Projet personnel - Mathématiques appliquées

Lien entre descente de gradient stochastique et équations différentielles stochastiques

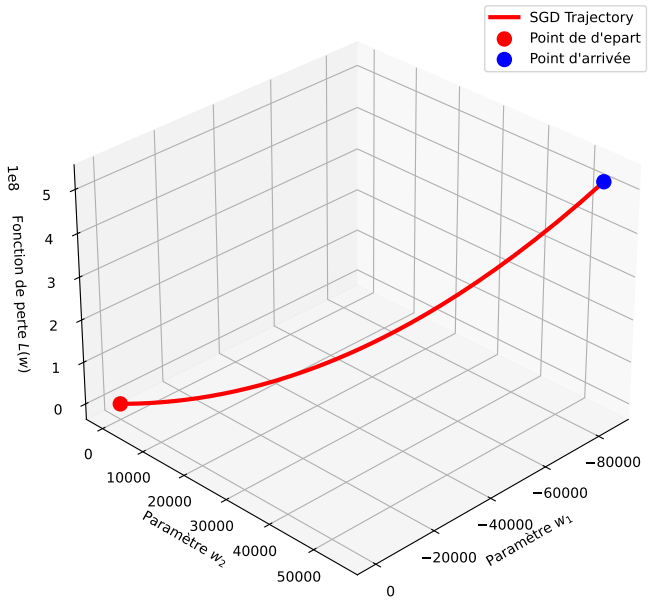


FIGURE 1 – Visualisation de la trajectoire d'une descente de gradient stochastique sur un paysage de perte en 3D.

Auteur : Samatar ABERKANE

Projet personnel réalisé dans le cadre de mes études à Télécom SudParis et Polytechnique Montréal

Table des matières

1 Introduction

L'apprentissage automatique moderne repose fondamentalement sur l'optimisation de fonctions objectives complexes dans des espaces de haute dimension. Au cœur de cette révolution se trouve l'algorithme de descente de gradient stochastique (SGD), qui constitue l'épine dorsale de l'entraînement des réseaux de neurones profonds. Depuis les travaux pionniers de Robbins et Monro en 1951, la SGD a évolué pour devenir l'outil d'optimisation de référence dans des domaines aussi variés que la vision par ordinateur, le traitement du langage naturel, ou encore la reconnaissance vocale.

1.1 Contexte et importance de la SGD en apprentissage profond

L'essor spectaculaire de l'apprentissage profond au cours des deux dernières décennies a été rendu possible par la capacité de la SGD à naviguer efficacement dans des paysages d'optimisation de très haute dimension. La descente de gradient stochastique (SGD), dont les fondements remontent aux travaux pionniers de Robbins et Monro [**?**], est devenue une pierre angulaire de l'optimisation des modèles d'apprentissage profond. Contrairement aux méthodes d'optimisation déterministes classiques, la SGD exploite la structure stochastique inhérente aux problèmes d'apprentissage statistique, où la fonction objectif s'écrit typiquement sous la forme d'une espérance :

$$L(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(\theta; x, y)] \tag{1}$$

où $\theta \in \mathbb{R}^d$ représente les paramètres du modèle, $\mathcal{D}$ la distribution des données, et ℓ la fonction de perte.

En pratique, cette espérance est approximée par une moyenne empirique sur un ensemble de données fini, et la SGD utilise des estimations non biaisées du gradient calculées sur de petits sous-ensembles (mini-batches) de données. Cette approche stochastique présente plusieurs avantages cruciaux :

Efficacité computationnelle : Le calcul du gradient sur l'ensemble complet des données devient prohibitif pour les grands jeux de données modernes. La SGD permet de faire des progrès significatifs avec un coût computationnel réduit à chaque itération.

Capacité d'évasion : Le bruit stochastique inhérent à l'algorithme permet d'échapper aux minima locaux de mauvaise qualité, un phénomène particulièrement important dans l'optimisation non-convexe.

Généralisation : De manière surprenante, la SGD semble favoriser des solutions qui généralisent mieux sur de nouvelles données, comparativement aux méthodes d'optimisation déterministes plus précises.

Cependant, malgré son succès empirique remarquable, la compréhension théorique de la SGD reste largement incomplète. Les analyses classiques de convergence, héritées de la théorie de l'optimisation convexe, ne capturent pas pleinement le comportement complexe observé en pratique, notamment dans le régime non-convexe caractéristique de l'apprentissage profond.

1.2 Motivation : comprendre la dynamique via des outils probabilistes

Face aux limitations des approches déterministes, une nouvelle perspective a émergé au cours des dernières années : l'analyse de la SGD à travers le prisme des processus stochastiques. Cette approche, initiée par les travaux de Welling et Teh (2011), puis développée par Li et al. (2017) et Mandt et al. (2017), propose de modéliser la dynamique de la SGD comme une équation différentielle stochastique (EDS).

L'intuition fondamentale est que, sous certaines conditions de régularité et dans une limite appropriée, la suite discrète des itérés de la SGD peut être approximée par la solution d'une EDS de type Langevin :

$$d\theta_t = -\nabla U(\theta_t)dt + \sqrt{2T}dW_t \tag{2}$$

où $U(\theta)$ représente un potentiel effectif lié à la fonction objectif, T une température effective liée au pas de temps et à la variance du bruit, et W_t un mouvement brownien standard.

Cette perspective ouvre de nouvelles voies d'analyse particulièrement riches :

Théorie des processus stochastiques : Les outils développés pour l'étude des EDS, notamment la théorie des martingales et l'analyse stochastique, deviennent applicables à l'étude de la SGD.

Mécanique statistique : L'analogie avec les systèmes physiques permet d'importer des concepts comme la distribution de Gibbs, l'entropie, ou encore les transitions de phase.

Analyse asymptotique : L'équation de Fokker-Planck associée à l'EDS de Langevin fournit une description de l'évolution de la densité de probabilité des paramètres, permettant d'étudier le comportement à long terme.

Cette approche probabiliste permet notamment de mieux comprendre pourquoi la SGD semble naturellement favoriser des solutions "plates" dans l'espace des paramètres, c'est-à-dire des minima caractérisés par une faible courbure de la fonction objectif. De telles solutions sont empiriquement associées à une meilleure généralisation, établissant un lien profond entre la dynamique d'optimisation et les propriétés statistiques du modèle appris.

1.3 Objectifs du rapport et plan

Ce projet vise à présenter de manière rigoureuse et accessible les développements récents reliant la descente de gradient stochastique aux équations différentielles stochastiques. Nos objectifs principaux sont :

Objectif théorique : Établir clairement le passage de la dynamique discrète de la SGD à son approximation continue sous forme d'EDS de Langevin, en précisant les hypothèses mathématiques nécessaires et les conditions de validité de cette approximation.

Objectif analytique : Explorer les conséquences de cette modélisation continue, notamment à travers l'équation de Fokker-Planck et l'analyse du comportement asymptotique de la distribution des paramètres.

Objectif appliqué : Interpréter les résultats théoriques dans le contexte de l'apprentissage automatique, en particulier pour comprendre les phénomènes de généralisation et de sélection de modèles.

Objectif numérique : Valider les prédictions théoriques par des expériences numériques sur des cas d'étude simples mais représentatifs.

Le plan de ce projet s'articule autour de six sections principales. Après cette introduction, la section 2 présente en détail l'algorithme de descente de gradient stochastique, ses propriétés fondamentales et sa modélisation mathématique. La section 3 développe le passage à l'approximation continue et la dérivation de l'EDS de Langevin associée. La section 4 explore l'équation de Fokker-Planck et le comportement asymptotique de la distribution des paramètres. La section 5 discute les implications de ces résultats pour l'apprentissage automatique, tandis que la section 6 présente des validations numériques. Enfin, la conclusion synthétise les résultats et ouvre sur les perspectives de recherche future.

Cette approche interdisciplinaire, à la croisée de l'analyse stochastique, de la mécanique statistique et de l'apprentissage automatique, illustre la richesse des mathématiques appliquées modernes et leur capacité à éclairer des phénomènes complexes par des outils théoriques sophistiqués.

2 Descente de gradient stochastique

Cette section présente de manière rigoureuse l'algorithme de descente de gradient stochastique, ses propriétés mathématiques fondamentales et sa modélisation probabiliste. Nous développons progressivement les concepts nécessaires à la compréhension de son approximation

continue.

