# Numerical analysis of highly oscillatory Stochastic PDEs

Charles-Edouard Bréhier

**THÈSE / ENS CACHAN - BRETAGNE**
*sous le sceau de l'Université européenne de Bretagne*
pour obtenir le titre de
**DOCTEUR DE L'ÉCOLE NORMALE SUPÉRIEURE DE CACHAN**
*Mention : Mathématiques*
**École doctorale MATISSE**


présentée par
## Charles-Edouard Bréhier
Préparée à l'Unité Mixte de Recherche n° 6625
Institut de recherche mathématique de Rennes


# Analyse numérique d'EDP Stochastiques hautement oscillantes

**Thèse soutenue le 27 novembre 2012**
devant le jury composé de :

**Sandra Cerrai**
Professeur à l'université du Maryland / *rapportrice*

**Tony Lelièvre**
Professeur à l'école des Ponts ParisTech / *rapporteur*

**Anne De Bouard**
Directrice de Recherches à l'École Polytechnique / *examinatrice*

**François Delarue**
Professeur à l'université de Nice Sophia-Antipolis / *examinateur*

**Florent Malrieu**
Maître de Conférences à l'université de Rennes 1 / *examinateur*

**Andreas Prohl**
Professeur à l'université de Tübingen / *examinateur*

**Arnaud Debussche**
Professeur à l'ENS Cachan-Bretagne / *directeur de thèse*

**Erwan Faou**
Directeur de Recherches à l'ENS Cachan-Bretagne / *directeur de thèse*

# ANALYSE NUMÉRIQUE D'EDP STOCHASTIQUES HAUTEMENT OSCILLANTES.

**THESE**

CHARLES-EDOUARD BRÉHIER
ENS CACHAN, ANTENNE DE BRETAGNE

# Remerciements

Mes premiers remerciements vont à Arnaud et à Erwan, sans qui l'écriture de cette thèse n'aurait pas été possible: pour m'avoir proposé des problèmes passionants; pour avoir été très disponibles et investis, toujours prêts à répondre à la moindre question, avec bonne humeur et enthousiasme; pour avoir relu ma "prose" (celle que vous tenez entre vos mains), et l'avoir nettement améliorée bien souvent...

Je voudrais aussi remercier mes deux rapporteurs, Sandra et Tony, pour avoir relu ma thèse, et pour leurs suggestions et remarques.

Merci également aux autres membres du jury: Anne, Andreas, Florent et François. C'est un honneur que vous ayez accepté de faire partie de ce jury.

Durant ces trois ans, j'ai passé d'excellents moments à travailler à l'ENS. C'est donc l'occasion de saluer comme il se doit l'ensemble des personnes que j'ai pu croiser: membres des équipes de mathématiques à Ker Lann et à l'université, et personnels qui nous facilitent la vie.

L'ambiance de travail au plateau de maths a toujours été excellente (parfois tellement qu'on en a oublié de travailler...), et ça a été très important. Plus particulièrement, ces trois années avec Guillaume et Thibaut ont été super: soirées, pauses café, poissons, discussions... et des tas d'autres choses. Merci également à Quentin, Marie, Sylvain, Martina, Katharina, Julia, Ludovic, Agnès, Shen-Shen, et aux autres doctorants de l'irmar.

Je remercie ma famille, et surtout mes parents, ma soeur, mon frère et leurs conjoints: pour le soutien, leur attention, et pour m'avoir supporté durant toutes ces années de mathématiques.

Je tiens à remercier chaleureusement David et Ronan, sur qui on peut toujours compter, malgré la distance et le temps qui nous manque pour se voir plus souvent...

Je salue mes amis Lyonnais, avec qui j'ai toujours eu plaisir à retourner passer quelques bons moments: merci à Alex B., Alex R., Arnaud, Damien, François, Lionel, Marc-Aurèle, Nicolas.

Je profite aussi de l'occasion pour adresser un message à tous mes anciens professeurs: notamment ceux de mathématiques, au lycée, en prépa, puis à l'ENS Lyon. C'est un peu grâce à eux que j'ai pu m'intéresser à ce qui est finalement devenu une passion.

Histoire de n'oublier personne, disons que je remercie les gens que j'ai pu rencontrer durant toutes ces années, notamment à l'occasion des conférences auxquelles j'ai pu assister.

Pour pouvoir travailler, j'ai besoin d'écouter de la musique, donc il faut que je remercie également tous les musiciens que j'ai pu écouter durant ces années. Plus particulièrement, merci à Eric Clapton, Jeff Beck, Carlos Santana et Marcus Miller pour leurs merveilleux concerts, et à David, Germain et Ronan pour m'y avoir accompagné. Et merci à Frank Zappa, Jimi Hendrix, Stevie Ray Vaughan et tous les autres (la liste est longue...) pour leurs albums.

*What is your conceptual continuity?* (F. Zappa)

*The crux of the biscuit is the apostrophe.* (F. Zappa)

*Round things are boring.*(F. Zappa)

*Our lives today are not conducted in linear terms. They are much more quantified; a stream of random events is taking place.* (J.G. Ballard)

*Some things take so long, but how do I explain?* (G.Harrison)

## Table des figures

# Introduction

Cette thèse contient des travaux portant sur l'analyse numérique et la théorie des probabilités. Dans un premier temps, on propose l'analyse d'une méthode numérique d'approximation d'un système d'équations aux dérivées partielles stochastiques hautement oscillant: le système présente deux composantes évoluant à des vitesses différentes, et la méthode proposée devient plus efficace qu'une méthode directe lorsque la séparation des échelles de temps devient plus importante. Du point de vue théorique, le système se situe alors dans un régime dans lequel la composante dite lente est approchée par un processus solution d'une équation dite moyennée; la moyenne est obtenue à partir du comportement asymptotique de la composante rapide, et on peut voir le résultat sous un angle "loi des grands nombres". Cet effet se retrouve au niveau de la discrétisation du système à l'aide d'un schéma numérique de type Euler, et il s'agit d'une part de définir une méthode de discrétisation qui tient compte de la séparation des échelles de temps, en suivant le principe de moyennisation en temps continu, et d'évaluer précisément toutes les sources d'erreur afin de prouver l'efficacité d'une telle méthode. Les trois premiers chapitres sont consacrés à ce problème, sous les différents aspects mentionnés.

