aantal iteraties volstaan om opnieuw een evenwichtstoestand te bereiken. Anderzijds zullen bij voldoende lange Markov-ketens grotere temperatuursprongen geoorloofd zijn. Afhankelijk van de manier om thermisch evenwicht vast te stellen worden twee klassen van afkoelingsschema's onderscheiden:

- klasse A: variabele lengte van de Markov-ketens en vaste temperatuurdaling;

- klasse B: vaste lengte van de Markov-ketens en variabele temperatuurdaling.

Hierbij hebben de begrippen "variabel" en "vast" betrekking op de respectievelijke afhankelijkheid en onafhankelijkheid van het verloop van de kostfunctiewaarde tijdens het algoritme. Het blijkt eenvoudiger te zijn om het concept van thermisch evenwicht om te zetten in een criterium voor de lengte van de Markovketens dan in een criterium voor de temperatuurdaling [VLaa87]. Daarom wordt door de meeste auteurs een afkoelingsschema van het type A gehanteerd. Ook voor dit onderzoek wordt geopteerd voor een afkoelingsschema van het eerste type.

Het door ons gehanteerde criterium voor de lengte van de Markov-ketens maakt gebruik van het concept "epoch". We definiëren een epoch als een vast aantal iteraties (bv. 100) en de kost van een epoch als de gemiddelde waarde van de kostfunctie gedurende een epoch. We stellen de kost van het i^{de} epoch voor door H_i. Na elk epoch wordt geëvalueerd of de Markov-keten thermisch evenwicht bereikt heeft. Hiervoor beschouwen we de laatste n epoch-kostwaarden, waarop we lineaire regressie toepassen. We bepalen de parameters a en b van de rechte

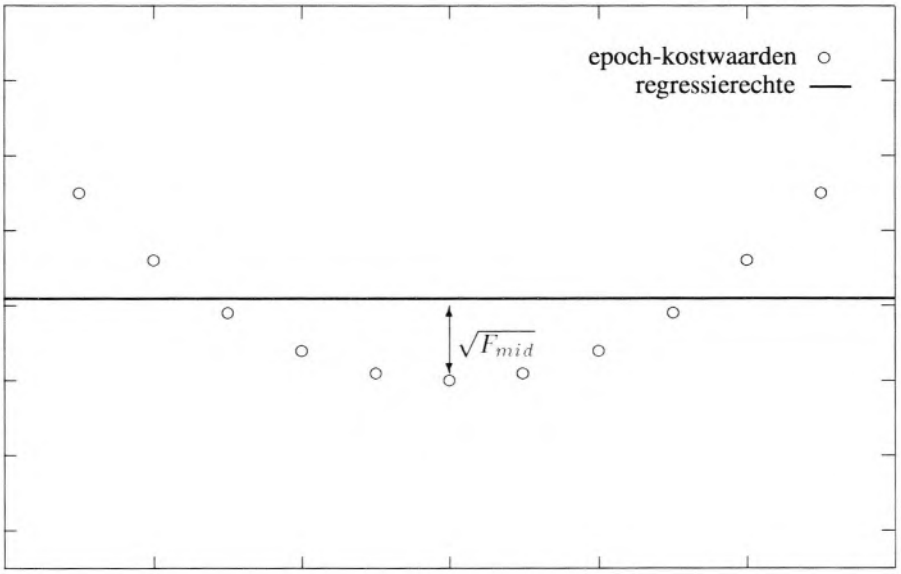
$$H(x) = H(x; a, b) = a + bx \tag{5.31}$$

door de gemiddelde kwadratische fout

$$F = F(a, b) = \frac{1}{n} \sum_{i=1}^{n} (H_i - a - bi)^2 \tag{5.32}$$

te minimaliseren. We evalueren drie criteria alvorens te besluiten dat evenwicht bereikt is. In eerste instantie controleren we of het verloop van de epoch-kost vlak is, t. t. z. of de richtingscoëfficiënt b van de regressierechte nul is. In praktijk betekent dit dat b binnen bepaalde grenzen moet liggen. Daarnaast evalueren we de grootte van de gemiddelde kwadratische fout F die gemaakt wordt door het epoch-kostverloop door een rechte te benaderen. In praktijk zullen we F/H_n evalueren i. p. v. F, aangezien de maximaal toegelaten afwijking daalt naarmate de kost daalt. Tenslotte vergelijken we de gemiddelde kwadratische fout F nog met de kwadratische fout voor het middelste epoch F_{mid}

$$F_{mid} = \left(H_{\frac{n+1}{2}} - a - b\frac{n+1}{2} \right)^2 . \tag{5.33}$$



Figuur 5.28: Illustratie van het mogelijke kostverloop waarbij aan de eerste twee criteria voldaan is en een derde criterium (op basis van de middelste waarde) nodig is opdat thermisch evenwicht verworpen zou worden.

Hiermee trachten we een kostverloop zoals schematisch voorgesteld in figuur 5.28 op te sporen. Een soortgelijk verloop kan zich nl. voordoen net voor (of net na) het optreden van een lokaal minimum. Hoewel in dit geval aan de eerste twee criteria voldaan kan zijn, is het toch wenselijk de Markov-keten voort te zetten. Wanneer aan alle drie de criteria voldaan is, besluiten we dat thermisch evenwicht bereikt is en dat de temperatuur verlaagd kan worden.