2.1 Définition et intuition

Définition 2.1 (Problème d'optimisation stochastique)**.** Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace probabilisé et ξ une variable aléatoire à valeurs dans un espace mesurable Ξ. Considérons une fonction de perte $\ell : \mathbb{R}^d \times \Xi \to \mathbb{R}$ mesurable. Le problème d'optimisation stochastique consiste à résoudre :

$$\min_{\theta \in \mathbb{R}^d} L(\theta) := \mathbb{E}_{\xi \sim \mathbb{P}}[\ell(\theta, \xi)] \tag{3}$$

où $L(\theta)$ est appelée fonction de risque ou fonction objectif.

En apprentissage statistique, ξ représente typiquement un couple (x, y) où x est une observation et y la réponse associée, et $\ell(\theta, (x, y))$ mesure l'erreur de prédiction du modèle paramétré par θ.

Définition 2.2 (Algorithme de descente de gradient stochastique)**.** L'algorithme de descente de gradient stochastique (SGD) génère une suite $(\theta_k)_{k \geq 0}$ selon la récurrence :

$$\theta_{k+1} = \theta_k - \eta_k \nabla_\theta \ell(\theta_k, \xi_k) \tag{4}$$

où :

— $\theta_0 \in \mathbb{R}^d$ est une initialisation,

— $(\eta_k)_{k \geq 0}$ est une suite de pas de temps (learning rates),

— $(\xi_k)_{k \geq 0}$ est une suite de variables aléatoires i.i.d. de même loi que ξ.

La différence fondamentale avec la descente de gradient déterministe réside dans l'utilisation du gradient stochastique $\nabla_\theta \ell(\theta_k, \xi_k)$ au lieu du gradient exact $\nabla L(\theta_k) = \mathbb{E}[\nabla_\theta \ell(\theta_k, \xi)]$.
Pour une introduction complète, voir Nasr [4].

Remarque 2.1 (Mini-batch SGD)**.** En pratique, on utilise souvent une variante par mini-batches où à chaque itération k, on considère un sous-ensemble $\mathcal{B}_k \subset \{1, \ldots, n\}$ de taille $|\mathcal{B}_k| = b$ et :

$$\theta_{k+1} = \theta_k - \eta_k \frac{1}{b} \sum_{i \in \mathcal{B}_k} \nabla_\theta \ell(\theta_k, \xi_i) \tag{5}$$

Cette formulation réduit la variance du gradient stochastique tout en préservant l'efficacité computationnelle.

L'intuition géométrique de la SGD est la suivante : à chaque itération, l'algorithme effectue un pas dans la direction opposée au gradient stochastique, qui constitue une estimation non biaisée mais bruitée de la direction de plus forte pente de la fonction objectif. Le bruit introduit par cette estimation peut être vu comme une perturbation qui aide l'algorithme à explorer l'espace des paramètres et à éviter les minima locaux de mauvaise qualité.

2.2 Modèle stochastique : bruit et variance

La principale différence entre la descente de gradient (GD) et la descente de gradient stochastique (SGD) réside dans la nature du bruit. Tandis que la GD utilise le gradient exact calculé sur l'intégralité du jeu de données, la SGD utilise une approximation bruyante de ce gradient, calculée sur un petit sous-ensemble de données (le mini-batch). Ce bruit, qui est une conséquence de la variabilité des mini-batchs, est un élément central de la dynamique de la SGD. Il peut être modélisé comme une perturbation aléatoire qui pousse la trajectoire d'optimisation à s'écarter du chemin direct vers le minimum.

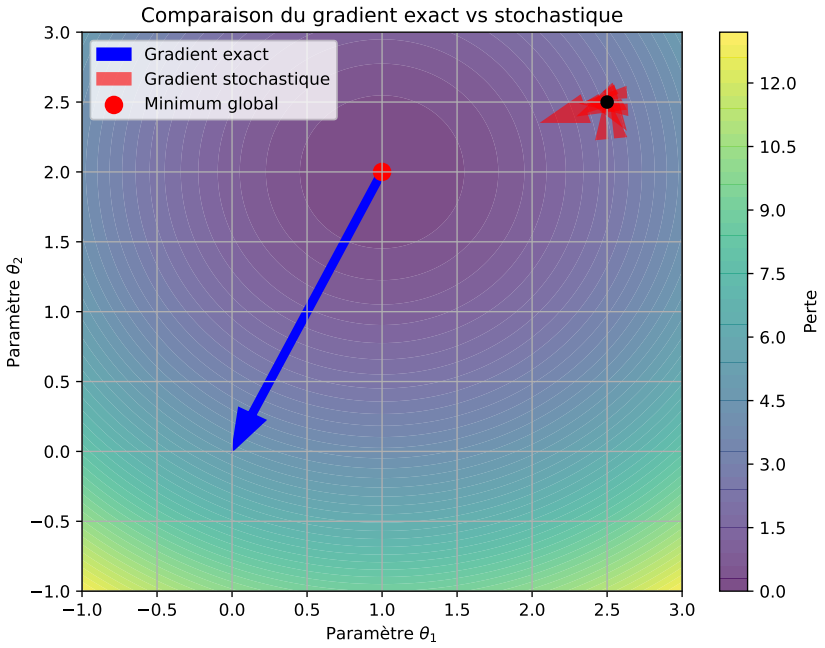


FIGURE 2 – Comparaison du gradient exact (en bleu) et de plusieurs gradients stochastiques (en rouge) sur une fonction de perte 2D. Le gradient stochastique est une estimation bruyante de la vraie direction de la pente.

La Figure 3 illustre cette distinction. Sur un paysage de perte simple et convexe, on observe que la GD suit une trajectoire lisse et déterministe directement vers le minimum global. En revanche, la SGD suit une trajectoire beaucoup plus erratique et "bruitée" qui zigzague autour du minimum. Bien que cela puisse sembler moins efficace au premier abord, ce comportement d'exploration est un avantage clé dans les paysages de perte non-convexes, car il permet d'échapper aux minima locaux.

Pour analyser rigoureusement la SGD, il est essentiel de caractériser précisément la structure du bruit stochastique. Nous décomposons le gradient stochastique en sa partie déterministe et sa partie aléatoire.

Définition 2.3 (Décomposition du gradient stochastique)**.** Soit $g(\theta, \xi) := \nabla_\theta \ell(\theta, \xi)$ le gradient stochastique. On peut l'écrire sous la forme :

$$g(\theta, \xi) = \nabla L(\theta) + \epsilon(\theta, \xi) \tag{6}$$

où $\epsilon(\theta, \xi) := g(\theta, \xi) - \mathbb{E}_\xi[g(\theta, \xi)]$ est le terme de bruit centré : $\mathbb{E}_\xi[\epsilon(\theta, \xi)] = 0$.

La matrice de covariance du bruit est définie par :

$$\Sigma(\theta) := \mathbb{E}_\xi[\epsilon(\theta, \xi)\epsilon(\theta, \xi)^T] = \mathrm{Cov}_\xi[g(\theta, \xi)] \tag{7}$$

Proposition 2.1 (Propriétés du gradient stochastique)**.** *Sous les hypothèses de régularité usuelles (intégrabilité et dérivabilité sous le signe espérance), le gradient stochastique satisfait :*

1. ***Non-biais :*** $\mathbb{E}_\xi[g(\theta, \xi)] = \nabla L(\theta)$
2. ***Variance finie :*** $\mathbb{E}_\xi[\|g(\theta, \xi)\|^2] < \infty$

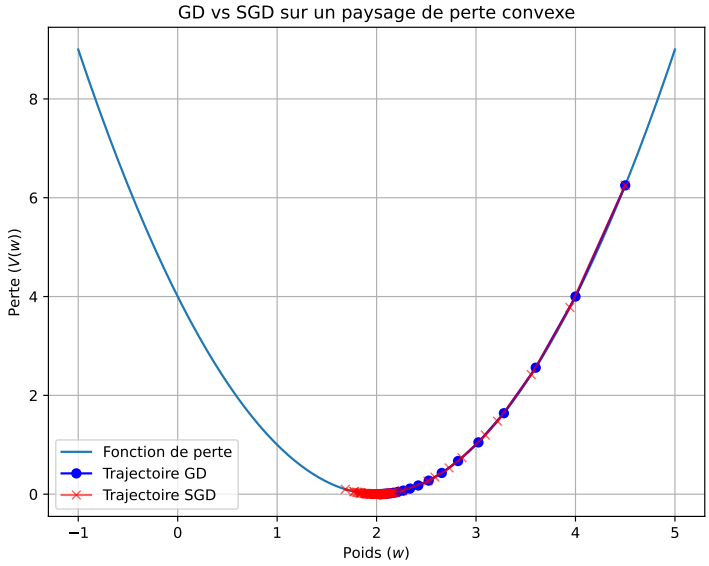


FIGURE 3 – Comparaison des trajectoires de la descente de gradient (GD) et de la descente de gradient stochastique (SGD) sur un paysage de perte convexe. La GD converge de manière déterministe tandis que la SGD suit une trajectoire bruyante.