Dans un second temps, on s'intéresse à un schéma de discrétisation d'équations aux dérivées partielles à l'aide d'une méthode de Monte-Carlo associée à une technique semi-lagrangienne. Celle-ci est construite à partir d'une formule de représentation probabiliste, permettant de relier la solution évaluée à des instants distincts. Du point de vue de la discrétisation, on approche la solution à des temps discrets et en des points appartenant à une grille en espace; afin d'utiliser une version de la formule de représentation, il est nécessaire de disposer d'une méthode d'interpolation. Dans le cas qui nous intéresse, la formule de représentation met en jeu un processus stochastique et le calcul de l'espérance de certaines variables aléatoires associées: une technique d'approximation possible est une méthode de Monte-Carlo. D'un point de vue théorique, on montre dans un cas simple mais susceptible de généralisation que la variance induite par cette approximation est mieux contrôlée que dans une méthode de Monte-Carlo traditionnelle: dans un régime liant les pas de discrétisation spatiale et temporelle, elle dépend du pas de temps utilisée. On propose également l'application de la méthode à différents problèmes en vue de mettre en avant son potentiel, en fournissant des simulations numériques.

Les deux parties présentées ci-dessus sont largement indépendantes.

Dans la suite de cette introduction, nous proposons une description des problèmes, des résultats obtenus et des idées de leur preuve et des techniques employées.

## Introduction

### 0.1. Analyse théorique et numérique d'un système d'EDPS hautement oscillant

Cette section présente le contenu des Chapitres 1, 2 et 3.

0.1.1. **Description du problème.** Le système d'équations aux dérivées partielles stochastiques qui est étudié est de la forme suivante:

$$\begin{aligned} \frac{\partial x^\epsilon(t,\xi)}{\partial t} &= \frac{\partial^2 x^\epsilon(t,\xi)}{\partial \xi^2} + f(\xi, x^\epsilon(t,\xi), y^\epsilon(t,\xi)), \\ \frac{\partial y^\epsilon(t,\xi)}{\partial t} &= \frac{1}{\epsilon}\frac{\partial^2 y^\epsilon(t,\xi)}{\partial \xi^2} + \frac{1}{\epsilon} g(\xi, x^\epsilon(t,\xi), y^\epsilon(t,\xi)) + \frac{1}{\sqrt{\epsilon}}\frac{\partial \omega(t,\xi)}{\partial t}, \end{aligned} \tag{0.1}$$

la variable d'espace $\xi$ appartenant au domaine $(0,1)$ et la variable de temps vérifiant $0 \leq t \leq T$, pour un temps final $T > 0$ donné. Les deux composantes $x^\epsilon$ et $y^\epsilon$ satisfont à des conditions au bord de type Dirichlet homogène: pour tout $t \geq 0$

$$x^\epsilon(t,0) = x^\epsilon(t,1) = y^\epsilon(t,0) = y^\epsilon(t,1) = 0.$$

Des conditions initiales $x$ et $y$, indépendantes du paramètre $\epsilon$ et déterministes sont aussi données: pour tout $\xi \in [0,1]$ $x^\epsilon(0,\xi) = x(\xi)$ et $y^\epsilon(0,\xi) = y(\xi)$.

La quantité $\frac{\partial \omega(t,\xi)}{\partial t}$ modélise le caractère aléatoire du comportement du système: il s'agit d'un bruit blanc en temps et en espace. Notons qu'une équation parabolique de type réaction-diffusion perturbée par un bruit additif blanc en temps et en espace n'est bien posée que lorsque la variable d'espace est unidimensionnelle, et que cette propriété justifie la restriction de l'étude à cette situation. Une généralisation de l'étude au cas multidimensionnel est possible sous certaines conditions en considérant un bruit blanc en temps et coloré en espace.

Enfin le paramètre $\epsilon > 0$ sert à modéliser la séparation des échelles de temps d'évolution des deux composantes, et on s'intéresse au comportement du système lorsque $\epsilon$ tend vers 0. Dans ce régime, l'évolution de la composante $y^\epsilon$ a lieu à l'échelle $\frac{t}{\epsilon}$ - en tenant compte des propriétés du bruit par changement de temps - tandis que la composante $x^\epsilon$ évolue selon le temps $t$. Ainsi $x^\epsilon$ est considérée comme la variable lente, et $y^\epsilon$ comme la variable rapide.

De tels systèmes d'équations avec évolution selon deux échelles de temps séparées servent à modéliser divers phénomènes. Les équations associées peuvent être déterministes ou stochastiques, des équations différentielles ordinaires ou des équations aux dérivées partielles. Un exemple en mécanique concerne le mouvement de planètes autour d'un astre - avec rotation lente autour de l'astre et rotation propre rapide. On peut aussi modéliser des phénomènes chimiques ou biologiques se produisant à des échelles de temps différentes.