De besproken methode heeft als voordeel dat de evaluatie van thermisch evenwicht relatief frequent gebeurt (t. t. z. met een periode van één epoch), terwijl de evaluatie zelf gebeurt over een groter aantal iteraties (n epochen). We zijn ons ervan bewust dat meer gesofisticeerde methoden denkbaar zijn om het thermisch evenwicht te evalueren. Zo zou bv. gebruik gemaakt kunnen worden van een best passende parabool (of meer algemeen een polynoom). Uit simulaties blijkt echter dat de besproken methode goede resultaten oplevert aan de hand van enkele eenvoudig berekenbare parameters. Een bekend probleem van een kleinstekwadratenrechte is de gevoeligheid voor sterk afwijkende waarden (zgn. outliers)

[Rous87]. Dit risico is hier echter gering, aangezien elk meetpunt een gemiddelde is over een aantal iteraties. Uit simulaties blijkt ook dat zich eventueel grote fluctuaties kunnen voordoen, maar vrijwel nooit geïsoleerde afwijkende waarden.

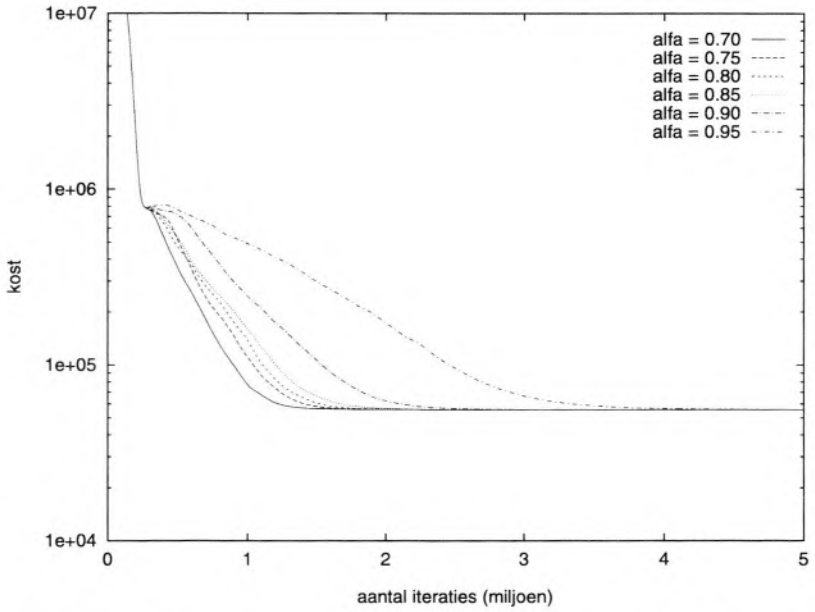
5.5.3 De methode voor temperatuurverlaging

We herhalen dat de controle van thermisch evenwicht geïncorporeerd is in het criterium voor de lengte van de Markov-ketens. Daarom kan een conceptueel eenvoudig schema gebruikt worden voor de temperatuurverlaging. Vrijwel alle auteurs gebruiken voor de temperatuur τ_k gedurende de k^{de} Markov-keten een schema van de vorm (zie ook paragraaf 4.8)

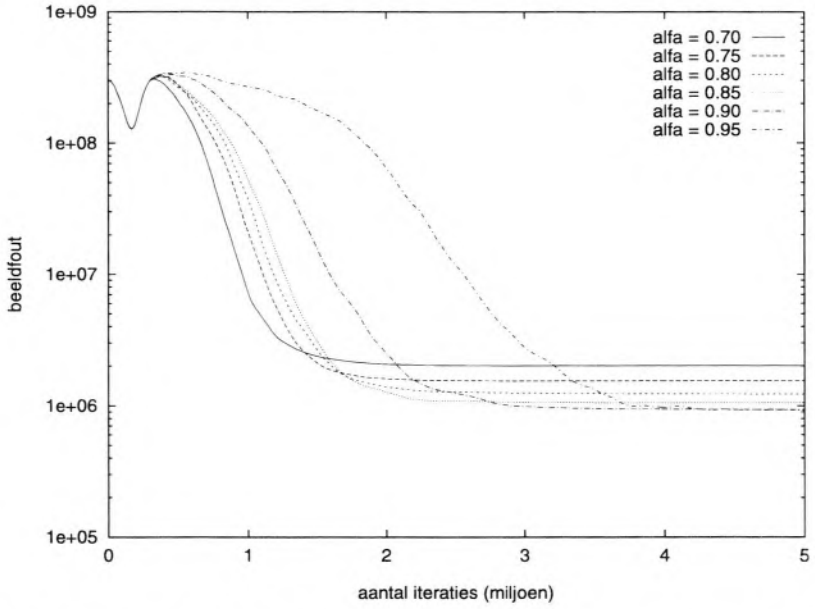
$$\tau_{k+1} = \alpha\, \tau_k. \tag{5.34}$$

Volgens zowel [VLaa87] als [Otte89] moeten de temperatuursprongen voldoende klein zijn, zodat korte Markov-ketens volstaan om opnieuw thermisch evenwicht te bereiken. De temperatuurdalingsfactor α wordt gedurende het verloop van het algoritme constant gehouden. Praktisch bruikbare waarden voor α zijn volgens [VLaa87] gelegen tussen 0.80 en 0.98.