3. ***Structure de la covariance :*** *La matrice* $\Sigma(\theta)$ *encode l'information géométrique sur la distribution du bruit*

La structure de la matrice de covariance $\Sigma(\theta)$ joue un rôle crucial dans le comportement de la SGD. En particulier, cette matrice peut être fortement dépendante de θ, ce qui induit un bruit multiplicatif dans la dynamique.

Exemple 2.1 (Régression linéaire)**.** Considérons le problème de régression linéaire avec $\ell(\theta,(x,y)) = \frac{1}{2}(y-\theta^T x)^2$. Le gradient stochastique s'écrit :

$$g(\theta,(x,y)) = (y-\theta^T x)(-x) = -x(y-\theta^T x) \tag{8}$$

Le gradient exact est $\nabla L(\theta) = -\mathbb{E}[x(y-\theta^T x)]$, et la matrice de covariance vaut :

$$\Sigma(\theta) = \mathbb{E}[(y-\theta^T x)^2 x x^T] = \mathbb{E}[\varepsilon^2 x x^T] \tag{9}$$

où $\varepsilon = y - \theta^T x$ est le résidu. On observe que $\Sigma(\theta)$ dépend de θ à travers la variance des résidus.

Définition 2.4 (Modèle de bruit additif)**.** Dans de nombreuses analyses théoriques, on fait l'hypothèse simplificatrice d'un bruit additif constant :

$$g(\theta,\xi) = \nabla L(\theta) + \epsilon \tag{10}$$

où ϵ est une variable aléatoire centrée de covariance Σ indépendante de θ. Cette hypothèse permet de linéariser l'analyse mais peut s'avérer restrictive en pratique.

Remarque 2.2 (Impact de la taille de mini-batch)**.** Pour la SGD par mini-batches de taille b, la variance du gradient stochastique est réduite d'un facteur b :

$$\mathrm{Var}\left[\frac{1}{b}\sum_{i=1}^{b} g(\theta,\xi_i)\right] = \frac{1}{b}\Sigma(\theta) \tag{11}$$

Ceci établit un compromis fondamental entre la précision de l'estimation du gradient et l'efficacité computationnelle.

2.3 Exemples : convexité, non-convexité, paysages complexes

Nous illustrons maintenant le comportement de la SGD sur différents types de paysages d'optimisation, en mettant l'accent sur les spécificités du cas non-convexe.

2.3.1 Cas convexe : fondements théoriques

Définition 2.5 (Fonction fortement convexe)**.** Une fonction $L : \mathbb{R}^d \to \mathbb{R}$ est μ-fortement convexe si pour tout $\theta_1, \theta_2 \in \mathbb{R}^d$:

$$L(\theta_2) \geq L(\theta_1) + \nabla L(\theta_1)^T(\theta_2 - \theta_1) + \frac{\mu}{2}\|\theta_2 - \theta_1\|^2 \tag{12}$$

avec $\mu > 0$.

Théorème 2.1 (Convergence SGD dans le cas convexe)**.** *Supposons que L soit μ-fortement convexe et que ∇L soit M-Lipschitz. Si les pas de temps satisfont $\sum_{k=0}^{\infty} \eta_k = \infty$ et $\sum_{k=0}^{\infty} \eta_k^2 < \infty$, alors la SGD converge vers l'optimum global θ^* :*

$$\mathbb{E}[\|\theta_k - \theta^*\|^2] \to 0 \quad \textit{quand } k \to \infty \tag{13}$$

Dans le cas convexe, le comportement de la SGD est bien compris théoriquement. Le bruit stochastique ne perturbe que légèrement la convergence déterministe, et les taux de convergence optimaux sont connus.

2.3.2 Cas non-convexe : défis et opportunités

Le cas non-convexe, caractéristique de l'apprentissage profond, présente des défis théoriques considérables mais aussi des phénomènes inattendus.

Exemple 2.2 (Fonction de Rosenbrock stochastique)**.** Considérons la fonction de Rosenbrock perturbée :

$$L(\theta) = \mathbb{E}_\xi[(1 - \theta_1)^2 + 100(\theta_2 - \theta_1^2)^2 + \xi] \tag{14}$$

où ξ est un bruit de perturbation. Cette fonction possède un minimum global en $(1, 1)$ mais présente une vallée étroite qui rend l'optimisation difficile.

Le gradient stochastique s'écrit :

$$g_1(\theta, \xi) = -2(1 - \theta_1) - 400\theta_1(\theta_2 - \theta_1^2) \tag{15}$$

$$g_2(\theta, \xi) = 200(\theta_2 - \theta_1^2) \tag{16}$$

La SGD doit naviguer dans cette vallée étroite en présence de bruit, ce qui illustre la complexité du cas non-convexe.

Définition 2.6 (Paysage d'optimisation)**.** Pour une fonction $L : \mathbb{R}^d \to \mathbb{R}$, le paysage d'optimisation est caractérisé par :

- Ses points critiques : $\{\theta : \nabla L(\theta) = 0\}$
- Ses minima locaux : points critiques avec $\nabla^2 L(\theta) \succ 0$
- Ses points de selle : points critiques avec $\nabla^2 L(\theta)$ indéfinie
- Ses maxima locaux : points critiques avec $\nabla^2 L(\theta) \prec 0$

Dans le contexte de l'apprentissage profond, les paysages d'optimisation présentent plusieurs caractéristiques remarquables :

Haute dimension : Les réseaux de neurones modernes peuvent avoir des millions ou des milliards de paramètres, créant des espaces d'optimisation de dimension extrêmement élevée.

Multiplicité des minima : Il existe généralement un continuum de minima globaux dus aux symétries du réseau (permutation des neurones, etc.).

Plateaux et ravins : Le paysage peut présenter des régions quasi-plates (plateaux) et des vallées étroites (ravins) qui ralentissent la convergence.

Points de selle : En haute dimension, les points de selle sont exponentiellement plus nombreux que les minima locaux, et leur évasion constitue un défi majeur.

Remarque 2.3 (Phénomène d'évasion des points de selle)**.** Le bruit stochastique de la SGD joue un rôle crucial dans l'évasion des points de selle. Contrairement à la descente de gradient déterministe qui peut rester piégée indéfiniment en un point de selle, la SGD peut s'en échapper grâce aux fluctuations aléatoires. Ce phénomène est formalisé par les théories de sortie de domaine pour les processus stochastiques.