L'étude des EDPS doit se faire dans un certain cadre théorique: ici il s'agit de celui des équations d'évolution stochastiques dans des espaces de Hilbert - suivant le formalisme de [15]. Les solutions sont alors vues comme des fonctions de la variable temporelle, à valeurs dans un espace fonctionnel par rapport à la variable spatiale. Le système prend alors la forme suivante, où les composantes $X^\epsilon$ et $Y^\epsilon$ sont des variables aléatoires à valeurs dans l'espace de Hilbert $H = L^2(0,1)$: pour tout $t \in (0,T]$

$$\begin{aligned} dX^\epsilon(t) &= \big(AX^\epsilon(t) + F(X^\epsilon(t), Y^\epsilon(t))\big)dt \\ dY^\epsilon(t) &= \frac{1}{\epsilon}\big(BY^\epsilon(t) + G(X^\epsilon(t), Y^\epsilon(t))\big)dt + \frac{1}{\sqrt{\epsilon}} dW(t), \end{aligned} \tag{0.2}$$

avec des conditions initiales $X^\epsilon(0) = x \in H$, $Y^\epsilon = y \in H$. Les coefficients $A$ et $B$ sont des opérateurs linéaires non bornés sur $H$, présentant un certain nombre de propriétés: par exemple, on peut considérer que $A$ et $B$ sont définis sur le domaine $H^2(0,1) \cap H_0^1(0,1)$ comme étant l'opérateur de dérivée seconde, les conditions au bord de type Dirichlet homogène étant prises en compte dans la définition du domaine. Les coefficients $F$ et $G$ vérifient certaines conditions de régularité; le cas des opérateurs de Nemytskii, permettant de traiter le cas du système (0.1), tel que $F(x,y)(\xi) = f(\xi, x(\xi), y(\xi))$ pour toutes fonctions $x$ et $y$ dans $L^2(0,1)$, entre dans le cadre, malgré de faibles propriétés de régularité (pas de classe $\mathcal{C}^2$ sur $L^2(0,1)$, mais réguliers selon des directions de dérivation appartenant à des espaces de Sobolev). Enfin le bruit dans l'équation rapide est représenté par un processus de Wiener cylindrique $(W(t))_{0\leq t\leq T}$ défini sur $H$. Les solutions sont définies dans un sens intégral du point de vue EDP.

L'étude du problème se divise en trois parties. Dans le premier chapitre, on prouve qu'un principe de moyennisation est satisfait: lorsque $\epsilon$ tend vers 0, pour tout instant $t \geq 0$ la composante lente $X^\epsilon(t)$ converge vers une quantité $\overline{X}(t)$, solution d'une équation dite moyennée, et obtenue en moyennant le coefficient $F$ par rapport à la variable rapide, selon une mesure de probabilité associée au comportement asymptotique de l'équation rapide. Par rapport à l'article [12] sur lequel on se base, on précise l'ordre de convergence par rapport à $\epsilon$: dans un sens fort il est égal à $1/2$ et à 1 dans un sens faible.

Dans le deuxième chapitre, on s'intéresse à la discrétisation temporelle du système, à travers un résultat essentiel dans la construction d'une méthode efficace pour le système multi-échelles. On montre qu'on peut approcher l'unique mesure de probabilité invariante associée à une EDPS comme celle gouvernant le comportement de la composante rapide de (0.2) grâce à un schéma d'Euler. L'ordre de la convergence associée est de $1/2$.

Dans le troisième chapitre, on construit une méthode numérique à l'aide d'un schéma de type Euler utilisant deux pas de temps différents, selon le cadre HMM (Heterogeneous Multiscale Methods); on généralise ainsi la construction et l'analyse de [25] pour des systèmes d'EDS. La définition et l'efficacité de la méthode sont des conséquences du résultat de moyennisation et d'approximation de mesure invariante. On prouve une estimée de convergence forte et de convergence faible; en comparant avec une méthode directe, on identifie des valeurs des paramètres qui rendent la méthode multi-échelle théoriquement plus performante qu'une méthode directe lorsque le paramètre $\epsilon$ tend vers 0.

### 0.1.2. **Le principe de moyennisation.**

Cette section présente le contenu du Chapitre 1.

Le principe de moyennisation énonce qu'on peut construire une équation dite moyennée, dont la solution est la limite de la composante lente du système (0.2) lorsque $\epsilon$ tend vers 0.

0.1.2.1. *Différents types de résultats.* De nombreux résultats sont associés à ce principe, suivant plusieurs critères de convergence et différentes techniques de preuves. Dans le cas des EDS, les livres [30] et [56] par exemple contiennent chacun un chapitre consacré entièrement au principe de moyennisation. Moins de travaux ont été publiés dans le cas des EDPS, mais certains articles - comme [10] et [12] - ont également constitué une base de travail essentielle.

Le résultat le plus général est une convergence en loi du processus $(X^\epsilon(t))_{0\leq t\leq T}$ vers $(\overline{X}(t))_{0\leq t\leq T}$, pour tout temps final $T \in (0, +\infty)$; dans ce cas, on impose des conditions très faibles sur les coefficients et sur le comportement asymptotique de l'équation rapide (essentiellement on requiert l'existence de limite aux moyennes temporelles pour définir les coefficients de dérive et de diffusion moyennés).

On peut aussi s'intéresser à la vitesse de convergence par rapport à $\epsilon$; pour cela, on peut d'une part considérer l'erreur au sens fort, par exemple en la contrôlant dans $L^1(\Omega)$, ou au sens faible, en contrôlant l'écart entre les lois à un instant fixé. L'objectif est de montrer que l'ordre faible (ici égal à 1) est strictement plus élevé que l'ordre fort (ici égal à $1/2$).

Il est enfin possible détudier les déviations des trajectoires réelles autour des solutions de l'équation moyennée. D'une part, on peut regarder les déviations dites normales, en montrant une convergence en loi non triviale de l'écart correctement renormalisé $\frac{X^\epsilon - \overline{X}}{\sqrt{\epsilon}}$. D'autre part, on peut montrer qu'un principe de grandes déviations est satisfait.