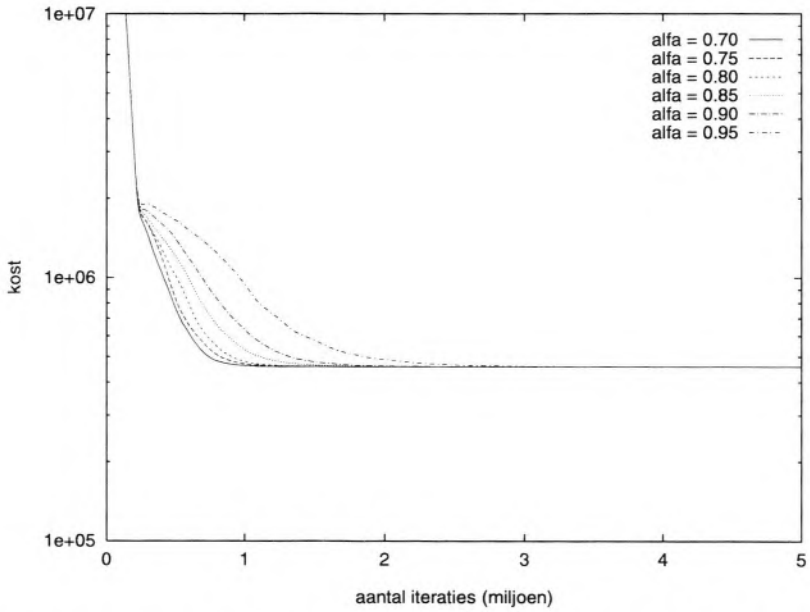
Tijdens simulaties hebben we de invloed onderzocht van de temperatuurverlagingsfactor α op het verloop van de kost en van de beeldfout (figuren 5.29 tot 5.32). Hieruit blijkt dat de optimale waarde voor α ongeveer 0.85 bedraagt. Een lagere waarde zal aanleiding geven tot te snelle afkoeling en bijgevolg tot kwalitatief slechtere beelden; een hogere waarde daarentegen zal aanleiding geven tot een te trage afkoeling (waardoor meer iteraties nodig zijn voor de reconstructie) zonder evenwel de kwaliteit van het eindbeeld te verbeteren. Vergelijken we figuur 5.30 en figuur 5.32 dan bemerken we bovendien dat de optimale α-waarde toeneemt naarmate het aantal tellen toeneemt. Dit kunnen we als volgt verklaren: bij een hoger aantal tellen is de invloed van ruis kleiner en dus zal de kwaliteit van de gereconstrueerde beelden hoger zijn. Kwalitatief betere beelden betekent lagere kostfunctiewaarden en daarom ook lagere minima. We hebben tijdens simulaties kunnen vaststellen dat daardoor de lokale minima ook meer uitgesproken ("dieper") zijn. Hierdoor stijgt het risico om in een lokaal minimum gevangen te geraken. Er moet bijgevolg langzamer afgekoeld worden, wat aanleiding geeft tot hogere waarden voor de temperatuurdalingsfactor. Deze ruisafhankelijkheid van α is echter klein: bij een laag aantal tellen neemt de beeldkwaliteit niet meer toe vanaf $\alpha = 0.80$, terwijl in vrijwel perfecte ruisomstandigheden (hoog aantal tellen) een α-waarde van 0.85 volstaat.



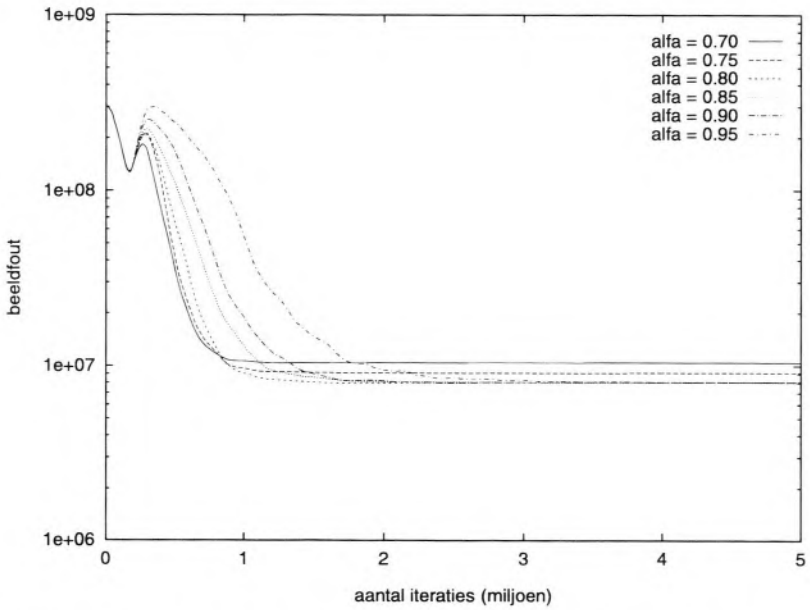
Figuur 5.29: Verloop van de totale kost i. f. v. het aantal iteraties voor verschillende (constante) waarden van de temperatuurdalingsfactor α (HC-fantoom, 6.10^6 tellen).



Figuur 5.30: Verloop van de beeldfout i. f. v. het aantal iteraties voor verschillende (constante) waarden van de temperatuurdalingsfactor α (HC-fantoom, 6.10^6 tellen).



Figuur 5.31: Verloop van de totale kost i. f. v. het aantal iteraties voor verschillende (constante) waarden van de temperatuurdalingsfactor α (HC-fantoom, 6.10^5 tellen).



Figuur 5.32: Verloop van de beeldfout i. f. v. het aantal iteraties voor verschillende (constante) waarden van de temperatuurdalingsfactor α (HC-fantoom, 6.10^5 tellen).