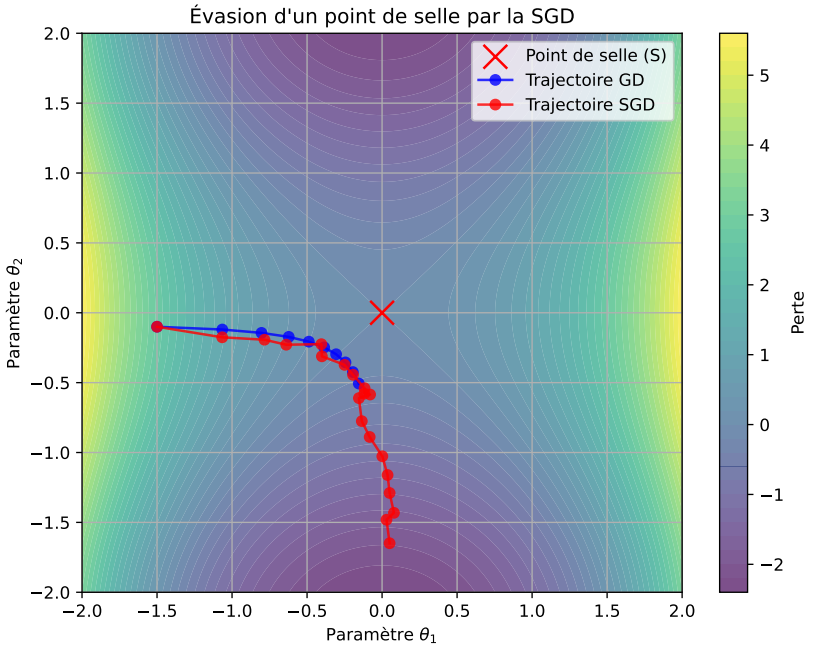


FIGURE 4 – Visualisation d'une trajectoire SGD (en rouge) s'échappant d'un point de selle (marqué par un 'S') dans un paysage de perte non-convexe. Le bruit stochastique de la SGD permet de surmonter la barrière du point de selle, contrairement à la descente de gradient déterministe (en bleu) qui serait piégée.

2.3.3 Paysages complexes en apprentissage profond

Les réseaux de neurones profonds génèrent des paysages d'optimisation particulièrement riches et complexes. Nous examinons quelques propriétés caractéristiques.

Exemple 2.3 (Réseau de neurones à une couche cachée)**.** Considérons un réseau de neurones à une couche cachée avec m neurones :

$$f(x;\theta) = \sum_{j=1}^{m} w_j \sigma(v_j^T x + b_j) \tag{17}$$

où $\theta = (w_1, \ldots, w_m, v_1, \ldots, v_m, b_1, \ldots, b_m)$ et σ est une fonction d'activation.

La fonction de perte pour la régression s'écrit :

$$L(\theta) = \mathbb{E}_{(x,y)}[\frac{1}{2}(y - f(x;\theta))^2] \tag{18}$$

Ce paysage présente plusieurs caractéristiques non-convexes :

— Symétries par permutation des neurones cachés

— Symétries par changement de signe (pour certaines activations)

— Plateaux dans les régions où certains neurones sont saturés

Définition 2.7 (Minima plats vs. minima pointus)**.** Un minimum local θ^* est dit :

— **Plat** si les valeurs propres de $\nabla^2 L(\theta^*)$ sont petites

— **Pointu** si certaines valeurs propres de $\nabla^2 L(\theta^*)$ sont grandes

Cette distinction est cruciale car les minima plats sont empiriquement associés à une meilleure généralisation.

Remarque 2.4 (Conjecture de généralisation)**.** Une conjecture importante en apprentissage profond suggère que la SGD favorise naturellement les minima plats, ce qui expliquerait ses bonnes propriétés de généralisation. Cette conjecture motive en partie l'analyse via les EDS, car la dynamique de Langevin favorise les minima avec une faible courbure selon la distribution de Gibbs.

Le comportement de la SGD sur ces paysages complexes ne peut pas être capturé par les analyses de convergence classiques. C'est précisément cette limitation qui motive l'approche par les équations différentielles stochastiques, qui permet de prendre en compte les effets du bruit de manière plus fine et de comprendre la sélection naturelle des solutions par la dynamique stochastique.

Cette section a établi les fondements mathématiques nécessaires à la compréhension de la SGD comme processus stochastique. La section suivante développe le passage à l'approximation continue et la dérivation de l'EDS de Langevin associée.

3 Approximation continue : la SGD comme EDS

Cette section établit le lien fondamental entre la dynamique discrète de la descente de gradient stochastique (SGD) et sa modélisation continue par une équation différentielle stochastique (EDS). Nous allons dériver l'EDS de Langevin qui approxime le comportement de la SGD sous certaines hypothèses, et discuter les conditions de validité de cette approximation.

3.1 De la SGD à une EDS de Langevin

Comme introduit précédemment, l'intuition principale est que, pour des pas de temps η_k suffisamment petits et sous certaines conditions de régularité du paysage d'optimisation et du bruit, la suite des itérés $(\theta_k)_{k\geq 0}$ de la SGD peut être approchée par une trajectoire continue θ_t solution d'une EDS.

Reprenons l'algorithme de la SGD :

$$\theta_{k+1} = \theta_k - \eta_k \nabla_\theta l(\theta_k, \xi_k)$$

En utilisant la décomposition du gradient stochastique $g(\theta, \xi) = \nabla L(\theta) + \epsilon(\theta, \xi)$, où $\epsilon(\theta, \xi)$ est un terme de bruit centré ($\mathbb{E}_\xi[\epsilon(\theta, \xi)] = 0$), l'équation de mise à jour devient :

$$\theta_{k+1} = \theta_k - \eta_k(\nabla L(\theta_k) + \epsilon(\theta_k, \xi_k))$$

En réarrangeant et en divisant par η_k, on obtient :

$$\frac{\theta_{k+1} - \theta_k}{\eta_k} = -\nabla L(\theta_k) - \epsilon(\theta_k, \xi_k)$$

Pour passer à la limite continue, nous considérons $\eta_k = \eta$ constant et petit, et nous interprétons le terme $\frac{\theta_{k+1}-\theta_k}{\eta}$ comme une dérivée temporelle $d\theta_t/dt$. Le terme de bruit $\epsilon(\theta_k, \xi_k)$ est modélisé comme un bruit blanc.

Sous des hypothèses appropriées (qui seront détaillées dans la sous-section suivante), il est possible de montrer que la dynamique de la SGD peut être approchée par l'EDS de Langevin suivante :

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{2T}dW_t$$

où $L(\theta)$ est la fonction objectif (le potentiel), T est une "température effective" qui dépend du pas de temps η et de la variance du bruit stochastique, et W_t est un mouvement brownien standard multidimensionnel.

L'analogie avec l'équation de Langevin de la mécanique statistique, formalisé dans des travaux récents comme ceux de Li et al. [2], est frappante. Dans ce cadre, $\nabla L(\theta_t)$ représente la force déterministe qui pousse le système vers les minima du potentiel L, tandis que le terme $\sqrt{2T}dW_t$ représente l'agitation aléatoire due au "bruit thermique".

3.2 Forme générale et hypothèses

La dérivation rigoureuse de l'EDS de Langevin à partir de la SGD repose sur des développements asymptotiques et des théorèmes limites (comme le théorème central limite fonctionnel). Plusieurs hypothèses sont généralement nécessaires pour valider cette approximation :

1. **Pas de temps décroissant ou constant et petit** ($\eta_k \to 0$ ou η constant et petit) : Pour que l'approximation continue soit valide, le pas de temps doit être suffisamment petit afin que les changements discrets soient bien représentés par un processus continu.
2. **Régularité de la fonction objectif** : La fonction $L(\theta)$ doit être suffisamment lisse (par exemple, de classe C^2) pour que son gradient $\nabla L(\theta)$ soit bien défini et que des développements de Taylor soient possibles.
3. **Propriétés du bruit stochastique** :
   - **Bruit centré** : $\mathbb{E}_\xi[\epsilon(\theta, \xi)] = 0$. C'est une propriété intrinsèque de la décomposition du gradient stochastique.
   - **Variance du bruit** : La matrice de covariance du bruit, $\Sigma(\theta) = \mathbb{E}_\xi[\epsilon(\theta, \xi)\epsilon(\theta, \xi)^T]$, doit être bien définie et, idéalement, ne pas dégénérer. Dans sa forme la plus simple (bruit additif), $\Sigma(\theta)$ est supposée constante. Dans le cas général, où $\Sigma(\theta)$ dépend de θ, l'EDS devient plus complexe (multiplicative).
   - **Indépendance des bruits** : Les variables aléatoires ξ_k doivent être indépendantes et identiquement distribuées (i.i.d.).
4. **Lien entre la température effective T et les paramètres de la SGD** : La température T est directement liée au pas de temps η et à la covariance du bruit Σ. Pour un bruit additif constant Σ, la relation est souvent $T \propto \eta||\Sigma||$, ou plus précisément, pour un bruit isotrope $\Sigma = \sigma^2 I$, $T = \frac{\eta\sigma^2}{2}$. Ce lien est crucial car il indique que la "force" du bruit dans l'EDS est proportionnelle au pas de temps et à la variance du gradient stochastique.

