Le Mans Université

Projet L3 Acoustique et Vibrations

Modélisation d'un circuit d'écrêtage à diodes dans les pédales de distorsion de guitare.

Eliot Deschang

Sous la supervision de

M. Lihoreau Bertrand, M. Novak Antonin

Licence 3 Acoustique et Vibrations

Année universitaire : 2022 - 2023

Remerciements

Je tiens tout d'abord à exprimer ma profonde gratitude envers mes deux encadrants, Bertrand Lihoreau, enseignant-chercheur et Antonin Novak, chercheur au laboratoire d'acoustique de l'université du Mans, pour leur soutien inestimable et leur encadrement tout au long de la rédaction de ce rapport. Bien qu'étant à l'impulsion de ce projet de recherche, la réalisation de celui-ci n'aurait pas été possible sans leur accord pour m'accompagner tout au long de ce projet.

Un grand merci à Bertrand Lihoreau, dont les conseils avisés et les discussions enrichissantes ont été une source constante d'inspiration et d'apprentissage. Cela a toujours été un plaisir d'échanger avec lui, tant sur mon sujet de recherche que sur notre passion commune pour le matériel "guitaristique". Je tiens à souligner l'implication et le dévouement dont il a fait preuve, consacrant une part importante de son temps à mon accompagnement.

Je remercie également chaleureusement Antonin Novak pour son soutien et son expertise précieuse. Ses mots bienveillant et sa passion ont été précieux pour moi tout au long de ce projet. Malgré un emploi du temps très chargé, il a su trouver le temps pour m'aider, m'expliquer différents concepts de manière limpide, et résoudre mes problèmes dans la réalisation de mon travail.

Table des matières

1 Introduction **4**

2 Circuit analogique **5**
2.1 Architecture d'une pédale de distorsion 5
2.2 Analyse superficielle du circuit 5

3 Mesures du circuit d'écrêtage **6**
3.1 Conditions expérimentales 6
3.2 Résultats des mesures 7

4 Mise en équation du circuit d'écrêtage **11**

5 Méthodes de résolution numériques en temps différé **13**
5.1 Euler explicite 13
5.2 Euler implicite 13
5.3 Runge-Kutta 14
5.4 Méthode des trapèzes 14
5.5 Newton-Raphson 14

6 Comparaison des méthodes de résolution **15**
6.1 Stabilité des méthodes 15
6.2 Comparaison des algorithmes avec LTspice 18
6.3 Comparaison des méthodes de résolution vis-à-vis du modèle expérimental 20

7 Méthode de résolution en temps réel **21**
7.1 Contexte 21
7.2 Approche Statique 22
7.3 Comparaison vis-à-vis de la méthode des trapèzes 23
7.4 Conception d'un plugin audio 25

8 Conclusion **26**

A Annexe **28**
A.1 Photo MXR Distortion + et DOD 250 28
A.2 Circuit d'alimentation de la pédale MXR Distortion + ... 28
A.3 Méthodes d'Euler 29
A.4 Euler explicite 29
A.4.1 Différence Avant 29
A.4.2 Intégration 30
A.4.3 Développement limité 30
A.4.4 Erreur locale 31
A.4.5 Erreur Globale 31
A.5 Euler implicite 32
A.5.1 Différence arrière 32
A.5.2 Intégration 33
A.5.3 Développement limité 34
A.5.4 Erreur locale 34
A.5.5 Erreur globale 34
A.6 Méthode des trapèzes 35
A.6.1 Intégration 35
A.6.2 Erreur locale 35

A.6.3 Erreur globale 36
A.7 Runge-Kutta 4 36
A.7.1 Erreur locale et globale 37
A.8 Newton-Raphson 37
A.8.1 Démonstration 37
A.8.2 Convergence numérique 38

1 Introduction

À l'aube des années 30, le monde assiste à la naissance d'un phénomène culturel majeur : la musique amplifiée. 20 ans plus tard, la guitare électrique et les amplificateurs à lampe se démocratisent. À cette époque, tout les studios d'enregistrement sont équipés de matériel analogique, conçu et construit avec une intention principale : produire un son aussi linéaire et fidèle que possible. Cependant, cette technologie, dans sa quête de perfection, n'était pas sans failles. Les caractéristiques intrinsèques des équipements, en particulier des amplificateurs à lampes, présentaient des zones de non-linéarité. Ces "imperfections", loin de déplaire aux musiciens, conféraient à leur musique une chaleur et une coloration sonore unique. Pour exploiter ces propriétés, certains musiciens n'hésitaient pas à pousser leurs amplificateurs à des niveaux de jeu très élevés, inconfortables voire dangereux pour l'audition, tout en exposant leur matériel à un risque de dommages.

Pour répondre à cette demande croissante d'un son distinctif et saturé, en permettant aux musiciens de jouer à des volumes plus raisonnables, des gens ont commencé à concevoir des boîtiers d'effets dédiés. Ces boîtiers, ou pédales d'effets, étaient conçus pour imiter et même amplifier les caractéristiques non linéaires de ces amplificateurs, en produisant leur propre distorsion. Ainsi naquit l'ère des pédales de distorsion, marquant une révolution dans la production de musique amplifiée et lançant une course à la création d'un son toujours plus distinctif et personnalisé.

Aujourd'hui, de par la démocratisation du matériel numérique, l'émulation du son produit par ces pédales est devenue possible. Les plugins, qui sont de petits programmes utilisés dans les logiciels de musique assistée par ordinateur, permettent à tout utilisateur équipé d'un ordinateur, d'une carte son et d'un instrument de jouer et de s'enregistrer en direct avec le son distinctif de ces dispositifs électroniques. Cependant, la fidélité de reproduction des nuances et des complexités de ces pédales d'effets représente un défi. En particulier, lorsqu'il s'agit d'implémenter ces solutions pour le traitement audio en temps réel, un algorithme efficace est nécessaire pour ne pas altérer l'effet original.

À travers ce projet, l'objectif est d'étudier de manière approfondie un circuit analogique d'écrêteur à diode, typique de ceux utilisés dans les pédales d'effets de distorsion. Cette analyse s'organise en plusieurs phases. Au commencement, une étude expérimentale du circuit est effectuée afin de saisir son comportement. Ensuite, une mise en équation du système est réalisée pour exprimer le lien entre l'entrée et la sortie du système. À partir de là, le système est résolu numériquement sans se soucier de l'aspect temps réel. Puis, à partir des conclusions tirées de la phase expérimentale, une méthode en temps réel est développée pour la résolution du système. L'ensemble de ces étapes permet une compréhension et une modélisation précises du circuit, avec pour objectif de développer un plugin émulant l'effet d'écrêtage dans ce type de pédale.

2 Circuit analogique

2.1 Architecture d'une pédale de distorsion

Le but de cette partie est de fournir au lecteur des explications quant au fonctionnement et à l'architecture d'une pédale de distorsion. Dans cette partie, le choix est fait de s'intéresser à la pédale MXR Distortion +. Une image de cette pédale est fournie en annexe [A.1]. Cette pédale a l'avantage de présenter une conception simple, et est représentatif du fonctionnement d'un grand nombre de pédale de distorsion, ce qui en fait un cas d'étude idéal. La Fig. 1 reprend la chaîne de traitement du signal de cette pédale.

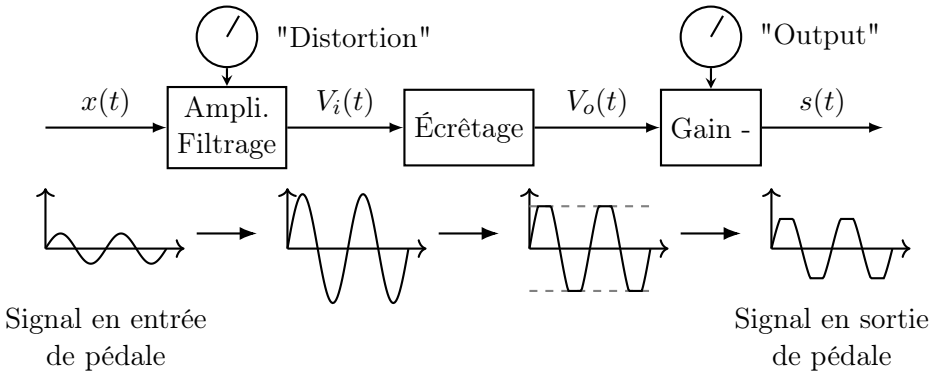


FIGURE 1 – Schématisation du traitement du signal dans une pédale de distorsion (label des potentiomètres de la MXR Distortion +).

La structure interne d'une pédale de distorsion s'articule généralement autour de deux étages principaux. Le premier est l'étage de filtrage actif. Ici, le signal d'entrée est amplifié et filtré. Cet étage est soutenu par un étage d'alimentation qui fournit à l'amplificateur opérationnel (ou op-amp) l'énergie nécessaire pour amplifier et filtrer le signal d'entrée.

Le signal, une fois amplifié et filtré, est ensuite dirigé vers le second étage principale de la pédale : l'étage d'écrêtage. À ce stade, toutes les parties du signal qui dépassent une certaine amplitude sont ramenées à cette amplitude maximale. L'objectif fondamental de cet étage est d'introduire une non-linéarité délibérée dans le signal audio, une manipulation qui, lorsqu'elle est traduite dans le domaine fréquentiel, génère une cascade harmonique. C'est ce processus d'écrêtage qui génère la distorsion caractéristique.

L'utilité de l'étage d'amplification devient évidente lorsqu'il est considéré comme le "régulateur de la force" avec laquelle le signal pénètre dans l'étage d'écrêtage : en ajustant le gain de l'étage d'amplification, il est possible de positionner le signal dans une zone plus ou moins linéaire de l'étage d'écrêtage, influençant ainsi le degré de distorsion produit. Le contrôle du gain sur l'étage du filtrage actif s'effectue à l'aide du potentiomètre nommé "Distortion" sur la MXR, et le contrôle du niveau de sortie à l'aide du potentiomètre nommé "Output".

2.2 Analyse superficielle du circuit

Le schéma électrique de la pédale MXR Distortion + est représenté Fig. 2 en utilisant les normes de représentations européennes (le B et le C sur les valeurs des potentiomètres représentent donc des courbes de variations respectivement logarithmiques et anti-logarithmique).

Sur la Fig. 2, l'amplificateur opérationnel est en configuration non inverseur, les résistances R_3, R_4, et le potentiomètre Distortion déterminent le gain en tension. Plusieurs condensateurs C_1, C_2, C_3, et C_4 filtrent le signal de la guitare. L'entrée + de l'amplificateur opérationnel est polarisée à

4.5V par l'intermédiaire de la résistance R_2, maintenant la masse virtuelle à 4.5V et permettant d'amplifier les signaux d'entrée bipolaires de la guitare.

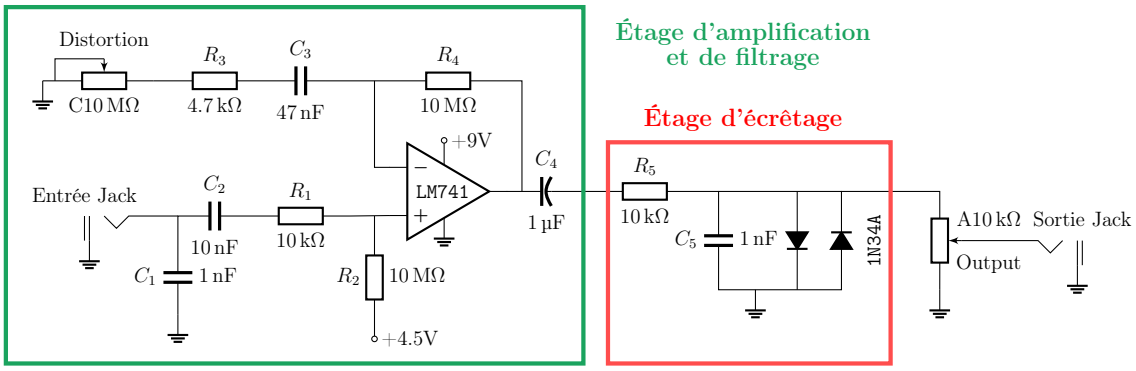


FIGURE 2 – Schéma électrique de la pédale d'effet MXR Distortion +. Le schéma électrique de l'étage d'alimentation de la pédale est fournit en annexe [A.2].

Là où la structure de l'étage de filtrage actif peut varier considérablement d'une pédale à l'autre, l'étage d'écrêtage est quant à lui beaucoup plus "standardisé". Celui-ci s'articule autour d'un filtre RC passe-bas, auquel deux diodes tête-bêche en parallèle sont liées. La résistance R_5 est nécessaire pour limiter la quantité de courant entrant dans les diodes (pour éviter le claquage). Le potentiomètre de sortie Output contrôle le volume de sortie en utilisant un potentiomètre logarithmique qui détourne une partie du signal d'entrée vers la masse.

La configuration anti-parallèle des diodes permet ici d'écrêter le signal, et ce peu importe le signe de sa tension. Si les deux diodes sont identiques, alors le seuil d'écrêtage est symétrique. Le seuil de tension à partir duquel la tension est écrêtée varie en fonction du modèle de diode utilisé dans le circuit d'écrêtage (diode silicium, germanium, schottky, électro-luminescente...), et ce choix va donc définir la "couleur" de la distorsion.

3 Mesures du circuit d'écrêtage

3.1 Conditions expérimentales

Le but est de recréer l'étage d'écrêtage typique de ces pédales, notamment à des fins de comparaison avec les modèles numériques simulant cette étage analogique. La fabrication du circuit est basée sur l'architecture de l'étage d'écrêtage de la MXR Distortion +. L'étage d'amplification ne fait pas l'objet de cette étude. Les diodes originels au germanium `1N341` n'ayant été trouvées, celles-ci sont remplacées par des diodes silicium `1N4148` dans l'étude expérimentale. Le schéma du montage est reporté Fig. 3.

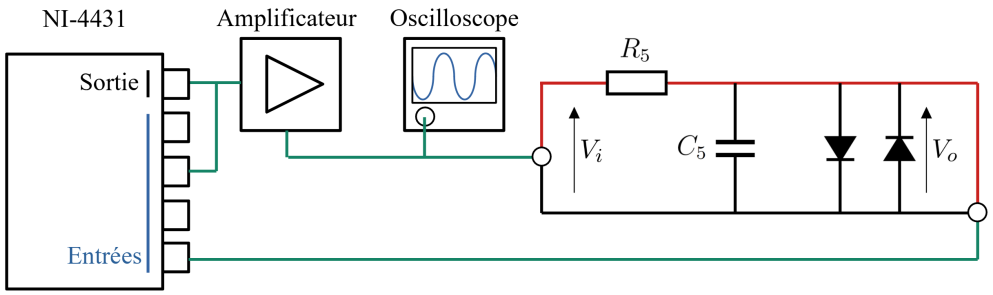


FIGURE 3 – Schématisation du montage expérimental.

Les différentes mesures sont effectuées à l'aide d'une carte National Instrument NI-4431 (5 voies), dont la sortie audio est reliée à un amplificateur qui envoie le signal une fois amplifiée aux bornes du circuit d'écrêtage. La tension en sortie d'amplificateur est également consulté sur un oscilloscope. La tension en sortie d'étage d'écrêtage V_o est ensuite transmise à une des entrées d'acquisition de la carte. Toutes les acquisition sont réalisées sur le logiciel INTAC avec l'outil *Temporal Acquisition*, avec des signaux d'excitation générées via python, enregistrés au format WAV (96 kHz, 24bits).

