Chapitre 2

Optimisation de la répartition de dose - bases théoriques

Le processus itératif de planification utilisé en radio-oncologie ou en radiochirurgie a pour but de déterminer le meilleur traitement à appliquer au patient. Il est représenté schématiquement sur la figure 2.1. Alors que l'ordinateur aide le dosimétriste dans deux étapes de ce processus (le calcul de dose et l'évaluation des résultats), la modification des paramètres de traitement dépend quant à elle de son expérience. Des algorithmes d'optimisation peuvent toutefois effectuer cette étape d'une manière automatique et plus objective. Ils sont encore très rarement utilisés. Dans ce chapitre, les étapes du processus (calcul, évaluation et optimisation) sont décrites.

2.1 Calcul de la répartition de dose

Dans la plupart des cas cliniques, une mesure directe de la dose délivrée au volume cible et aux tissus sains n'est pas envisageable. C'est pourquoi le radio-oncologue doit se fier au calcul prévisionnel de la répartition de dose pour évaluer la qualité d'un traitement. Des études cliniques ont montré qu'un écart de $\pm 5\%$ dans la dose délivrée peut résulter en une différence significative pour le contrôle local sans complication (voir ICRU-Report [1976, 1999], Dische et al. [1993]).

Pour atteindre la précision de 5% dans la dose délivrée, il est nécessaire de garantir une grande précision dans toutes les étapes du traitement. Ces dernières sont : la mesure des caractéristiques de base des faisceaux, le calcul du plan de traitement à partir des données spécifiques au patient, la calibration de la machine, le positionnement du patient et l'application du traitement. Dans ces

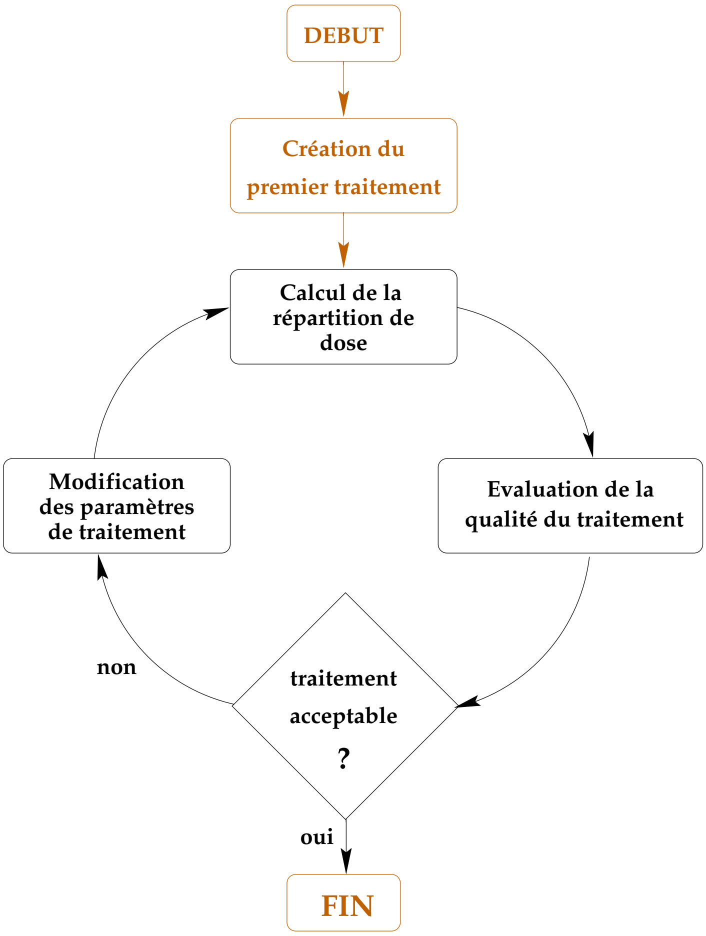


FIG. 2.1 – Processus itératif de planification d'un traitement.

conditions, la précision dans le calcul de la répartition de dose doit alors être de 2 à 3%. Cette exigence nécessite des algorithmes relativement complexes et n'est pas respectée par tous les systèmes disponibles.

2.1.1 Calcul de la dose produite par le rayonnement primaire

Lors de l'irradiation avec un faisceau externe de rayons X, l'énergie transmise par le rayonnement X aux particules chargées du milieu par unité de masse, à une profondeur z ($Kerma_{prim}(z)$), se calcule en intégrant sur tout le spectre d'énergie le produit de la fluence à la profondeur z par le coefficient de transfert d'énergie :

$$Kerma_{prim}(z) = \int_0^{E_{max}} \Phi_{prim}(E, z) \left(\frac{\mu_k(E)}{\rho} \right) dE, \tag{2.1}$$

où $\Phi_{prim}(E,z)$ est le spectre de fluence énergétique à la profondeur z,
E_{max} est le maximum d'énergie du rayonnement de freinage et
$\frac{\mu_k(E)}{\rho}$ est le coefficient de transfert d'énergie pour le matériau irradié et l'énergie E.

La fluence énergétique ($\Phi_{prim}(E,z)$) à la profondeur z se calcule à l'aide de la fluence énergétique en surface ($\Phi_{prim}(E,0)$), corrigée par l'atténuation le long de l'axe z :

$$\Phi_{prim}(E,z)dE = \Phi_{prim}(E,0)\exp\left[-\left(\int_0^z \frac{\mu_a(E)}{\rho}\rho(x)dx\right)\right]dE, \qquad (2.2)$$

où $\Phi_{prim}(E,0)dE$ est la fluence énergétique des photons à l'énergie E dans l'air,
$\frac{\mu_a(E)}{\rho}$ est le coefficient d'atténuation à l'énergie E et
$\rho(x)dx$ est le chemin radiologique infinitésimal traversé par le rayon dans le milieu.
L'intégration de $\frac{\mu_a(E)}{\rho}\rho(x)dx$ permet de tenir compte des hétérogénéités présentes dans le milieu.

Même dans les régions où l'équilibre électronique est atteint, la dose absorbée à la profondeur z ($D_{prim}(z)$) produite par les photons primaires n'est pas identique au kerma car la production de rayonnement de freinage par les particules chargées n'est pas comptabilisée dans la dose absorbée. Ceci s'exprime par le paramètre g dépendant de l'énergie :

$$D_{prim}(z) = Kerma_{prim}(z)\cdot(1-g). \qquad (2.3)$$

Le paramètre g vaut $3,6\cdot 10^{-3}$ pour un faisceau de 1 MeV dans l'eau (pour d'autres valeurs de g voir ICRU-Report [1984]).

2.1.2 Calcul de la dose produite par le rayonnement diffusé

La dose absorbée associée aux photons Compton diffusés peut être séparée en deux composantes : la première diffusion et les diffusions multiples. La dose engendrée par la première diffusion Compton peut être calculée par une méthode directe. La fluence énergétique produite par les premières diffusions Compton d'un élément de volume dV vers le point de mesure (direction Ω) se calcule à l'aide de la probabilité de Klein-Nishina :

$$\begin{aligned}\Phi_{diff}(E,\theta,z)dEd\Omega &= \Phi_{prim}(E_0,z)\cdot(\frac{r_0^2}{2})\cdot(1+cos^2\theta)\cdot \\ &\quad F_{KN}\cdot e^{-\left(\frac{\mu_a(E_{diff})}{\rho}\int^r\rho(x')dx'\right)}\,dE\,d\Omega \qquad (2.4)\end{aligned}$$

où $r_0 = \frac{e^2}{mc^2}$ est le rayon classique de l'électron,
F_{KN} est la probabilité de Klein-Nishina qu'un photon d'énergie E soit diffusé à un angle θ par unité d'angle solide $d\Omega = 2\pi sin\theta d\theta$,
$E_{diff} = \frac{E_0}{1+\alpha(1-cos\theta)}$ est l'énergie du photon Compton diffusé,
E_0 est l'énergie initiale du photon,
$\alpha = \frac{E_0}{m_0 c^2}$,
$\frac{\mu_a(E_{diff})}{\rho}$ est le coefficient d'atténuation pour le photon Compton diffusé à l'énergie E_{diff}.

L'énergie transférée aux électrons de recul par les premières diffusions Compton $Kerma_{diff1}(p)$ se calcule en intégrant sur tout le spectre et tout le volume V la fluence énergétique :

$$Kerma_{diff1}(p) = \int^{Volume} \int_{E_{min}}^{E_{max}} \Phi_{diff}(E,z) \cdot \left(\frac{\mu_k(E_{diff})}{\rho}\right) dEdV \tag{2.5}$$

Ces équations relativement complexes, nécessitant diverses intégrations, ne tiennent pas compte des diffusions multiples, et il n'y a pas de solution analytique à ce problème. L'écart entre la dose mesurée et calculée en se limitant à la première diffusion est de l'ordre de 10% pour un faisceau de 1 MeV avec un champ carré de 10 cm de côté à 15 cm de profondeur. La précision ainsi obtenue est insuffisante. D'autre part, la présence des hétérogénéités doit également être prise en compte. Ainsi différentes techniques d'approximation ont été développées pour atteindre la précision requise (Bloch and Wallace [1979], Elsberry [1979], Siddon et al. [1985], Bateman and Procter [1988], Murray et al. [1989], Redpath [1995], Storchi and Woudstra [1996]). Notons parmi celles-ci la dernière en date : les cônes de convolution de Ahnesjö [1989], Ahnesjo et al. [1995], Weber et al. [1996] qui s'appuient sur l'utilisation de noyaux de diffusion préalablement calculés par une méthode de simulation de Monte-Carlo. Outre ces méthodes analytiques, la simulation directe par Monte-Carlo, rendue possible par la puissance des ordinateurs actuels, a fait son apparition ces dernières années (Weaver et al. [1996], Bohm et al. [1999][1]).

2.1.3 Calcul de la dose en radiochirurgie

Technique par thérapie en pendulaire

En radiochirurgie par thérapie pendulaire, l'utilisation de faisceaux étroits produisant peu de rayonnement diffusé et irradiant un volume relativement homogène, rend possible la simplification de l'algorithme de calcul de la répartition

[1]Projet PEREGRINE du *Lawrence Livermore National Laboratory* : http ://www-phys.llnl.gov/peregrine

de dose, ce qui réduit grandement le temps de calcul et facilite l'application ultérieure d'algorithmes d'optimisation. La dose produite par un faisceau statique à une profondeur z à une distance radiale r de l'axe du faisceau pour un collimateur circulaire d'ouverture S_c à une distance source-peau DSP se calcule, dans le cas de cette approximation, par le produit de cinq facteurs :

$$\begin{aligned} Dose\left(DSP, S_c, z, r\right) &= D_{ref} \cdot FOC\left(S_c\right) \cdot \left(\frac{DSP_{ref} + z_{ref}}{DSP + z_{ref}}\right)^2 \\ &\cdot RP\left(DSP, S_c, z\right) \cdot Profil\left(DSP, S_c, z, r\right), \end{aligned} \quad (2.6)$$

où D_{ref} est la dose de référence mesurée pour un champ carré 10 cm $\times$ 10 cm
$FOC(S_c)$ est le facteur d'ouverture de champ en fonction du collimateur S_c, cette courbe est normée de sorte qu'une unité moniteur (UM) demandée au pupitre de l'accélérateur linéaire, délivre 1 cGy à la profondeur de référence pour un champ carré de 10 cm de côté,
DSP_{ref} est la distance source-peau de référence, en principe 100 cm,
le terme quadratique permet de corriger la dose au 100% du rendement en fonction de la distance source-peau (DSP),
$RP(DSP, S_c, z)$ est le rendement en profondeur de la dose (voir fig. 2.2) et
$Profil(DSP, S_c, z, r)$ est le profil du faisceau (voir fig. 2.2).