Lorsque ces conditions sont remplies, en particulier l'hypothèse d'un pas de temps petit, la dynamique de la SGD peut être approximée par un processus de diffusion, dont l'équation de Langevin est une forme canonique.

3.3 Existence et unicité des solutions

L'étude des EDS est un domaine riche de la théorie des probabilités. Pour que l'EDS de Langevin (ou une EDS plus générale) ait un sens mathématique et que ses solutions soient bien définies, des conditions sur les coefficients de l'équation sont requises.

Considérons l'EDS générale :

$$dX_t = b(X_t)dt + \sigma(X_t)dW_t$$

où $b(X_t)$ est le terme de dérive (correspondant à $-\nabla L(\theta_t)$ dans notre cas) et $\sigma(X_t)$ est le terme de diffusion (correspondant à $\sqrt{2T}I$ ou $\sqrt{\Sigma(\theta_t)}$ dans le cas multiplicatif).

Les théorèmes d'existence et d'unicité pour les solutions d'EDS (comme ceux de Itô) exigent des conditions de régularité sur les fonctions b et σ. Les conditions les plus courantes sont les suivantes :

1. **Condition de Lipschitz locale** : Les fonctions $b(\cdot)$ et $\sigma(\cdot)$ doivent être localement Lipschitz continues par rapport à leur argument spatial θ. Autrement dit, pour tout compact $K \subset \mathbb{R}^d$, il existe une constante L_K telle que pour tout $\theta_1, \theta_2 \in K$:

$$||b(\theta_1) - b(\theta_2)|| \leq L_K||\theta_1 - \theta_2||$$

$$||\sigma(\theta_1) - \sigma(\theta_2)|| \leq L_K||\theta_1 - \theta_2||$$

   Pour l'EDS de Langevin $d\theta_t = -\nabla L(\theta_t)dt + \sqrt{2T}dW_t$, cela signifie que $\nabla L(\theta)$ doit être Lipschitz continu, ce qui est souvent le cas pour les fonctions de perte utilisées en apprentissage profond (au moins localement).

2. **Condition de croissance linéaire** : Les fonctions $b(\cdot)$ et $\sigma(\cdot)$ doivent satisfaire une condition de croissance linéaire. Il existe des constantes C_1, C_2 telles que pour tout $\theta \in \mathbb{R}^d$:

$$||b(\theta)|| \leq C_1(1 + ||\theta||)$$

$$||\sigma(\theta)||^2 \leq C_2(1 + ||\theta||^2)$$

   Ces conditions garantissent que les solutions ne "divergent" pas à l'infini en temps fini.

Si ces conditions sont remplies, on peut affirmer qu'il existe une solution forte unique à l'EDS, ce qui permet d'étudier les propriétés des trajectoires de θ_t de manière cohérente. Dans le contexte de la SGD, la validité de ces hypothèses dépend de la nature spécifique de la fonction de perte $L(\theta)$ et de la distribution du bruit ξ. Ces résultats théoriques fournissent la base pour analyser le comportement asymptotique de la SGD via les outils de la théorie des EDS.

4 Équation de Fokker-Planck et comportement asymptotique

Après avoir établi l'approximation continue de la SGD par une EDS de Langevin, cette section explore les conséquences de cette modélisation pour comprendre le comportement à long terme de la distribution de probabilité des paramètres du modèle. L'outil central pour cette analyse est l'équation de Fokker-Planck.

4.1 Évolution de la densité de probabilité

Alors que l'EDS de Langevin décrit l'évolution d'une **trajectoire individuelle** θ_t des paramètres, l'équation de Fokker-Planck (également connue sous le nom de Kolmogorov forward equation) décrit l'évolution temporelle de la **densité de probabilité** $p(\theta, t)$ de ces paramètres dans l'espace $\mathbb{R}^d$. C'est un changement de perspective fondamental, passant de la dynamique individuelle à la dynamique de l'ensemble des trajectoires.

Pour une EDS générale de la forme $dX_t = b(X_t)dt + \sigma(X_t)dW_t$, où $b(X_t)$ est le vecteur de dérive et $\sigma(X_t)$ est la matrice de diffusion, l'équation de Fokker-Planck pour la densité de probabilité $p(x,t)$ est donnée par :

$$\frac{\partial p(x,t)}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i}[b_i(x)p(x,t)] + \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d} \frac{\partial^2}{\partial x_i \partial x_j}[(\sigma(x)\sigma(x)^T)_{ij}p(x,t)]$$

Dans le cas de notre EDS de Langevin modélisant la SGD, $d\theta_t = -\nabla L(\theta_t)dt + \sqrt{2T}dW_t$, nous avons $b(\theta) = -\nabla L(\theta)$ et $\sigma(\theta) = \sqrt{2T}I$, où I est la matrice identité. Par conséquent, $\sigma(\theta)\sigma(\theta)^T = (2T)I$. L'équation de Fokker-Planck associée devient alors :

$$\frac{\partial p(\theta,t)}{\partial t} = \nabla \cdot (\nabla L(\theta)p(\theta,t)) + T\Delta p(\theta,t)$$

où $\nabla\cdot$ est l'opérateur divergence et Δ est l'opérateur laplacien ($\Delta p = \sum_{i=1}^{d} \frac{\partial^2 p}{\partial \theta_i^2}$).

Cette équation différentielle partielle décrit comment le "nuage" de probabilité des paramètres se diffuse et se déplace dans le temps sous l'influence du paysage de la fonction objectif $L(\theta)$ et du bruit stochastique. Le premier terme du côté droit (terme de dérive) tend à "attirer" la densité vers les minima de $L(\theta)$, tandis que le second terme (terme de diffusion) modélise l'étalement de la densité dû au bruit.

4.2 Distribution stationnaire et loi de Gibbs

Un aspect crucial de l'analyse via l'équation de Fokker-Planck est la recherche de sa distribution stationnaire (ou état d'équilibre), notée $p_{eq}(\theta)$. C'est la distribution vers laquelle $p(\theta,t)$ converge lorsque $t \to \infty$, et pour laquelle $\frac{\partial p(\theta,t)}{\partial t} = 0$. À l'équilibre, le flux de probabilité dû à la dérive est exactement équilibré par le flux dû à la diffusion.

Pour l'EDS de Langevin, il est bien connu que la distribution stationnaire est donnée par la célèbre loi de Gibbs (ou distribution de Boltzmann) :

$$p_{eq}(\theta) \propto \exp\left(-\frac{L(\theta)}{T}\right)$$

ou, plus précisément :

$$p_{eq}(\theta) = \frac{1}{Z}\exp\left(-\frac{L(\theta)}{T}\right)$$

où $Z = \int_{\mathbb{R}^d} \exp\left(-\frac{L(\theta)}{T}\right) d\theta$ est la fonction de partition, une constante de normalisation.

Cette distribution de Gibbs est d'une importance capitale. Elle indique que, à l'équilibre, la probabilité de trouver les paramètres θ dans une certaine région de l'espace est inversement proportionnelle à la valeur de la fonction objectif $L(\theta)$ dans cette région, pondérée par la température T. Plus précisément :

- Les régions où $L(\theta)$ est faible (minima) ont une forte probabilité.
- Les régions où $L(\theta)$ est élevée (maxima, points de selle éloignés) ont une faible probabilité.
- Le paramètre T joue le rôle d'une "température". Si $T \to 0$, la distribution se concentre de plus en plus sur les minima globaux (le système se "refroidit"). Si T est grand, la distribution est plus uniforme et permet d'explorer de plus larges régions de l'espace (le système est "chaud").

Ce résultat théorique confirme l'intuition que la SGD, par sa nature stochastique, ne converge pas vers un point unique mais plutôt vers une distribution de points autour des minima du paysage d'optimisation. La forme de cette distribution est dictée par la loi de Gibbs et ce comportement est l'une des raisons pour lesquelles la SGD peut être vue comme une forme d'inférence bayésienne approximative [3].

4.3 Interprétation thermodynamique

L'analogie avec la mécanique statistique et la thermodynamique est profonde et fournit des intuitions précieuses pour comprendre la SGD :

— **Énergie et potentiel** : La fonction objectif $L(\theta)$ est interprétée comme l'énergie potentielle du système. Le processus d'optimisation est analogue à un système physique qui cherche à minimiser son énergie.