Les valeurs de la résistance et du condensateur ont été mesurées en effectuant dans un premier temps la mesure de la résistance avec un ohmmètre et dans un second temps une réponse en fréquence du filtre RC (sans les deux diodes), de manière à identifier par lecture graphique la fréquence de coupure à -3dB f_c du filtre passe-bas et ainsi en déduire indirectement la valeur du condensateur. Ici $R = R_5 = 9.90 \pm 0.01$ kΩ, et $f_c = 14615 \pm 2$ Hz. La valeur de C, vaut donc :

$$f_c = \frac{1}{2\pi RC} \Longrightarrow C = \frac{1}{2\pi \times 9900 \times 14615} = 1.1 \pm 0.1 \text{ nF}.$$

3.2 Résultats des mesures

Le choix est fait d'envoyer dans un premier temps un sinus de même fréquence f à différentes amplitudes, de manière à constater l'effet de l'écrêtage en fonction de cette amplitude A du signal d'entrée $V_i(t) = A\sin(2\pi ft)$. L'amplitude A est associée au signal d'excitation en sortie d'amplificateur/à l'entrée du circuit. Celle-ci est déterminée à l'aide de l'oscilloscope, en ajustant le gain de l'amplificateur. Le signal temporel V_o, pour une fréquence d'excitation $f = 100$ Hz et différentes valeurs de A est tracé Fig. 4, sur deux périodes du sinus d'excitation.

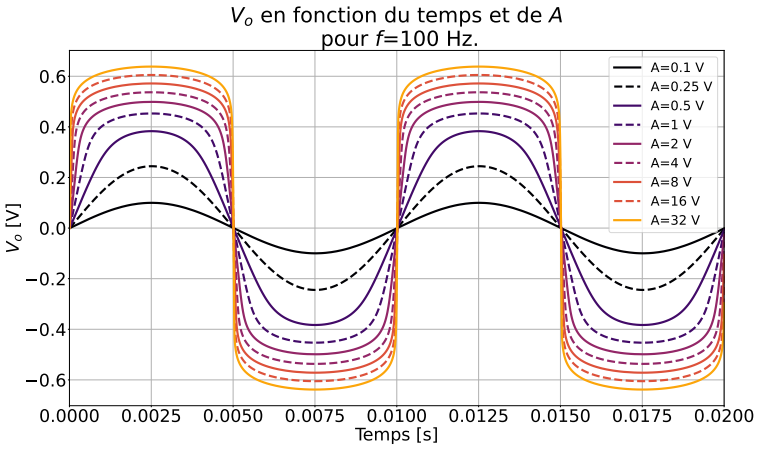


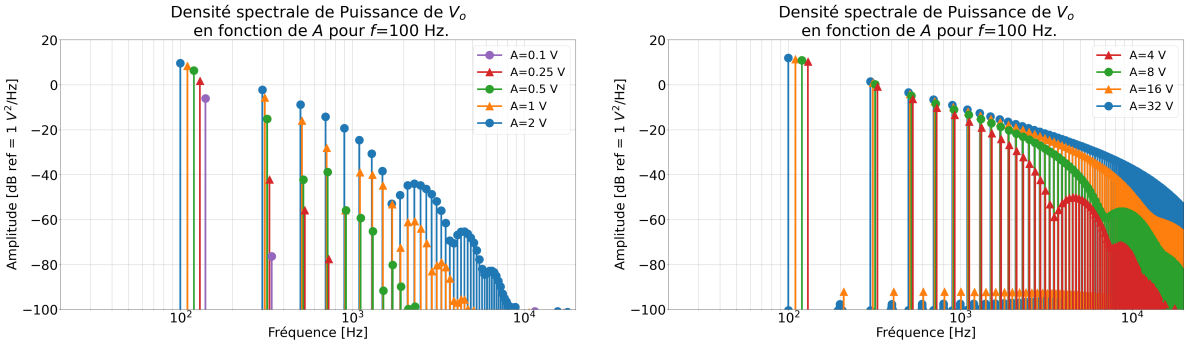
FIGURE 4 – Tracés de V_o en fonction du temps t et de l'amplitude A du sinus d'excitation de fréquence $f = 100$ Hz.

Le graphique met en évidence un aspect crucial : le seuil d'écrêtage n'est pas constant et est en fait influencé par l'amplitude du signal d'entrée de la pédale. En d'autres termes, la tension maximale que le signal de sortie peut atteindre fluctue en fonction de l'amplitude du signal entrant. Toutefois, la tension maximale une valeur limite, correspondant à la tension de seuil des diodes, qui est d'environ $v_{seuil} \approx 0,7$V pour und diode au silicum). Pour des amplitudes d'entrée relativement faibles (A=0.1 V et 0.25 V), le signal ne semble pas subir d'écrêtage, suggérant l'existence d'une zone linéaire. Dans cette zone, le signal reste exempt de modifications attribuables à l'écrêtage, mais peut subir des altérations du fait du filtre passe-bas (ce ne sont que des hypothèses). L'analyse ultérieure des spectres de puissance de ces deux signaux permettra de

confirmer ou d'infirmer l'hypothèse selon laquelle ces deux signaux (A=0.1 V et 0.25 V) subissent où non un effet de distorsion.

Les densités spectrales de puissances (DSP) associées aux signaux temporels de la Fig. 4 sont reportées Fig. 5. L'acquisition de ces signaux est réalisée sur une durée de 10 s pour obtenir une précision fréquentielle importante sur les DSP et observer les raies spectrales associées aux harmoniques générées de manière plus précise. Le code couleur diffère de celui de la Fig. 4 et les raies spectrales sont décalées successivement de 10 Hz pour chacun des tracés afin de faciliter la visualisation des raies en basses-fréquences.

Comme escompté, à mesure que l'amplitude du signal d'amplification augmente (cf. Fig. 4), le signal est de plus en plus écrêté, et donc distordu. Dans le domaine fréquentiel, cette distorsion se traduit par la génération d'harmonique impaires de rang de plus en plus élevées et d'amplitude de plus en plus importante à mesure que l'amplitude du signal temporel d'excitation est importante (cf. Fig. 5).



(a) Tracé pour des valeurs de $A \in [0.1 - 2]$ V. (b) Tracé pour des valeurs de $A \in [4 - 32]$ V.

FIGURE 5 – Tracé des DSP associées aux signaux temporels de la Fig. 4

Pour les densités spectrales de puissance (DSP) des signaux avec des amplitudes de A=0.1 V et 0.25 V, une légère distorsion est observable. Cela suggère que même à ces amplitudes, les signaux d'entrée semblent subir un effet d'écrêtage, bien que très modéré. Cependant, il n'est pas certain que la première harmonique générée sur le signal avec A=0.1 V soit réellement due à la distorsion dans le circuit, cell-ci pourrait être attribuable à d'autres facteurs, tels que l'amplificateur ou le format d'acquisition des données.

À noter que le rang des harmoniques générées est ici purement impaire. De manière intuitive, cela peut s'expliquer par le fait qu'à mesure que le signal d'entrée est d'amplitude élevée, celui-ci est de de plus en plus écrêté et la forme d'onde résultante se rapproche de l'allure d'une onde carrée, dont la décomposition en série de fourrier ne comprend que des harmoniques d'ordre impair.

Une exception est cependant faite pour A=16 et 32 V sur la Fig. 5(b), dont les DSP respectives révèlent des harmoniques de rang paires. Ces harmoniques sont néanmoins d'amplitude extrêmement faible. À ce stade, il est difficile d'identifier avec certitude la source de ces raies spectrales. Cela peut-être dû à la distorsion intrinsèque à l'amplificateur, ou bien à une multitudes d'autres facteurs au sein de la chaîne de mesure.

Le signal temporel V_o est maintenant tracé pour un signal d'excitation sinusoïdale de fréquence $f = 4$ kHz et différentes amplitudes A sur la Fig. 6.

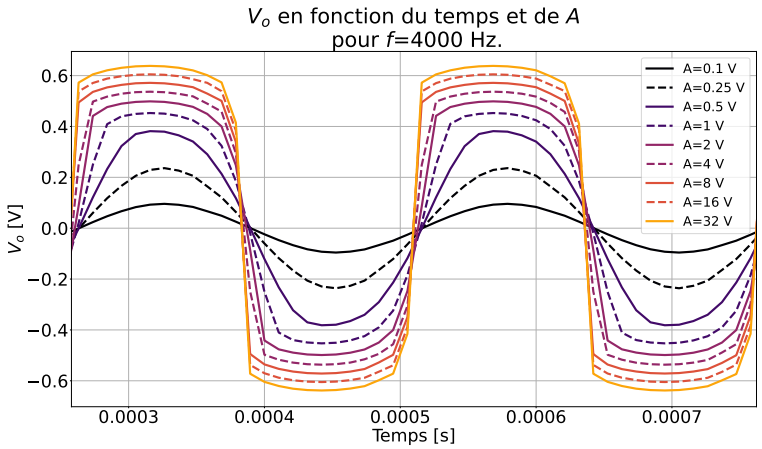


FIGURE 6 – Tracés de V_o en fonction du temps t et de l'amplitude A du sinus d'excitation de fréquence $f = 4$ kHz.

Ce tracé permet de constater que les valeurs des seuils d'écrêtage en fonction de l'amplitude A du signal d'entrée sont quasi-identiques à celles visualisées sur le tracé Fig. 4. À cette fréquence, il est intéressant de noter que chacun des lobes du signal décrit par V_o, pour des grandes valeurs de A, ne semble symétrique par rapport au moment ou celui-ci atteint son niveau maximal. Ce phénomène est dû au processus de charge et de décharge du condensateur qui n'est pas instantané et introduit une sorte de retard ou de "temps de réponse" dans le circuit. Dans le domaine fréquentiel, pour un filtre RC, ce retard se produit autour et au delà de la fréquence de coupure f_c du filtre : la phase de la fonction de transfert du filtre vaut 0 en dessous de f_c, $-\pi/4$ à la fréquence de coupure, et tend vers $-\pi/2$ dépassé f_c. Une analyse à une fréquence plus élevée, au-delà de la fréquence de coupure du filtre RC composant le circuit, permet de mettre en évidence ce phénomène de manière plus explicite. Le choix est fait de prendre une fréquence $f = 16$ kHz pour avoir 6 points de mesure par période avec une fréquence d'échantillonnage de 96 kHz, et obtenir un motif se répétant parfaitement au bout d'une seule période. Le signal d'entrée V_i pour $A = 2$V et $f = 16$ kHz est représenté Fig. 7.

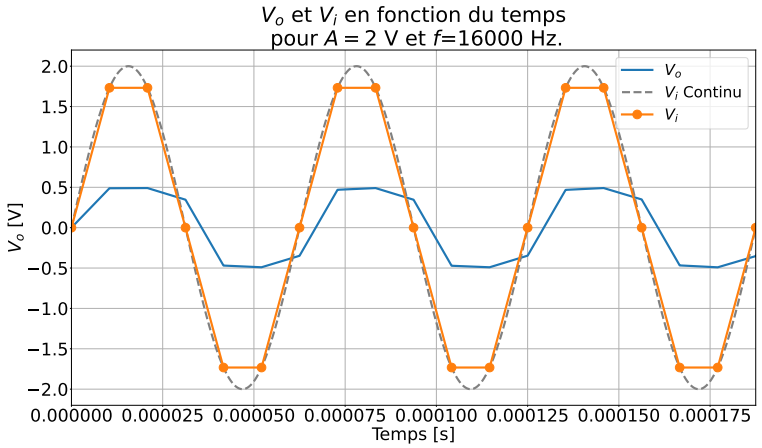


FIGURE 7 – Tracés de V_o et V_i en fonction du temps t pour $A = 2$ V et $f = 16$ kHz.

Sur ce graphique, le retard temporel de V_o par rapport au signal d'excitation V_i est d'autant plus visible. Non seulement le signal de sortie est écrêté, mais il semble que l'amplitude maximale de celui-ci soit plus petite comparée aux situations où la fréquence est plus basse avec la même amplitude d'excitation (voir 4 et 6). Le processus de charge et de décharge du condensateur

devient plus perceptible dans ce contexte. En effet, la forme d'onde observée se rapproche de celle typiquement visible lors de l'étude d'un circuit RC sous l'influence d'un signal carré de fréquence faible devant la fréquence de coupure du circuit. Cette similitude permet de mettre en évidence les phénomènes de régime transitoire et de régime permanent qui caractérisent le comportement du condensateur.

Les tracés temporels de V_o et de leur DSP associée, pour une excitation sinusoïdale V_i avec A=2 V, et plusieurs fréquences f d'excitations sont maintenant tracés Fig. 8. L'objectif de ce tracé est de montrer que pour une même amplitude et différentes fréquences d'excitations, les harmoniques générées pour un même rang donné ne sont pas identiques. La DSP est représentée en échelle semi-logarithmique sur l'axe des fréquences afin de garantir, indépendamment de la fréquence, un espacement constant entre deux harmoniques successives.

Sur la Fig. 8, ce principe est bien vérifié : l'amplitude des harmoniques de même rang n'est pas identique en fonction de la fréquence, même si les signaux temporels présentent une ressemblance très prononcée. En d'autres termes, même si les signaux à la sortie de l'étage d'écrêtage présentent une forme similaire dans le temps (relativement à la taille de leur période), leur densité spectrale de puissance peut varier significativement, la cascade harmonique n'a pas la même allure, révélant des nuances subtiles de la distorsion qui ne sont pas directement perceptibles dans le domaine temporel.

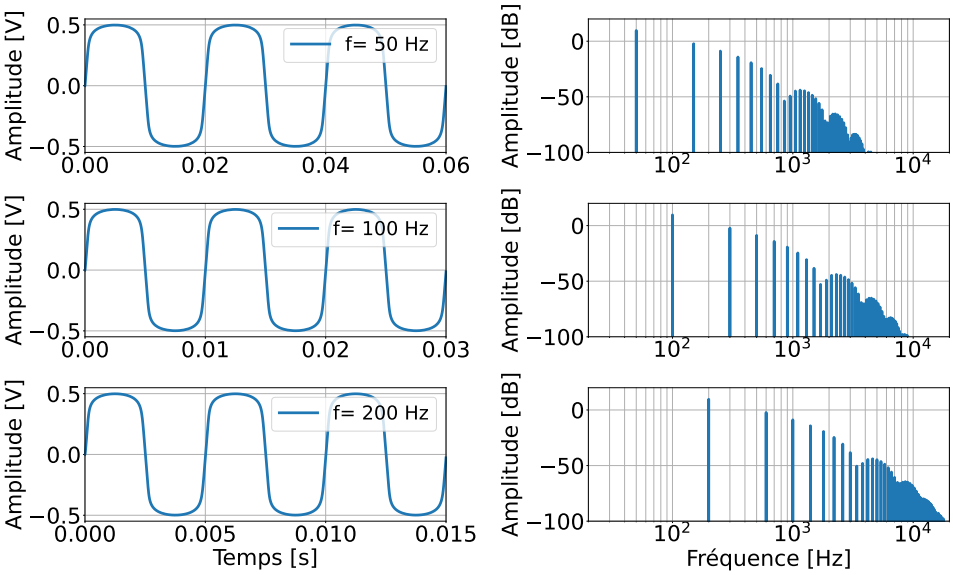


FIGURE 8 – Tracés temporels de V_o et de sa DSP associée, pour une excitation sinusoïdale de V_i avec A=2 V, et plusieurs fréquences f d'excitations.

Il est important de souligner qu'une constance semble se dessiner au niveau des amplitudes des premières harmoniques de rang identique pour une même tension d'excitation à différentes fréquences. C'est-à-dire que même si les fréquences d'excitation varient, l'amplitude des premières harmoniques de rang équivalent semble être la même. Ce constat pourrait suggérer que, pour ces premières harmoniques, l'effet de la distorsion ne varie pas avec la fréquence d'excitation, ce qui serait une information précieuse dans l'analyse du comportement du système. Cette observation nécessite néanmoins une validation supplémentaire, car elle se base uniquement sur une tendance visible dans les données actuelles. Pour visualiser ce constat d'une meilleure manière, la Fig. 9 affiche les tracés temporels de V_o et de leur DSP associée, pour une excitation sinusoïdale V_i avec A=2 V, et des fréquences f d'excitations plus élevées (1kHz, 2kHz et 4kHz).