Dans cette thèse, on prouve des résultats du deuxième type: la connaissance de l'ordre de convergence est en effet cruciale pour estimer la performance d'une méthode de résolution numérique.

Examinons quelques techniques de preuve, en mettant en évidence le type de résultat associé, et les difficultés rencontrées, notamment concernant leur transposition du cas des EDS au cas des EDPS.

Pour obtenir la convergence forte, la technique la plus utile consiste à découper l'intervalle de temps $[0, T]$, pour construire une subdivision de taille $\delta(\epsilon)$; cette technique apparaît dans [37] . On introduit alors un système d'évolution auxiliaire, tel que sur chaque sous-intervalle on considère que la composante lente est constante dans l'évolution de la composante rapide. Cette décomposition est très naturelle par rapport à l'heuristique du principe de moyennisation: la composante rapide converge vers un équilibre dont dépend le coefficient moyenné. L'erreur entre les systèmes d'évolution initial et auxiliaire dépend du paramètre $\delta(\epsilon)$ notamment à travers la régularité des trajectoires. L'ajustement de $\delta(\epsilon)$ en fonction de $\epsilon$ fournit la vitesse de convergence. Notons que cette technique de subdivision est également utile pour démontrer des résultats de convergence faible, notamment dans [10].

Une autre méthode de preuve de la convergence forte repose sur un développement de l'erreur utilisant la solution d'une équation de Poisson associée au générateur $L_1$ de la composante rapide, et dont le second membre est la différence entre le coefficient de l'équation lente et le coefficient moyenné associé. Cette preuve est assez naturelle si on interprète le principe de moyennisation comme une loi des grands nombres. Plus précisément, il s'agit d'écrire $F(x, y) - \overline{F}(x) = L_1\phi(x, y)$ pour une certaine fonction $\phi$, et d'utiliser la formule d'Itô pour développer $\phi(X^\epsilon(T), Y^\epsilon(T))$, afin de borner $X^\epsilon(T) - \overline{X}(T)$. Dans le cas des EDS, cette méthode est développée dans [56]; une adaptation au cas des EDPS est proposée dans [12], avec des changements significatifs: une équation de Poisson modifiée est utilisée pour y prouver une convergence, a priori uniquement en loi; le processus limite étant déterministe, cette convergence a en fait lieu en probabilité, mais sans ordre de convergence.

Pour obtenir un résultat de convergence faible avec une vitesse de convergence, à tout instant $T$ fixé, une stratégie consiste à interpréter les termes de l'erreur comme solutions d'équations de Kolmogorov associées aux évolutions du système et de l'équation moyennée. On construit alors un développement asymptotique en $\epsilon$, par l'analyse des générateurs. Notons que pour obtenir le résultat de moyennisation un développement à l'ordre 1 suffit; dans le cas d'un système d'EDS, il est possible d'obtenir un développement à tout ordre, qui fait intervenir des fonctions dépendant des deux échelles de temps $t$ et $t/\epsilon$ - voir [38]. Dans le cas des EDPS, on n'a pu obtenir que le développement à l'ordre 1, dans le cas d'un bruit additif dans l'équation rapide et sans bruit dans l'équation lente. Un point technique notable concerne l'hypothèse de dissipativité de l'équation rapide considérée, qui assure des propriétés de convergence exponentielle sur les lois et non sur les trajectoires. Si on ajoute une bruit dans l'équation lente, les termes ajoutés ne sont pas contrôlables par les méthodes que nous avons utilisées; on détaille ces problèmes dans la section 0.1.2.5.

Une autre technique pour obtenir une convergence en loi des processus, sous des hypothèses assez générales de régularité et de dépendance des coefficients de dérive et de diffusion par rapport aux composantes, repose sur une interprétation en termes de problème de martingale. L'identification de l'unique valeur d'adhérence possible, comme étant l'unique solution du problème de martingale associé à l'équation moyennée, repose sur l'utilisation de fonctions test perturbées $\phi^\epsilon = \phi + \epsilon\phi + \ldots$, contruites à partir de la vraie fonction test $\phi$, afin de simplifier l'expression et d'annuler le termes singuliers par rapport à $\epsilon$. Cette méthode est bien adaptée pour prouver la convergence en loi, mais ne fournit pas de vitesses de convergence. Pour le cas des EDS, une présentation de la méthode est fournie dans le livre [29]; son adaptation aux EDPS est présentée dans [10].

Dans ce travail, on ne montre pas de résultat de déviations. Bien que dans le cas des EDS le résultat de déviations normales et le principe de grandes déviations sont connus, la situation concernant les EDPS a fait jusqu'ici l'objet de moins d'études: à notre connaissance aucun principe de grandes déviations n'a été

démontré. D'autre part, le principe de déviations normales a été prouvé dans la situation simple du système (0.2), où la nonlinéarité $G$ ne dépend pas de la composante lente $x$, dans [11]: il y a convergence de $\frac{X^\epsilon - \overline{X}}{\sqrt{\epsilon}}$, en loi dans $\mathcal{C}([0,T],H)$, vers une loi dépendant d'un processus gaussien, dont la définition est plus compliquée que pour la situation en dimension finie.