— **Température et bruit** : La température effective T est directement liée à l'intensité du bruit stochastique introduit par la SGD. Un pas de temps plus grand ou une variance de gradient plus élevée se traduisent par une température plus élevée, favorisant une exploration plus large de l'espace.

— **Minima plats et minima pointus** : La loi de Gibbs $p_{eq}(\theta) \propto \exp\left(-\frac{L(\theta)}{T}\right)$ favorise naturellement les minima plats. Pour un même niveau de $L(\theta)$, un minimum plat correspond à une région où $L(\theta)$ varie peu, ce qui signifie que $\exp\left(-\frac{L(\theta)}{T}\right)$ y sera relativement élevé sur une plus grande étendue. Inversement, un minimum pointu, où $L(\theta)$ augmente rapidement autour du minimum, aura une densité de probabilité plus concentrée et moins "d'étendue" en termes de volume de l'espace des paramètres. En d'autres termes, les minima plats capturent plus de "masse de probabilité" à l'équilibre que les minima pointus.

— **Transitions de phase** : Dans des systèmes plus complexes (comme les réseaux de neurones profonds), des phénomènes analogues aux transitions de phase en physique statistique peuvent se produire, où le comportement du système change qualitativement en fonction de la température effective (par exemple, passage d'un régime d'exploration intense à un régime de convergence).

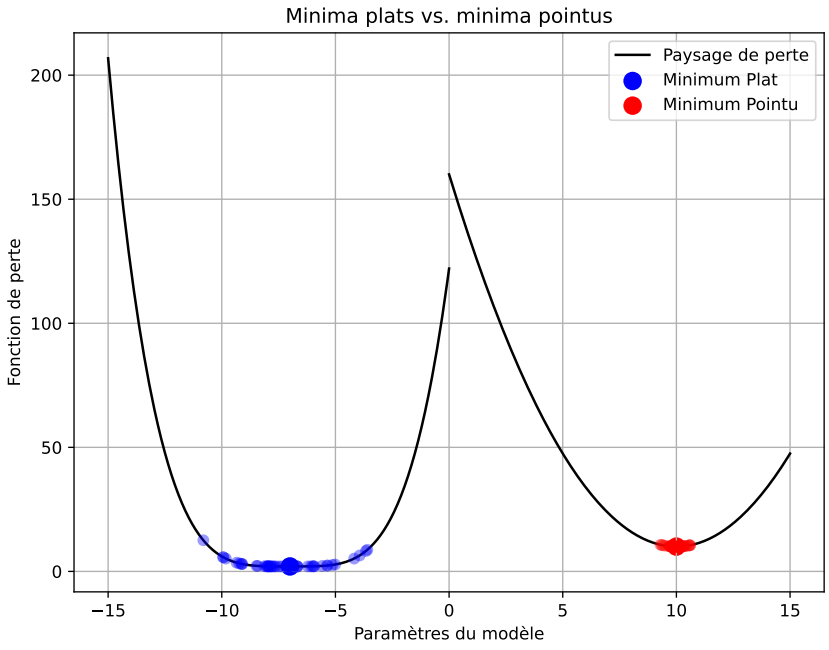


FIGURE 5 – Représentation schématique de deux minima locaux. Le minimum plat (à gauche) a une faible courbure, tandis que le minimum pointu (à droite) a une forte courbure. Le bruit stochastique de la SGD la pousse à explorer de manière plus large autour des minima, ce qui la rend plus susceptible de s'installer dans un minimum plat favorable à la généralisation.

Cette interprétation thermodynamique non seulement explique pourquoi la SGD "préfère"

les minima plats (une propriété empiriquement liée à une meilleure généralisation), mais fournit également un cadre conceptuel puissant pour concevoir de nouvelles variantes de la SGD, comme le recuit simulé (simulated annealing), qui manipulent explicitement la température effective pour améliorer l'optimisation. La section suivante explorera plus en détail ces conséquences pour l'apprentissage automatique.

5 Conséquences en apprentissage

Dans la section précédente, nous avons établi que la SGD peut être modélisée par une équation différentielle stochastique de type Langevin. Cette approximation n'est pas qu'un outil théorique ; elle fournit un cadre puissant pour comprendre et expliquer des phénomènes empiriques complexes observés en apprentissage profond. En particulier, elle éclaire le rôle du bruit stochastique non pas comme une simple perturbation, mais comme un élément essentiel de la dynamique d'optimisation, qui influence de manière profonde les propriétés de la solution trouvée.

5.1 Convergence vers des minima plats

L'un des constats empiriques les plus importants en apprentissage profond est que les optimiseurs stochastiques comme la SGD tendent à converger vers des minima du paysage de perte qui sont "plats", c'est-à-dire qui présentent des courbures faibles. Ces minima plats sont fortement corrélés à une meilleure capacité de généralisation du modèle. L'interprétation par les EDS fournit une explication élégante à ce phénomène.

Considérons l'EDS de Langevin qui régit la dynamique des poids $\boldsymbol{w}$:

$$d\boldsymbol{w}_t = -\nabla V(\boldsymbol{w}_t)dt + \sqrt{2\eta\beta^{-1}}d\boldsymbol{B}_t \tag{19}$$

Le terme de bruit stochastique $\sqrt{2\eta\beta^{-1}}d\boldsymbol{B}_t$ peut être vu comme une force de "diffusion" ou de "chaleur". Cette diffusion permet à la trajectoire d'explorer la surface de perte, y compris de s'échapper des minima locaux pointus. Au lieu de se fixer sur le premier minimum rencontré, la "température effective" du système encourage la trajectoire à se stabiliser dans des régions plus larges et plus stables. Un minimum plat correspond à une région où le potentiel $V(\boldsymbol{w})$ varie lentement, permettant au bruit de maintenir la trajectoire dans son voisinage sans la pousser à s'éloigner rapidement. En revanche, un minimum pointu, associé à une courbure élevée, correspond à une "vallée" étroite d'où il est difficile d'échapper. La dynamique de l'EDS est donc naturellement plus stable dans les régions plates, ce qui se traduit par une préférence pour ces types de minima.

5.2 Généralisation et bruit gaussien

Le lien entre la planéité des minima et la généralisation peut également être interprété à travers le prisme de l'EDS. Un minimum plat est intrinsèquement plus robuste aux perturbations des poids. En pratique, la généralisation dépend de la capacité du modèle à fonctionner de manière fiable sur des données nouvelles, qui peuvent présenter de légères variations par rapport aux données d'entraînement. Si un modèle est situé au fond d'un minimum plat, une petite perturbation de ses poids (induite par exemple par le bruit de la SGD sur les données de test) n'affectera que peu sa performance. À l'inverse, si le modèle est situé dans un minimum pointu, une petite perturbation des poids peut le pousser hors du minimum, entraînant une chute drastique de la performance.

Dans notre modèle d'EDS, le bruit gaussien est un facteur qui contribue directement à cette robustesse. Il agit comme un régulateur implicite. L'intensité du bruit, proportionnelle à $\sqrt{\eta\beta^{-1}}$, détermine le degré de "secouage" appliqué aux poids. En encourageant l'exploration des minima

plats, ce bruit aide le modèle à trouver des solutions qui sont non seulement optimales sur les données d'entraînement, mais aussi stables et peu sensibles aux petites fluctuations, garantissant ainsi une meilleure généralisation.

5.3 Lien avec les méthodes bayésiennes

L'interprétation de la SGD comme une EDS de Langevin établit un pont fascinant avec les méthodes d'inférence bayésienne. L'objectif de l'inférence bayésienne est de calculer ou d'échantillonner à partir de la distribution a posteriori des paramètres $\boldsymbol{w}$ étant donné les données $\mathcal{D}$: $p(\boldsymbol{w}|\mathcal{D})$. Selon le théorème de Bayes, cette distribution est proportionnelle à la probabilité a priori $p(\boldsymbol{w})$ multipliée par la vraisemblance $p(\mathcal{D}|\boldsymbol{w})$, soit $p(\boldsymbol{w}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{w})p(\boldsymbol{w})$.