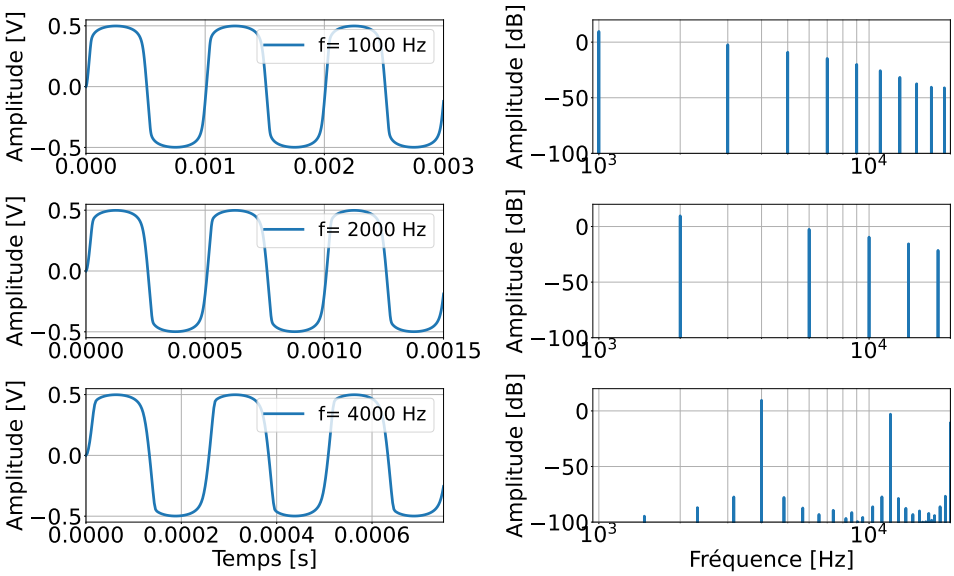


FIGURE 9 – Tracés temporels de V_o et de sa DSP associée, pour une excitation sinusoïdale de V_i avec A=2 V, et plusieurs fréquences f d'excitations.

En conclusion sur cette partie expérimentale, le signal d'écrêtage a le comportement attendu : il distord le signal d'entrée en générant une cascade harmonique. L'amplitude des harmoniques générées dépend fortement de l'amplitude du signal d'entrée. Les harmoniques générées sont de rang impair et le signal de sortie tend vers un signal carré lorsque le signal d'entrée est de grande amplitude. Par ailleurs, l'étude expérimentale a montré que la réponse du circuit dépend très peu de la fréquence du signal d'entrée hormis lorsque celle-ci est proche ou supérieure à la fréquence de coupure f_c du filtre RC ($f_c = 1/(2\pi RC) \approx 15kHz$). Ce dernier constat sera utile pour faire des simplifications du problème permettant une résolution en temps réel (cf. partie 7).

4 Mise en équation du circuit d'écrêtage

L'objectif est ici de modéliser le circuit analogique spécialement réalisé par l'intermédiaire d'une équation mathématique exprimant la relation entre la tension d'entrée, notée V_i, et la tension de sortie, notée V_o, de l'étage d'écrêtage. L'obtention de cette équation est démontrée ci-dessous. La Fig. 10 reprend le circuit d'écrêtage de la pédale MXR Distortion +.

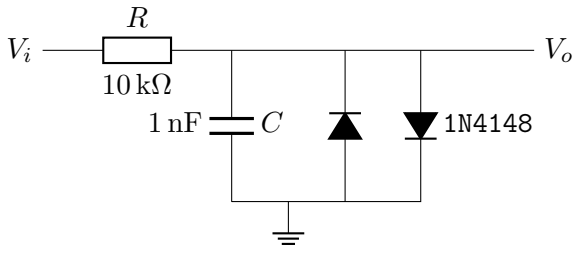


FIGURE 10 – Schéma électrique de l'étage d'écrêtage de la pédale d'effet MXR Distortion +.

À partir du schéma électrique Fig. 10 représenté ci-dessus, il est possible d'exprimer la tension de sortie V_o de ce circuit en fonction de la tension V_i à l'entrée de celui-ci, supposée connue. Pour ce faire, un modèle caractérisant une diode de jonction PN est nécessaire (l'anode correspondant au semi-conducteur dopé P et la cathode au semi-conducteur dopé N). Une approximation traditionnelle du premier ordre de la diode est une fonction linéaire par morceaux, qui est équivalente

à un modèle de commutateur. Un modèle d'ordre supérieur est choisi ici. L'équation de la diode de Shockley traduit cette jonction par la relation suivante :

$$I_d = I_s \left(e^{V/\eta V_t} - 1 \right), \tag{1}$$

où le courant I_d traversant la diode est fonction de la tension V à ses bornes. Le reste des variables, à savoir le courant de saturation I_s, la tension thermique V_t et le facteur de qualité η de la diode sont des paramètres qui dépendent du modèle de diode et pouvant être déterminés par mesure et/ou fournis par la documentation constructeur.

À l'aide de la loi de noeuds de Kirchhoff et de l'équation (1), les deux diodes tête-bêche peuvent être assimilés à un générateur de courant. Ce raisonnement est visualisé Fig. 11 et est démontré dans la partie suivante.

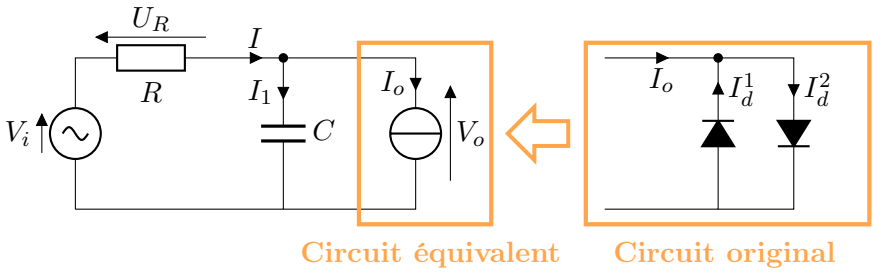


FIGURE 11 – Schéma électrique équivalent d'un étage d'écrêtage de pédale de distorsion.

En appliquant la loi des noeuds au circuit original au niveau des diodes tête-bêche (Fig. 11) et à partir de l'équation (1), les deux diodes peuvent être identifiées à un générateur de courant I_o :

$$I_o = I_d^2 - I_d^1 = I_s \left(e^{V_o/\eta V_t} - e^{-V_o/\eta V_t} \right) = 2 I_s \sinh \left(\frac{V_o}{\eta V_t} \right). \tag{2}$$

La loi des noeuds appliquée cette fois-ci au circuit équivalent (Fig. 11) permet ainsi d'obtenir l'équation :

$$I = I_o + I_1. \tag{3}$$

En injectant l'équation (2) et les expressions des courants affichées ci-dessous :

$$\begin{cases} I = \dfrac{U_R}{R} \text{ (loi d'Ohm)}, \\ I_1 = C \dfrac{\mathrm{d}V_o}{\mathrm{d}t} \text{ (relation aux bornes de } C), \end{cases}$$

puis en exprimant à l'aide de la loi des mailles la tension aux bornes de la résistance, soit $U_R = V_i - V_o$ au sein de l'équation (3), l'équation différentielle ordinaire (EDO) non-linéaire exprimant le lien entre les grandeurs V_o et V_i est ainsi obtenue :

$$\frac{\mathrm{d}V_o}{\mathrm{d}t} = \frac{V_i - V_o}{RC} - 2\frac{I_s}{C} \sinh \left(\frac{V_o}{\eta V_t} \right). \tag{4}$$

Cette EDO ne faisant pas partie des EDO non-linéaires pouvant être résolues de manière algébrique, seul des méthodes de résolutions numériques permettent la résolution de cette équation, afin d'exprimer le signal de sortie V_o en fonction des paramètres du système et du signal d'entrée V_i. Ces différentes méthodes sont abordées plus en détail dans la partie suivante. Il est important de noter qu'il est question ici d'équation différentielle ordinaire, c'est-à-dire ne dépendant que d'une seule variable, ici V_o.

5 Méthodes de résolution numériques en temps différé

Dans le but de résoudre l'EDO (4), plusieurs méthodes de résolution numériques, telles que Forward Euler, Backward Euler où Runge-Kutta peuvent être utilisées. Ces solveurs d'ODE utilisent l'intégration numérique pour résoudre l'EDO (4) de la forme :

$$\frac{\mathrm{d}v}{\mathrm{d}t} = v' = f(t, u(t), v(t)), \tag{5}$$

où $v(t)$ représente la sortie du système - la quantité à modéliser - en fonction du temps noté t, et où $f(t, u(t), v(t))$ est une fonction non linéaire déterminant la dérivée temporelle de $v(t)$, dépendante de l'état actuel de $v(t)$ et englobant l'entrée $u(t)$ du système. $u(t)$ est indépendant de $v(t)$ et supposée connue pour tout instant t. Dans le cas de l'équation (4), la tension de sortie du système V_o est représentée par la quantité $v(t)$ et la tension V_i à l'entrée du système est notée $u(t)$.

Les méthodes de résolutions numériques étant, par définition, des méthodes à temps discret, il convient d'exprimer les variables du système sous une forme discrétisée. Ici, les méthodes sont présentées en utilisant la notation en indice pour indiquer l'index temporel $n \in \mathbb{N}$ et la taille du pas d'échantillonnage est notée T, soit la version discrétisée du temps tel que $t_n = nT$, avec u_n la version discrétisée de $u(t)$ tel que $u_n = u(t_n) = u(nT)$ et $v_n = v(t_n) = v(nT)$.

Les sous-parties suivantes abordent différentes méthodes de résolution numériques de l'EDO (4) mais ne traitent pas en détail de l'obtention de celles-ci, et de l'expression de l'erreur de troncature locale et globale propre à chacune de ces méthodes. Pour plus de détails, se référer à l'annexe [A.3].

5.1 Euler explicite

En analyse numérique, les différences finies sont largement utilisées pour approximer les dérivées, et le terme "différence finie" est souvent utilisé comme abréviation de "approximation par différences finies de dérivées". La méthode d'Euler explicite (EE) peut être vue comme l'approximation par différence "en avant" de la dérivée (cf. annexe [A.4]), lorsque $T \ll 1$, conduisant à la relation :

$$v_{n+1} = v_n + T v'_n, \tag{6}$$

où :

$$v'_n = f(t_n,\, u_n,\, v_n) = \frac{u_n - v_n}{RC} - 2\frac{I_s}{C}\sinh\left(\frac{v_n}{\eta V_t}\right),$$

est la formulation discrétisée de l'équation (5) à l'échantillon n.

La méthode d'Euler explicite est une méthode explicite de premier ordre, puisque la sortie (v_{n+1}) dépend uniquement de l'état des étapes temporelles précédentes. À noter que la méthode d'Euler explicite peut être obtenue par différentes moyens (cf. annexe [A.4]). Cette méthode est l'une des plus simple à mettre en place, mais possède également la plus grande erreur de troncature, et présente une stabilité moindre pour le problème étudié, en comparaison avec les autres méthodes utilisées (cela sera vu par la suite).

5.2 Euler implicite

La méthode de Euler implicite (EI) peut être vu comme une approximation par différences finies "en arrière" de la dérivée (cf. annexe [A.5]), débouchant sur la relation :

$$v_{n+1} = v_n + T v'_{n+1}. \tag{7}$$

À noter que dans ce cas, v_{n+1} est présent dans le terme de gauche et celui de droite. Contrairement à la méthode d'Euler explicite, la grandeur recherchée v_{n+1} est reliée à une fonction qui dépend de cette même grandeur. Autrement dit, v_{n+1} est définie implicitement, d'où le nom de la méthode. Un solveur de racine comme Newthon-Raphson (cf. partie [5.5]) permet d'exprimer la tension de sortie v à l'échantillon $n+1$ pour laquelle l'équation (7) est vérifiée, en passant tous les termes de l'équation (7) d'un côté de l'équation, de façon à obtenir une fonction de type $y(v_{n+1}) = 0$ à résoudre. Cela signifie qu'il y a une boucle de résolution d'équation à l'intérieur de la boucle faisant avancer l'équation différentielle.

5.3 Runge-Kutta

La méthode de Runge-Kutta d'ordre 4 (RK4) est une des méthodes les plus répandues lorsqu'il s'agit de résoudre des EDO. En plus de présenter une précision satisfaisante, cette méthode est explicite et ne nécessite pas de ce fait l'utilisation d'un solveur de racine pour déterminer la valeur de v au n-ième échantillon. L'expression de v_{n+1} à l'aide de la méthode de Runge-Kutta d'ordre 4 est donnée par :

$$v_{n+1} = v_n + \frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6}, \tag{8}$$

avec :

$$\begin{cases} k_1 = T\, g(n,\, v_n), \\ k_2 = T\, g(n + \frac{1}{2},\, v_n + \frac{k_1}{2}), \\ k_3 = T\, g(n + \frac{1}{2},\, v_n + \frac{k_2}{2}), \\ k_4 = T\, g(n + 1,\, v_n + k_3), \end{cases}$$

où $g(m, v) = f(t_m, u_m, v)$ et $f(t, u(t), v(t))$ est telle que définie dans (5), donnant ainsi :

$$g(m, v) = \frac{u_m - v}{RC} - 2\frac{I_s}{C} \sinh\left(\frac{v}{\eta V_t}\right).$$

Il convient de noter que la méthode RK4 requiert des évaluations de fonctions à intervalles réguliers de demi-échantillons. En conséquence, l'entrée doit être fournie à un taux d'échantillonnage au minimum deux fois supérieur à celui de la sortie.

5.4 Méthode des trapèzes

L'expression de v_{n+1} à l'aide de la méthode des trapèzes implicite est donnée par :

$$v_{n+1} = v_n + \frac{T}{2}(v'_n + v'_{n+1}). \tag{9}$$

L'avantage de la méthode des trapèzes implicite est sa stabilité, puisque celle-ci produit des solutions numériques stables même pour des pas de temps relativement grands (une méthode est dite stable si l'accumulation d'erreurs dû aux erreurs successives permet tout de même de converger vers la solution). Cette stabilité est particulièrement utile pour les problèmes "raides", c'est-à-dire les problèmes où les solutions varient rapidement. L'erreur de troncature globale est également plus petite que pour les méthodes d'Euler explicite et implicite (cf. annexe [A.6]).

5.5 Newton-Raphson

La méthode de Newton-Raphson constitue dans son application la plus simple, un algorithme efficace pour trouver numériquement une approximation précise d'une racine d'une fonction réelle d'une variable réelle, autrement dit, une valeur x pour laquelle une fonction $f(x) = 0$. Cette méthode implique de partir d'une estimation initiale de la racine (valeur initiale) et de

déterminer la tangente à la courbe de la fonction à cet endroit. L'intersection de cette tangente avec l'axe des abscisses produit une meilleure approximation de la racine. Cette démarche est itérée jusqu'à ce que la racine soit obtenue avec une précision satisfaisante de sorte à ce que :

$$\begin{cases} x_0 = \text{valeur initiale}, \\ x_{i+1} = x_i - \dfrac{f(x_i)}{f'(x_i)}. \end{cases} \tag{10}$$

où x_{i+1} est la $(i+1)$-ième approximation de la racine. Ce schéma itératif est répété jusqu'à satisfaire un certain critère d'arrêt, qui peut être défini par rapport à une tolérance sur la valeur de la solution (si la valeur absolue de la fonction en l'estimation courante de la racine est inférieure à une tolérance prédéfinie), sur la mise à jour de la solution (si la différence entre deux itérations successives de la solution est inférieure à une certaine tolérance) ou sur le nombre maximum d'itérations (cf. annexe A.8).

6 Comparaison des méthodes de résolution

L'objectif de cette partie porte sur la pertinence de l'utilisation de différentes méthodes de résolution, dans un premier temps et de manière générale pour des simulations ou l'intérêt est porté sur la fidélité du résultat vis-à-vis du modèle physique, au détriment de la rapidité de calcul. Les simulations sont réalisées en utilisant **les valeurs expérimentales** du circuit d'écrêtage de la MXR Distortion +, soit $R = 9.9$ kΩ et $C = 1.1$ nF. La diode sélectionnée est la diode `1N4148`, dont le facteur de qualité est $\eta = 1.752$, et dont le courant de saturation vaut $I_s = 2.52$ nA. La température extérieure est fixée à 27°C, et la tension thermique vaut donc $V_t = 25.85$ mV.