Door een aantal auteurs wordt gewezen op het optreden van cruciale stadia tijdens het simulated-annealingproces. Zo zullen de algemene kenmerken van de eindoplossing reeds bij vrij hoge temperaturen gevormd worden, terwijl de fijnere details bij lagere temperatuurwaarden ontstaan. We verwijzen hierbij naar de analogie met de statistische mechanica uit paragraaf 4.9. Deze cruciale stadia worden faseovergangen genoemd en corresponderen met een hoge waarde voor de soortelijke warmte $c(T)$:

$$c(T) = \frac{d\langle H(T)\rangle}{dT}. \tag{5.35}$$

Een verandering van de soortelijke warmte kan bijgevolg beschouwd worden als een indicatie voor een cruciaal stadium tijdens de optimalisatie, zodat trage afkoeling noodzakelijk is.

In praktijk blijkt de berekening van een soortelijke warmte echter moeilijk te zijn, omdat de temperatuur sprongsgewijze verlaagd wordt en dus (5.35) niet bruikbaar is. Catthoor *et al.* gebruiken een alternatieve methode om cruciale stadia op te sporen aan de hand van wijzigingen in de lengte van de Markov-ketens [Catt88a, Catt88b]. Een toename van de lengte van de Markov-ketens betekent dat meer iteraties nodig zijn alvorens evenwicht bereikt wordt. Dit impliceert meestal dat de kost sterker daalt dan tijdens voorgaande Markov-ketens, hoewel de temperatuurdaling gelijk blijft (cfr. grotere "warmteafvoer"). Dit wijst op zijn beurt op een faseovergang, zodat tragere afkoeling en dus een hogere α-waarde aangewezen zijn. Anderzijds betekent een afname van de lengte van de Markov-ketens dat evenwicht sneller bereikt wordt. Dit doet vermoeden dat de "inwendige toestand" van het systeem weinig verandert bij dalende temperatuur, zodat grotere temperatuurdalingen mogelijk zijn. Dit leidt tot een adaptieve waarde voor α, binnen de grenzen α_{min} en α_{max}. De nieuwe waarde α_{nieuw} wordt bepaald op basis van de oude waarde α_{oud} en de lengte L_{nieuw} van de huidige en L_{oud} van de vorige Markov-keten.

- als $L_{nieuw} < L_{oud}$:

$$\alpha_{nieuw} = \alpha_{oud} - (\alpha_{oud} - \alpha_{min})\frac{L_{oud} - L_{nieuw}}{L_{oud}}; \tag{5.36}$$

- als $L_{nieuw} > L_{oud}$:

$$\alpha_{nieuw} = \alpha_{oud} + (\alpha_{max} - \alpha_{oud})\frac{L_{nieuw} - L_{oud}}{L_{nieuw}}. \tag{5.37}$$

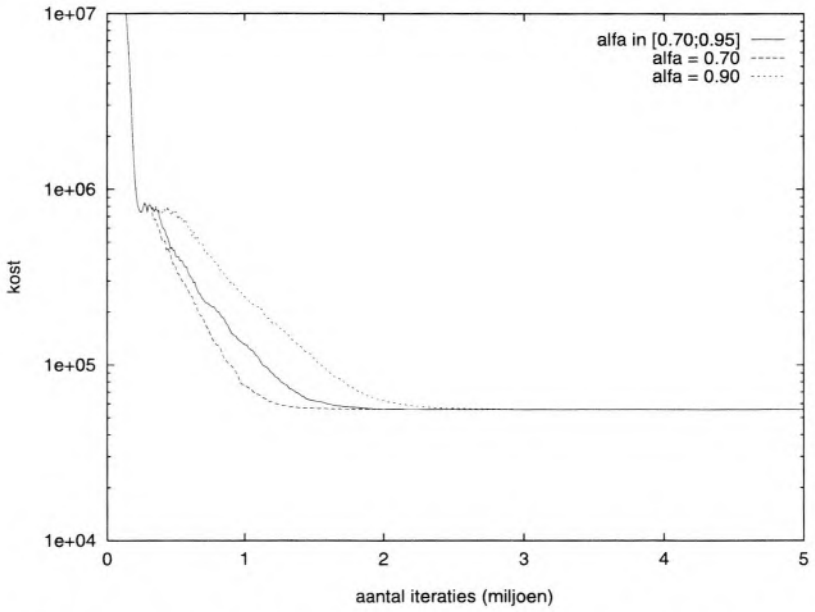
Uit simulaties blijkt dat het gebruik van een adaptief schema de convergentie van het algoritme verbetert. Deze resultaten worden weergegeven in figuren 5.33 en 5.34. Vergelijken we de adaptieve methode met een algoritme met constante temperatuurdaling, dan zien we enerzijds dat de eindkwaliteit van de beelden beter is dan bij een algoritme met gelijkaardige snelheid. Anderzijds is de nieuwe methode beduidend sneller dan een algoritme dat een gelijkaardige beeldkwaliteit oplevert.