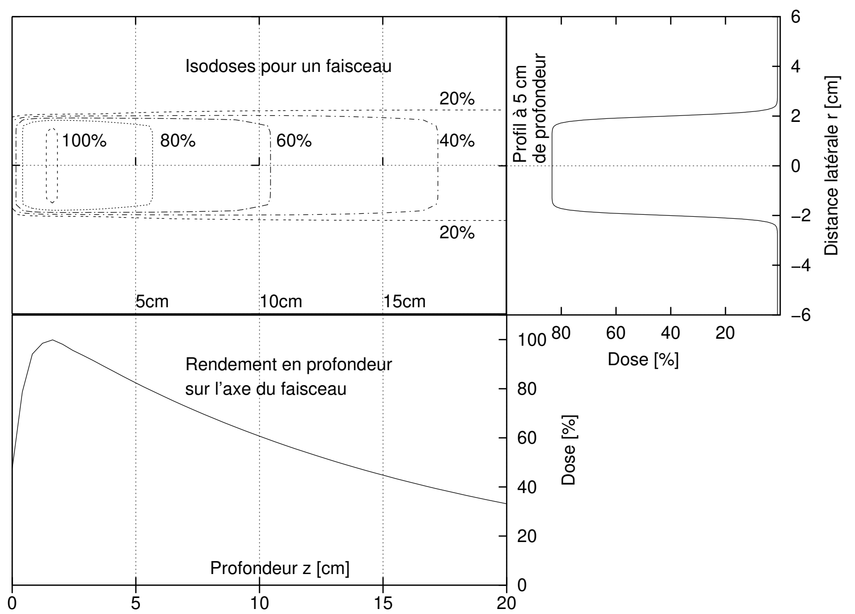


FIG. 2.2 – Isodoses produites par un faisceau. Dans ce cas, la DSP est de 100 cm, le collimateur cylindrique est le plus grand disponible (i.e. 3,8 cm à la DSP). Le profil et le rendement en profondeur correspondent aux mêmes conditions.

Les facteurs $FOC(S_c)$, $RP(DSP, S_c, z)$ et $Profil(DSP, S_c, z, r)$ sont mesurés (voir section 3.3) dans une cuve d'eau pour chaque collimateur S_c selon un réseau de points permettant des interpolations précises pour chaque distance source-peau (DSP), profondeur z ou distance radiale r. L'obliquité et la courbure de la surface d'entrée ne sont pas prises en compte. De même, la boîte crânienne est considérée comme entièrement équivalente à de l'eau. Ces approximations se justifient car l'erreur apportée est négligeable (Pike et al. [1987], Bardash et al. [1992], Lam and Zhu [1998]). L'irradiation dynamique de la thérapie par arcs est simulée par une série de faisceaux statiques suffisamment rapprochés pour éviter des artefacts sur la répartition de dose.

Technique par collimateurs multilames

Le calcul de la répartition de dose pour des faisceaux produits par CMML se base sur la même formule que l'équation (2.6). La différence réside dans le fait que l'ouverture circulaire S_c est remplacée par le champ défini par le collimateur multilames (voir fig. 2.3). Comme il n'est pas envisageable de mesurer pour toutes les formes de champs possibles les valeurs de facteurs d'ouverture de champ et de rendements en profondeur, les positions des lames sont prises en compte de façon globale par l'intermédiaire d'un champ effectif S_{eff}. Dong et al. [1998] calcule ce champ effectif en deux étapes : la somme des largeurs des lames ouvertes donne la largeur effective W_{eff} et la moyenne des ouvertures des paires de lames donne la longueur effective L_{eff}, d'où un champ effectif rectangulaire. La seconde étape consiste à déterminer le carré équivalent au champ rectangulaire ce qui est une démarche classique en radiothérapie (Day and Aird [1996]). Dong et al. [1998] suggère de calculer le profil en intégrant un noyau gaussien n_g sur tout le champ ouvert par le CMML et de normer le tout par l'intégrale du noyau sur le champ effectif (voir fig. 2.4) :

$$Profil_{CMML}(DSP, S, p, r) = \frac{\iint_{CMML} n_g(x', y', x, y, z)dx'dy'}{\iint_{S_{eff}} n_g(x', y', 0, 0, z)dx'dy'} \qquad (2.7)$$

Le noyau gaussien n_g est défini par la somme de trois courbes de Gauss.

$$n_g(x', y', x, y, z) = \sum_{i=1}^{3} \frac{A_i(z)}{2\pi\sigma_i^2(z)} \cdot \exp\left(-\frac{(x-x')^2 + (y-y')^2}{2\sigma_i^2(z)}\right) \qquad (2.8)$$

Où $A_i(z)$ est l'amplitude de la i^egaussienne déterminée par un ajustement de courbe, avec la contrainte que $\sum_{i=1}^{3} A_i(z) = 1$,
$\sigma_i^2(z)$ est la variance de la i^egaussienne, également déterminée à partir des mesures.

On peut calculer analytiquement le profil du faisceau. En effet, il suffit de décomposer le champ CMML en petits rectangles définis par les paires de lames

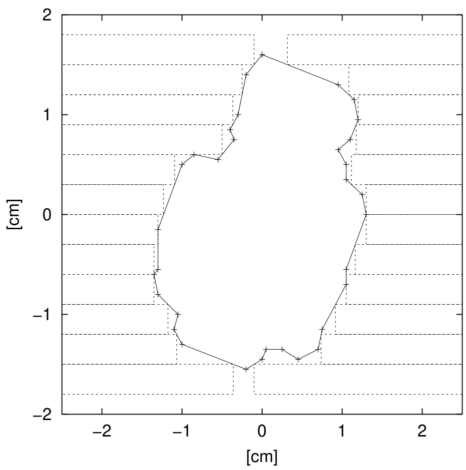
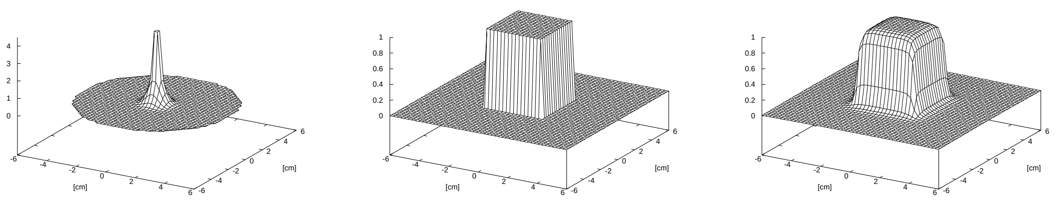


FIG. 2.3 – Position des micro-multilames par rapport au champ d'irradiation désiré.



(a) (b) (c)

FIG. 2.4 – Principe de la convolution pour générer le profil d'un faisceau. Pour ces illustrations, le champ choisi est un carré de 4 cm × 4 cm, la profondeur du profil est de 5 cm ; (a) est le noyau triple-gaussien, (b) représente le champ ouvert et (c) le profil généré par la convolution de (a) et (b).

en vis-à-vis. Par exemple, si $N_j(x,y,z)$ est le noyau intégré sur le rectangle défini par la j^{e}paire de lames en vis-à-vis, l'intégration de $N_j(x,y,z)$ s'écrit :

$$\begin{aligned} N_j(x,y,z) &= \iint_{rectangle_j} n_g(x',y',x,y,z)dx'dy' \\ &= \sum_{i=1}^{3} \frac{A_i(z)}{4} \\ &\quad \cdot \left(\operatorname{erf}\left(\frac{f \cdot X_j^{max} - x}{\sqrt{2}\sigma_i(z)} \right) - \operatorname{erf}\left(\frac{f \cdot X_j^{min} - x}{\sqrt{2}\sigma_i(z)} \right) \right) \\ &\quad \cdot \left(\operatorname{erf}\left(\frac{f \cdot Y_j^{max} - y}{\sqrt{2}\sigma_i(z)} \right) - \operatorname{erf}\left(\frac{f \cdot Y_j^{min} - y}{\sqrt{2}\sigma_i(z)} \right) \right), \end{aligned} \tag{2.9}$$

Où X_j^{min}, Y_j^{min}, X_j^{max} et Y_j^{max} sont les coordonnées des coins du rectangle j formé par la j^{e}paire de lames,
$f = \frac{DSP+z}{DSP_{ref}}$ permet de corriger la taille du champ par rapport à l'agrandissement géométrique.

Ainsi, il ne reste plus qu'à sommer cette expression sur toutes les paires de lames ouvertes pour obtenir la valeur du numérateur de l'équation (2.7). De même le dénominateur de cette équation peut se transformer en une expression similaire. Il suffit d'intégrer directement sur le rectangle correspondant au champ effectif :

$$\begin{aligned} \iint_{S_{eff}} n_g(x',y',0,0,z)dx'dy' &= \sum_{i=1}^{3} \frac{A_i(z)}{4} \\ &\quad \cdot \operatorname{erf}\left(\frac{f \cdot L_{eff}}{2\sqrt{2}\sigma_i(z)} \right) \\ &\quad \cdot \operatorname{erf}\left(\frac{f \cdot W_{eff}}{2\sqrt{2}\sigma_i(z)} \right) \end{aligned} \tag{2.10}$$

Et finalement, pour les CMML, l'équation (2.6) devient

$$\begin{aligned} Dose\,(DSP,S,z,r) &= D_{ref} \cdot FOC\,(S_{eff}) \cdot \left(\frac{DSP_{ref} + z_{ref}}{DSP + z_{ref}} \right)^2 \\ &\quad \cdot RP\,(DSP, S_{eff}, z) \\ &\quad \cdot \frac{\iint_{CMML} n_g(x',y',x,y,z)dx'dy'}{\iint_{S_{eff}} n_g(x',y',0,0,z)dx'dy'} \end{aligned} \tag{2.11}$$

2.2 Moyens d'évaluation des plans de traitement

2.2.1 Représentation par isodoses

La représentation d'une coupe TDM avec les isodoses correspondantes sous forme de lignes (voir fig. 2.5) permet au dosimétriste de retrouver aisément les surdosages ou les sous-dosages. Cette information n'est accessible que coupe par coupe. Par exemple, pour découvrir la valeur maximum de la répartition de dose, il est nécessaire de feuilleter toutes les coupes TDM concernées et sur chacune de repérer les surdosages pour finalement ne retenir que le plus élevé. Ce genre d'opération étant le pain quotidien du dosimétriste, il est souvent préférable d'utiliser une représentation tridimensionnelle de la répartition de dose.