6.1 Stabilité des méthodes

La stabilité d'une méthode de résolution numérique, essentielle pour prévenir l'accumulation d'erreurs numériques, est définie comme sa réaction aux perturbations. Si de légères perturbations de l'état initial ou sources d'erreur n'induisent que de faibles perturbations de la solution à tout autre moment, la méthode est qualifiée de stable. Cette stabilité peut cependant varier selon le problème à résoudre, soulignant l'importance du choix d'une méthode adaptée.

Les méthodes de résolution explicites, bien que faciles à mettre en œuvre et moins gourmandes en ressources numériques par itération que leurs homologues implicites, sont cependant plus limitées par des conditions de stabilité strictes. Ces dernières contrôlent la taille maximale du pas de temps utilisable. Pour des problèmes particulièrement non linéaires, ces conditions peuvent rendre l'application de ces méthodes explicites impossible. La vérification de ce principe est l'objet de cette sous-partie. L'intérêt est porté sur l'utilisation des méthodes explicites, et l'inconvénient que ces méthodes présentent dans notre cas, pour représenter fidèlement le signal en sortie du système, noté v_n.

Pour ce faire, un sinus d'amplitude A et de fréquence f est appliqué à l'EDO discrétisé 5, de manière à obtenir le système suivant à résoudre pour obtenir v_n, définit comme étant l'approximation numérique du signal de sortie analogique $V_o(t)$:

$$\begin{cases} u_n = A\sin(2\pi f t_n), \\ v_0 = u_0 = 0, \\ v'_n = f(t_n,\, u_n,\, v_n) = \dfrac{u_n - v_n}{RC} - 2\dfrac{I_s}{C}\sinh\left(\dfrac{v_n}{\eta V_t}\right), \end{cases} \tag{11}$$

où $t_n \in [0 \;-\; T(N-1)]$, avec $T = 1/F_s$, où F_s représente la fréquence d'échantillonage et N le nombre total d'échantillon. Ce système peut être résolu à l'aide des méthodes évoquées

précédemment partie 5. Le signal d'excitation choisi noté u_n est un sinus d'amplitude A=4.5 V, de fréquence $f = 192$ Hz. Ce signal est échantillonné à $F_s = 48$ kHz.

La Fig. 12 affiche les résultats des tracés temporels de la simulation sur un intervalle de temps $t \in [0\ s - 0.015\ s]$, et les DSP associées à ces signaux, sur une échelle fréquentielle linéaire. Ces DSP sont estimées sur un intervalle de temps de 10 s. Les simulations sont effectuées sur une version suréchantillonnée du signal originel, obtenue en utilisant la fonction `signal.resample` de `scipy` sur Python. Ce taux de suréchantillonnage est noté "Os" sur la Fig. 12.

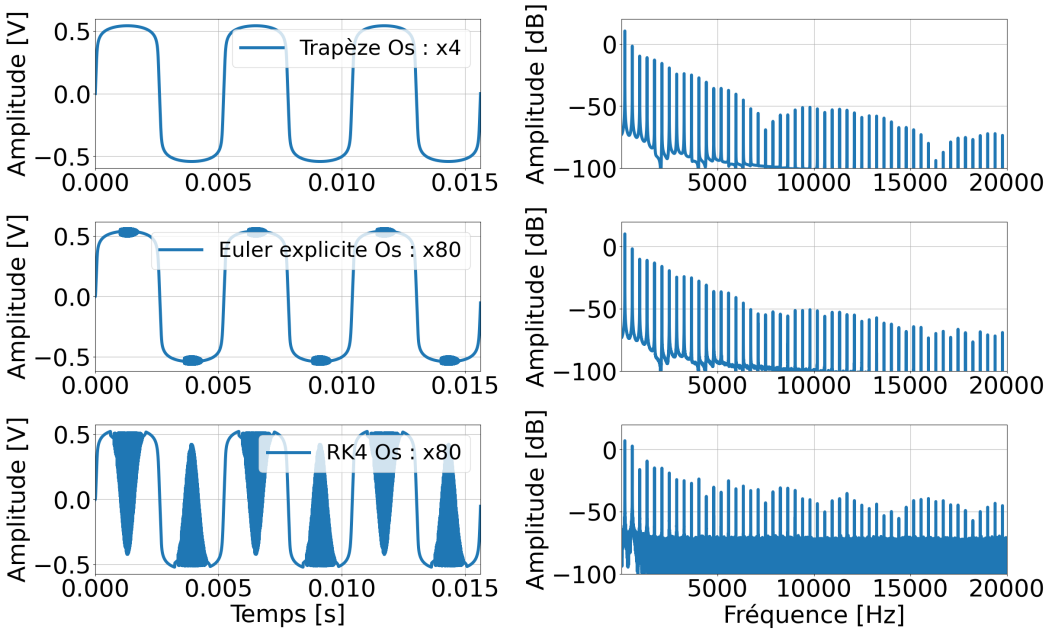


FIGURE 12 – Tracés temporels de v_n et DSP associée pour une excitation sinusoïdale de u_n, avec $A = 4.5$ V, $f = 192$ Hz, obtenues à l'aide de la méthode des trapèzes, de Euler explicite et RK4.

Sur cette figure, le suréchantillonnage est dans un premier temps utilisé pour limiter l'impact du repliement fréquentiel dû aux harmoniques générées au-delà de la fréquence maximale respectant le critère de Shannon ($Fs/2$). En effet, la méthode de résolution numérique est ici effectuée dans le domaine temporel, ce qui rend impossible l'application d'un filtre anti-repliement sur le signal simulé, puisque celui-ci contient déjà des artefacts lié au repliement fréquentiel en son sein. Une solution consiste donc à suréchantillonner le signal d'entrée et à effectuer la simulation sur ce nouveau signal, de manière à "repousser" $F_s/2$ plus loin et ainsi obtenir un repliement qui survienne plus tard sur l'échelle fréquentielle.

Ici, un suréchantillonnage de 80 fois est utiliser sur les méthodes explicites pour démontrer le haut niveau de suréchantillonnage nécessaire pour une simulation présentant un semblant de stabilité du système.

La Fig. 12 montre les résultats obtenues avec l'utilisation des méthodes explicites, comparés à l'utilisation de la méthode du trapèze, avec un suréchantillonnage de 4 fois. Pour les méthodes explicites, la forme d'onde temporelle suit la courbe de la solution mais devient instable au moment du processus d'écrêtage, lorsque la diode a une conductance élevée. Il est probable que cette instabilité soit due à la nature plate du signal à ce moment précis. En effet, la dérivée du signal de sortie tend vers zéro, ce qui semble poser problème pour les méthodes explicites. Les spectres des méthodes explicites ressemblent à ceux de la solution obtenue avec la méthode des

trapèzes dans les basses fréquences, mais l'instabilité se manifeste sous la forme d'un plancher de bruit à large bande qui sonne particulièrement désagréable pour la méthode RK4 et qui ne peut pas être éliminé par un simple filtrage. Pour la méthode d'Euler explicite, cette instabilité se manifeste sous la forme d'un excédent d'amplitude de certaines des harmoniques visibles en plus hautes fréquences, comparativement à l'amplitude des harmoniques calculées par la méthode des trapèzes.

Au vu de ce schéma, il ne faut pas supposer d'emblée que les solutions obtenues avec RK4 sont moins stables que celles obtenues avec Euler explicite, puisque cela dépend en grande partie du type de signal d'excitation qui entre dans le système. Il est également important de noter que malgré un suréchantillonnage de 80 fois avec la méthode RK4, la sortie ne contient que 40 fois le nombre d'échantillons initiaux. Cela s'explique par le fait que RK4 doit évaluer le signal à des demi-échantillons. Pour ce faire, les demi-échantillons sont obtenues en appliquant la méthode de RK4 sur un intervalle avec deux fois moins de points que le signal originel, de manière à prendre tout les échantillons impaires du signal originel pour constituer les valeurs à $n + \frac{1}{2}$, et les échantillons paires du signal originel pour constituer les valeurs à $n + 1$ (cf. équation 8).

La Fig. 13 affiche les tracés temporels de v_n ainsi que leur DSP associée pour une excitation sinusoïdale de u_n, avec $A = 4.5$ V, $f = 192$ Hz, obtenues à l'aide de la méthode d'Euler explicite, pour différents facteurs de suréchantillonage.

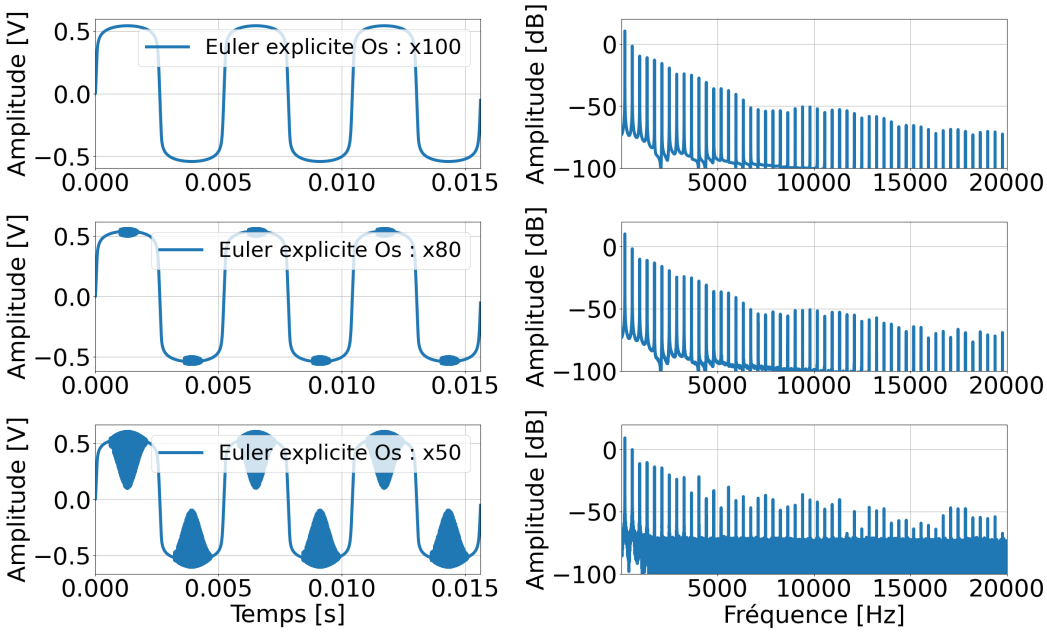


FIGURE 13 – Tracés temporels de v_n et DSP associée pour une excitation sinusoïdale de u_n, avec $A = 4.5$ V, $f = 192$ Hz, obtenues à l'aide de la méthode d'Euler explicite, pour différents facteurs de suréchantillonage.

Cette figure illustre le fait que de manière générale, la stabilité d'une méthode dépend grandement du taux d'échantillonnage. Plus le pas est faible, et plus l'erreur de troncature commise à chaque itération de l'algorithme est faible, et moins celle-ci est susceptible de perturber la stabilité de la méthode utilisée.

À la suite de ce travail, il a été conclu que les méthodes de résolution explicite n'était pas assez stables pour satisfaire les besoins de simulations à des taux de suréchantillonnage raisonnable ($\times 2$, $\times 4$, ..., $\times 32$), et pour des fréquences d'excitation élevées. Par contraste, les deux méthodes

de résolution implicites examinées ont démontré une stabilité des solutions pour des fréquences couvrant l'ensemble du spectre audible, sans nécessiter de suréchantillonnage.

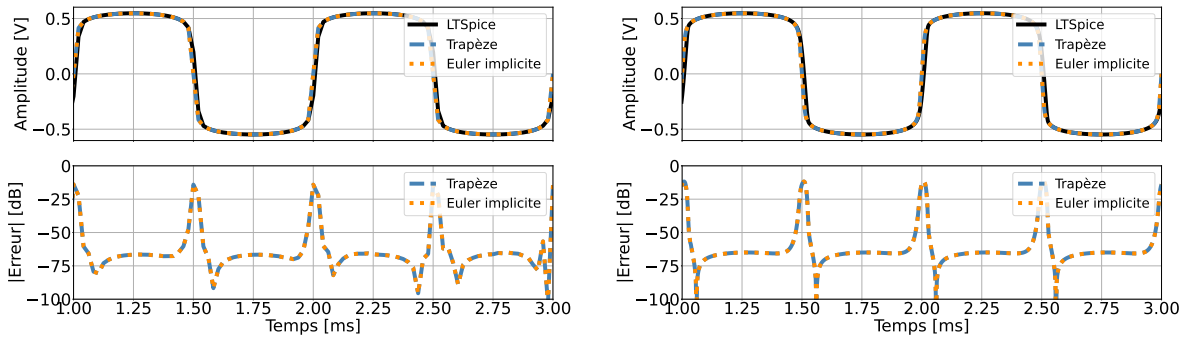
6.2 Comparaison des algorithmes avec LTspice

Dans la suite de cette partie, le logiciel LTspice XVII est utilisé. LTspice est un outil de simulation de circuits électroniques analogiques haute performance, utilisé par un grand nombre d'entreprises dans le monde de l'ingénierie électronique. Dans cette étude, le logiciel sert de référentiel quant à l'exactitude des méthodes numériques utilisées pour simuler le circuit d'écrêtage. Cet outil a la capacité d'importer et d'exporter des fichiers WAV contenant les données souhaitées par l'utilisateur (tensions, courants...) en utilisant l'intégration par une méthode des trapèze modifiée, ou à l'aide de la méthode GEAR (qui correspond aux formules de différentiation vers l'arrière). La tension aux bornes du générateur branché au circuit d'écrêtage est défini dans LTspice, et un fichier WAV encode en 16 bits la tension en sortie de circuit.

Cette comparaison vérifie les méthodes utilisées dans ce travail. Une correspondance exacte dans le domaine temporel avec LTspice ne peut être attendue en raison de différences dans les critères de convergence et dans la manipulation numérique. En effet, LTspice utilise une taille de pas adaptative pour contrôler l'erreur, et applique une interpolation linéaire pour trouver les valeurs manquante lors de l'encodage du fichier sonore WAV.

Le choix est fait de s'intéresser à des signaux sinusoïdaux dont les fréquences pourrait être celles de fondamentales de notes de guitare. Les fréquences fondamentales dans la plage typique jouable par une guitare électrique couvrent une plage d'environ 80 Hz à 1200 Hz.

La Fig. 14 présente différentes versions de $v[n]$ obtenues avec la méthode d'Euler explicite, la méthode des trapèzes et LTspice. Ces résultats sont produits pour un signal d'excitation $u[n]$ sinusoïdal de fréquence $f = 1000$ Hz et d'amplitude $A = 5$V, échantillonné à 48 kHz. Les simulations ont été réalisées sans suréchantillonnage et avec un facteur de suréchantillonnage Os $= \times 8$. Cette figure illustre également l'erreur absolue commise à chaque itération par les méthodes par rapport à la simulation effectuée avec LTspice.



(a) Tracé sans suréchantillonnage.

(b) Tracé avec Os $= \times 8$.

FIGURE 14 – Tracés temporels de v_n obtenues à l'aide des méthodes de résolution implicites et erreur absolue associée aux méthodes d'Euler implicite et du trapèze vis-à-vis de la simulation sur LTspice.

En termes qualitatifs, les tracés temporels obtenus par la méthode des trapèzes et la méthode d'Euler implicite semblent presque identiques, que le suréchantillonnage soit utilisé ou non. L'erreur absolue par rapport à la simulation fournie par LTspice est plus petite avec l'utilisation d'un signal d'entrée suréchantillonné pour les deux méthodes. Les différences mineures entre les deux tracés temporels peuvent être discernées en examinant l'erreur : les algorithmes implémentés

pour ce projet présentent une erreur plus marquée par rapport à LTspice au niveau du passage du sinus à chaque demi-alternance subséquente (passage au zéro).