0.1.2.2. *Construction du processus moyenné.* La première étape dans la construction de cette équation est l'étude du comportement asymptotique de l'équation rapide dans laquelle on fixe la composante lente. En effet, pour tout $t > 0$, on a $\frac{t}{\epsilon} \to +\infty$: sur un intervalle $[t, t+\delta t]$ avec un $\delta t > 0$ choisi assez petit, on peut considérer que la composante lente n'évolue pas, tandis que la composante rapide atteint un équilibre - dans un sens à préciser. Selon cette heuristique, on doit donc analyser le comportement asymptotique de l'équation rapide avec composante lente figée en $x \in H$ - après un changement de temps éliminant le paramètre $\epsilon$:

$$\begin{aligned} dY_x(t,y) &= \big(BY_x(t,y) + G(x, Y_x(t,y))\big)dt + dW(t), \\ Y_x(0,y) &= y. \end{aligned} \tag{0.3}$$

On impose aux coefficients de cette équation des conditions assurant l'ergodicité du processus $Y_x$; ces conditions dépendent toutefois du critère de convergence choisi - convergence forte ou faible.

On peut tout d'abord imposer une condition de dissipativité forte: si $\mu > 0$ est la plus petite valeur propre de l'opérateur $-B$ - qui est supposé autoadjoint et d'inverse compact donc diagonalisable dans une base hilbertienne de $H$ - et si $L_g$ désigne un majorant de la constante de Lipschitz de $G(x,.)$, uniformément par rapport à $x \in H$, on suppose

$$L_g < \mu. \tag{SD}$$

Dans cette situation, les trajectoires issues de deux conditions initiales $y_1$ et $y_2$ différentes, et soumises à la même perturbation stochastique $(W(t))$, se rapprochent exponentiellement vite, presque sûrement: pour tout $t \geq 0$

$$|Y_x(t,y_1) - Y_x(t,y_2)|_H \leq e^{-\frac{(\mu - L_g)}{2}t}|y_1 - y_2|_H.$$

Cette contraction est une propriété de l'équation déterministe, qui est préservée par perturbation par bruit additif; elle assure l'unicité d'une mesure de probabilité invariante. Quant à elle, l'existence d'une telle mesure est une conséquence des propriétés du bruit, blanc en temps et en espace; notamment la propriété de Feller est vérifiée.

La condition de dissipativité peut être affaiblie, auquel cas la contraction des trajectoires n'est plus valable; en revanche, grâce à des techniques de couplage on obtient une contraction sur les lois. Plus précisément, on suppose qu'il existe deux constantes $c > 0$ et $C > 0$ telles que pour tout $x \in H$ et $y \in D(B)$ le domaine de $B$, on a

$$< By + G(x,y), y > \leq -c|y|^2 + C. \tag{WD}$$

Il suffit pour cela de supposer que $G$ est une fonction bornée, et la dissipation est assurée par la positivité de $-B$. La preuve d'existence d'une mesure de probabilité invariante n'est pas modifiée par rapport à la situation de dissipativité forte; néanmoins celle de l'unicité devient plus complexe, et nécessite aussi des hypothèses sur la nature du bruit.

On obtient alors la propriété de contraction suivante, qui fournit le résultat d'unicité: il existe deux constantes $c > 0$ et $C > 0$, telles que si $\phi$ est une fonction test bornée, on a pour toute composante lente $x \in H$, toutes conditions initiales $y_1$ et $y_2$ et tout $t \geq 0$

$$\big|\mathbb{E}\phi(Y_x(t,y_1)) - \mathbb{E}\phi(Y_x(t,y_2))\big| \leq C\|\phi\|_\infty(1 + |y_1|^2 + |y_2|^2)e^{-ct}. \tag{0.4}$$

L'unique mesure de probabilité invariante du processus $Y_x$ est alors notée $\mu^x$. Une hypothèse supplémentaire sur le coefficient $G$ permet par ailleurs d'en donner une expression explicite:

$$\mu^x(dy) = \frac{1}{Z(x)}e^{2U(x,y)}\nu(dy), \tag{0.5}$$

où on suppose que le potentiel $U$ est tel que $G(x,y) = D_yU(x,y)$, $\nu = \mathcal{N}(0, (-B)^{-1}/2)$ est une mesure gaussienne sur $H$ - correspondant à l'unique mesure invariante dans le cas $G = 0$ - et $Z(x)$ est une constante

de normalisation. La dépendance de $\mu^x$ par rapport à $x \in H$ est donc bien connue, et est un outil important dans la preuve du résultat.

On peut finalement définir le coefficient moyenné: pour tout $x \in H$

$$\overline{F}(x) = \int_H F(x,y)\mu^x(dy). \tag{0.6}$$

On remarque que $\overline{F}$ vérifie de bonnes propriétés. Notamment on constate que la dépendance en $x$ est double, à travers $F$ et $\mu^x$. Le passage aux équations d'évolution dans $H$ permet de définir correctement ce coefficient: on constate que même lorsque $F$ et $G$ sont des coefficients de Nemytskii ce n'est pas nécessairement vrai pour $\overline{F}$.

Le processus dit moyenné est l'unique solution intégrale de l'équation d'évolution suivante: pour tout $t \in (0,T]$

$$d\overline{X}(t) = \big(A\overline{X}(t) + \overline{F}(\overline{X}(t))\big)dt, \tag{0.7}$$

avec la condition initiale $\overline{X}(0) = x \in H$.

0.1.2.3. *Résultats de convergence.* Comme la convergence met en jeu des processus stochastiques, différents critères sont disponibles. Le résultat le plus faible et le plus général est le suivant: étant donnés un temps final $T > 0$ et des conditions initiales $x$ et $y$ dans des espaces de Sobolev, le processus $(X^\epsilon(t))_{0\le t\le T}$ converge en loi et en probabilité vers le processus (déterministe) $(\overline{X}(t))_{0\le t\le T}$ lorsque $\epsilon$ tend vers 0.

Pour connaître la vitesse de convergence, on étudie l'erreur selon différents critères.