5.5.4 Het stopcriterium

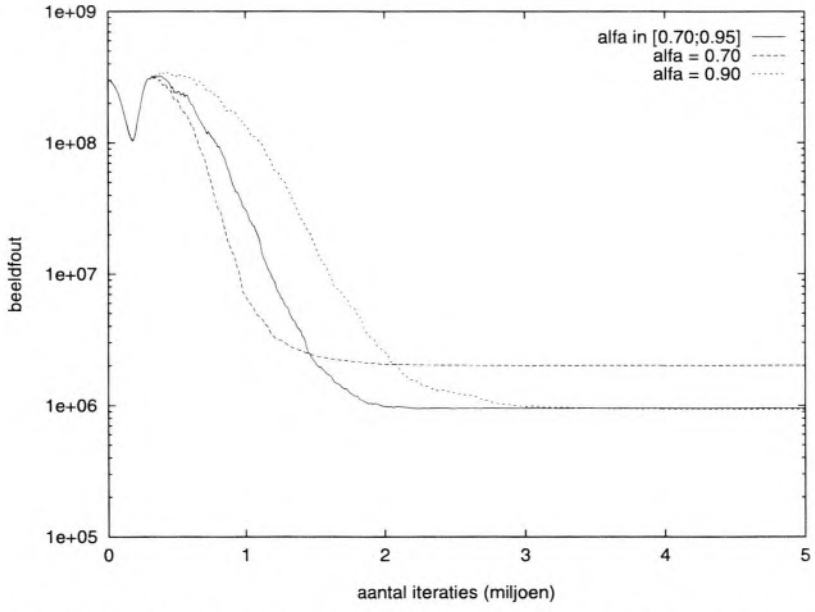
Het uitgevoerde onderzoek heeft hoofdzakelijk tot doel de kwaliteit van de gereconstrueerde beelden bij gebruik van simulated annealing te onderzoeken. De toevoeging van een a priori-term aan de kostfunctie heeft voor gevolg dat de beeldkwaliteit steeds blijft toenemen bij verdere iteraties. We worden dus niet geconfronteerd met het probleem van deterioratie van de beeldkwaliteit bij te ver doorgedreven reconstructie (zie paragraaf 2.6). We lopen enkel het risico dat tijdens de laatste iteratiestappen de verdere toename van de beeldkwaliteit visueel niet waarneembaar is. De formulering van een stopcriterium is bijgevolg van minder cruciaal belang. Bij de simulatieresultaten die voorgesteld worden, werd geen gebruik gemaakt van een adaptief stopcriterium. Er werden steeds een vast aantal iteratiestappen uitgevoerd (meestal 10 miljoen). Dit aantal is zó gekozen dat het beeld onder alle omstandigheden volledig geconvergeerd is.

Voor een toepassing van de simulated-annealingreconstructiemethode wensen we echter een zo efficiënt mogelijke benutting van de rekentijd. Daarom is het noodzakelijk om een bruikbaar stopcriterium te ontwikkelen. We trachten de toestand te vinden waarna de kostfunctie niet significant meer afneemt. We kunnen deze toestand bv. beschrijven door te stellen dat de evenwichtsdistributie van opeenvolgende Markov-ketens niet verandert. Dit impliceert dat de evenwichtstemperatuur constant blijft bij verder itereren. We zouden daarom gebruik kunnen maken van een gelijkaardig criterium als de statische methode uit paragraaf 5.5.1, hetgeen echter zou betekenen dat de hieraan verbonden berekeningen steeds herhaald moeten worden. Dit zou aanleiding geven tot een onaanvaardbare toename van de reconstructietijd.

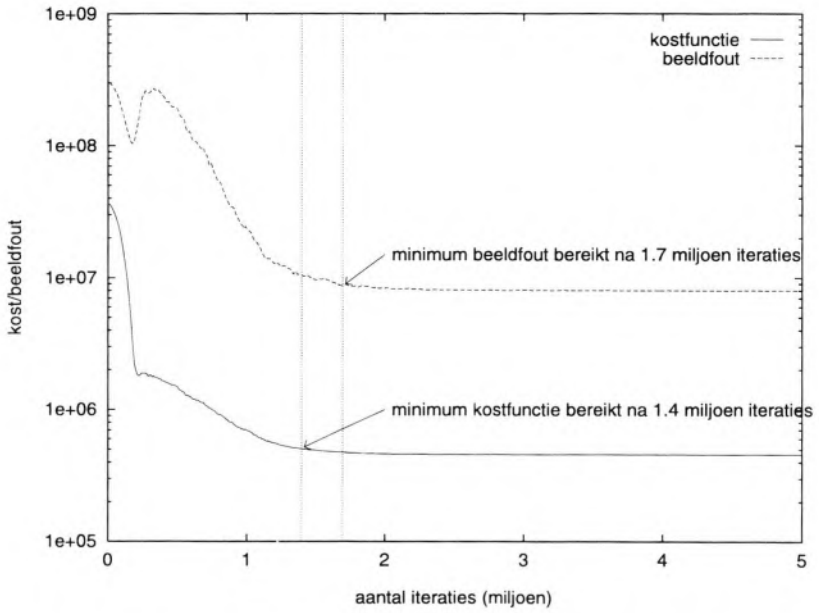
Een tweede en minder rekenintensieve mogelijkheid is gebruik te maken van een gelijkaardig criterium als voor de bepaling van de lengte van de Markovketens, waarbij we de epoch-kostwaarden vervangen door de kostwaarden op het einde van de Markov-ketens. We merken echter bij alle simulaties op dat de kost-



Figuur 5.33: Verloop van de totale kost i. f. v. het aantal iteraties voor een adaptieve temperatuurdalingsfactor α (HC-fantoom, 6.10^6 tellen).



Figuur 5.34: Verloop van de beeldfout i. f. v. het aantal iteraties voor een adaptieve temperatuurdalingsfactor α (HC-fantoom, 6.10^6 tellen).