Pour cette fréquence spécifique, l'application d'un facteur de suréchantillonnage ne semble pas particulièrement bénéfique, compte tenu du peu de différences observables dans la variation du tracé de l'erreur avant et après suréchantillonnage.

Pour les fréquences plus élevées, l'application du suréchantillonnage apparaît nécessaire afin de minimiser autant que possible l'impact du repliement spectral dans le spectre audible. Cette notion est mise en évidence à la Fig. 15, où les densités spectrales de puissance des signaux obtenus avec Euler implicite, la méthode des trapèzes et LTspice sont présentées pour un signal d'excitation sinusoïdale $u[n]$ de 5 V à 15000 Hz.

La figure 15 révèle que le suréchantillonnage atténue effectivement certaines raies spectrales indésirables, résultantes du repliement des harmoniques générées au-delà de la fréquence de Nyquist.

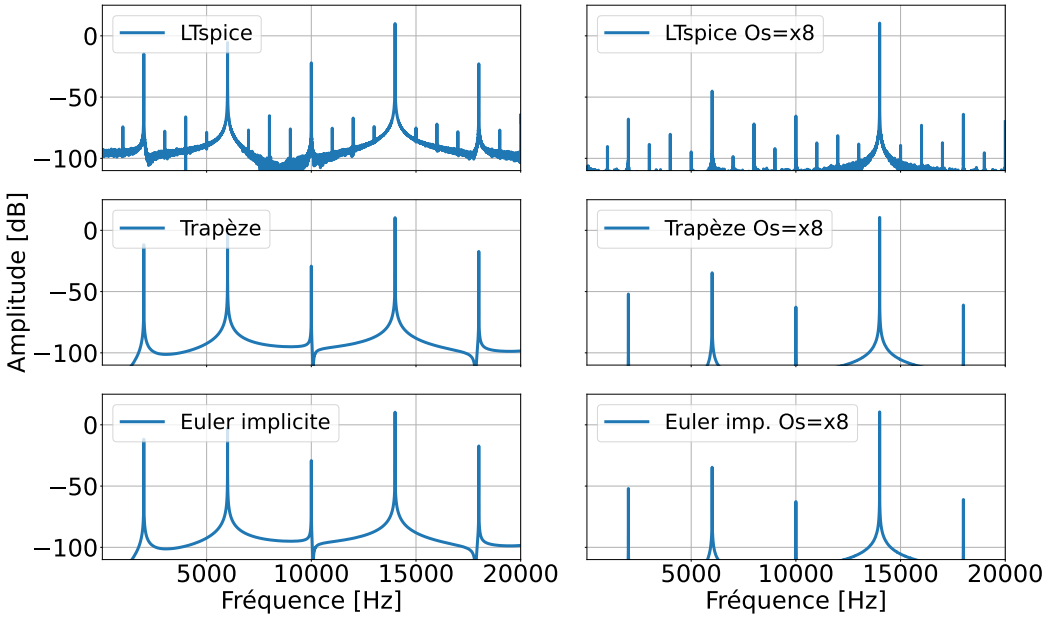


FIGURE 15 – DSP associées aux signaux $v[n]$ obtenus par Euler implicite, la méthode des trapèzes et la simulation LTspice, pour un signal d'entrée $u[n]$ d'une amplitude $A = 5$V et de fréquence $f = 14$ kHz, avec et sans suréchantillonage (Os = $\times 8$), pour $Fs = 48$ kHz.

Les tracés des DSP obtenues avec la méthode d'Euler implicite et la méthodes des trapèzes sont visuellement identiques peut importe la fréquence d'échantillonnage. En réalité, la superposition des tracés indique des différences au niveau de la hauteurs des différentes raies spectrales, mais celle-ci sont de l'ordre du centième de dB.

Un autre point important est mis en évidence : les simulations sur LTspice exposent davantage de raies spectrales à des fréquences qui ne peuvent être associées au repliement des harmoniques de $f = 14000$ kHz, et du bruit est observable sur les spectres. Bien que les niveaux impliqués pour ces deux phénomènes soient mineurs, cela démontre une différence notable entre les algorithmes développés pour ce projet et la manière dont LTspice résout le système. Il est possible que LTspice n'utilise pas le modèle de diode de Schockley, préférant un modèle plus avancé. À ce stade, il est difficile de déterminer précisément l'origine de cette différence.

En outre, les méthodes d'Euler et du trapèze offrent des approches stables et faciles pour résoudre l'équation différentielle (EDO) 5 dans un contexte discret. Une comparaison avec LTspice met en évidence une correspondance remarquablement proche entre les résultats obtenus par les méthodes du trapèze et d'Euler implicite, tant dans le domaine temporel (voir Fig. 14) que fréquentiel (voir Fig. 15). La prochaine section de cette étude s'appuiera sur des résultats expérimentaux pour comparer avec ceux obtenus par la méthode des trapèzes. L'objectif est de détecter d'éventuelles divergences significatives entre la théorie et l'expérience, spécifiquement dans l'étude du circuit d'écrêtage présenté dans ce projet

6.3 Comparaison des méthodes de résolution vis-à-vis du modèle expérimental

Les comparaisons sont effectuées entre la méthode des trapèze, et les données expérimentales. Les paramètres de simulations sont ici fixées par le contexte d'expérimentation. Comme précédemment, les simulations sont effectuées avec les mêmes valeurs des composants mesurées dans la partie expérimentale.

Différents tracés sont affichés, pour des fréquences et tensions arbitraires. Les signaux expérimentaux sont relevés selon le même protocole expérimental qu'évoqué dans la partie 3.1, à une fréquence de 96 kHz. Les signaux théoriques sont estimés à partir de la méthode des trapèze à une fréquence d'échantillonnage de 394 kHz (4×96kHz). La Figure 16 compare à la fois le tracé temporel et la DSP du signal de sortie V_o de la pédale, à la simulation du même signal $v[n]$. Ceci est réalisé pour une excitation sinusoïdale de fréquence $f = 200$ Hz et d'amplitude $A = 2$ V. Les raies des DSP associées aux signal expérimentaux sont volontairement décalées de 100 Hz pour $f = 200$ Hz et de 200 Hz pour $f = 800$ Hz afin de faciliter la lecture.

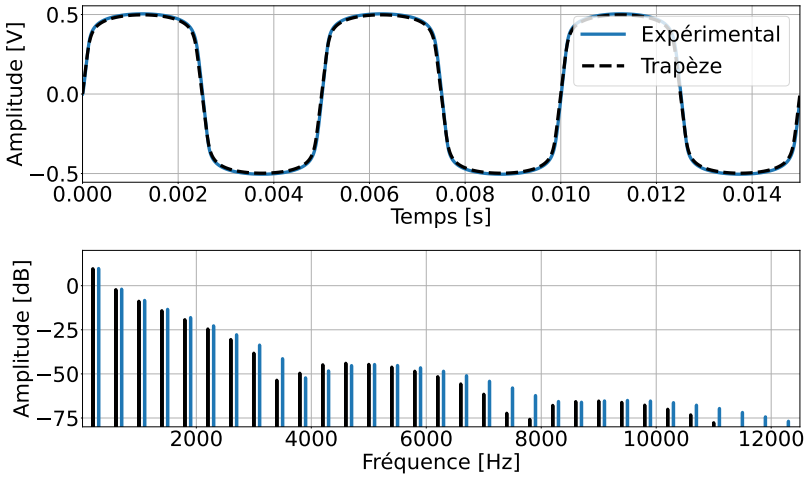


FIGURE 16 – Comparaison du signal de sortie expérimental V_o (tracé temporel et DSP) avec sa simulation $v[n]$ pour une entrée sinusoïdale de fréquence $f = 200$ Hz et d'amplitude $A = 2$ V.

Le tracé 17 met en comparaison le signal en sortie de pédale V_o par rapport à la simulation de ce même signal notée $v[n]$, pour une excitation sinusoïdale de fréquence $f = 800$ Hz et d'amplitude $A = 1$ V.
De manière générale, les graphiques temporels et spectraux obtenus expérimentalement présentent de nombreuses similitudes avec ceux obtenus par simulation. Des divergences mineures sont toutefois observables sur les tracés temporels pour $f = 200$ Hz et $f = 800$ Hz. Dans le cas de $f = 200$ Hz, les lobes formés par les raies spectrales sont reproduits, mais avec une largeur différente. Pour $f = 800$ Hz, la tension maximale excède celle obtenue en simulation. Cette différence pourrait résulter d'une gestion incorrecte du gain sur l'amplificateur à cette fréquence

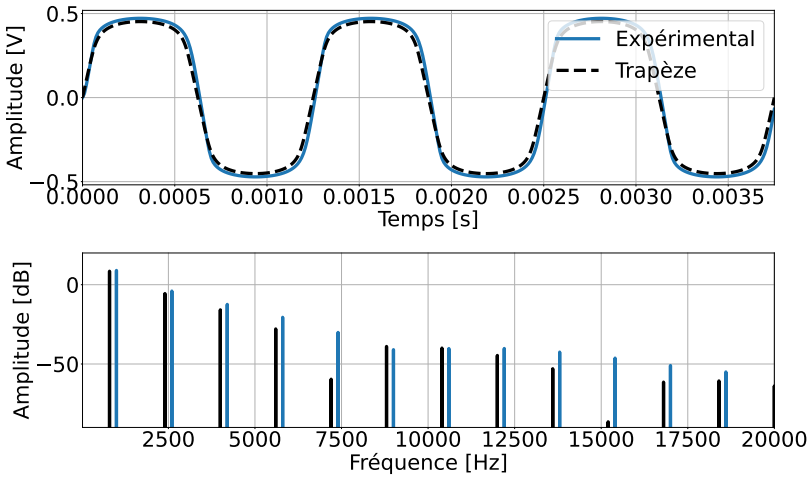


FIGURE 17 – Comparaison du signal de sortie expérimental V_o (tracé temporel et DSP) avec sa simulation $v[n]$ pour une entrée sinusoïdale de fréquence $f = 800$ Hz et d'amplitude $A = 1$ V.

spécifique. En effet, l'analyse expérimentale a révélé que les amplitudes de seuil ne semblaient pas dépendre de la fréquence à basse fréquence puisque le filtre RC n'agit pas sur le système. Plusieurs sources d'incertitude peuvent influencer les résultats obtenues pour ces deux figures, à savoir la précision de la mesure de la tension de sortie de l'amplificateur V_i par l'oscilloscope, la température ambiante qui détermine la tension thermique V_t, et enfin, les valeurs des composants déterminées expérimentalement.

7 Méthode de résolution en temps réel

7.1 Contexte

Dans le monde de la musique assistée par ordinateur, les plugins, des micro-programmes utilisables au sein de logiciels de production de musique, ont bouleversé la manière de créer et de produire cell-ci. Ces plugins ont la capacité d'émuler les caractéristiques analogiques de divers appareils électroniques, tels que les pédales d'effets de guitare. Par conséquent, un musicien équipé d'un ordinateur, d'une carte son et d'un instrument peut jouer et s'enregistrer en direct avec le son typique de ces appareils, sans avoir à se préoccuper de l'achat et de la maintenance de matériel physique souvent coûteux.

Malgré ces progrès technologiques, reproduire fidèlement les nuances et les complexités des pédales d'effets demeure un défi, en particulier dans le contexte du traitement audio en temps réel. Les méthodes de résolution numériques classiques utilisées au cours de ce projet, bien qu'ayant prouvé leur efficacité dans la simulation de ces effets, se révèlent coûteuses en termes de calculs et ne sont pas adaptées à une utilisation en temps réel.

En effet, pour offrir une expérience utilisateur optimale, l'émulation de l'effet de la pédale doit se faire en temps réel, sans latence perceptible entre le moment où l'action de jouer est réalisée par le musicien et le moment où le son est retransmis. Ceci nécessite un algorithme qui non seulement reproduit fidèlement l'effet de la pédale, mais qui le fait également de manière efficace et rapide.

Ainsi, l'objectif est de développer un algorithme capable de fournir des résultats satisfaisants, simulant fidèlement l'effet d'écrêtage de la pédale de distortion, tout en minimisant le coût en calculs numériques. Cela passe par une simplification du modèle étudié précédemment.

7.2 Approche Statique

À partir des constats posées dans la partie expérimental, il a été constaté que l'amplitude des harmoniques générées dépend fortement de l'amplitude du signal d'entrée et que ces harmoniques sont de rang impair. De plus, l'étude expérimentale a révélé que la réponse du circuit dépend très peu de la fréquence du signal d'entrée, sauf lorsque celle-ci est proche ou supérieure à la fréquence de coupure f_c du filtre RC. En se basant sur ce dernier postulat, une première simplification du problème initial peut être envisagée, visant à négliger la dépendance fréquentielle dans le circuit. Négliger la dépendance revient à négliger l'effet mémoire du circuit, et la fonction résultante de cette simplification du modèle doit donc être une simple fonction à appliquer au signal temporel d'entrée pour obtenir le signal de sortie.

La non-linéarité utilisée est l'approximation en courant continu (DC) de la non-linéarité réelle, générée à partir de l'ODE en fixant la dérivée temporelle de la tension de sortie de l'équation 5 à zéro [5], tel que :

$$\frac{\mathrm{d}V_o}{\mathrm{d}t} = \frac{V_i - V_o}{RC} - 2\frac{I_s}{C}\sinh\left(\frac{V_o}{\eta V_t}\right) = 0. \tag{12}$$

Dans ce contexte, l'approche statique fait référence à l'annulation de la dérivée. Cette équation est résolue à l'aide de la méthode de Newton-Raphson pour différentes valeurs discrètes de V_i, ce qui permet de construire une fonction discrète reliant la tension de sortie V_o à la tension d'entrée V_i.

La Fig. 18 présente cette fonction discrète obtenue à l'aide de Newton-Raphson, pour une plage de valeurs de tension d'entrée allant de -5 V à 5 V.

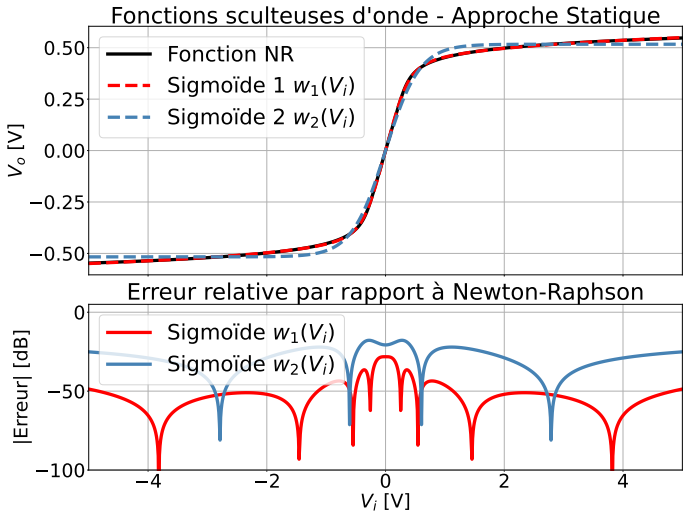


FIGURE 18 – Fonctions sculteuses de l'onde obtenues à l'aide de Newton Raphson (NR), et à l'aide de la méthode d'ajustement de la courbe avec des fonctions sigmoïdes. Représentaion de l'erreur de ces ajustement vis-à-vis de la courbe de référence obtenue avec NR.

D'un point de vue pratique, la manipulation de cette fonction discrète se révèle complexe lorsqu'il s'agit de déterminer des valeurs d'échantillons de sortie non préalablement définies. Pour pallier ce problème, une option est d'utiliser une interpolation entre les différents points pour transformer cette courbe en une version continue. Une autre stratégie, choisie dans ce contexte, est de rendre la fonction continue en ajustant la courbe à une fonction prédéfinie et en modifiant les coefficients de cette fonction théorique afin qu'elle s'aligne le plus possible sur la courbe obtenue à l'aide de Newton-Raphson.