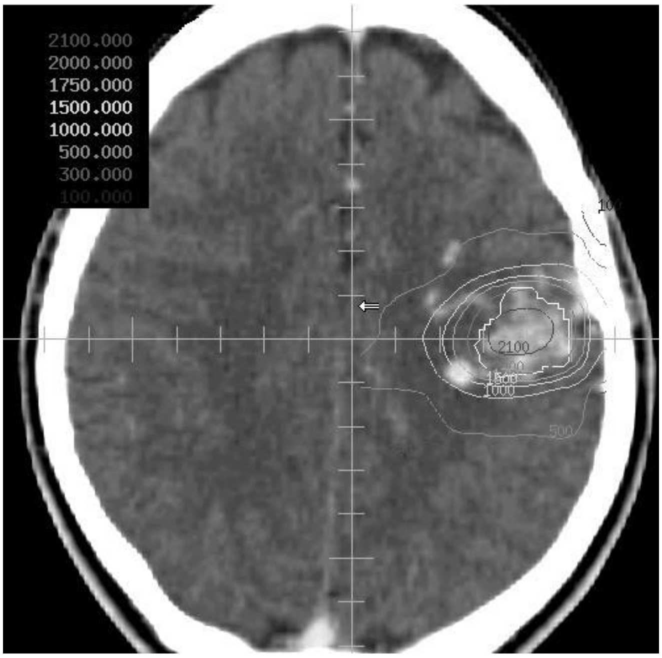
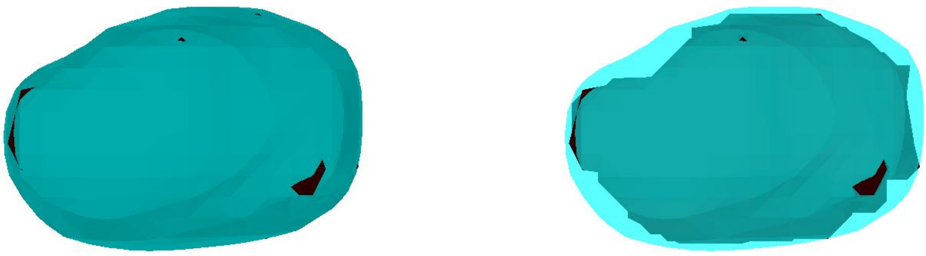


FIG. 2.5 – Isodoses dans le plan transverse.

En effet, la représentation par une isodose sous forme de surface tridimensionnelle est un moyen d'évaluation plus global (voir fig. 2.6). Pour juger si une isodose particulière recouvre bien une lésion, on les représente par des surfaces opaques et l'on recherche si la lésion s'étend hors de l'isodose. Puisque la représentation à l'écran est bidimensionnelle, il est nécessaire de faire tourner le volume de manière à l'observer sous tous les angles. Avec une isodose transparente et une lésion opaque, on peut évaluer le degré de conformation du traitement. Au niveau de l'appréciation visuelle, ces représentations tridimensionnelles sont très efficaces, il en est tout autrement pour la dosimétrie. Pour chaque valeur de dose intéressante, il faut représenter l'isodose correspondante de manière isolée, car plusieurs isodoses 3D surchargent la représentation.



(a) (b)

FIG. 2.6 – Isodose tridimensionnelle. Le diamètre principale de cette lésion est de 3 cm. (a) L'isodose est opaque et permet de juger de la couverture de la lésion : ici les parties sombres appartiennent à la cible et ne sont pas incluses dans l'isodose. (b) La même isodose transparente montre la conformation du traitement, autrement dit, la marge entre l'isodose et la lésion.

La représentation des isodoses à l'avantage d'être complète. Cependant, l'évaluation et la comparaison de traitements sont délicates. D'autres outils plus synthétiques comme les histogrammes dose-volume, ont été créés pour aider le dosimétriste dans ces comparaisons.

2.2.2 Histogrammes dose-volume

Un histogramme dose-volume (voir la définition page 6) permet de représenter la répartition de dose à l'intérieur d'une structure d'intérêt (voir fig. 2.7(a)). Afin de comparer des traitements de manière plus aisée, on utilise ces histogrammes sous une forme cumulative (voir fig. 2.7(b)). Ce type de graphique permet d'apprécier directement les doses minimum et maximum dans une structure.

Pour comparer des traitements, il suffit dans bien des cas de tracer, sur le même graphique, les histogrammes correspondant à chaque structure d'intérêt. Le radio-oncologue choisit le meilleur cas en tenant compte de la couverture du volume cible et de l'atteinte des organes à risque. Toutefois, certaines situations ne permettent pas de prendre une décision en se basant uniquement sur les HDV. En particulier, lorsque les courbes se croisent (voir fig. 2.7) : préfère-t-on un traitement avec un grand volume de l'organe à risque recevant une faible dose ou un petit volume recevant une dose élevée ? Un tel choix nécessite l'introduction d'un modèle radiobiologique.

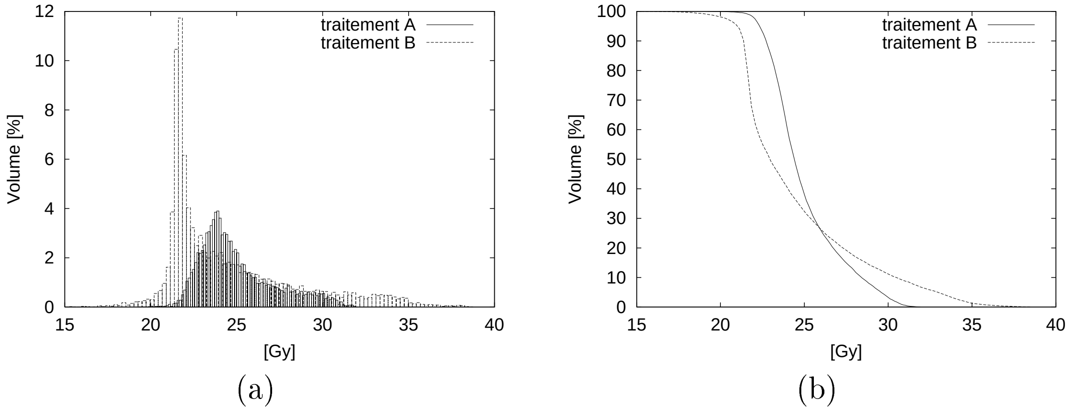


FIG. 2.7 – Histogrammes dose-volume. (a) Histogramme dose-volume (différentiel). (b) Histogramme dose-volume (cumulatif). C'est sous cette forme qu'on les utilise en radio-oncologie pour les comparaisons de traitements.

2.2.3 Probabilité de réponse à l'irradiation

La modélisation des courbes dose-réponse est fondamentale pour une évaluation précise de la réponse au traitement et pour la détermination de la répartition de dose optimale. Dans ce contexte, on peut définir la répartition de dose optimale comme la répartition de dose qui va éradiquer la lésion en provoquant le minimum de dégâts dans les tissus sains. En radio-oncologie, on parle de la probabilité de contrôle de la tumeur et de la probabilité de complication dans les tissus sains (PCTS). Pour ce qui concerne la radiochirurgie, la probabilité de contrôle de la tumeur est remplacée par la probabilité d'oblitération (PO).

Les auteurs (Archambeau and Shymko [1988], Niemierko and Goitein [1991], Schultheiss et al. [1983], Withers et al. [1988], Yorke et al. [1993]) décrivent les tissus en termes de sous-unités sérielles et parallèles. D'un point de vue structurel, la tumeur est un tissu à structure parallèle. En effet, il faut éliminer toutes les cellules clonogéniques pour pouvoir la stériliser. Les tissus sains peuvent être de nature sérielle ou parallèle. Ainsi la moelle épinière est un organe à structure sérielle, une destruction de la moelle sur un tronçon de 1 cm détruisant le fonctionnement de la partie distale à la zone irradiée. Par contre, le poumon est un organe à structure parallèle, car composé de sous-unités fonctionnelles relativement indépendantes. Dans la pratique, un organe présente souvent un mélange des deux types.

Dans la littérature, la relation dose-réponse pour les tumeurs et tissus sains est décrite sous différentes formes. Källman et al. [1992] présente une analyse de la situation et les relations entre les différents formalismes adoptés pour les courbes dose-réponse :

1. le modèle probabilistique "probit" :

$$P(D) = \frac{1}{2}\left(1 - \text{erf}\left[\sqrt{\pi}\gamma(1 - \frac{D}{D_{50}})\right]\right),$$

2. le modèle "logit" :

$$P(D) = \left[1 + \left(\frac{D_{50}}{D}\right)^{4\gamma}\right]^{-1},$$

3. le modèle de Poisson :

$$P(D) = 2^{-\exp(e\gamma(1-D/D_{50}))}, \tag{2.12}$$

où $P(D)$ est la probabilité de réponse,
D est la dose à laquelle est irradié toute la structure d'intérêt,
D_{50} est la dose correspondant à une probabilité de réponse de 0,5,
$\frac{\gamma}{D_{50}}$ est le gradient maximal sur la courbe. Dans le cas du modèle "probit", le gradient est maximal juste au-dessus de D_{50}, pour le modèle "logit", exactement à D_{50} et pour le modèle de Poisson, lorsque la probabilité de réponse vaut $\frac{1}{e} \approx 37\%$.

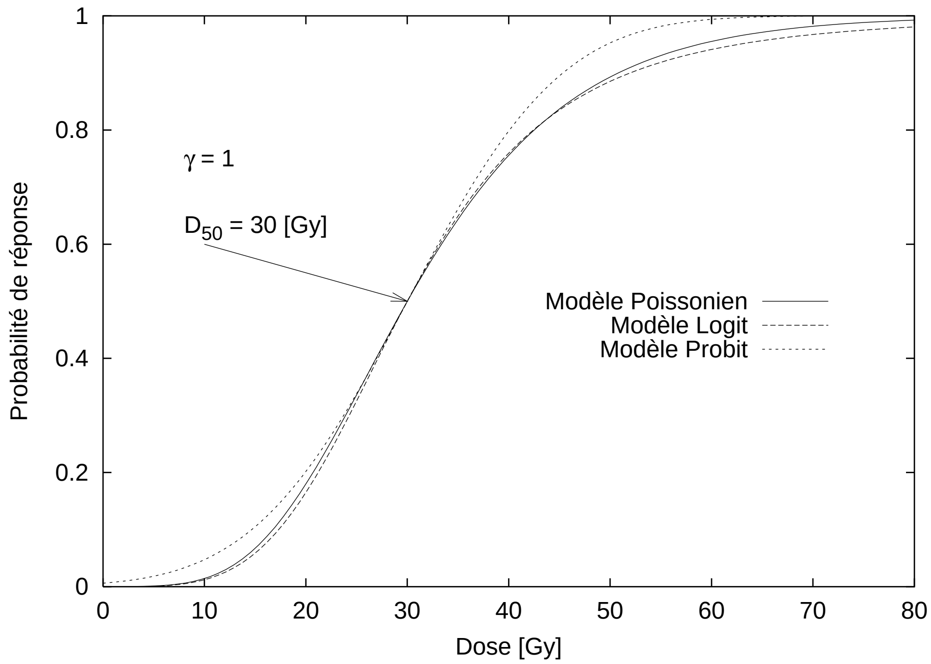


FIG. 2.8 – Comparaison des courbes de réponses

On trouvera à la figure 2.8 les trois courbes représentées pour une même dose $D_{50} = 30$ Gy et $\gamma = 1$. Comme on peut le constater, ces courbes sont très proches et les valeurs cliniques disponibles n'ont pas une précision suffisante pour fixer le choix sur l'une d'elles.