Figuur 5.35: Verschil in convergentie tussen het verloop van de kostfunctie en van de beeldfout (HC-fantoom, 6.10^5 tellen).

functie sneller naar haar minimale waarde convergeert dan de beeldfout. In figuur 5.35 is dit duidelijk zichtbaar. Als maat voor de convergentie gebruiken we het punt waarop 110% van de minimale waarde bereikt wordt. Dit betekent dat de beeldkwaliteit nog verder toeneemt hoewel de kost reeds vrijwel constant blijft. We moeten bijgevolg een criterium gebruiken dat strengere eisen oplegt wat betreft de richtingscoëfficiënt en de benaderingsfout bij lineaire regressie van de kostwaarden.

Als evaluatie van dit stopcriterium voeren we de simulaties van paragraaf 5.5.1 (tabel 5.3) opnieuw uit, ditmaal met toepassing van een stopcriterium. We zien in tabel 5.4 dat dit criterium het gewenste aantal iteraties (t. t. z. het aantal iteraties waarop ook de beeldfout geconvergeerd is) goed benadert. We stellen bovendien vast dat het aantal iteraties bij stopzetting steeds groter is dan de maat voor convergentie van de beeldfout. Hieruit volgt dat de afwijkingen van de beeldkwaliteit t. g. v. het stopcriterium steeds kleiner dan 10% zullen zijn. Deze afwijkingen zijn in praktijk visueel niet of nauwelijks zichtbaar.

beginbeeld	convergentie kost	convergentie beeldfout	stopcriterium
blanco, 6.10^6 tellen	2.1 miljoen	2.4 miljoen	2.6 miljoen
ML-EM, 6.10^6 tellen	800.000	1.4 miljoen	1.7 miljoen
blanco, 6.10^5 tellen	1.4 miljoen	1.7 miljoen	1.9 miljoen
ML-EM, 6.10^5 tellen	600.000	800.000	950.000

Tabel 5.4: Evaluatie van het stopcriterium: vergelijking van het aantal iteraties bij stopzetting met het aantal iteraties voor convergentie van de kostfunctie en van de beeldfout.

5.6 Conclusie

In dit hoofdstuk hebben we de diverse aspecten van het generatiemechanisme en het afkoelingsschema in detail bestudeerd. We herhalen de belangrijkste opmerkingen en resultaten. Deze resultaten zijn het gevolg van simulaties m. b. v. fantoombeelden. De kwaliteit van de gereconstrueerde beelden wordt geëvalueerd m. b. v. een gemiddelde kwadratische afwijking. Wat de modellering van het detectieproces betreft stellen we vast dat het model van vierkanten met variabele breedte een goede benadering vormt voor het geometrisch exacte model.

De bespreking van het generatiemechanisme kan onderverdeeld worden in de pixelkeuze, de aanpassingsmethode en de korrelgrootte. Bij gesimuleerde data kan de efficiëntie van de pixelkeuze sterk verhoogd worden door gebruik te maken van voorafgaande contourdetectie a. h. v. nulelementen in de sinogrammatrix. Wat de aanpassingsmethode betreft merken we op dat in alle praktisch bruikbare gevallen een aantal lokale minima geïntrodiceerd worden. We constateren dat het vrijwel onmogelijk is deze lokale minima te vermijden door de aanpassingsmethode te optimaliseren. Het gebruik van een gepaste a priori-term in de kostfunctie is meer doeltreffend. Daarom wordt gekozen voor de (eenvoudige) GAM als aanpassingsmethode. Voor de korrelgrootte hebben we een criterium opgesteld voor de automatische bepaling van de intervalgrenzen van een uniforme verdeling. Dit resulteert in een hogere efficiëntie per iteratiestap.

Wat het afkoelingsschema betreft worden achtereenvolgens de begintemperatuur, de lengte van de Markov-ketens, de methode voor temperatuurverlaging en het stopcriterium besproken. Op theoretische gronden worden twee criteria opgesteld voor de bepaling van de begintemperatuur, nl. één voor een willekeurig en één voor een zinvol beginbeeld. De nieuwe methode voor een willekeurige begin-

beeld resulteert in temperaturen die beter voldoen aan de vooropgestelde eis (nl. 80% aanvaarding van positieve overgangen) dan de bestaande methoden. We benadrukken dat vooral de bepaling van de temperatuur voor een zinvol beginbeeld tot nog toe vrijwel niet bestudeerd werd. Deze nieuwe methode laat toe het reconstructieproces aanzienlijk te versnellen door gebruik te maken van andere (snellere) reconstructietechnieken als initiële benadering. Verder wordt de lengte van de Markov-ketens bepaald m. b. v. een criterium dat gebaseerd is op het concept van thermisch evenwicht. De temperatuur wordt tussen opeenvolgende Markov-ketens verlaagd m. b. v. een adaptief criterium dat toelaat voldoende traag af te koelen tijdens cruciale fasen van de reconstructie. We hebben ook de invloed van de temperatuurdalingsafctor op de kwaliteit van de gereconstrueerde beelden onderzocht. Als stopcriterium wordt een aangepaste versie van het criterium voor de lengte van de Markov-ketens gebruikt. We vestigen er tenslotte de aandacht op dat deze resultaten bekomen werden tijdens de ontwikkeling van een reconstructiealgoritme voor PET-beelden, maar vrijwel allemaal toepasbaar zijn voor een algemeen probleem van combinatorische optimalisatie.