En évaluant l'erreur en norme $L^1$, on montre que la convergence est d'ordre 1/2. Plus précisément, on démontre dans le premier chapitre de cette thèse le résultat suivant:

**Théorème 1** (Ordre fort)**.** *Supposons la stricte dissipativité* (SD) *vérifiée. Alors pour tout* $0 < r < 1/2$, $T > 0$, $x \in H$, $y \in H$, *il existe une constante* $C > 0$ *telle que pour tout* $\epsilon > 0$ *et* $0 \le t \le T$

$$\mathbb{E}\big|X^\epsilon(t) - \overline{X}(t)\big|_H \le C\epsilon^{1/2-r}. \tag{0.8}$$

Ainsi, on compare les trajectoires des processus $X^\epsilon$ et $\overline{X}$: l'hypothèse de dissipativité forte est par conséquent requise.

On peut obtenir un ordre de convergence plus élevé en affaiblissant le critère. Comparant à présent les lois des processus à chaque instant, on obtient le résultat suivant:

**Théorème 2** (Ordre faible)**.** *Supposons la dissipativité faible* (WD) *vérifiée. Alors pour tout* $0 < r < 1$, $T > 0$, $0 < \theta \le 1$, $x \in D(-A)^\theta$, $y \in H$, $\phi \in \mathcal{C}_b^2(H)$, *il existe une constante* $C > 0$, *dépendant de* $r$, $T$, $\phi$, $|x|_{(-A)^\theta}$, $|y|$, *telle que pour tout* $\epsilon > 0$ *et* $0 \le t \le T$

$$\big|\mathbb{E}[\phi(X^\epsilon(t))] - \mathbb{E}[\phi(\overline{X}(t))]\big| \le C\epsilon^{1-r}. \tag{0.9}$$

La notation $\mathcal{C}_b^2(H)$ désigne l'espace des fonctions définies sur $H$, à valeurs dans $\mathbb{R}$, de classe $\mathcal{C}^2$, bornées et à dérivées bornées; ces fonctions étant en particulier lipschitziennes une conséquence immédiate du Théorème 1 est l'estimation suivante: $|\mathbb{E}[\phi(X^\epsilon(t))] - \mathbb{E}[\phi(\overline{X}(t))]| \le C\epsilon^{1/2-r}$. La signification du Théorème 2 consiste en une majoration stricte de l'ordre fort 1/2 par l'ordre faible 1, qui va donc demander des techniques de preuve plus élaborées. Notons que dans les deux théorèmes précédents le paramètre $r$ est choisi strictement positif, mais aussi petit que souhaité.

L'espace $D(-A)^\theta$ avec $\theta > 0$ est le domaine d'un opérateur défini à partir de $A$, de type Sobolev: on impose une très faible restriction de régularité sur la condition initiale $x$.

L'hypothèse de dissipativité faible suffit pour obtenir le Théorème 2: la preuve n'utilise que des comparaisons entre lois de processus, pas entre trajectoires.

Les ordres des Théorèmes 1 et 2 sont les mêmes que pour le principe de moyennisation en dimension finie (i.e. pour des systèmes d'EDS). On les obtient par des généralisations des preuves correspondantes pour les EDS, avec des difficultés supplémentaires dues au cadre infini-dimensionnel. Les Théorèmes 1 et 2 sont prouvés dans le Chapitre 1. On donne quelques éléments essentiels de leur preuve dans la Section 0.1.2.4; dans la section 0.1.2.5, on expose les difficultés rencontrées lorsqu'on essaye d'étendre ces résultats si l'équation lente est bruitée.

0.1.2.4. *Éléments de preuves.*

Preuve du Théorème 1

Pour prouver le Théorème 1, l'idée fondamentale, comme dans [37], est de décomposer l'intervalle $[0,T]$ en sous-intervalles de taille $\delta(\epsilon)$ (un paramètre à optimiser en fonction de $\epsilon$), afin d'introduire un système auxiliaire dont la solution $(\tilde{X}^\epsilon, \tilde{Y}^\epsilon)$ vérifie les propriétés suivantes:

- à l'instant initial on a $\tilde{X}^\epsilon(0) = x = X^\epsilon(0)$ et $\tilde{Y}^\epsilon(0) = y = Y^\epsilon(0)$.
- sur chaque sous-intervalle $[k\delta(\epsilon), (k+1)\delta(\epsilon)]$, les deux composantes évoluent selon des équations découplées:

$$\begin{aligned} d\tilde{X}^\epsilon(t) &= (A\tilde{X}^\epsilon(t) + F(X^\epsilon(k\delta(\epsilon)), \tilde{Y}^\epsilon(t)))dt, \\ d\tilde{Y}^\epsilon(t) &= \frac{1}{\epsilon}(B\tilde{Y}^\epsilon(t) + G(X^\epsilon(k\delta(\epsilon)), \tilde{Y}^\epsilon(t)))dt + \frac{1}{\sqrt{\epsilon}}dW(t). \end{aligned} \tag{0.10}$$

On se ramène alors à une situation plus simple: les coefficients de l'équation rapide ne dépendent pas de la composante lente, si bien qu'on peut utiliser les propriétés d'ergodicité et de convergence exponentielle à l'équilibre pour obtenir le résultat. Dans le cas général, un tel argument peut-être utilisé sur chaque sous-intervalle, avec un contrôle de l'erreur entre $\tilde{X}^\epsilon$ et $\overline{X}$ dépendant de $\epsilon$ et de $\delta(\epsilon)$.