L'observation de la courbe discrète issue de la méthode Newton-Raphson révèle une allure de sigmoïde. Ainsi, l'ajustement de cette courbe à l'aide de fonctions sigmoïdes est envisagé, en introduisant des coefficients aux termes de ces fonctions sigmoïdes afin de déterminer quelle combinaison de coefficients offre le meilleur ajustement. Les fonctions sigmoïdes utilisées sont inspirées du développement limité à l'ordre 1 de la fonction tangente hyperbolique et de la fonction logistique tels que :

$$w_1(x) = \frac{ax}{(1+|cx|^n)^{\frac{b}{n}}}, \tag{13}$$

et :

$$w_2(x) = d + \frac{f}{1+e^{-k(x-x_0)}}, \tag{14}$$

où a, b, c, n, d, f et x_0 sont les coefficients déterminés par ajustement de la courbe. Ces coefficients sont déterminés à l'aide de la fonction `curve_fit` de la librairie `scipy.optimize`. Ainsi, les ajustements de ces deux fonctions w_1 et w_2 par rapport à la courbe discrète obtenue avec Newton-Raphson sont tracés Fig. 18, avec les erreurs relatives de ces fonctions par rapport à la courbe obtenue avec Newton-Raphson. Les valeurs des coefficients déterminés avec `scipy` sont reportées Tab. 1.

	w_1		w_2
a	1.0377426	d	1.0336885
b	0.8923303	f	-3.699939e-7
c	2.4771343	x_0	3.509624
n	3.1341134		

TABLE 1 – Valeurs des coefficients associées aux sigmoïdes pour l'ajustement à la courbe obtenue à l'aide de Newton-Raphson.

Au vu des résultats de la Fig. 18, la sigmoïde w_1 semble davantage coïncider avec la courbe obtenue par Newton-Raphson. Cette fonction est utilisée par la suite pour étudier le comportement du sculpteur d'onde vis-à-vis de la méthode des trapèzes, qui servira de référence.

7.3 Comparaison vis-à-vis de la méthode des trapèzes

Les comparaisons des tracés temporels et fréquentiels, entre la méthode du sculpteur d'onde obtenue par l'approche statique et la méthode des trapèzes, sont réalisées pour deux fréquences et deux tensions d'excitation différentes. L'objectif est de déterminer si l'approche statique est fidèle au modèle théorique développé dans la partie 4, résolu de manière numérique avec la méthode des trapèzes. Le choix est fait de sélectionner des fréquences cohérentes avec la plage fréquentielle sur laquelle la plupart des instruments émettent une fondamentale [50 Hz - 4000 Hz]. Les comparaisons sont effectuées pour $f = 200$ Hz sur la Fig. 19, et $f = 3$ kHz sur la Fig. 20. Les simulations sont effectuées à une fréquence d'échantillonnage de $F_s = 8 \times 48000$=384 kHz. Les raies spectrales pour la méthode des trapèzes sont décalées de 200 Hz pour faciliter la lecture.

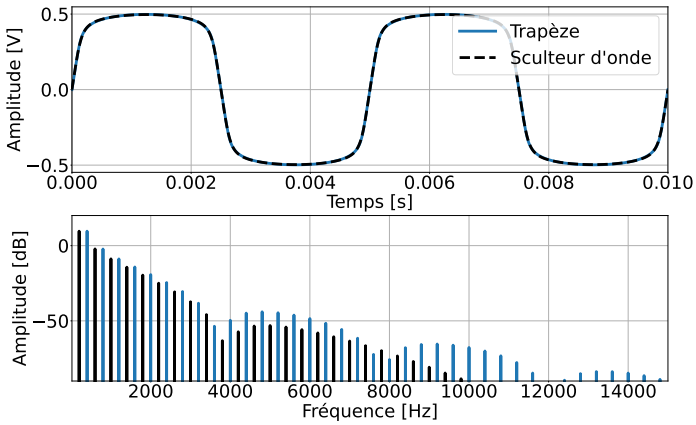


FIGURE 19 – Comparaison du signal de sortie théorique obtenue par la méthode des trapèze avec la méthode du sculpteur d'onde pour une entrée sinusoïdale de fréquence $f = 200$ Hz et d'amplitude $A = 2$ V.

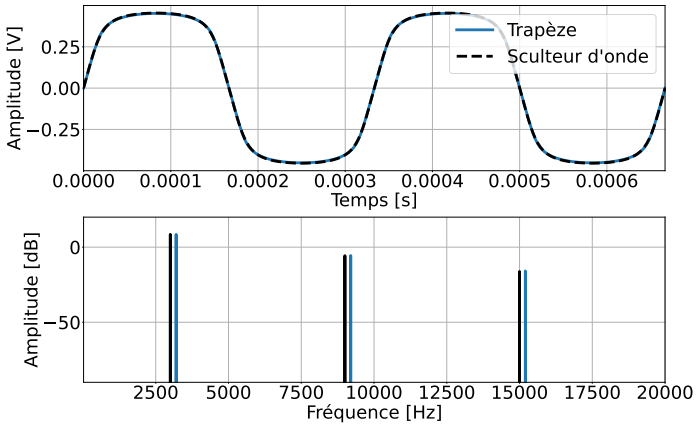


FIGURE 20 – Comparaison du signal de sortie théorique obtenue par la méthode des trapèze avec la méthode du sculpteur d'onde pour une entrée sinusoïdale de fréquence $f = 3$ kHz et d'amplitude $A = 1$ V.

Comme illustré dans les deux figures ci-dessus, les formes temporelles obtenues avec le sculpteur d'ondes coïncident presque parfaitement avec celles obtenues à l'aide de la méthode des trapèzes pour ces deux fréquences. Les profils spectraux sont également similaires, bien que certaines différences apparaissent en termes d'amplitude des harmoniques de rang supérieur, comme illustré sur la Fig. 19, pour des harmoniques impaires de rang supérieur à 6.

Cependant, à 3000 kHz (voir Fig. 20), seulement deux harmoniques impaires sont audibles. Étant donné qu'il s'agit d'harmoniques de rang faible, le sculpteur d'onde reproduit ces harmoniques avec une amplitude similaire à celle obtenue avec la méthode des trapèzes.

En général, l'approche statique montre ses limites lorsqu'il s'agit de reproduire fidèlement l'amplitude des harmoniques de rang élevé. Cependant, puisque l'amplitude de ces harmoniques est généralement assez faible, cela ne pose pas de problème significatif d'un point de vue perceptif. Il semblerait, de ce fait, impossible de différencier à l'oreille un son émulé par l'une ou l'autre des méthodes. L'utilisation du sculpteur d'onde pour émuler en temps réel le système d'écrêtage d'une pédale de distorsion semble être une excellente stratégie.

7.4 Conception d'un plugin audio

Dans le cadre de cette recherche sur une méthode de résolution en temps réel, un plugin dédié a été spécifiquement conçu et développé pour ce projet. Ce plugin est l'incarnation pratique de l'étude menée en utilisant une approche statique : il utilise un sculpteur d'onde pour modéliser le signal de sortie de l'étage d'écrêtage. L'interface utilisateur du plugin est présentée Fig. 21.

FIGURE 21 – Image du plugin StaticClipper.

Pour des question d'intégration aux logiciels de production musicale actuels (Pro Tools, FL Studio, Ableton), le logiciel a été intégralement codé en C++, en utilisant la librairie JUCE Framework, conçu pour le développement de ce genre d'outil. Le code source du plugin est disponible sur un dépot github [3].

Le plugin est composé de plusieurs éléments interactifs, dont des boutons rotatifs et des menus déroulants. Deux boutons rotatifs, "Input Gain" et "Output Gain", permettent de contrôler les gains appliqués aux signaux avant et après l'effet de distorsion, respectivement. Le menu "Oversampling" offre la possibilité de régler le taux de suréchantillonnage. Ce paramètre peut être désactivé ou réglé sur des valeurs multiples de deux, allant de $\times 2$ à $\times 8$. De plus, un curseur "Mix" permet de régler la proportion du mélange entre le signal d'entrée et le signal de sortie. Ceci offre à l'utilisateur la possibilité d'ajuster la quantité d'effet de distorsion appliqué au signal original. Enfin, un menu déroulant situé en haut du plugin permet de sélectionner le type de sculpteur d'onde à utiliser. Plusieurs options sont disponibles, y compris le sculpteur d'onde w_1 (cf. équation 14) et d'autres formes de sigmoïdes plus ou moins conventionnelles.

8 Conclusion

Dans le cadre de ce travail, une étude approfondie de la modélisation d'un circuit d'écrêtage à diode présent dans les pédales de distorsion de guitare a été réalisée. L'objectif initial était d'élucider les propriétés fondamentales de ces circuits de manière expérimentale et de déterminer une approche appropriée pour leur modélisation. Dans un premier lieu, le but a été de définir la méthode de résolution numérique la plus à même de résoudre l'EDO caractérisant la relation entre l'entrée et la sortie du système de la manière la plus fidèle possible, vis-à-vis du modèle expérimental. En second lieu, et à partir des postulats déduits des expérimentations, des simplifications du système ont été conduitent de manière à prendre en compte les contraintes pratiques liées à l'usage en temps réel d'une simulation du système.

L'étude expérimental du circuit d'écrêtage à diode a permis dans un premier temps d'identifier différents comportements. Le circuit d'écrêtage distord le signal d'entrée en générant une cascade harmonique. L'amplitude des harmoniques générées dépend fortement de l'amplitude du signal d'entrée. Les harmoniques générées sont de rang impair et le signal de sortie tend vers un signal carré lorsque le signal d'entrée est de grande amplitude. La réponse du circuit est en grande partie indépendante de la fréquence du signal d'entrée, sauf lorsque cette fréquence s'approche de, ou dépasse la fréquence de coupure du filtre passe-bas intégré au circuit. Cette observation a été essentielle pour l'élaboration d'une méthode de résolution adaptée à une utilisation en temps réel.

À la suite de cette étude expérimentale, un modèle basé sur une EDO a été proposé pour rendre compte des caractéristiques intrinsèques de ce type de circuit. Ce modèle a été comparé aux données expérimentales et au logiciel de simulation LTspice, et il a été démontré qu'il reproduisait fidèlement les propriétés clés du circuit d'écrêtage, à condition que l'EDO soit résolu avec une méthode de résolution suffisamment stable, et que le calcul s'effectue à une fréquence d'échantillonnage relativement élevée pour éviter le repliement fréquentiel d'harmoniques générées au-delà de la fréquence de Nyquist. Dans le cas étudié, les méthodes des trapèzes et d'Euler implicite ont prouvé leur efficacité pour résoudre ce type d'EDO. Les méthodes explicites ne garantissent pas ici la stabilité du système.

En dernier lieu, l'analyse des résultats expérimentaux a permis de développer une méthode fonctionnant en temps réel. Cette méthode repose sur l'approche statique, qui consiste à fixer la valeur de la dérivée à zéro dans l'EDO originelle à résoudre. Le système est alors résolu à l'aide d'un calculateur de racines, ce qui permet d'établir une relation simple entre l'entrée et la sortie du système. Les résultats obtenus, lorsqu'ils ont été comparés aux méthodes de résolution en temps différé, ont démontré un haut degré de satisfaction en termes de fidélité sonore et d'efficacité vis-à-vis de la méthode des trapèzes utilisée en temps différé. Cette méthode a finalement pu être implémenté au sein d'un plugin, constituant le livrable de projet.

Enfin, bien que ce travail ait permis de montrer un moyen rudimentaire mais convenable de modéliser en temps réel des circuits d'écrêtage à diode, il reste encore des aspects qui pourraient être explorés pour enrichir et améliorer la simulation. La modélisation de l'étage de filtrage actif de la pédale de distorsion pourrait constituer une extension naturelle de cette étude. En effet, l'intégration d'un modèle précis du circuit de filtrage pourrait permettre d'obtenir une simulation complète du son produit par un modèle particulier de pédale, en tenant compte non seulement de l'écrêtage du signal, mais aussi de la manière dont celui-ci est filtré et amplifié dans l'étage de filtrage actif.

Bibliographie

[1] Kendall E. Atkinson. *An introduction to numerical analysis, 2^{nd} edition.* John Wiley & Sons, 1989.

[2] Kendall E. Atkinson, Weimin Han, and David Stewart. *Numerical Solution of Ordinary Differential Equations.* Pure and Applied Mathematics : A Wiley Series of Texts, Monographs and Tracts. John Wiley & Sons, 2011.

[3] Eliot Deschang. *StaticClipper source code.* `https://github.com/eliot-des/StaticClipper/tree/main`, 2023.

[4] Timothy Sauer. *Numerical Analysis, 2^{nd} edition.* Pearson Addison Wesley, 2006.

[5] David T. Yeh, Jonathan S. Abel, and Julius O. Smith III. Simplified, physically-informed models of distortion and overdrive guitar effects pedals. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-07)*, pages 10–14, 2007.

A Annexe

A.1 Photo MXR Distortion + et DOD 250

FIGURE 22 – Photo de la pédale MXR Distortion +

A.2 Circuit d'alimentation de la pédale MXR Distortion +

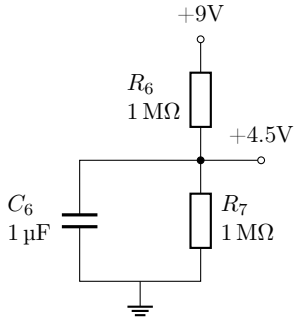


FIGURE 23 – Étage d'alimentation de la pédale MXR Distorsion +.

A.3 Méthodes d'Euler

Dans les annexes A.4, A.5, A.6 et A.7 les démonstrations sont effectuées sur une équation différentielle de la forme :

$$\frac{\mathrm{d}x}{\mathrm{d}t} = x'(t) = f(t,\ x(t)), \quad x(t_0) = x_0,$$

avec t_0 un temps pour lequel la valeur de la fonction x est connue (condition initiale). Sous forme discrète, en utilisant les conventions de notations énoncées partie 5, l'équation différentielle s'exprime comme :

$$x'_n = f(t_n,\ x_n), \quad x(t_0) = x_0,$$

avec $x_n = x(t_n)$, $t_n = nT$, où $n \in \mathbb{N}$ correspond à l'index de discrétisation et T au pas d'échantillonnage. Les méthodes abordées permettant d'approcher la solution exacte de la fonction $x(t)$, le choix est fait de noter la fonction $\tilde{x}_n \approx x_n$ la fonction discrète approchant la fonction x_n, le symbole $\tilde{}$ étant utilisé pour indiquer qu'il s'agit d'une approximation de la "vraie" valeur de x à l'échantillon n.

Cette partie aborde les méthodes de Forward et Backward Euler pour résoudre ce type d'équations différentielles.

A.4 Euler explicite

A.4.1 Différence Avant

Dans un premier temps, la méthode d'Euler explicite peut être déterminée par approximation de la dérivée par différence finie. Considérant la fonction $x(t)$, l'expression de sa dérivée $x'(t) = f(t,\ x(t))$ par "différence avant" est donnée par :

$$x'(t) = \frac{\mathrm{d}}{\mathrm{d}t}x(t) = \lim_{\delta \to 0} \frac{x(t+\delta) - x(t)}{\delta}.$$

Pour une pas d'échantillonnage T suffisamment petit, il advient :

$$x'(t) \approx \frac{x[(n+1)T] - x(nT)}{T} = \tilde{x}'(t) = \tilde{x}'(nT),$$

où $\tilde{x}'(nT)$ est la valeur approchée de la dérivée de x au n-ième échantillon. Si l'approximation précédente est utilisée pour approcher l'expression de la dérivée en x_n, alors en utilisant la notation en indice pour dénoter l'index temporel n (cf. partie 5), tel que $x_n = x(t_n) = x(nT)$, l'approximation de la dérivée de x au n-ième échantillon vaut :

$$\tilde{x}'_n = \frac{x_{n+1} - x_n}{T},$$

en supposant x_{n+1} et x_n connues. Cela n'est pas le cas en pratique puisque seulement la valeur du premier échantillon x_0 est supposée connue. En isolant x_{n+1} dans l'équation précédente, la méthode d'Euler explicite apparaît telle que :

$$\begin{cases} \tilde{x}_0 &= x(t_0), \\ \tilde{x}_{n+1} &= \tilde{x}_n + T\tilde{x}'_n,\ \forall n > 0. \end{cases} \tag{15}$$

Puisque $f(t_n,\ x_n) = x'_n$ est donnée par l'expression de l'équation différentielle, la méthode d'Euler explicite permettant d'exprimer l'approximation de $x_{n+1} = x(t_n + T)$ à partir de l'échantillon déterminée $x_n = x(t_n)$ et de l'expression de sa dérivée x'_n.