Seule la dernière équation a une signification radiobiologique, elle est fondée sur le modèle statistique de Poisson. Toutefois, son développement est basé sur la probabilité de contrôle d'une tumeur et sur le comptage (statistique de Poisson) de cellules clonogéniques. La courbe obtenue est ensuite exprimée en D_{50} et γ pour être généralisée, les sous-unités fonctionnelles de l'organe à risque étant alors comptées à la place des cellules clonogéniques. Källman et al. [1992] traite de ce problème et discute de la pertinence des trois modèles de dose-réponse. Pour ce travail, c'est le modèle de Poisson qui a été retenu pour exprimer la PCTS.

Probabilité de complication dans les tissus sains

L'équation (2.12) représente l'effet d'une dose homogène sur tout l'organe à risque considéré. Cette situation idéale n'est jamais atteinte dans la pratique. Il s'agit donc d'exprimer à partir de l'équation (2.12), l'effet produit par une irradiation inhomogène. Pour déterminer la probabilité de réponse d'une structure irradiée de façon inhomogène, on procède en deux étapes. D'abord on décrit l'effet de l'irradiation homogène sur une structure, puis, en morcelant la structure en autant de parties que nécessaire, on exprime l'irradiation inhomogène comme la juxtaposition de petites fractions de volume où la dose peut être considérée comme homogène (Schultheiss et al. [1983]).

Dans le cas d'un organe entièrement sériel (de m unités), la probabilité de complication s'exprime par 1 moins la probabilité qu'aucune sous-unité ne soit détruite :

$$P = 1 - \prod_{i=1}^{m}(1 - P_i), \tag{2.13}$$

où P_i est la probabilité de destruction de la i$^{\text{e}}$sous-unité.

Dans le cas d'un organe entièrement parallèle de n unités, la probabilité de complication correspond à la probabilité d'éliminer toutes les sous-unités fonctionnelles :

$$P = \prod_{j=1}^{n} P_j, \tag{2.14}$$

où P_j est la probabilité de destruction de la j$^{\text{e}}$sous-unité.

Dans le cas d'un organe composé de n structures parallèles de m unités sérielles (voir fig. 2.9), la probabilité de complication est :

$$P = \prod_{j=1}^{n}\left[1 - \prod_{i=1}^{m}(1 - P_{ij})\right], \tag{2.15}$$

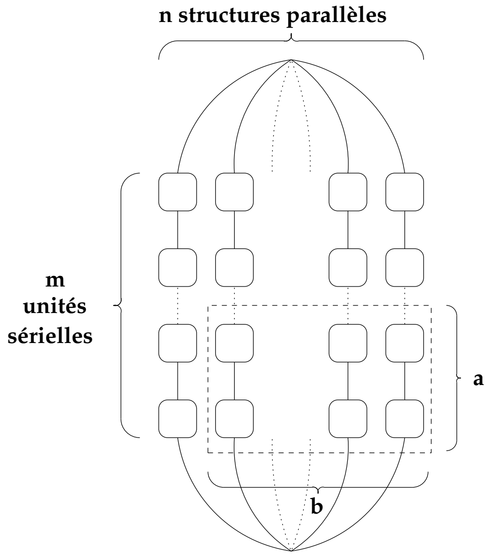


FIG. 2.9 – Structure d'un organe présentant un mélange de sous-unités sérielles et parallèles.

où P_{ij} est la probabilité de destruction d'une sous-unité fonctionnelle.

En supposant les sensibilités de toutes les sous-unités identiques $P_{ij} \equiv P_{\Delta}$, la réponse devient :

$$P = [1 - (1 - P_{\Delta})^m]^n . \tag{2.16}$$

Ce qui donne :

$$P_{\Delta} = 1 - \left[1 - P^{1/n}\right]^{1/m} . \tag{2.17}$$

Si a désigne la fraction relative des sous-unités sérielles irradiés à une dose homogène et b la fraction relative des structures parallèles, la probabilité d'induire un effet dans un organe dont une fraction $a \cdot b$ est irradiée ($0 \leq a \leq 1, 0 \leq b \leq 1$), se calcule par :

$$P_{ab} = [1 - (1 - P_{\Delta})^{a \cdot m}]^{b \cdot n} \tag{2.18}$$

et, en utilisant l'équation (2.17) pour substituer P_{Δ},

$$P_{ab} = \left[1 - \left[1 - P^{1/n}\right]^a\right]^{b \cdot n} \tag{2.19}$$

Un tel modèle résout bien l'effet de volume partiel. Pour des irradiations inhomogènes, il faut introduire une simplification supplémentaire : toutes les structures parallèles sont irradiées (i.e. $b := 1$). On définit la sérialité s par le rapport du nombre d'unités sérielles sur le nombre total d'unités :

$$s = \frac{m}{n \cdot m} = \frac{1}{n}. \tag{2.20}$$

Considérons qu'une fraction ν du volume de l'organe est irradiée à la dose homogène D. La fraction ν est donc le rapport du volume irradié V au volume de l'organe V_{ref} :

$$\nu = \frac{V}{V_{ref}} = a \cdot b \overbrace{= a}^{car\ b:=1} . \tag{2.21}$$

Ainsi, $P_{ab} = P_\nu$ s'écrit :

$$P_\nu = [1 - (1 - P^s)^\nu]^{1/s} . \tag{2.22}$$

En inversant cette équation et en supposant toujours une irradiation homogène, on obtient :

$$P = \left[1 - (1 - P_\nu^s)^{1/\nu}\right]^{1/s} . \tag{2.23}$$

La dernière étape consiste à calculer la réponse du tissu lorsque l'irradiation n'est pas homogène. Dans ce cas, la fraction de volume ν correspondant à un voxel et pour un organe composé de M voxels est égale à l'inverse du nombre de voxels. Ainsi, on obtient :

$$P = \left[1 - (1 - (P_\nu)^s)^M\right]^{1/s} \tag{2.24}$$

On peut généraliser cette expression pour l'appliquer à une irradiation non-homogène, la probabilité de complication dans les tissus sains devient :

$$PCTS = \left[1 - \prod_{i=1}^{M} [1 - (P_\nu(D_i))^s]\right]^{1/s} . \tag{2.25}$$

En exprimant P_ν en fonction de P, on obtient :

$$PCTS = \left[1 - \prod_{i=1}^{M} [1 - (P(D_i))^s]^\nu\right]^{1/s} \tag{2.26}$$

Cette formule s'applique très bien au cas des tissus sains car, dans ce cas, la radiosensibilité est relativement homogène.

Probabilité de complication dans le cas des MAV

Dans le contexte de la radiochirurgie, Ingmar and Karlsson [1996] appellent *complication radio-induite* chaque symptôme ou signe neurologique nouveau ou aggravé, transitoire ou permanent apparaissant comme lié à l'irradiation (œdème ou radio-nécrose). Ils calculent les probabilités de complication (PCTS) dans le cerveau résultant d'une utilisation thérapeutique des radiations selon le même formalisme que les équations (2.26) et (2.12). Une étude ultérieure (Karlsson et al. [1997a]), basée sur un millier de patients, a permis d'affiner encore les probabilités de réponse en ventilant les cas selon que la malformation est centrale, périphérique avec ou sans hémorragie. Une MAV est considérée comme centrale lorsqu'elle est dans le mésencéphale, le thalamus, le ganglion basal, les régions intra ou paraventriculaire ou dans le corps calleux. Comme cette étude a démontré que la sérialité n'est pas significativement différente de un, pour ce travail de thèse, le cerveau est considéré comme un organe sériel (avec $s = 1$) et la définition d'une complication radio-induite, décrite ci-dessus, est appliquée. Les valeurs des paramètres sont indiquées dans la table 2.1.

TAB. 2.1 – Modèle sériel ($s = 1$). Les cas trop rares des localisations centrales avec hémorragie n'ont pas été analysés par Karlsson et al. [1997a]. Le volume de référence du cerveau V_{ref}, nécessaire à l'ajustement de ces paramètres vaut 1136 cm^3. On remarque que la pente γ est plus élevée pour les cas ventilés, ce qui confirme le bien-fondé de cette séparation. D'autre part, notons que les tissus sains sont plus résistants dans le cas des MAV ayant présenté une hémorragie, car leur fonction a déjà été endommagée par l'épisode de saignement. Ainsi la probabilité d'une nouvelle complication diminue.

Paramètres des MAV	D_{50} **[Gy]**	γ
sans distinction	*7,63*	*0,85*
localisation centrale	7,17	1,46
localisation périphérique avec hémorragie	13,0	1,45
localisation périphérique sans hémorragie	9,22	1,48

Probabilité d'oblitération d'une malformation artérioveineuse

Karlsson et al. [1997c] traite de la probabilité d'oblitération d'une MAV suite à l'irradiation. Il rend compte de résultats surprenants si on les compare à nos connaissances de radio-oncologie. En effet, contrairement aux notions qui veulent que plus le volume cible augmente, plus la dose curative doit augmenter, dans le cas des MAV, la dose diminue lorsque le volume cible augmente. L'explication avancée serait que plus les vaisseaux sanguins que l'on désire obturer sont longs, plus la probabilité de créer une réaction radio-induite augmente. Ainsi on peut diminuer la dose prescrite sans perte d'efficacité. L'auteur définit un indice de caractérisation du traitement K. Cet indice se calcule de la façon suivante :

$$K[Gy \cdot cm] = (dose_{min}[Gy])(volume[cm^3])^{1/3}, \tag{2.27}$$

où $dose_{min}$ est la dose minimum dans la cible et
$volume$ est le volume de la cible.

La figure (2.10) représente la probabilité d'oblitération en fonction du paramètre K. Les droites ajustées aux valeurs expérimentales proviennent de l'article original. C'est à l'intersection des droites que, selon Karlsson, se trouve la valeur optimale de K ($K = 27$ [Gy·cm]). Pour des valeurs de K inférieures à 27, la probabilité d'oblitération de la lésion diminue. Pour des valeurs plus grandes, la dose délivrée est inutile, car elle n'augmente plus la probabilité d'oblitérer la MAV. La courbe quadratique semble tout aussi raisonnable sur cet intervalle et elle a été utilisée pour cette thèse. Les résultats de Karlsson et al. [1997c] sont confirmés par Touboul et al. [1998] et Flickinger et al. [1996], avec des collectifs de patients plus faibles et une relation basée uniquement sur la dose minimum au volume cible et non pas l'indice K.