Hoofdstuk 6

Analyse van de kostfunctie

6.1 Inleiding

In dit hoofdstuk analyseren we de kostfunctie en bestuderen we de invloed van de diverse parameters op de kwaliteit van de gereconstrueerde beelden. We merken op dat de MAP-oplossing voor het reconstructieprobleem het beeld X is waarvoor de a posteriori-waarschijnlijkheid $p(X|Y)$ maximaal is; we herhalen daarom een aantal notaties en begrippen uit paragraaf 3.7 i. v. m. de a posteriori-waarschijnlijkheidsdistributie. Maximalisatie van $p(X|Y)$ impliceert minimalisatie van de a posteriori-Gibbs-energie $H(X,Y)$. Deze Gibbs-energie vervult bijgevolg de rol van kostfunctie tijdens de reconstructie. Volgens (3.27) kunnen we de Gibbs-energie schrijven als

$$H(X,Y) = H_D(X,Y) + \beta\, H_P(X). \tag{6.1}$$

De kostfunctie bestaat dus uit twee termen. Enerzijds een term $H_D(X,Y)$ die we de dataterm noemen en die overeenstemt met $-\ln p(Y|X)$, en anderzijds een a priori-term $H_P(X)$ die overeenstemt met $-\ln p(X)$. De regularisatieparameter β bepaalt het relatieve aandeel van de beide termen tot de totale kostfunctie. We zullen in wat volgt de beide termen meer in detail bespreken. De regularisatieparameter zal gelijktijdig met de a priori-term behandeld worden.

6.2 Bespreking van de dataterm

De dataterm $H_D(X,Y)$ komt overeen met het aandeel van de directe waarschijnlijkheidsdistributie $P(Y|X)$ tot de a posteriori-waarschijnlijkheidsdistributie

$P(X|Y)$. Deze term wordt volledig bepaald door de fysische karakteristieken van het beeldvormingssysteem. Algemeen kunnen we stellen dat het verband tussen het beeld X en de metingen Y gegeven wordt door

$$Y = \Psi X + N, \tag{6.2}$$

waarbij Ψ staat voor de transfertmatrix van het beeldvormingssysteem en N voor de ruis. De manier waarop we de ruis N modelleren bepaalt de vorm van $p(Y|X)$ en dus ook van $H_D(X, Y)$. Vandaar ook dat vaak de benaming "ruisterm" gebruikt wordt.

6.2.1 Gaussiaanse ruis met constante standaardafwijking

In eerste instantie onderstellen we dat de ruis tijdens het detectieproces gemodelleerd kan worden als Gaussiaanse ruis met gemiddelde 0:

$$p(Y|X) = \frac{1}{Z_N} \exp\left(-\frac{\|Y - \Psi X\|^2}{2\sigma^2}\right). \tag{6.3}$$

$$-\ln p(Y|X) = \ln Z_N + \frac{\|Y - \Psi X\|^2}{2\sigma^2}. \tag{6.4}$$

Hierin is Z_N de normalisatiefactor en σ de voor alle pixels constant onderstelde standaardafwijking van de Gaussiaanse ruis. We zien dat de maximalisatie van $p(Y|X)$ overeenkomt met de minimalisatie van $\|Y - \Psi X\|^2$. We definiëren daarom

$$\begin{aligned} H_D(X, Y) &= \|Y - \Psi X\|^2 \\ &= \sum_{i=1}^{M} [Y_i - (\Psi X)_i]^2. \end{aligned} \tag{6.5}$$

We zien dat de onderstelling van Gaussiaanse ruis aanleiding geeft tot een kleinste-kwadratenkostfunctie tussen de metingen Y en de pseudodata ΨX die corresponderen met het beeld X.

Deze keuze voor $H_D(X, Y)$, die we de kleinste-kwadratendataterm noemen, wordt in praktijk vaak gebruikt [Webb89, Kear90]. Dit is voornamelijk het geval vanwege de computationele eenvoud. De onderstelling van een Gaussiaanse verdeling met constante standaardafwijking voor de sinogramelementen is echter intuïtief onaanvaardbaar, vooral bij een laag aantal tellen. We voeren deze dataterm dan ook enkel in als tussenstap naar paragraaf 6.2.3.

6.2.2 Poisson-ruis

Wanneer $E[N] = 0$ volgt uit (6.2) dat

$$E[Y] = \Psi X. \tag{6.6}$$

Maken we voor N de veronderstelling van Poisson-ruis dan vinden we dat

$$\begin{aligned} p(Y|X) &= \prod_{i=1}^{M} p(Y_i|X) \\ &= \prod_{i=1}^{M} \exp(-(\Psi X)_i)\ \frac{(\Psi X)_i^{Y_i}}{Y_i!}. \end{aligned} \tag{6.7}$$

Hierbij hebben we gebruik gemaakt van de eigenschap van statistische onafhankelijkheid van metingen in PET. Voor een bewijs hiervan refereren we naar [Desm95]. Overgaand naar de logaritmische vorm vinden we dat

$$\begin{aligned} \ln p(Y|X) &= \sum_{i=1}^{M} \ln p(Y_i|X) \\ &= \sum_{i=1}^{M} [-(\Psi X)_i + Y_i \ln(\Psi X)_i - \ln Y_i!]. \end{aligned} \tag{6.8}$$

Deze laatste term $\ln Y_i!$ is onafhankelijk van het beeld X en mag dus buiten beschouwing gelaten worden voor de maximalisatie van $p(Y|X)$. Definiëren we

$$H_D(X, Y) = \sum_{i=1}^{M} [(\Psi X)_i - Y_i \ln(\Psi X)_i], \tag{6.9}$$

dan vinden we opnieuw dat maximalisatie van $p(Y|X)$ overeenkomt met de minimalisatie van $H_D(X, Y)$.