À noter que la valeur approximative du $(n+1)$-ième échantillon de x, notée $\tilde{x}_{n+1}$, est déterminée à partir de la valeur approximative de l'échantillon précédent, $\tilde{x}_n$, et de sa dérivée approximative, $\tilde{x}'_n$. Cette dernière est calculée en utilisant l'équation différentielle au temps t_n, c'est-à-dire $\tilde{x}'_n = f(t_n,\ \tilde{x}_n)$. Par conséquent, les valeurs approximatives de $\tilde{x}$, obtenues à chaque étape d'échantillonnage, sont basées sur des approximations successives dérivées des valeurs précédentes.

A.4.2 Intégration

Une autre possibilité consiste à intégrer l'équation différentielle $x'(t) = f(t,\ x(t))$ sur l'intervalle $[t_n,\ t_{n+1}]$ afin d'obtenir le lien entre x et l'intégrale de sa dérivée :

$$\underbrace{x(t_{n+1}) - x(t_n)}_{x_{n+1} - x_n} = \int_{t_n}^{t_{n+1}} f(t,\ x(t))\, dt.$$

En approximant maintenant l'intégrale par la méthode des rectangles à gauche avec un seul rectangle (Fig. 24), il advient :

$$\int_{t_n}^{t_{n+1}} f(t,\ x(t))\, dt \approx T \underbrace{f\left(t_n,\ x(t_n)\right)}_{x'_n = f(t_n,\ x_n)}.$$

En combinant ces deux équation, la relation permettant d'estimer x_{n+1} à partir de x_n et de sa dérivée est obtenue :

$$\begin{aligned} x_{n+1} - x_n &\approx T x'_n, \\ x_{n+1} &\approx x_n + T x'_n, \end{aligned}$$

aboutissant à la méthode d'Euler explicite (cf. équation 15), lorsque les valeurs de x_n et de x'_n sont considérées comme étant des approximations obtenues par récurrence.

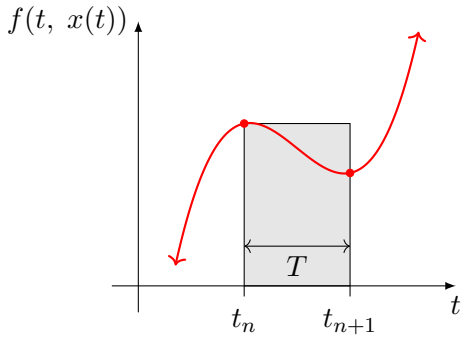


FIGURE 24 – Méthode d'intégration des rectangles à gauche sur l'intervalle $[t_n, t_{n+1}]$.

A.4.3 Développement limité

La méthode d'Euler explicite peut également être démontrée par développement limité, et donc par troncature de la série de Taylor de la fonction $x(t)$. L'expression de $x(t)$ par série de Taylor, infiniement dérivable en un point a est donnée par :

$$x(t) = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{\partial^k x(t)}{\partial t^k}\bigg|_{t=a} \times (t-a)^k \tag{16}$$

En exprimant la série de Taylor de $x(t)$ tronquée à l'ordre 1 (développement limité) au voisinage de t_n, soit à $t = t_n + T$, la méthode d'Euler explicite apparaît telle que :

$$x(t_n + T) \approx x(t_n) + \underbrace{\left.\frac{\partial x(t)}{\partial t}\right|_{t=t_n}}_{f(t_n,\, x_n)=x'_n} \times \big((t_n + T) - t_n\big)\,,$$
$$x_{n+1} \approx x_n + T x'_n,$$

aboutissant à la méthode d'Euler explicite (cf. équation (15)), lorsque les valeurs de x_n et de x'_n sont considérées comme étant des approximations obtenues par récurrence.

A.4.4 Erreur locale

Cette erreur est liée à la différence entre la solution exacte et la solution approximée à un instant donné, en considérant un seul pas d'échantillonnage et en supposant que la solution exacte est connue à l'échantillon précédent, et est utilisé pour déterminer la valeur de l'échantillon suivant, sur lequel on estime l'erreur. L'erreur de troncature locale est donc l'erreur commise pour obtenir $\tilde{x}_{n+1}$ lorsque les données précédentes (x_n, etc.) sont connues exactement. Elle est "locale" dans le sens où elle n'inclut pas les erreurs créées lors des étapes précédentes. Cette erreur se quantifie algébriquement par la relation :

$$e_{n+1} = |\tilde{x}_{n+1} - x_{n+1}|, \tag{17}$$

sous réserve que $\tilde{x}_n = x(t_n)$.

En supposant x'' continu, et en utilisant le théorème de Taylor (cf. équation (16)), la solution exacte en $t_{n+1} = t_n + T$ est :

$$x(t_{n+1}) = x(t_n) + T x'(t_n) + \underbrace{\frac{T^2}{2} x''(\xi_n)}_{\mathcal{O}(T^2)},\ t_n < \xi_n < t_{n+1}.$$

En considérant $\tilde{x}_n = x(t_n)$ et $x'(t_n) = f(t_n,\ \tilde{x}_n)$, l'équation précédente peut être écrite comme :

$$x(t_{n+1}) = \tilde{x}_n + T f(t_n,\ \tilde{x}_n) + \frac{T^2}{2} x''(\xi_n).$$

Sachant que la méthode d'Euler est donnée par l'équation (15), telle que :

$$\tilde{x}_{n+1} = \tilde{x}_n + T\tilde{x}'_n,$$

l'erreur de troncature locale e définie équation (17) est obtenue par la soustraction des deux expressions :

$$e_{n+1} = |\tilde{x}_{n+1} - x_{n+1}| = \frac{T^2}{2}|x''(\xi_n)| \leq \frac{T^2}{2} C,$$

où $C = \max(|x''(\xi_n)|)$.

A.4.5 Erreur Globale

L'erreur locale étant seulement identifiable à la premier itération de l'algorithme - puisqu'elle nécessite que $\tilde{x}_n = x(t_n)$, ce qui est uniquement valable à $n = 0$ (condition initiale) - la limite d'exploitation de cette erreur est vite atteinte. L'erreur globale est quant à elle un bon indicateur

de l'évolution de l'erreur à mesure que les itérations sont effectués pour estimer $x(t_n)$ de manière numérique, puisque celle-ci est définie de la manière suivante :

$$g_n = |\tilde{x}_{n+1} - x_{n+1}|.$$

Le théorème suivant est le théorème principal sur la convergence des solveurs d'équations différentielles en un seul pas. La dépendance de l'erreur globale par rapport à T montre qu'il est envisageable de s'attendre à ce que l'erreur diminue à mesure que T diminue, de sorte que (au moins en arithmétique exacte) l'erreur peut être rendue aussi petite que souhaité. Cela amène à un autre point important : la dépendance exponentielle de l'erreur globale sur t_N. À mesure que le temps augmente, la limite d'erreur globale peut devenir extrêmement grande. Pour des valeurs de t_n importantes, la taille du pas T nécessaire pour maintenir une petite erreur globale peut être si petite qu'elle en devient peu pratique.

Théorème :

En supposant que $f(t, x(t))$ a une constante de Lipschitz L pour la variable x et que la solution exacte du problème $x(t_n)$, de valeur initiale $x(t_0) = x_0$ à t_n est approximée par $\tilde{x}_n$ à partir d'un algorithme avec erreur de troncature locale $e_n \leq CT^{k+1} = \mathcal{O}(T^{k+1})$, pour une certaine constance C et $k \geq 0$, alors, pour chaque $t_0 \leq t_n < t_N$, l'algorithme a une erreur de troncature globale :

$$g_n \leq \frac{CT^k}{L}(e^{L(t_n - t_0)} - 1) = \mathcal{O}(T^k). \tag{18}$$

Pour l'algorithme d'Euler, $e_n = \mathcal{O}(T^2)$, par conséquent :

$$g_n = \mathcal{O}(T)$$

Un autre moyen de voir cet erreur globale consiste à remarquer que, sur chaque intervalle, l'erreur associée à la méthode du rectangle à gauche est $e_n = \mathcal{O}(T^2)$. L'erreur totale sur l'intégrale entre t_0 et t fixé est la somme des erreurs $g_n = \sum_{n=0}^{N-1} e_n = Ne_n = N\mathcal{O}(T^2) = \mathcal{O}(1/N)$=$\mathcal{O}(T)$.

L'erreur de la méthode tend vers zéro lorsque le pas de temps tend vers zéro. L'ordre est généralement déterminé par l'analyse de l'erreur de troncature locale de la méthode. Pour une méthode d'ordre p, si la taille du pas de temps est réduit par un facteur de 2, l'erreur devrait être réduite par un facteur de 2^p. Autrement dit, une méthode d'ordre p présentera une réduction de l'erreur d'un facteur de T^p lorsque le pas de temps h tend vers zéro. Au vue de l'erreur de troncature globale de la méthode d'Euler explicite, celle-ci est une méthode d'ordre 1, ce qui signifie que si la taille du pas de temps est réduit par un facteur de 2, l'erreur devrait être réduite de moitié.

Il est important de noter que l'ordre de la méthode donne une indication sur la précision de la méthode pour des petits pas de temps, mais il ne dit rien sur la stabilité de la méthode, qui est un autre aspect important à considérer lors du choix d'une méthode numérique pour résoudre un problème particulier.

A.5 Euler implicite

A.5.1 Différence arrière

Tout comme Forward Euler, la méthode d'Euler implicite peut être obtenue par approximation de la dérivée par différence finie, mais par l'expression de sa dérivée par "différence arrière". Considérant la fonction $x(t)$, l'expression de sa dérivée $x'(t) = f(t,\ x(t))$ par "différence arrière" est donnée par :

$$x'(t) = \frac{\mathrm{d}}{\mathrm{d}t}x(t) = \lim_{\delta \to 0} \frac{x(t) - x(t - \delta)}{\delta}.$$

Pour une pas d'échantillonage T suffisamment petit, il advient :

$$x'(t) \approx \frac{x(nT) - x[(n-1)T]}{T} = \tilde{x}'(nT).$$

En considérant les mêmes changement de notations que dans l'annexe A.4, l'approximation de la dérivée de x au n-ième échantillon conduit à :

$$\tilde{x}'_n = \frac{x_n - x_{n-1}}{T}.$$

En isolant $\tilde{x}'_n$ dans l'équation précédente, et en effectuant le changement de variable $n \to n+1$ la méthode de Euler implicite apparaît telle que :

$$\begin{cases} \tilde{x}_0 = x(t_0), \\ \tilde{x}_{n+1} = \tilde{x}_n + T\tilde{x}'_{n+1}, \ \forall \geq > 0. \end{cases} \tag{19}$$

A.5.2 Intégration

Une autre possibilité consiste à intégrer l'équation différentielle $x'(t) = f(t,\ x(t))$ sur l'intervalle $[t_n,\ t_{n+1}]$ afin d'obtenir le lien entre x et l'intégrale de sa dérivée :

$$\underbrace{x(t_{n+1}) - x(t_n)}_{x_{n+1} - x_n} = \int_{t_n}^{t_{n+1}} f(t,\ x(t))\, dt.$$

En approximant maintenant l'intégrale par la méthode des rectangles à droite avec un seul rectangle (Fig. 25), il advient :

$$\int_{t_n}^{t_{n+1}} f(t,\ x(t))\, dt \approx T \underbrace{f\left(t_{n+1},\ x(t_{n+1})\right)}_{x'_{n+1} = f(t_{n+1},\, x_{n+1})}.$$

En combinant ces deux équation, la relation permettant d'estimer x_{n+1} à partir de x_n et de sa dérivée est obtenue :

$$\begin{aligned} x_{n+1} - x_n &\approx T x'_{n+1}, \\ x_{n+1} &\approx x_n + T x'_{n+1}, \end{aligned}$$

aboutissant à la méthode d'Euler implicite (cf. équation (19)), lorsque les valeurs de x_n et de x'_n sont considérées comme étant des approximations obtenues par récurrence.

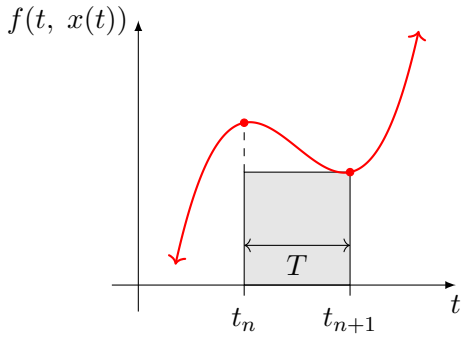


FIGURE 25 – Méthode d'intégration des rectangles à droite sur l'intervalle $[t_n, t_{n+1}]$.

A.5.3 Développement limité

Un autre moyen de démontrer la méthode d'Euler implicite consiste à exprimer la série de Taylor (cf. équation (16)) tronqué à l'ordre 1 à l'instant t_n pour obtenir la valeur à l'instant $t_n - T$:

$$x(t_n - T) \approx x(t_n) + \underbrace{\left.\frac{\partial x(t)}{\partial t}\right|_{t=t_n}}_{f(t_n,\, x_n)=x'_n} \times \big((t_n - T) - t_n\big)\,.$$

Soit :

$$x_{n-1} \approx x_n - T x'_n.$$

En isolant $x_n = x(t_n)$ d'un côté de l'équation, la méthode d'Euler implicite permettant d'estimer x_n à partir de sa dérivée et de x_{n-1} apparaît tel que :

$$x_n \approx x_{n-1} + T x'_n,$$

conduisant à la relation de l'équation (19), en effectuant le changement de variable $n \to n+1$.

A.5.4 Erreur locale

L'erreur locale avec la méthode d'Euler rétrograde est la même que celle commise avec Euler en avant, en raison du fait que la source d'erreur sur l'approximation de l'intégrale par la méthode du rectangle à gauche ou à droite est la même. Celle-ci est donc exprimée comme :

$$e_{n+1} = |\tilde{x}_{n+1} - x_{n+1}| = \underbrace{\frac{T^2}{2}|x''(\xi_n)\,|}_{\mathcal{O}(T^2)} \leq \frac{T^2}{C},$$

où $C = \max|x''(\xi_n)|$.

A.5.5 Erreur globale

Par application du théorème A.4.5, l'ordre de l'erreur globale commise est $\mathcal{O}(T)$. La méthode d'Euler implicite est donc également une méthode d'ordre 1. Cette erreur peut également être obtenue par sommation des erreurs locales successivement commise à chaque itération de l'algorithme même si une manière plus "correcte" d'obtenir cette erreur reste par application du théorème énoncé précédemment.

A.6 Méthode des trapèzes

A.6.1 Intégration

En voyant la méthode d'Euler explicite (ou implicite), il est légitime de se dire qu'une meilleur approximation de l'intégrale peut-être conduite, par exemple avec la méthode des trapèzes. Cette sous-section montre pourquoi cette méthode peut-être vu comme une méthode d'intégration par trapèze (Fig. 26).