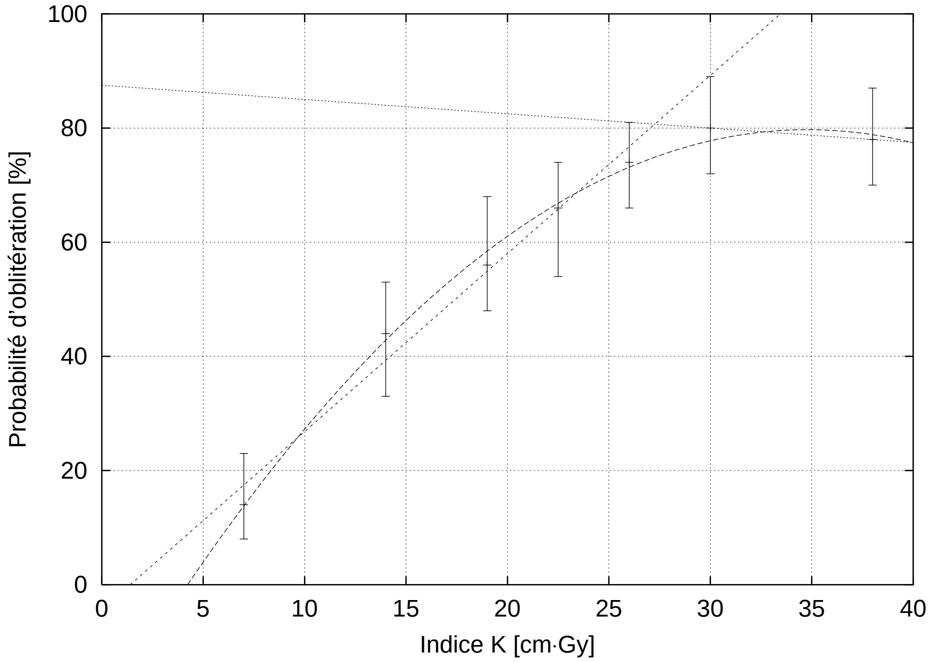


FIG. 2.10 – Probabilité d'oblitération pour les MAV en fonction de l'indice K.

Probabilité de succès du traitement

L'étape ultime dans la synthétisation des moyens d'évaluation d'un traitement réside dans sa probabilité de succès. La probabilité de succès (PS) se calcule comme la probabilité d'oblitération de la MAV (voir fig. 2.10) multiplié par la probabilité de ne pas avoir de complication dans les tissus sains (équation 2.26) :

$$PS = PO \cdot (1 - PCTS) \tag{2.28}$$

La valeur de PS permet de comparer aisément les différents traitements et sera le critère d'optimisation préféré pour les méthodes d'optimisation les plus sophistiquées.

2.3 Algorithmes d'optimisation

Le dosimétriste, qui procède par itération, doit décider des modifications à apporter au plan de traitement en vue de l'améliorer. Cette étape est, dans la pratique, effectuée sans l'aide d'un outil informatique. Pourtant, des algorithmes d'optimisation automatique sont introduits depuis peu dans les services de radio-oncologie.

Le problème général de l'optimisation s'énonce de la façon suivante : Soit une fonction f dépendant de multiples variables, trouvons les valeurs de ces variables de sorte que f soit maximum ou minimum. Les algorithmes d'optimisation appliqués à la radiochirurgie sont de trois types : les algorithmes matriciels, les méthodes de descente et les algorithmes stochastiques. Pour chaque type d'algorithme le problème général d'optimisation est formulé d'une manière particulière.

2.3.1 Méthodes matricielles

Formulation mathématique du problème

Pour les méthodes matricielles, le problème s'énonce comme une prescription de dose. En effet, le radio-oncologue décide de la dose structure par structure : par exemple, 15 Gy dans la lésion, et 0 Gy dans les yeux et le cerveau. Les variables du problème sont les poids des faisceaux. Elles sont liées à la dose prescrite de manière linéaire. Poursuivons notre exemple : Supposons que nous possédions deux faisceaux et quatre points où la dose est prescrite : un dans la MAV, un par oeil et un dans le cerveau. Dans ce cas, le système d'équations linéaire s'écrit :

$$\begin{aligned} A_{11} \cdot p_1 + A_{12} \cdot p_2 &= 15\,Gy, \text{ dans la MAV,} \\ A_{21} \cdot p_1 + A_{22} \cdot p_2 &= 0\,Gy, \text{ dans l'oeil gauche,} \\ A_{31} \cdot p_1 + A_{32} \cdot p_2 &= 0\,Gy, \text{ dans l'oeil droit,} \\ A_{41} \cdot p_1 + A_{42} \cdot p_2 &= 0\,Gy, \text{ dans le cerveau,} \end{aligned} \tag{2.29}$$

où A_{ij} est la dose délivrée par le j^{e}faisceau au i^{e}point et
p_j est le poids du j^{e}faisceau.

Ce système d'équations s'exprime sous forme matricielle par :

$$A \cdot p = D_0, \tag{2.30}$$

où A est la matrice des A_{ij},
p est le vecteur des poids et
D_0 est le vecteur correspondant à la prescription de dose.

Cet exemple comporte plus d'équations que d'inconnues. En général, un tel système n'a pas de solution p pour la prescription de dose D_0. On peut transformer l'équation en

$$A \cdot p - D_0 = 0. \tag{2.31}$$

La meilleure approximation de la solution correspond à minimiser la distance entre $A \cdot p$ et D_0, par la méthode des moindres carrés :

$$min_p \parallel Ap - D_0 \parallel^2 . \tag{2.32}$$

La décomposition en valeurs singulières, présentée ci-dessous, et la méthode des moindres carrés avec contraintes, décrite plus loin, sont les deux méthodes matricielles étudiées qui résolvent ce type de problème (équation 2.30).

Décomposition en valeurs singulières

La décomposition en valeurs singulières (DVS) est utilisée dans le cadre de la radiochirugie par Grandjean et al. [1998]. Elle permet de diagonaliser une matrice quelconque et de calculer, par la suite, sa matrice pseudo-inverse (voir annexe A). Soit A une matrice quelconque $m \times n$ et A^T sa transposée, posons :

$$A_1 = A \cdot A^T, \tag{2.33}$$

$$A_2 = A^T \cdot A. \tag{2.34}$$

A_1 est une matrice $m \times m$ et A_2 une matrice $n \times n$. Remarquons d'abord que A_1 et A_2 sont symétriques. La démonstration s'écrit :

$$(A_1)_{i,j} = \sum_{k=1}^{m} A_{i,k} \cdot A^T_{k,j} = \sum_{k=1}^{m} A_{i,k} \cdot A_{j,k}, \tag{2.35}$$

$$(A_1)_{j,i} = \sum_{k=1}^{m} A_{j,k} \cdot A^T_{k,i} = \sum_{k=1}^{m} A_{j,k} \cdot A_{i,k}. \tag{2.36}$$

Par conséquent :

$$(A_1)_{i,j} = (A_1)_{j,i} . \tag{2.37}$$

La démarche est strictement identique pour A_2. De plus A_1 et A_2 sont semi-définies positives. La démonstration pour A_1 est la suivante :

$$\begin{aligned} x^T A_1 x &= x^T A A^T x \\ &= (x^T A)(A^T x) \\ &= (A^T x)^T (A^T x) \\ &= \parallel A^T x \parallel^2 \geq 0, \forall x \in \Re^*. \end{aligned} \tag{2.38}$$

Il s'ensuit que les valeurs propres de A_1 sont toutes supérieures ou égales à zéro. En effet, si x_i est un vecteur propre de A_1 :

$$x_i^T A_1 x_i = \lambda_i \|x_i\| \geq 0. \tag{2.39}$$

La démonstration est strictement identique pour A_2. Sous ces conditions, le théorème spectral indique qu'il existe une matrice U telle que :

$$A_1 = U D_1 U^T,$$

avec $UU^T = U^T U = Id_m$, $U^{-1} = U^T$ et D_1 est une matrice de dimension $m \times m$, diagonale dont tous les termes sont positifs ou nuls. De même pour A_2, il existe une matrice V telle que :

$$A_2 = V D_2 V^T,$$

avec $VV^T = V^T V = Id_n$, $V^{-1} = V^T$ et D_2 est une matrice de dimension $n \times n$, diagonale dont tous les termes sont positifs ou nuls.

Théorème 1 (Décomposition en valeurs singulières) *Soit A une matrice $m \times n$, il existe une matrice orthogonale U $m \times m$, une matrice orthogonale V $n \times n$ et une matrice diagonale W $m \times n$ (les valeurs de la diagonale de W sont appelées valeurs singulières) avec :*

$$A = UWV^T \tag{2.40}$$

Démonstration : Une matrice comme $A_2 = A^T A$ possède un ensemble complet (base) de vecteurs propres orthogonaux notés x_j, que l'on place en colonnes dans V :

$$A_2 x_j = A^T A x_j = \lambda_j x_j \tag{2.41}$$

avec $x_i^T x_j = \delta_{ij}$ et $\lambda_j \geq 0$, $\forall i, j = 1 \cdots n$ et $V = (x_1, \cdots, x_r, \cdots, x_n)$. Supposons que les valeurs propres $\lambda_1, \cdots, \lambda_r$ soient strictement positives et les valeurs propres $\lambda_{r+1}, \cdots, , \lambda_m$ soient nulles. Pour les valeurs positives, on définit :

$$\sigma_j = \sqrt{\lambda_j} \;\; et \tag{2.42}$$

$$q_j = \frac{1}{\sigma_j} \cdot A x_j \tag{2.43}$$

Ces r vecteurs orthonormés $q_1, \cdots, q_r$ peuvent être complétés par l'algorithme de Gram-Schmidt pour former une base $q_1, \cdots, q_m$ de m vecteurs orthonormés. Ces vecteurs constituent les colonnes de la matrice U. Il reste à calculer le produit $U^T AV$ et à démontrer qu'il constitue la matrice W diagonale réelle :

$$\left(U^T AV\right)_{ij} = q_i^T A x_j = 0, \;\; si \; j > r \tag{2.44}$$

En effet :

$$\begin{aligned} q_i^T A x_j &= \left(\frac{A x_i}{\sigma_i}\right)^T A x_j \\ &= \frac{1}{\sigma_i} x_i^T A^T A x_j \\ &= \frac{1}{\sigma_i} x_i^T \left(A_2 x_j\right) = 0 \end{aligned}$$

Car, si $j > r$, $A_2 x_j = 0$. De plus,

$$\left(U^T A V\right)_{ij} = q_i^T A x_j = \sigma_j \delta_{ij},\ si\ j \leq r \tag{2.45}$$