We merken hierbij op dat het model van Poisson-ruis algemeen aanvaard is voor het detectieproces bij PET (hoofdstuk 2). Deze keuze voor de dataterm zal dan ook het beste overeenstemmen met de realiteit. Het optreden van een logaritme in de uitdrukking voor $H_D(X, Y)$ betekent echter ook dat deze kostfunctie aanzienlijk meer rekentijd zal vergen dan de bovenvermelde kleinstekwadratenkostfunctie. Aangezien deze vorm voor $H_D(X, Y)$ overeenstemt met klassieke maximum-likelihoodmethodes voor PET (zoals bv. ML-EM reconstructie), zullen we deze dataterm de ML-dataterm noemen.

We merken ook op dat $H_D(X, Y)$ minimaal wordt als $\Psi X = Y$. De minimale waarde bedraagt dan

$$H_{N,min} = \sum_{i=1}^{M} [Y_i(1 - \ln Y_i)]. \tag{6.10}$$

Het is duidelijk dat deze minimale waarde in de praktijk meestal een negatieve waarde zal aannemen. De waarden Y_i, die het aantal tellen van een detectorpaar voorstellen, zullen namelijk in de meeste omstandigheden groter zijn dan 3 (zodat $\ln Y_i > 1$). Omdat we er de voorkeur aan geven te beschikken over een kostfunctie die enkel positieve waarden aanneemt en waarvan de minimale waarde 0 bedraagt, zullen we tijdens het reconstructieproces de volgende uitdrukking gebruiken:

$$\begin{aligned} H_D(X, Y) &= \sum_{i=1}^{M} [(\Psi X)_i - Y_i \ln(\Psi X)_i] + H_0 \\ H_0 &= \sum_{i=1}^{M} [Y_i(\ln Y_i - 1)]. \end{aligned} \tag{6.11}$$

Verder vestigen we er ook de aandacht op dat de term $Y_i \ln(\Psi X)_i$ eventueel voor problemen kan zorgen. Dit is bijvoorbeeld het geval wanneer een element uit de pseudo-datamatrix $(\Psi X)_i$ nul wordt zonder dat het overeenkomstig gemeten matrixelement Y_i nul is. We zullen er tijdens de reconstructie dan ook zorg voor moeten dragen dat deze toestand zich niet kan voordoen.

Omdat we tijdens de reconstructie nooit de volledige kostfunctie zullen willen berekenen, maar enkel veranderingen van de kostfunctiewaarde ten gevolge van de aanpassing van een of enkele beeldpixels, ontwikkelen we hiervoor een uitdrukking.

$$\begin{aligned} \Delta H_D(X, Y) &= H_D(X + \Delta X, Y) - H_D(X, Y) \\ &= \sum_{i=1}^{M} [(\Psi X)_i + \Delta(\Psi X)_i - Y_i \ln((\Psi X)_i + \Delta(\Psi X)_i)] \\ &\quad - \sum_{i=1}^{M} [(\Psi X)_i - Y_i \ln(\Psi X)_i] \\ &= \sum_{i=1}^{M} [\Delta(\Psi X)_i - Y_i \ln(1 + \frac{\Delta(\Psi X)_i}{(\Psi X)_i})]. \end{aligned} \tag{6.12}$$

6.2.3 Gaussiaanse ruis met Poisson-standaardafwijking

We hernemen de veronderstelling van Gaussiaanse ruis uit paragraaf 6.2.1, maar zullen trachten een meer zinvolle hypothese te maken voor de standaardafwijking σ. Gebruik makend van het feit dat Poisson-ruis fysisch de meest correcte onderstelling is, leiden we een schatting voor de standaardafwijking af. Voor de variantie van een Poisson-verdeelde toevalsgrootheid z geldt algemeen dat

$$\sigma^2[z] \equiv E[z^2] - (E[z])^2 = E[z]. \tag{6.13}$$

Om hiervan gebruik te kunnen maken in een Gaussiaanse benadering, hebben we de verwachtingswaarden $E[Y_i]$ nodig. Aangezien deze waarden niet gekend zijn, benaderen we ze door de waarden van Y_i zelf [Tsui91]. Substitueren we σ^2 hierdoor in (6.4), dan vinden we dat

$$\begin{aligned} -\ln p(Y|X) &= -\sum_{i=1}^{M} \ln p(Y_i|X) \\ &= \ln Z_N + \sum_{i=1}^{M} \frac{[Y_i - (\Psi X)_i]^2}{2Y_i}. \end{aligned} \tag{6.14}$$

We definiëren daarom nu

$$H_D(X,Y) = \sum_{i=1}^{M} \frac{[Y_i - (\Psi X)_i]^2}{Y_i}. \tag{6.15}$$

Deze dataterm leidt echter tot problemen indien Y_i nul wordt. Daarom voegen we in de noemer van (6.15) een additieve constante η toe, zodat

$$H_D(X,Y) = \sum_{i=1}^{M} \frac{[Y_i - (\Psi X)_i]^2}{Y_i + \eta}. \tag{6.16}$$

Kleine waarden voor η zullen het relatief belang van nulelementen in het gemeten sinogram sterk benadrukken, zodat de overeenstemmende elementen uit het pseudosinogram ook snel naar nul zullen convergeren [Kauf93, Fess94]. Wij zullen in wat volgt voor η steeds de waarde 0.5 kiezen (wat betekent dat $\eta \ll Y_i$).

We zien dat de onderstelling van Gaussiaanse ruis met een van de Poissondistributie afgeleide benadering voor de standaardafwijking aanleiding geeft tot een gewogen kleinste-kwadratendataterm (GKK-dataterm). Deze vorm bezit nog steeds de computationele eenvoud van een kleinste-kwadratenterm, maar geeft een