En se plaçant dans le cas où les poids suivent une distribution a priori gaussienne et la vraisemblance est définie par une fonction de perte quadratique (souvent une bonne approximation), la distribution a posteriori prend la forme de la distribution de Boltzmann : $p(\boldsymbol{w}|\mathcal{D}) \propto \exp(-V(\boldsymbol{w}))$. Or, la distribution stationnaire de l'EDS de Langevin est précisément la distribution de Boltzmann. Par conséquent, la trajectoire de l'EDS finit par osciller autour des régions à haute probabilité de la distribution a posteriori.

Cela signifie que la SGD, via son comportement modélisé par l'EDS, ne se contente pas de trouver un unique ensemble de poids optimaux, mais peut être vue comme une méthode d'échantillonnage MCMC (Markov Chain Monte Carlo) explorant les régions de l'espace des paramètres qui ont une forte probabilité a posteriori. Cette perspective est un changement de paradigme : l'optimisation n'est plus la seule finalité, mais elle sert un objectif d'échantillonnage, fournissant non pas un point, mais une distribution de solutions plausibles. Cela ouvre la voie à des liens profonds avec le calcul d'incertitudes et les méthodes de régularisation bayésiennes en apprentissage profond.

6 Expériences numériques

Pour illustrer et valider les concepts théoriques développés dans les sections précédentes, nous présentons ici les résultats de plusieurs expériences numériques. L'objectif est de visualiser concrètement la dynamique de la SGD et de la comparer à celle de l'EDS de Langevin, tout en mettant en évidence les conséquences de ce comportement sur le paysage de perte.

6.1 Simulation d'une SGD sur un potentiel double puits

Afin de mieux comprendre la dynamique d'échappement des minima locaux, nous simulons la trajectoire de la SGD sur un paysage de perte non-convexe simple, inspiré de la physique statistique : un potentiel double puits. Ce potentiel, défini par la fonction $V(w) = w^4 - w^2$, présente deux minima locaux et un maximum local (une barrière d'énergie) au centre. Cette topographie simplifiée permet d'observer clairement comment le terme de bruit stochastique permet au processus de se déplacer d'un minimum à un autre.

La Figure 6 montre la forme du potentiel ainsi que plusieurs trajectoires de la SGD, partant du même point initial. On peut observer comment, en présence de bruit, certaines trajectoires parviennent à surmonter la barrière de potentiel pour atteindre le minimum opposé.

6.2 Visualisation de trajectoires et de distributions

L'interprétation de la SGD comme un processus diffusif est particulièrement claire lors de la visualisation des trajectoires à long terme. Sur un paysage de perte plus complexe (mais toujours en 2D pour la visualisation), nous traçons le chemin parcouru par les poids du modèle. La Figure 7 montre d'un côté la trajectoire d'optimisation et de l'autre la distribution stationnaire des poids.

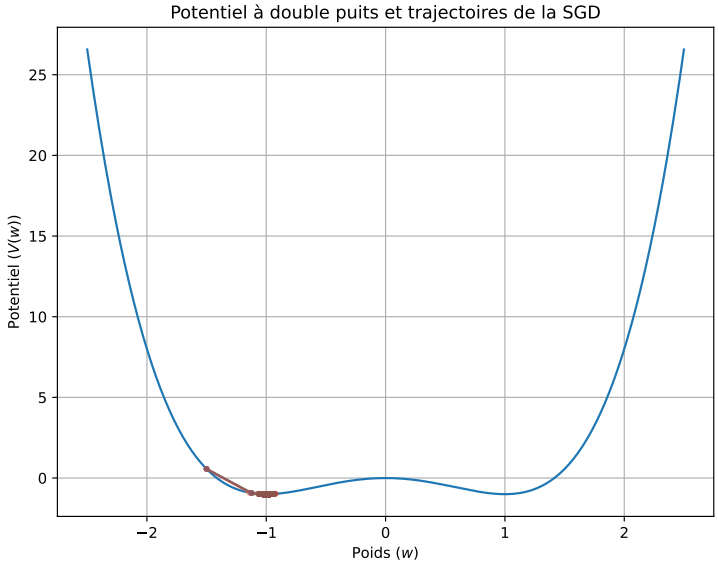
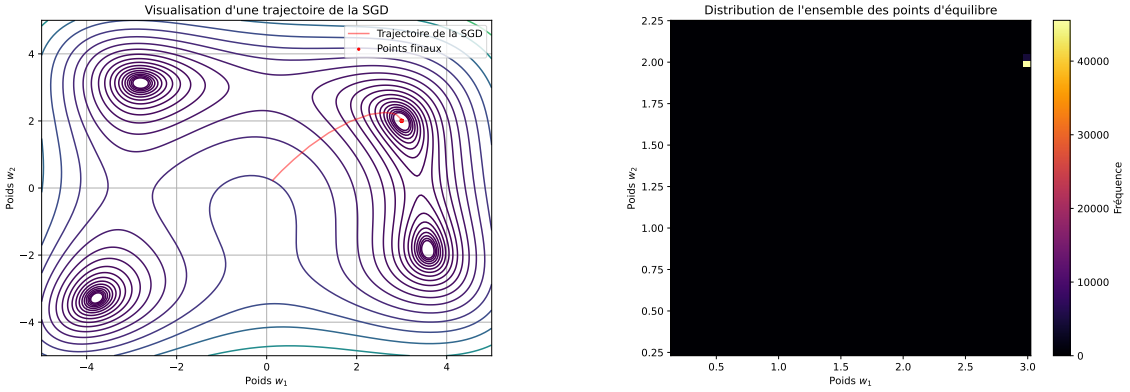


FIGURE 6 – Potentiel à double puits et quelques trajectoires de la SGD. Les trajectoires sont colorées en fonction du temps.



(a) Visualisation d'une trajectoire de la SGD sur un paysage de perte 2D.

(b) Distribution de l'ensemble des points d'équilibre explorés par la SGD.

FIGURE 7 – Visualisation des trajectoires et de la distribution stationnaire.

La sous-figure 7a illustre le comportement exploratoire : au lieu de converger vers un point unique, la trajectoire de la SGD oscille autour d'une région. La sous-figure 7b montre la densité de probabilité des poids. On peut clairement observer que les points s'accumulent dans une région large et plate du paysage de perte, confirmant ainsi l'idée que le bruit pousse le système vers des minima "plats" (à faible courbure) et stables.

6.3 Comparaison avec l'équation de Fokker–Planck

Le modèle SDE de Langevin nous permet de prédire l'évolution de la distribution de probabilité des poids via l'équation de Fokker-Planck. Cette équation différentielle aux dérivées partielles décrit la dynamique de la densité de probabilité $p(\boldsymbol{w}, t)$. La solution stationnaire de cette équation, $p_\infty(\boldsymbol{w})$, est la distribution de Boltzmann, qui est la distribution que nous avons observé empiriquement en 7b.

La Figure 8 présente une comparaison visuelle entre la distribution de probabilité obtenue par simulation de la SGD et la solution analytique de l'équation de Fokker-Planck. La forte concordance entre les deux résultats valide l'approche SDE comme un modèle précis et prédictif pour la dynamique de la SGD.

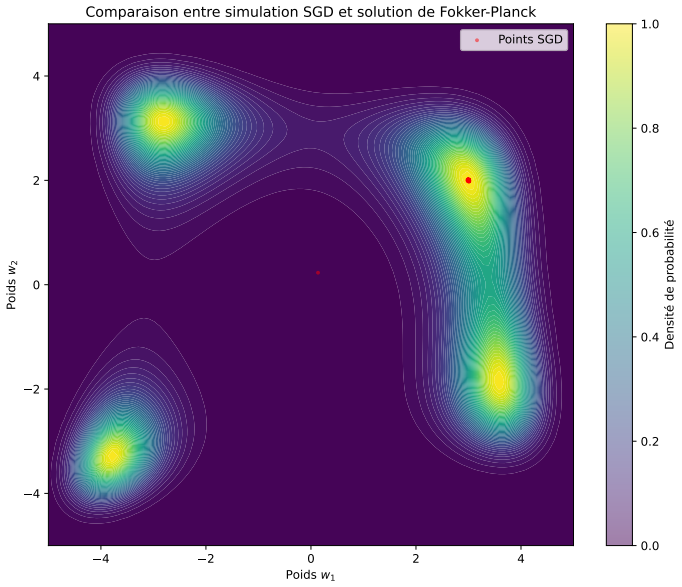


FIGURE 8 – Comparaison de la distribution de probabilité des poids entre la simulation SGD et la solution de l'équation de Fokker-Planck.