En effet :

$$\begin{aligned} q_i^T A x_j &= q_i^T \sigma_j \left(\frac{A x_j}{\sigma_j}\right) \\ &= \sigma_j q_i^T q_j \\ &= \sigma_j \delta_{ij} \end{aligned}$$

Ainsi, les seules valeurs $\left(U^T A V\right)_{ij} \neq 0$ sont les valeurs pour lesquelles $i = j$ et $i \leq r$. Donc on peut écrire $W = U^T A V$ et dans ce cas W est diagonale réelle. Comme U et V sont orthonormales, $U^T U = U U^T = Id_m$, $V^T V = V V^T = Id_n$, on a :

$$W = U^T A V \tag{2.46}$$

$$UW = U U^T A V = A V$$

$$U W V^T = A V V^T = A$$

Et finalement :

$$A = U W V^T \tag{2.47}$$

On vient de démontrer, par construction, l'existence de la décomposition en valeurs singulières d'une matrice quelconque A. Il reste à montrer la relation avec la matrice pseudo-inverse présentée en annexe. Le problème d'optimisation à résoudre est le suivant :

$$\begin{aligned} min_p \parallel Ap - D_0 \parallel &= min_p \parallel \left(U W V^T\right) p - D_0 \parallel \\ &= min_p \parallel \left(W V^T\right) p - U^T D_0 \parallel . \end{aligned} \tag{2.48}$$

On définit y (de même norme que p) :

$$y := V^T p. \tag{2.49}$$

Ce qui permet d'écrire :

$$min_p \parallel \left(WV^T\right) p - U^T D_0 \parallel = min_y \parallel Wy - U^T D_0 \parallel . \tag{2.50}$$

Ainsi, minimiser $\parallel Ap - D_0 \parallel$ revient au même que minimiser $\parallel Wy - U^T D_0 \parallel$, et ceci correspond au cas particulier présenté précédemment puisque W est diagonale. Donc, on sait calculer y^+ :

$$y^+ = W^+ \left(U^T D_0\right) . \tag{2.51}$$

Il suffit de calculer p^+ en utilisant l'équation (2.49) pour faire apparaître la matrice pseudo-inverse :

$$p^+ = Vy^+ = \overbrace{VW^+U^T}^{A^+} D_0 = A^+ D_0. \tag{2.52}$$

A^+ est définie comme la matrice pseudo-inverse de A.

$$A^+ = VW^+U^T. \tag{2.53}$$

La routine effectuant une décomposition en valeurs singulières occupant plusieurs pages, elle n'est pas présentée ici (voir Press et al. [1992a] pour un listage de l'algorithme in extenso). Dans la pratique, les valeurs singulières sont triées par ordre décroissant pour constituer la diagonale de la matrice W. Si elles couvrent plusieurs ordres de grandeurs, le système d'équations est mal conditionné. En ne retenant que les valeurs singulières les plus élevées, il est possible de reconditionner le problème.

La solution p^+ fournie par la décomposition en valeurs singulières de l'équation (2.52) est la plus proche au sens des moindres carrés de la prescription de dose D_0. Toutefois cette solution n'est pas toujours physiquement exploitable, car elle peut aboutir à des poids négatifs qui n'ont pas de signification physique. La solution adoptée lorsque la méthode DVS est appliquée à la radiothérapie (ou à la radiochirurgie comme Grandjean et al. [1998]) est de mettre à zéro les poids qui ont une valeur négative. Ceci ne fournit pas nécessairement la meilleure solution au sens des moindres carrés du problème soumis à la contrainte que tous les poids soient positifs ou nuls.

Méthode des moindres carrés avec contraintes

La méthode NNLS (*non negative least square fit*) a reçu son nom de Lawson and Hanson [1974] qui ont basé leur algorithme sur le travail de Stoer [1971].

Elle permet de trouver la solution la plus proche (au sens des moindres carrés) de la répartition de dose prescrite en tenant compte du fait que l'on ne peut pas avoir de poids négatifs pour des faisceaux d'irradiation. Elle améliore le principal défaut de la méthode de la décomposition en valeurs singulières. D'autre part, elle est en général plus rapide et dans tous les cas moins gourmande en mémoire.

On définit le problème NNLS comme la recherche du minimum $\hat{p}$ de $\| Ap - D_0 \|$ sous la contrainte que toutes les composantes de p soient positives. Avec les mêmes définitions que pour l'équation (2.32), on note :

$$\hat{p} = min_p f(p) = min_p \frac{1}{2} \| Ap - D_0 \|^2 \;\; avec\, p_i \geq 0, \forall i \tag{2.54}$$

Théorème 2 (Conditions de Khun-Tucker) *On obtient une solution $\hat{p}$ au problème NNLS si et seulement s'il existe :*

$$\begin{aligned} \hat{w} &= A^T(Ap - D_0) && (2.55) \\ \hat{p}_i &= 0,\ \forall i \in Z,\ \textit{(définition de Z)} && (2.56) \\ \hat{p}_i &> 0,\ \forall i \in P,\ \textit{(définition de P)} && (2.57) \\ \hat{w}_i &\geq 0,\ \forall i \in Z && (2.58) \\ \hat{w}_i &= 0,\ \forall i \in P && (2.59) \end{aligned}$$

Ces équations définissent les conditions de Kuhn-Tucker suffisantes pour obtenir une solution $\hat{p}$.

Démonstration : $p_i \geq 0$ définit un hyperplan dont e_i (i^{e}vecteur de la base canonique) est la normale pointant à l'intérieur de l'espace possible. $\hat{p}$ est à l'intérieur des hyperplans $\forall i \in P$ et sur la limite des hyperplans $\forall i \in Z$. À noter que $\hat{w} = A^T(Ap - D_0)$ est le vecteur gradient de $f(p) = \frac{1}{2} \| Ap - D_0 \|^2$ en $p = \hat{p}$. Puisque $\hat{w}_i = 0$, $\forall i \notin Z$, l'équation (2.55) devient :

$$\hat{w} = \sum_{i \in Z} \left(\hat{w}^T e_i \right) \cdot e_i = \sum_{i \in Z} \hat{w}_i e_i = A^T(A\hat{p} - D_0) \tag{2.60}$$

Ce qui indique que le vecteur gradient de $f(p)$ s'écrit comme une combinaison linéaire de termes positifs ($\hat{w}_i \geq 0$) avec les vecteurs e_i. Autrement dit, pour un i donné, soit on est sur une frontière ($\hat{x}_i = 0$) parce que l'optimum se trouve dans la zone interdite par la contrainte ($\hat{x}_i$ ne peut pas être négatif) et dans ce cas, $\hat{w}_i \geq 0$ car aller dans le sens de e_i augmente $f(p)$, soit on est à l'optimum dans la direction e_i et dans ce cas $\hat{w}_i = 0$ (le gradient est nul dans la direction de e_i). Il reste à vérifier que $\hat{p}$ est la solution et qu'il n'en existe pas d'autres. Par l'absurde, si $(\hat{p}+u)$ est une solution alors u doit satisfaire $u_i \geq 0, \forall i \in Z$. Ce qui impose que $u^T\hat{w} \geq 0$. Or :

$$\begin{aligned}
f(\hat{p}+u) &= \frac{1}{2} \parallel A(\hat{p}+u) - D_0 \parallel^2 \\
&= \frac{1}{2}\left(A\hat{p} + Au - D_0\ \right)^T \left(A\hat{p} + Au - D_0\right) \\
&= \frac{1}{2}\left(A\hat{p} - D_0 + Au\right)^T\ \left(A\hat{p} - D_0 + Au\right) \\
&= \frac{1}{2} \parallel A\hat{p} - D_0 \parallel^2\ + \frac{2}{2}\left(Au\right)^T \left(A\hat{p} - D_0\ \right) + \frac{1}{2} \parallel Au \parallel^2 \\
&= f(\hat{p}) + \underbrace{\left(u^T\hat{w}\right)}_{u^T\hat{w}\geq 0} + \frac{1}{2} \parallel Au \parallel^2
\end{aligned}$$

Donc $f(\hat{p}+u) \geq f(\hat{p})$ et $\hat{p}$ est bien le minimum de $f(p) = \frac{1}{2} \parallel Ap - D_0 \parallel^2$ avec les contraintes $p_i \geq 0, \forall i$.

La routine à utiliser est décrite par l'algorithme 1. À la fin de la procédure, on obtient exactement $\hat{p}$ soumis aux conditions de Kuhn-Tucker (2.55)–(2.59). Donc $\hat{p}$ est solution du problème NNLS.

Algorithme 1 NNLS

1: $P \leftarrow \emptyset$, $Z \leftarrow \{1, 2, .., n\}$, $p \leftarrow 0$
2: $w \leftarrow A^T\left(Ap - D_0\right)$
3: Si $(Z = \emptyset)$ ou $(\forall j \in Z,\ w_j \geq 0)$ alors **FIN**
4: Sinon on cherche $t \in Z \mid w_t = min\left\{w_j,\ \forall j \in Z\right\}$
5: On transfère t de Z dans P
6: Soit A_P, la matrice $m \times n$ définie par toutes les colonnes j de A telles que $j \in P$, des colonnes j nulles si $j \in Z$
On résout en z : $A_p z \approx D_0$, $z_j \leftarrow 0\ \forall j \in Z$
7: Si $\forall j \in P$, $z_j \geq 0$ alors $p \leftarrow z$ et on continue en **2**
8: Sinon on cherche $q \in P \left| \frac{p_q}{p_q - z_q} = min\left\{\frac{p_j}{p_j - z_j} \mid z_j \leq 0,\ j \in P\right\}\right.$
9: $\alpha = \frac{p_q}{p_q - z_q}$ et $p = p + \alpha\left(z - p\right)$
10: On transfère j de P dans Z, $\forall j \in P$ avec $p_j = 0$, on continue en **6**

2.3.2 Méthodes de descente

Formulation mathématique du problème

L'idée des méthodes de descente consiste à calculer le gradient de la fonction à optimiser et à descendre dans une direction calculée à partir du gradient afin de se rapprocher du minimum. Dans ce travail, la fonction à optimiser est choisie de sorte à exprimer une prescription de dose. Toutefois, au lieu de fixer une valeur de dose pour chaque structure, comme dans le cas des algorithmes matriciels, le radio-oncologue décide des valeurs minimum et maximum : par exemple, entre 14,5 et 15,5 Gy dans la lésion et inférieur à 2 Gy dans les yeux et le cerveau. D'autre part, il peut également influencer l'algorithme en fixant des poids à ces contraintes. Dans notre exemple, un poids de 1 pour la lésion et les yeux, et de 0,01 pour le cerveau permet au programme de "sacrifier" un peu du cerveau en protégeant mieux les yeux. Les poids des contraintes minimum et maximum peuvent même être différents si cela s'avère nécessaire.