La distance entre $\tilde{X}^\epsilon$ et $X^\epsilon$ dépend des propriétés de régularité temporelle des trajectoires de ces processus: elles sont d'exposant de Holder $1-r$, pour tout $0 < r < 1$. De plus, il est nécessaire d'évaluer l'erreur entre les trajectoires de $Y^\epsilon$ et $\tilde{Y}^\epsilon$: l'hypothèse de dissipativité forte (SD) est nécessaire. On adapte notamment au cas des EDPS un argument présent dans [48], qui concerne la convergence exponentielle en temps d'une constante de Lipschitz par rapport à la composante lente, en plus de la convergence en norme infinie.

Pour conclure, il suffit de choisir $\delta(\epsilon) = \sqrt{\epsilon}$. On remarque que le fait que l'ordre de convergence $1/2$ ne soit pas exactement atteint est une conséquence du cadre de dimension infinie: le paramètre $r > 0$ est non nul, mais aussi petit que souhaité.

Preuve du Théorème 2

L'idée principale de la preuve consiste à interpréter chaque terme de l'erreur comme la solution d'une équation de Kolmogorov rétrograde à l'instant $t$, comme dans [38]. Il s'agit d'une technique très classique quand on démontre des résultats de convergence faible - on la retrouve notamment lorsqu'on explicite l'ordre de convergence faible d'un schéma numérique pour les EDS ou les EDPS. Dans ce travail, elle intervient également dans les démonstrations des Théorèmes 3 et 5.

Afin de rendre le raisonnement rigoureux, il est nécessaire d'approcher le problème à l'aide d'une méthode de Galerkin: on remplace le systême (0.2) par un systême d'équations, dont les solutions sont à valeurs dans des sous-espaces de dimension finie, et tel qu'il y a convergence de l'approximation lorsque la dimension tend vers $+\infty$. On ne mentionne pas la dimension dans les expressions: la construction et la convergence des quantités qui apparaissent sont assurées, et il suffit de contrôler les termes d'erreur uniformément par rapport à la dimension pour obtenir le résultat.

D'une part, la fonction $u^\epsilon : (t,x,y) \mapsto E[\phi(X^\epsilon(t,x,y))]$ - en notant explicitement la dépendance du processus par rapport à la condition initiale - est solution de l'EDP

$$\begin{aligned} \frac{\partial u^\epsilon}{\partial t}(t,x) &= L^\epsilon u^\epsilon(t,x), \\ u^\epsilon(0,x) &= \phi(x), \end{aligned} \tag{0.11}$$

où $L^\epsilon$ est le générateur infinitésimal du processus de diffusion $(X^\epsilon, Y^\epsilon)$ dans $H \times H$. La singularité par rapport au paramètre d'échelle de temps $\epsilon$ se traduit par la décomposition

$$L^\epsilon = \frac{1}{\epsilon}L_1 + L_2, \tag{0.12}$$

$L_1$ correspondant à la composante rapide $Y^\epsilon$ et $L_2$ à la composante lente $X^\epsilon$. Comme on s'intéresse uniquement à la convergence de la composante lente, on a choisi une fonction test $\phi$ ne dépendant que de la composante lente $x$, et pas de $y$. Néanmoins la dépendance par rapport à la composante rapide $Y^\epsilon$ de $X^\epsilon$ entraîne l'apparition de la partie singulière du générateur dans l'EDP.

D'autre part, la fonction $\overline{u} : (t,x) \mapsto \phi(\overline{X}(t,x))$ est solution de

$$\begin{aligned} \frac{\partial \overline{u}}{\partial t} &= \overline{L}\overline{u}, \\ \overline{u}(0,.) &= \phi, \end{aligned} \tag{0.13}$$

où formellement l'opérateur associé vérifie $\overline{L} = \int_H L_2 \mu^x(dy)$.

La stratégie de la preuve réside dans un développement asymptotique par rapport au paramètre $\epsilon$ de la forme

$$u^\epsilon = u_0 + \epsilon u_1 + v^\epsilon, \tag{0.14}$$

avec $u_0$ et $u_1$ à construire et $v^\epsilon$ un reste à contrôler.

On peut voir le premier terme comme un moyen d'identifier la limite de $X^\epsilon$: le principe de moyennisation se traduit par $u_0 = \overline{u}$. La construction de $u_1$ repose sur l'étude de l'équation de Poisson en la variable $y$, avec paramètre $x$ fixé:

$$\begin{aligned} L_1 v &= -\Psi, \\ \int_H v(y)\mu^x(dy) &= 0, \end{aligned}$$

pour tout second membre $\Psi$ suffisamment régulier, tel que $\int_H \Psi(y)\mu^x(dy) = 0$. En utilisant les propriétés de dissipativité de l'équation rapide, on dispose de la formule suivante:

$$v(y) = \int_0^{+\infty} \mathbb{E}[\Psi(Y_x(s,y))]ds.$$

On constate à ce moment de la preuve que seule l'hypothèse (WD) est nécessaire, et il s'agit d'une amélioration notable par rapport aux résultats de moyennisation mentionnés dans la Section 0.1.2.1. Pour assurer la convergence de l'intégrale et vérifier qu'on a bien défini une solution de l'EDP, on étend le résultat de convergence exponentielle en temps pour $(s,y) \mapsto \mathbb{E}[\Psi(Y_x(s,y))]$ à ses dérivées spatiales grâce à un argument de type régularisation utilisant la dérivée de Malliavin.

On dispose alors d'une expression explicite de $v^\epsilon$, et la fin de la preuve consiste en des estimations techniques de différents termes.

0.1.2.5. *Et pour une équation lente bruitée?* Dans le système (0.1), l'équation lente concernant l'évolution de la composante $x^\epsilon$ est une EDP à coefficients aléatoires, couplée à l'équation rapide qui est une EDP stochastique. On a vu que l'équation moyennée est alors une EDP déterministe.