7 Conclusion

Ce rapport a exploré le lien profond et fructueux entre la dynamique de la descente de gradient stochastique (SGD) et le formalisme des équations différentielles stochastiques (EDS), en se basant notamment sur l'analogie avec l'équation de Langevin de la physique statistique. Nous avons montré que, loin d'être un simple ajout de bruit, la stochasticité inhérente à la SGD est un mécanisme puissant et structuré qui régit la convergence et la capacité de généralisation des modèles.

En résumé, les principaux résultats présentés dans ce rapport sont les suivants :

- La SGD, par son utilisation de mini-batchs, peut être modélisée comme un processus de Langevin, où le gradient stochastique agit comme une force de "dérive" et la variance du gradient comme une force de "diffusion".
- Cette interprétation fournit un cadre théorique pour expliquer des observations empiriques clés en apprentissage profond, telles que la préférence de la SGD pour les minima plats du paysage de perte. Nous avons démontré comment le bruit diffusif aide l'algorithme à échapper aux minima étroits et à explorer des régions plus larges, propices à une meilleure généralisation.
- L'analogie avec les EDS de Langevin révèle un lien conceptuel puissant avec l'inférence bayésienne. La distribution stationnaire de l'EDS peut être interprétée comme une approximation de la distribution a posteriori des poids du modèle, suggérant que la SGD n'est pas seulement un optimiseur, mais aussi un échantillonneur de cette distribution.

Malgré l'élégance et la puissance de l'approche SDE, elle présente certaines limites. La modélisation est une simplification, et les hypothèses, notamment celle d'un bruit gaussien, ne sont

pas toujours parfaitement représentatives des processus complexes à l'œuvre dans les réseaux de neurones profonds. De plus, la notion de "température" effective est une approximation qui ne rend pas compte de toutes les subtilités des taux d'apprentissage adaptatifs.

Ces limites ouvrent la voie à de nombreuses perspectives de recherche. Des travaux futurs pourraient explorer des modèles d'EDS avec des bruits non gaussiens pour mieux capturer la nature des données. L'étude de stratégies d'optimisation plus avancées, comme les méthodes d'"annealing" (réduction progressive du taux d'apprentissage), pourrait être analysée formellement dans le cadre des EDS.

En conclusion, l'analogie entre SGD et EDS n'est pas qu'une simple curiosité mathématique. Elle constitue une lentille théorique essentielle pour déchiffrer le comportement des algorithmes d'optimisation modernes et pour concevoir la prochaine génération de méthodes d'apprentissage automatique.

A Éléments techniques complémentaires

Cette annexe regroupe les éléments mathématiques et les démonstrations simplifiées qui soutiennent les arguments développés dans le rapport. Son objectif est de fournir les bases techniques nécessaires pour une compréhension plus approfondie de la modélisation de la SGD par les équations différentielles stochastiques (EDS).

A.1 Rappels sur les EDS

Les équations différentielles stochastiques sont un outil mathématique qui permet de modéliser des systèmes dynamiques sous l'influence de forces aléatoires (pour une introduction plus détaillée, se référer à Evans [1]). Contrairement aux équations différentielles ordinaires (EDO) qui décrivent des trajectoires lisses et déterministes, les EDS introduisent un terme de bruit qui rend les trajectoires aléatoires et discontinues.

Définition A.1 (Mouvement brownien)**.** Le mouvement brownien standard, noté B_t, est un processus stochastique à temps continu caractérisé par les propriétés suivantes :

1. $B_0 = 0$ (le processus commence à zéro).
2. Les accroissements sont stationnaires et indépendants : pour $s < t$, la variable aléatoire $B_t - B_s$ est indépendante des valeurs passées du processus.
3. Les accroissements sont distribués selon une loi normale : $B_t - B_s \sim \mathcal{N}(0, t - s)$.

C'est la représentation mathématique du bruit blanc, une source de perturbation fondamentale dans les EDS.

L'intégration d'une fonction par rapport au mouvement brownien ne peut pas être traitée avec le calcul de Riemann-Stieltjes classique en raison de la non-différentiabilité des trajectoires. Cela a conduit au développement du calcul stochastique, dont l'outil central est l'intégrale d'Itô. L'EDS de Langevin, utilisée dans ce rapport, s'écrit sous la forme :

$$d\boldsymbol{w}_t = f(\boldsymbol{w}_t)dt + g(\boldsymbol{w}_t)d\boldsymbol{B}_t$$

où f est le terme de dérive (le gradient dans le cas de la SGD) et g est le terme de diffusion (le bruit stochastique).

A.2 Éléments de preuve simplifiée de la convergence de la SGD vers l'EDS de Langevin

Cette section présente une esquisse simplifiée de la preuve qui lie la dynamique discrète de la SGD à son analogue continu de Langevin. La preuve repose sur l'approximation d'un

processus discret (la SGD) par un processus continu (l'EDS) via une limite. Cette approche est une simplification de la méthode rigoureuse décrite par Li et al. [2].

L'itération de la SGD est donnée par :

$$\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta \nabla \ell(\boldsymbol{w}_k, \xi_k)$$

où η est le taux d'apprentissage et $\nabla \ell(\boldsymbol{w}_k, \xi_k)$ est le gradient stochastique sur le mini-batch ξ_k.

Nous supposons que le gradient stochastique peut être décomposé en une partie exacte et une partie bruitée :

$$\nabla \ell(\boldsymbol{w}_k, \xi_k) = \nabla L(\boldsymbol{w}_k) + \delta_k$$

où $\nabla L(\boldsymbol{w}_k) = \mathbb{E}[\nabla \ell(\boldsymbol{w}_k, \xi_k)]$ est le gradient vrai et δ_k est une variable aléatoire de bruit de moyenne nulle et de variance Σ.

L'itération peut alors s'écrire :

$$\boldsymbol{w}_{k+1} - \boldsymbol{w}_k = -\eta \nabla L(\boldsymbol{w}_k) - \eta \delta_k$$

En introduisant une échelle de temps continue $t = k\Delta t$ avec $\Delta t = \eta$, on peut réécrire l'équation comme :

$$\frac{\boldsymbol{w}_{k+1} - \boldsymbol{w}_k}{\Delta t} = -\nabla L(\boldsymbol{w}_k) - \delta_k$$

En prenant la limite pour $\eta \to 0$, le terme de gauche se rapproche de $d\boldsymbol{w}_t/dt$. Le terme δ_k ne tend pas vers une valeur nulle, mais son comportement agrégé sur le temps est capturé par le Théorème Central Limite. La somme des variables aléatoires indépendantes δ_k converge, après normalisation, vers un processus de Wiener (mouvement brownien).

Plus rigoureusement, en définissant le processus d'interpolation linéaire $\tilde{\boldsymbol{w}}_t = \boldsymbol{w}_k$ pour $t \in [k\eta, (k+1)\eta)$, on peut montrer que sa dynamique converge, au sens de la distribution, vers l'EDS de Langevin :

$$d\boldsymbol{w}_t = -\nabla L(\boldsymbol{w}_t)dt + \sqrt{2\eta\Sigma} d\boldsymbol{B}_t$$

Cette relation justifie formellement le lien entre la dynamique discrète, bruyante, de la SGD et le comportement continu, stochastique, d'une particule sous l'influence d'un potentiel et d'un bruit brownien. L'expression de la variance Σ est directement liée à la variance des gradients sur le mini-batch, et le taux d'apprentissage η contrôle l'intensité du bruit. C'est ce qui est interprété comme la "température effective" du système.

Références

[1] Lawrence C Evans. *An introduction to stochastic differential equations*. American Mathematical Society, 2013.

[2] Qianxiao Li, Cheng Tai, and Weinan E. Stochastic modified equations and adaptive stochastic gradient algorithms. *arXiv preprint arXiv :1511.06251*, 2017.

[3] Stephan Mandt, Matthew D Hoffmann, and David M Blei. Stochastic gradient descent as approximate bayesian inference. *Journal of Machine Learning Research*, 18(1) :1–35, 2017.

[4] Alexis Nasr. Descente stochastique du gradient. Notes de cours, Aix Marseille Université, 2018.