La fonction à optimiser $f(p)$ est de la forme :

$$f(p) = \sum_{k=1}^{k_{max}} \left[\; w_k^{min} \left[D_k^{min} - (Ap)_k\right]_+^2 + \; w_k^{max} \left[(Ap)_k - D_k^{max}\right]_+^2 \right], \tag{2.61}$$

où l'indice k parcourt tous les voxels,
A est la matrice des A_{ij}, comme à l'équation (2.30),
p est le vecteur des poids,
$[\cdots]_+$ vaut zéro si le contenu des crochets est négatif et la valeur du contenu dans les autres cas,
D_k^{min} et D_k^{max} sont les doses minimum et maximum à délivrer à la structure auquel appartient le voxel k et
w_k^{min} et w_k^{max} sont les pondérations qui varient comme D_k^{min} et D_k^{max}, c.-à-d. différentes pour chaque organe.

Comme la fonction $f(p)$ décrite par l'équation (2.61) est convexe, tous les algorithmes de descente vont aboutir à l'unique minimum global.

Méthode de Newton

Le principe de la méthode de Newton pour l'optimisation est de minimiser successivement les approximations au second ordre de la fonction $f(p)$. L'approximation au second ordre $f_0(p)$ s'écrit :

$$f_0(p) = f(p_0) + \nabla f(p_0)(p - p_0) + \frac{1}{2}(p - p_0)^T f''(p_0)(p - p_0) \tag{2.62}$$

On minimise cette fonction quadratique $f_0(p)$ (i.e. on annule le gradient) :

$$\nabla f_0(p) = \nabla f(p_0) + f''(p_0)(p - p_0) = 0. \tag{2.63}$$

Ce qui fournit un vecteur p_1 solution du système linéaire

$$f''(p_0)p_1 = f''(p_0)p_0 - \nabla f(p_0), \tag{2.64}$$

dont la solution s'écrit

$$p_1 = p_0 - [f''(p_0)]^{-1} \nabla f(p_0). \tag{2.65}$$

A l'itération k, on construit f_k, approximation quadratique de f au voisinage p_k, que l'on minimise pour obtenir p_{k+1}, défini par

$$p_{k+1} = p_k - [f''(p_k)]^{-1} \nabla f(p_k) \tag{2.66}$$

On pourrait démontrer la convergence de l'algorithme de Newton dans tout le voisinage d'un minimum local, ainsi qu'une vitesse de convergence quadratique. Cette méthode fonctionne très bien pour des problèmes de petites dimensions (quelques dizaines de variables), lorsque le calcul du Hessien f'' est facile. Dans les autres cas, par exemple en radiochirurgie, on préférera une méthode de quasi-Newton.

Méthode de quasi-Newton

La méthode de quasi-Newton consiste à imiter l'algorithme de Newton, mais sans calculer le Hessien de f ni son inverse. Au lieu de procéder à l'itération de l'équation (2.66), on calcule

$$p_{k+1} = p_k - \alpha_k S_k \nabla f(p_k) \tag{2.67}$$

avec S_k une approximation symétrique définie positive de $[f''(p_k)]^{-1}$ et α_k un paramètre positif. Évidemment plus S_k sera proche (au sens d'une norme matricielle) de $[f''(p_k)]^{-1}$, plus l'algorithme convergera rapidement. En l'occurence, le Hessien f'' est remplacé par sa diagonale, ce qui en facilite l'inversion. De plus, le paramètre α de l'équation (2.67) doit, dans ce cas, uniquement réduire le pas effectué à chaque itération afin que l'on n'atteigne pas un régime oscillatoire divergent. Comme chaque faisceau considéré augmente la dose proportionnellement à son poids, on prend $\alpha_k = \frac{1}{N_f}$, où N_f est le nombre de faisceaux. Ainsi l'équation (2.67) devient

$$p_{k+1} = p_k - \frac{1}{N_f} \left[\frac{\partial^2 f}{\partial^2 p_k}(p_k)\right]^{-1} \nabla f(p_k). \tag{2.68}$$

La i^{e}composante du gradient $\nabla f(p)$ se calcule par ($\frac{dA_j}{dp_i} = A_{ij}$)

$$\nabla f\left(p\right)_i = 2\sum_{j=1}^{j_{max}}\left[\ -A_{ij}w_j^{min}\ \left(D_j^{min} - (Ap)_j\right)_+ \ + A_{ij}w_j^{max}\ \left((Ap)_j - D_j^{max}\right)_+\right]. \tag{2.69}$$

Le i^{e}élément de la diagonale de $\frac{\partial^2 f(p)}{\partial^2 p}$ se calcule par

$$\left(\frac{\partial^2 f\left(p\right)}{\partial^2 p}\right)_i = \sum_{j=1}^{j_{max}}\left[\ A_{ij}^2 w_j^{min}\ H\left(D_j^{min} - (Ap)_j\right)\ + A_{ij}^2 w_j^{max}\ H\left((Ap)_j - D_j^{max}\right)\right], \tag{2.70}$$

où $\quad H$ est la fonction de Heaviside :

$$H(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ 1 & \text{si } x > 0 \end{cases}.$$

Comme rien n'empêche le vecteur p_k d'avoir des composantes négatives, à chaque itération, elles sont annulées avant le calcul de p_{k+1}.

Bortfeld [1999] utilise cet algorithme de quasi-Newton pour optimiser des traitements de radiothérapie par faisceaux à modulation d'intensité. Son travail a démontré que l'algorithme est très efficace pour ce type d'application puisque qu'à chaque itération elle produit une diminution d'un facteur 100 de la norme du gradient. Toutefois, en radiochirugie, les fonctions à optimiser présentent souvent l'aspect de vallées plus étroites qui font converger l'algorithme très lentement. En effet, cette méthode se déplace dans l'espace de phase des poids du point p_k vers le point p_{k+1} en suivant plus ou moins la direction de $-\nabla f(p_k)$. En minimisant $f(p)$ sur cette direction, la direction suivante $-\nabla f(p_{k+1})$ est pratiquement perpendiculaire à $-\nabla f(p_k)$ (voir Annexe B). Ainsi, il faut un grand nombre d'étapes dans des directions quasi-perpendiculaires pour suivre la direction générale du fond de la vallée de $f(p)$.

Algorithme des gradients conjugués

Les méthodes qui permettent de minimiser $f(p)$ dans des directions successives en prenant en compte les directions de minimisations déjà rencontrées s'appellent les méthodes de gradients conjugués (Press et al. [1992b], Nash and Sofer [1996]). Elles améliorent très sensiblement la vitesse de convergence lorsque $f(p)$ présente des "vallées étroites" (i.e. mathématiquement, lorsque le Hessien est mal conditionné).

Supposons que $f(p)$ est remplacée par l'approximation $f_0(p)$ comme dans l'équation (2.62). On définit une matrice M correspondant au Hessien et un vecteur g correspondant au gradient :

$$M := f''(p_0) \; et \tag{2.71}$$

$$g := -\nabla f(p_0). \tag{2.72}$$

De cette manière, $f_0(p)$ s'exprime par :

$$f_0(p) = f(p_0) - g \cdot (p - p_0) + \frac{1}{2}(p - p_0)^T M (p - p_0). \tag{2.73}$$

Si le but est d'optimiser $f_0(p)$, on peut la remplacer par une forme quadratique plus simple (notée Q), sans perte de la généralité. On définit $u := p - p_0$ et puisque $f(p_0)$ est une constante qui n'intervient pas dans l'optimisation, on peut l'omettre :

$$Q := \frac{1}{2}u^T \cdot M \cdot u - g \cdot u. \tag{2.74}$$

Pour optimiser cette forme quadratique, il faut annuler son gradient :

$$\nabla Q = Mu - g = 0. \tag{2.75}$$

A noter que le Hessien M est une matrice symétrique définie positive. Pour une telle matrice, deux directions d_0 et d_1 sont dites conjuguées par rapport à M si $d_0^T M d_1 = 0$.

Si une matrice M est symétrique définie positive, et si les vecteurs non nuls $d_1, ..., d_k$ sont conjugués deux à deux par rapport à M, alors ils forment une liste de vecteurs indépendants (base vectorielle). La démonstration s'effectue en supposant qu'il existe une combinaison linéaire nulle non triviale et en aboutissant à une contradiction, lors de l'application de la définition de la conjugaison.

Ainsi, si on note u^* la solution de $Mu = g$ et si l'on dispose d'une liste de n vecteurs conjugués par rapport à M, alors la liste étant une base de $\Re^n$, on peut exprimer u^* dans cette base :

$$u^* = \sum_{k=0}^{n-1} \alpha_k d_k. \tag{2.76}$$

On calcule aisément les coefficients α_i en multipliant cette équation à gauche par $d_i^T M$, d'où

$$d_i^T M u^* = \alpha_i \left(d_i^T M d_i \right) = \alpha_i \parallel d_i \parallel_M^2 . \tag{2.77}$$

Il ne reste qu'à trouver la liste des d_k. Les vecteurs propres de M sont une solution, mais elle est trop onéreuse. On applique une autre méthode, qui optimise Q en même temps qu'elle construit les directions d_k :

– On choisit un u_0 et on défini $d_0 := \nabla Q(u_0)$.

- Notons u_1 le vecteur obtenu par application de la méthode de gradient à pas optimal (voir annexe B), le long de d_0, $u_1 = u_0 - \alpha_0 d_0$. α_0 est donné par (équation B.3) :

$$\alpha_0 = \frac{\| \nabla Q(u_0) \|^2}{\| \nabla Q(u_0) \|_M^2} = \underbrace{\frac{d_0^T \nabla Q(u_0)}{\| d_0 \|_M^2}}_{\text{équation (B.2)}} .$$

On dispose ainsi de $\nabla Q(u_1)$ qui est orthogonal à $\nabla Q(u_0)$ (propriété établie par l'équation (B.1) .

- Recherchons une nouvelle direction d_1 comme combinaison linéaire de d_0 et $\nabla Q(u_1)$, qui soit également conjuguée à M :

$$d_1 = \nabla Q(u_1) - \beta_0 d_0.$$

Comme on désire que d_1 et d_0 soient conjugués, on choisit β_0 :

$$0 = d_0^T M d_1 = d_0^T M \nabla Q(u_1) - \beta_0 d_0^T M d_0.$$

Ce qui donne

$$\beta_0 = \frac{d_0^T M \nabla Q(u_1)}{d_0^T M d_0}.$$

- On peut maintenant optimiser Q le long de la direction d_1, ce qui fournit

$$u_2 = u_1 - \alpha_1 d_1,$$

avec

$$\alpha_1 = \frac{d_1^T \nabla Q(u_1)}{\| d_1 \|_M^2}.$$

et on recherche une nouvelle direction d_2, de forme

$$d_2 = \nabla Q(u_2) - \beta_1 d_1,$$

qui soit conjuguée par rapport à M avec d_0 et d_1, etc.