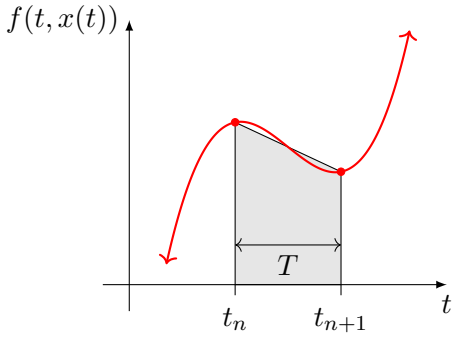


FIGURE 26 – Méthode d'intégration des trapèzes sur l'intervalle $[t_n,\ t_{n+1}]$.

Une possibilité afin de déterminer la méthode des trapèze implicite, consiste à intégrer l'équation différentielle $x'(t) = f(t, x(t))$ entre un instant t_n et un instant $t_n + T = t_{n+1}$ afin d'obtenir le lien entre x et l'intégrale de sa dérivée :

$$x(t_{n+1}) - x(t_n) = \int_{t_n}^{t_{n+1}} f(t,\ x(t))\, dt.$$

En approximant maintenant l'intégrale par la méthode des trapèzes avec un seul trapèze (Fig. 26), il advient :

$$\int_{t_n}^{t_{n+1}} f(t, x(t))\, dt \approx \frac{T}{2} \underbrace{\left\{ f\left(t_n,\ x(t_n)\right) + f\left(t_{n+1},\ x(t_{n+1})\right) \right\}}_{x'_n + x'_{n+1}},$$

où $f\left(t_n,\ x(t_n)\right)$ et $f\left(t_{n+1},\ x(t_{n+1})\right)$ représentes les dimensions de la grande et petite base du trapèze, et où T représente la hauteur de celui-ci. Ici, l'intégrale dépend des valeurs de x_n et x_{n+1} ce qui donne lieu à une méthode implicite. La méthode des trapèzes est donc définie telle que :

$$\begin{cases} \tilde{x}_0 = x(t_0), \\ \tilde{x}_{n+1} = \tilde{x}_n + \dfrac{T}{2}(\tilde{x}'_n + \tilde{x}'_{n+1}),\ \forall n \geq 0. \end{cases} \tag{20}$$

A.6.2 Erreur locale

L'erreur locale de la méthode des trapèzes peut être obtenue en exprimant l'erreur de troncature propre à l'approximation de l'intégrale avec la méthode des trapèzes. Pour une fonction à valeurs réelles, deux fois continuement dérivable sur le sergent $[a, b]$, l'erreur est de la forme :

$$\int_a^b g(x)dx - (b-a)\frac{f(a)+f(b)}{2} = -\frac{(b-a)^3}{12} g''(\xi), \tag{21}$$

pour un certain $\xi \in [a,\ b]$, et en faisant l'intégrale avec la méthode des trapèzes sur l'intervalle $[a, b]$ avec un seul trapèze [2]. Dans le cas étudié, l'expression de l'erreur de troncature locale sur

l'intervalle $[t_n, t_n + T]$ avec la méthode des trapèze vaut :

$$e_{n+1} = \underbrace{-\frac{T^3}{12}x''(\xi_n)}_{\mathcal{O}(T^3)}$$

sous réserve que $\tilde{x}_n = x(t_n)$ et où $\xi \in [t_n,\ t_n + T]$.

A.6.3 Erreur globale

Par application du théorème A.4.5, l'erreur de troncature globale de la méthode des trapèze est $g_{n+1} = \mathcal{O}(T^2)$, résultant en une méthode de résolution d'ordre 2.

A.7 Runge-Kutta 4

La méthode de Runge-Kutta d'ordre 4 est une étape supplémentaire dans le raffinement du calcul de l'intégrale (1). Au lieu d'utiliser la méthode des trapèzes, la méthode de Simpson est utilisé . Celle-ci consiste à remplacer la fonction intégrée par une parabole passant par les points extrêmes et le point milieu. L'approximation de l'intégrale d'une fonction $g(x)$ sur un intervalle $x \in [a,\ b]$ avec la méthode de Simpson s'exprime comme :

$$\int_a^b g(x)\mathrm{d}x \simeq \frac{b-a}{6}\left[g(a) + 4f\left(\frac{a+b}{2}\right) + g(b)\right]$$

Appliquée à la fonction $f(t, x(t)$ sur l'intervalle $t \in [t_n,\ t_n + T]$, cela donne :

$$\int_{t_n}^{t_{n+T}} f(t, x(t))\mathrm{d}t \approx \frac{T}{6}\left[f\left(t_n, x\left(t_n\right)\right) + 4f\left(t_{n+1/2}, x\left(t_{n+1/2}\right)\right) + f\left(t_{n+1}, x\left(t_{n+1}\right)\right)\right]$$

d'où la relation :

$$x_{n+1} = x_n + \frac{T}{6}\left[f\left(t_n, x_n\right) + 4f\left(t_{n+1/2}, x_{n+1/2}\right) + f\left(t_{n+1}, x_{n+1}\right)\right]. \tag{22}$$

Ici, une difficulté apparaît car l'équation présente deux inconnues : $x_{n+1/2}$ et x_{n+1}. Pour rendre le schéma explicite, il faut estimer $4f\left(t_{n+1/2}, x_{n+1/2}\right)$ et $f\left(t_{n+1}, x_{n+1}\right)$ à partir de x_n, t_n et T.

Le terme $4f\left(t_{n+1/2}, x\left(t_{n+1/2}\right)\right)$ peut-être décomposer en deux termes identiques tel que :

$$4f\left(t_{n+1/2}, x\left(t_{n+1/2}\right)\right) = 2f\left(t_{n+1/2}, x_{n+1/2}\right) + 2f\left(t_{n+1/2}, x_{n+1/2}\right).$$

Dans le premier, $x_{n+1/2}$ est remplacé par sa valeur déduite de la méthode d'Euler explicite, à savoir $x^a_{n+1/2} = x_n + \frac{T}{2}f\left(t_n, x_n\right)$. Dans le deuxième terme, $x_{n+1/2}$ est remplacé par sa valeur déduite de la méthode d'Euler implicite : $x^b_{n+1/2} = x_n + \frac{T}{2}f\left(t_{n+1/2}, x_{n+1/2}\right)$ qui peut être approché par $x_n + \frac{T}{2}f\left(t_{n+1/2}, x^a_{n+1/2}\right)$. Les méthodes d'Euler implicite et explicite produisant des erreurs quasi opposées, cela offre l'opportunité de minimiser l'erreur sur le calcul de $4f\left(t_{n+1/2}, x_{n+1/2}\right)$. Pour résumer, il peut être écrit :

$$4f\left(t_{n+1/2}, x_{n+1/2}\right) \approx 2k_2 + 2k_3, \quad \text{avec} \begin{cases} k_1 &= f\left(t_n,\ x_n\right), \\ k_2 &= f\left(t_n + \frac{T}{2},\ x_n + \frac{T}{2}k_1\right), \\ k_3 &= f\left(t_n + \frac{T}{2},\ x_n + \frac{T}{2}k_2\right). \end{cases}$$

Quant au terme $f(t_{n+1}, x_{n+1})$ de l'équation (22), celui-ci est approché en x_{n+1} par la méthode du point milieu, c'est-à-dire en appliquant la méthode du rectangle au milieu :

$$x_{n+1} \approx x_n + Tf\left(t_{n+1/2}, x_{n+1/2}\right) \approx x_n + Tf\left(t_{n+1/2}, x^b_{n+1/2}\right).$$

Finalement le schéma explicite, dit de Runge-Kutta d'ordre 4 est obtenu :

$$\tilde{x}_{n+1} = \tilde{x}_n + \frac{k_1}{6} + \frac{k_2}{3} + \frac{k_3}{3} + \frac{k_4}{6} \quad \text{avec} \begin{cases} k_1 &= Tf\left(t_n,\ \tilde{x}_n\right), \\ k_2 &= Tf\left(t_n + \frac{T}{2},\ \tilde{x}_n + \frac{T}{2}k_1\right), \\ k_3 &= Tf\left(t_n + \frac{T}{2},\ \tilde{x}_n + \frac{T}{2}k_2\right), \\ k_4 &= Tf\left(t_n + T,\ \tilde{x}_n + Tk_3\right). \end{cases} \tag{23}$$

Par rapport à la méthode RK2, ce schéma numérique exige deux fois plus de calculs à chaque pas et nécessite donc un temps de calcul plus long (il nécessite des évaluations de la fonction f à des "demi-échantillons"). Cependant, ce défaut est compensé par un gain de précision car cette méthode est une méthode d'ordre 4.

A.7.1 Erreur locale et globale

L'estimation de l'erreur locale pour la méthode de Runge-Kutta d'ordre 4 se révèle plus complexe. En effet, elle ne se limite pas à l'erreur de troncature inhérente à l'intégration avec la méthode de Simpson. La méthode de Runge-Kutta d'ordre 4 constitue elle-même une approximation de l'intégrale de Simpson, reposant sur des équations explicites. Plusieurs ouvrages fournissent des démonstrations de l'ordre de l'erreur locale et globale générée [4],[1]. L'erreur locale s'avère être de l'ordre de $\mathcal{O}(T^5)$, tandis que l'erreur globale est de l'ordre de $\mathcal{O}(T^4)$. Cela confirme bien que la méthode de Runge-Kutta d'ordre 4 est une méthode d'ordre 4.

A.8 Newton-Raphson

L'algorithme de Newton-Raphson constitue une méthode itérative pour déterminer les racines d'une fonction en utilisant la dérivée de cette fonction. Le principe de cette méthode est de générer une suite de valeurs de l'inconnue qui convergent vers la racine cherchée, à partir d'une valeur initiale choisie. L'algorithme de Newton-Raphson se fonde sur la notion de la tangente à la courbe d'une fonction pour approximer la valeur de la racine. La méthode consiste à estimer une valeur initiale de la racine, puis à déterminer la pente de la tangente à la courbe en ce point. L'intersection de la tangente ainsi obtenue avec l'axe des abscisses permet alors de déterminer une nouvelle approximation de la racine. Cette démarche est répétée jusqu'à ce que la précision souhaitée soit atteinte.

A.8.1 Démonstration

La méthode de Newton-Raphson peut être obtenue en utilisant le développement en série de Taylor de la fonction $f(x)$ autour de la valeur initiale x_0. Si $f(x)$ est une fonction suffisamment régulière, indéfiniment dérivable en un point x_0, la série de Taylor de f en ce point s'écrit $f(x)$:

$$f(x) = \sum_{k=0}^{\infty} \frac{1}{k!} \frac{\partial^k f(x)}{\partial x^k}\bigg|_{x=x_0} \times (x - x_0)^k. \tag{24}$$

Lorsque il est souhaité de trouver la racine de cette fonction, c'est-à-dire le point x pour lequel $f(x) = 0$, les termes d'ordre supérieur dans la série de Taylor peuvent être négligés résultant en l'équation suivante :

$$f(x) = 0 \approx f(x_0) + f'(x_0)(x - x_0).$$

En résolvant cette équation pour x, l'équation suivant est obtenue :

$$x \approx x_0 - \frac{f(x_0)}{f'(x_0)},$$

où x correspond à la première estimation notée x_1 de la racine obtenue à partir de la valeur initiale x_0, qui correspond à l'estimation initiale de la racine. L'ensemble du processus est donc répété, en commençant par x_1, pour produire x_2, et ainsi de suite. Par récurrence, l'algorithme de Newton-Raphson apparaît tel que :

$$\begin{cases} x_0 = \text{valeur initiale}, \\ x_{i+1} = x_i - \dfrac{f(x_i)}{f'(x_i)}, \end{cases} \tag{25}$$

où x_{i+1} est la $(i+1)$-ième approximation de la valeur de la racine.

La solution exacte n'étant pas censé être connue, les conditions d'arrêt sont déterminées par rapport à la valeur précédente de l'approximation de la racine, soit avec une erreur relative ou absolue tel que :

$$\text{Sortie de la boucle si } \begin{cases} \left| \dfrac{x_i - x_{i-1}}{x_i} \right| < \text{erreur relative}, \\ \\ |x_i - x_{i-1}| < \text{erreur absolue}. \end{cases}$$

À noter que ce type de condition de sortie de boucle, à partir d'un critère d'erreur relative pose problème pour $x_i = 0$, auquel cas la fraction ne peut être définie de manière mathématique (division par 0 impossible), ce qui s'avère définie d'un point de vue informatique comme une valeur "NaN" (pour *Not a Number*) ou "Inf" (pour infini). Pour éviter ce problème, la fonction `np.isclose()` est utilisée. Cette fonction de la librairie `numpy` de Python permet de tester la proximité de deux nombres avec une tolérance relative ou absolue tout évitant les problèmes de division par 0.

A.8.2 Convergence numérique

Soit $e_i = |r - x_i|$ l'erreur après l'étape i d'une méthode itérative. L'itération converge de manière quadratique si :

$$M = \lim_{i \to \infty} \frac{e_{i+1}}{e_i^2} < \infty.$$

Théorème :

Soit f deux fois continûment dérivable et $f(r) = 0$. Si $f'(r) \neq 0$, alors la méthode de Newton converge localement et quadratiquement vers r. L'erreur e_i à l'étape i vérifie :

$$\lim_{i \to \infty} \frac{e_{i+1}}{e_i^2} = M,$$

où :

$$M = \frac{f''(r)}{2f'(r)}.$$

La convergence quadratique pour la méthode de Newton-Raphson signifie que l'erreur entre la solution approximée à chaque itération et la solution exacte diminue de manière quadratique. En d'autres termes, pour une estimation de la racine x_i, l'erreur entre la prochain estimation x_{i+1} et la solution exacte sera approximativement le carré de l'erreur entre x_i et la solution exacte r.

Démonstration - Preuve de la convergence quadratique

Pour prouver la convergence quadratique, la méthode de Newton est dérivée une deuxième fois, en gardant un œil attentif sur l'erreur $e_i = |r - x_i|$ commise à chaque itération. La formule de Taylor indique la différence entre les valeurs d'une fonction en un point donné et un autre point proche. Pour les deux points, la racine r et l'estimation actuelle x_i après i étapes est utilisée :

$$f(r) = f(x_i) + (r - x_i) f'(x_i) + \frac{(r - x_i)^2}{2} f''(\xi_i)$$

Ici, ξ_i est compris entre x_i et r. Parce que r est la racine, alors $f(r) = 0$ et donc :

$$\begin{aligned} 0 &= f(x_i) + (r - x_i) f'(x_i) + \frac{(r - x_i)^2}{2} f''(\xi_i), \\ -\frac{f(x_i)}{f'(x_i)} &= r - x_i + \frac{(r - x_i)^2}{2} \frac{f''(\xi_i)}{f'(x_i)}, \end{aligned}$$

en supposant que $f'(x_i) \neq 0$. En réarrangeant cette expression, la prochaine itération de Newton-Raphson peut être comparée à la racine :

$$\begin{aligned} x_i - \frac{f(x_i)}{f'(x_i)} - r &= \frac{(r - x_i)^2}{2} \frac{f''(\xi_i)}{f'(x_i)} \\ x_{i+1} - r &= e_i^2 \frac{f''(\xi_i)}{2f'(x_i)} \\ e_{i+1} &= e_i^2 \left| \frac{f''(\xi_i)}{2f'(x_i)} \right|. \end{aligned} \tag{26}$$

Dans cette équation, l'erreur à l'étape i est définie par $e_i = |x_i - r|$. Puisque ξ_i se situe entre r et x_i, il converge vers r tout comme x_i, et :

$$\lim_{i \to \infty} \frac{e_{i+1}}{e_i^2} = \left| \frac{f''(r)}{2f'(r)} \right|,$$

ce qui est la définition de la convergence quadratique.

La formule d'erreur 26 développée peut être considérée comme :

$$e_{i+1} \approx M e_i^2,$$

où $M = \left| f''(r)/2f'(r) \right|$, en supposant que $f'(r) \neq 0$. L'approximation s'améliore à mesure que la méthode de Newton converge, puisque les estimations x_i se rapprochent de r et que ξ_i est pris entre x_i et r.