- Le fait que d_2 soit conjuguée avec d_1 s'impose naturellement, comme précédemment en choisissant β_1 :

$$\beta_1 = \frac{d_1^T M \nabla Q(u_2)}{d_1^T M d_1}.$$

Il est plus étonnant que d_2 ainsi définie soit aussi conjuguée avec d_0. Cette propriété provient du fait que Q est quadratique. Démontrons-le :

1. Notons que $\nabla Q(u_2)$ est orthogonal à $\nabla Q(u_1)$ et $\nabla Q(u_0)$. En effet, comme $u_2 = u_1 - \alpha_1 d_1$, on a $Mu_2 = Mu_1 - \alpha_1 Md_1$. Or Q est quadratique, et donc
$$\begin{aligned}\nabla Q(u_2) &= Mu_2 - g\\ &= Mu_1 - \alpha_1 Md_1 - g\\ &= \nabla Q(u_1) - \alpha_1 Md_1.\end{aligned}$$
Maintenant, $\nabla Q(u_0)$ est orthogonal à $\nabla Q(u_1)$ par suite de la recherche linéaire (définition de α_0), et $\nabla Q(u_0) = d_0$ est conjugué avec d_1. Donc $\nabla Q(u_0)$ est orthogonal à $\nabla Q(u_2)$. D'autre part, par définition de d_1, on a $\nabla Q(u_1) = d_1 + \beta_0 d_0$, et $\nabla Q(u_2)$ est orthogonal à d_1 (définition de α_1) et - comme on vient de le montrer - à $\nabla Q(u_0) =: d_0$.
2. Écrivons $d_0 = \frac{u_0 - u_1}{\alpha_0}$. Alors, $d_2^T Md_0$ s'écrit
$$\begin{aligned}d_2^T Md_0 &= \frac{d_2^T(Mu_0 - Mu_1)}{\alpha_0}\\ &= \frac{d_2^T(\nabla Q(u_0) - \nabla Q(u_1))}{\alpha_0}\\ &= \frac{(\nabla Q(u_2) - \beta_1 d_1)^T(\nabla Q(u_0) - \nabla Q(u_1))}{\alpha_0}\end{aligned}$$
(et comme $\nabla Q(u_2) \perp (\nabla Q(u_1) - \nabla Q(u_0)))$
$$\begin{aligned}d_2^T Md_0 &= -\frac{\beta_1}{\alpha_0} d_1^T(\nabla Q(u_0) - \nabla Q(u_1)\\ &= -\frac{\beta_1}{\alpha_0} d_1^T M(u_0 - u_1)\\ &= -\beta_1 \underbrace{d_1^T Md_0}_{=0}\\ &= 0.\end{aligned}$$

Pour résumer, les optimisations successives se font dans des espaces de dimensions croissantes, de bases $d_0, ..., d_k$ ou $\nabla Q(u_0), ..., \nabla Q(u_k)$.

L'algorithme général, qui peut aussi s'appliquer à des formes non-quadratiques, est présenté à la page 42. Il s'agit d'une variante qui évite de passer par le Hessien M pour le calcul des β_k. Cette variante utilise le fait que pour une forme quadratique, on a l'égalité
$$\beta_k = \frac{d_k^T M \nabla Q(u_{k+1})}{d_k^T Md_1} = \frac{\parallel \nabla Q(u_{k+1}) \parallel^2}{\parallel \nabla Q(u_k) \parallel^2}.$$
La démonstration se base sur la définition de u_{k+1} :
$$u_{k+1} = u_k - \alpha_k d_k,$$

d'où

$$\nabla Q(u_{k+1}) = \nabla Q(u_k) - \alpha_k M d_k.$$

Or,

$$\beta_k = \frac{d_k^T M \nabla Q(u_{k+1})}{d_k^T M d_k} = \frac{\nabla Q(u_{k+1})^T (\nabla Q(u_k) - \nabla Q(u_{k+1}))}{d_k^T (\nabla Q(u_k) - \nabla Q(u_{k+1}))},$$

après avoir remplacé Md_k par sa valeur. Or, d'une part $\nabla Q(u_{k+1})^T \nabla Q(u_k) = 0$, d'autre part, $\nabla Q(u_{k+1})^T d_k = 0$. D'où

$$\beta_k = -\frac{\| \nabla Q(u_{k+1}) \|^2}{d_k^T \nabla Q(u_k)}.$$

Enfin, comme $d_k = \nabla Q(u_k) - \beta_{k-1} d_{k-1}$ et $d_{k-1}^T \nabla Q(u_k) = 0$, le dénominateur vaut bien $\| \nabla Q(u_k) \|^2$.

Lorsqu'on minimise une fonction plus générale qu'une forme quadratique, les directions d_k sont pondérées sur trop d'itérations pour être efficace localement. C'est pourquoi on force périodiquement un β_k à zéro. Par exemple, un choix judicieux est d'effectuer une mise à zéro de β_k toutes les n itérations, où n est le nombre de composantes du vecteur p_k.

Algorithme 2 Gradients conjugués

initialisation de l'indice : $k \leftarrow 0$.
choix du vecteur des poids initial p_0 et de la tolérance de convergence ϵ.
initialisation de la direction de descente : $d_0 \leftarrow \nabla f(p_0)$.
répéter
  recherche dichotomique de α_k ($\alpha_k = 1, \frac{1}{2}, \frac{1}{4}, ...$) pour déterminer

$$p_{k+1} = p_k - \alpha_k \cdot d_k$$

  de sorte que

$$f(p_{k+1}) < f(p_k)$$

  calcul du paramètre β_k :

$$\beta_k \leftarrow \frac{\| \nabla f(p_{k+1}) \|^2}{\| \nabla f(p_k) \|^2},$$

  calcul de la nouvelle direction de descente d_k :

$$d_{k+1} \leftarrow \nabla f(p_{k+1}) + \beta_k \cdot d_k,$$

  incrémentation de k :

$$k \leftarrow k + 1.$$

jusqu'à ce que $\| \nabla f(p_k) \| < \epsilon$.
FIN

2.3.3 Algorithmes stochastiques

Pour ces méthodes, l'optimisation peut se faire d'une manière plus raffinée, car on n'est pas obligé de construire une fonction convexe qui ne présente qu'un minimum global. En effet, les algorithmes stochastiques sont construits de sorte à parcourir l'espace de phase avec une stratégie qui assure que la probabilité de trouver l'optimum global est élevée par rapport aux nombres d'évaluation de la fonction à optimiser. Dans notre cas, cette dernière est constituée par la probabilité de succès du traitement en tenant compte de la probabilité d'oblitération et de complication dans le cerveau. Un seul algorithme stochastique a été étudié, il s'agit du recuit simulé.

Recuit simulé

La méthode du recuit simulé s'apparente à la façon dont la nature retrouve la configuration optimale (avec l'énergie la plus faible) dans le refroidissement d'un métal. L'analogie peut se prolonger : lorsqu'on soumet un métal à la trempe, il n'atteint pas une configuration suffisamment optimale, le refroidissement est trop rapide. Il est possible de réchauffer le métal pour qu'il atteigne une configuration plus optimale. On peut aider l'algorithme de recuit simulé à atteindre l'optimum global lorsqu'il est piégé dans un optimum local. L'image qui se prête le mieux pour expliquer l'algorithme utilisé est celle d'une balle qui rebondit sur un relief. A chaque saut, elle est susceptible d'atteindre un autre endroit. Comme son énergie diminue à chaque étape, elle finira par être arrêtée dans un minimum, autrement dit au fond d'une vallée. Si l'on est certain qu'il s'agit du minimum global, on arrête l'algorithme. Dans le cas contraire, il est possible de lui redonner de l'énergie pour que la balle tente d'atteindre après une nouvelle série de déplacements, un endroit d'altitude plus faible.

L'algorithme recherche la solution par essais successifs en se déplaçant dans l'espace de phase des paramètres à analyser (coordonnées, poids des faisceaux, etc...). Il mémorise les meilleures solutions. Un nouvel état $k+1$ est déterminé à partir de l'état k en modifiant tous les paramètres décrivant l'état k. Par exemple, le paramètre à varier p, ou mieux p_{k+1} est calculé à partir de p_k par

$$p_{k+1} = p_k + x_T\left(p_{max} - p_{min}\right), \tag{2.78}$$

où p_{min} et p_{max} sont les bornes inférieure et supérieure du paramètre p,
T représente la température,
$x_T \in [-1, 1]$ est une variable aléatoire de densité de probabilité $g(x_T)$

avec

$$g(x_T) = \frac{1}{2\left(\mid x_T \mid +T\right) \cdot \ln\left(1 + \frac{1}{T}\right)} \tag{2.79}$$

$g(x_T)$ est une fonction dont la largeur autour de 0 diminue lorsque T diminue (voir figure 2.11). Ainsi plus la température associée à l'état k diminue, plus les nouveaux paramètres de l'état $k+1$ se rapprochent de l'état k. Cette fonction remplit les conditions nécessaires afin d'assurer qu'en probabilité, tout l'espace de phase est parcouru. Ceci a été vérifié par Ingber and Rosen [1992], Ingber [1993].

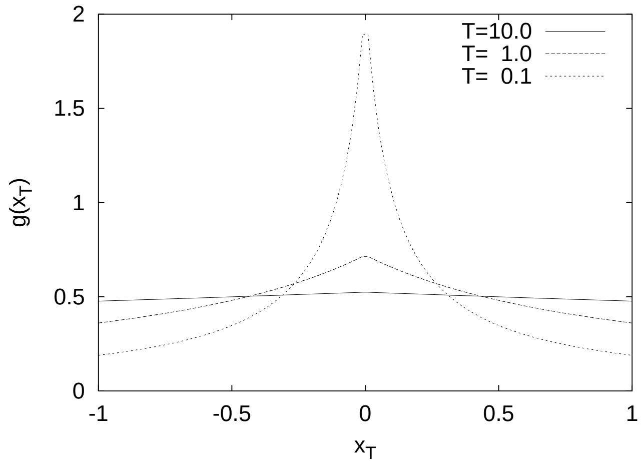


FIG. 2.11 – Densité de probabilité $g(x_T)$

Le coût associé à l'état $k+1$, C_{k+1} est ensuite calculé comme la probabilité d'échec du traitement, c'est-à-dire un moins la probabilité de succès calculée selon l'équation (2.28). Si la probabilité d'échec est plus faible que la précédente, le nouvel état est mémorisé et utilisé comme point de départ pour l'étape suivante $k+2$. Dans le cas contraire, le nouvel état $k+1$ n'est pas mémorisé, mais il est tout de même utilisé pour poursuivre la recherche de l'optimum si la condition suivante est remplie (u est une variable aléatoire tirée dans l'intervalle $[0,1]$) :

$$\exp\left(-\frac{C_{k+1}-C_k}{T}\right) > u, \tag{2.80}$$

où C_{k+1} est la probabilité d'échec du traitement correspondant aux paramètres p_{k+1},
C_k est la probabilité d'échec du traitement correspondant aux paramètres p_k.

Dans le cas où cette condition n'est pas remplie, l'état $k+2$ sera calculé à partir de l'état k. Lorsque le nombre d'états acceptés par rapport au nombre d'états générés est trop faible, on augmente le paramètre T en le multipliant par un facteur. C'est de là que provient la désignation de "recuit simulé" donnée à cet algorithme.