Le principe de moyennisation se généralise à des systèmes où chacune des équations est stochastique. Dans cette situation, l'équation moyennée devient une EDP stochastique; la définition du coefficient de diffusion associé passe par l'étude du générateur infinitésimal du système, en suivant l'approche ayant permis de montrer le résultat de convergence faible précédent.

Sous des hypothèses de régularité des coefficients et de dissipativité de l'équation rapide, la convergence en loi de la composante lente vers la solution de l'équation moyennée a été obtenue dans [10]. Néanmoins, l'ordre de convergence n'y est pas précisé, et l'extension de nos résultats n'est pas aussi générale. On notera aussi qu'un résultat de convergence forte a été prouvé dans [31], mais qu'il concerne des EDPS avec bruit scalaire.

Pour examiner les difficultés générées par cet ajout d'un bruit dans l'équation lente, on considère le problème suivant:

$$\begin{aligned} dX^\epsilon(t) &= \big(AX^\epsilon(t) + F(X^\epsilon(t), Y^\epsilon(t))\big)dt + \sigma(X^\epsilon(t), Y^\epsilon(t))dw(t), \\ dY^\epsilon(t) &= \frac{1}{\epsilon}\big(BY^\epsilon(t) + G(X^\epsilon(t), Y^\epsilon(t))\big)dt + \frac{1}{\sqrt{\epsilon}}dW(t), \end{aligned} \tag{0.15}$$

pour $0 \leq t \leq T$, et des conditions initiales $X^\epsilon(0) = x$ et $Y^\epsilon(0) = y$.

Le coefficient de diffusion $\sigma : H \times H \to \mathcal{L}(H)$ vérifie certaines hypothèses de régularité; $(w(t))_{0\leq t\leq T}$ est un processus de Wiener cylindrique sur $H$, indépendant de $W$.

L'équation rapide n'est pas modifiée par rapport à (0.2), si bien que les propriétés concernant le comportement asymptotique sont préservées sous hypothèse de dissipativité, forte ou faible.

L'équation moyennée (0.7) devient une EDPS

$$d\overline{X}(t) = \big(A\overline{X}(t) + \overline{F}(\overline{X}(t))\big)dt + \overline{\sigma}(\overline{X}(t))dw(t), \tag{0.16}$$

pour $0 \leq t \leq T$, avec condition initiale $\overline{X}(0) = x$. Les coefficients moyennés $\overline{F}$ et $\overline{\sigma}$ vérifient

$$\overline{F}(x) = \int_H F(x,y)\mu^x(dy),$$
$$\overline{\sigma}(x)\overline{\sigma}^T(x) = \int_H \sigma(x,y)\sigma(x,y)^T\mu^x(dy).$$

Afin d'obtenir un résultat de convergence forte, il est nécessaire de supposer que le coefficient de diffusion $\sigma$ ne dépend pas de la variable rapide $y$, auquel cas on a pour tout $x, y \in H$ $\sigma(x,y) = \overline{\sigma}(x)$. En effet, il est possible de construire un contre-exemple à la convergence forte, avec un système d'équations différentielles stochastiques en dimension 1:

$$\begin{aligned} dx^\epsilon(t) &= y^\epsilon(t)dB^x(t), \\ dy^\epsilon(t) &= -\frac{1}{\epsilon}y^\epsilon(t) + \frac{1}{\sqrt{\epsilon}}dB^y(t), \end{aligned} \tag{0.17}$$

pour tout $0 \leq t \leq T$, avec conditions initiales $x^\epsilon(0) = x$ et $y^\epsilon(0) = y$. Les processus $(B^x(t))_{0\leq t\leq T}$ et $(B^y(t))_{0\leq t\leq T}$ sont deux mouvements Browniens standard, supposés indépendants. La composante rapide est un processus d'Ornstein-Uhlenbeck, dont l'unique mesure de probabilité invariante est une mesure gaussienne centrée, de telle sorte que l'équation moyennée est

$$d\overline{x}(t) = \frac{1}{2}dB^x(t),$$

avec $\overline{x}(0) = x$.

On constate alors que la convergence forte n'a pas lieu:

$$\mathbb{E}|x^\epsilon(t) - \overline{x}(t)|^2 = \mathbb{E}|\int_0^t (y^\epsilon(s) - \frac{1}{2})dB^x(s)|^2 = \int_0^t \mathbb{E}|y^\epsilon(s) - \frac{1}{2}|^2 ds,$$

cette dernière quantité ne convergeant pas vers 0 quand $\epsilon$ tend vers 0.

Si on s'intéresse à la généralisation du Théorème 1, on fait donc l'hypothèse que $\sigma$ ne dépend pas de $y$, et on se place sous condition de dissipativité forte (SD). La modification principale concerne la régularité temporelle des trajectoires de $X^\epsilon$ et $\overline{X}$: selon la régularité du bruit (qui dépend de $\sigma$), les trajectoires vont être d'exposant de Holder $1/2 - r$ ou $1/4 - r$, pour tout $r > 0$. Les ordres de convergence correspondants, obtenus par optimisation du paramètre $\delta(\epsilon)$, sont respectivement $1/3$ et $1/5$. En dimension finie, la même situation apparaît, puisque les trajectoires de solutions d'EDS sont d'exposant de Holder $1/2 - r$, pour tout $0 \leq r < 1/2$. Pour montrer que l'ordre de la convergence est plus élevé, on peut essayer d'analyser l'erreur en faisant apparaître des développements plus précis: la difficulté principale réside dans le fait qu'en général les solutions d'EDPS ne sont pas des semi-martingales, si bien que le développement n'est pas possible. De plus, des hypothèses de régularité très restrictives devraient être imposées aux coefficients $F$ et $G$, dans le but d'étendre certains résultats de convergence exponentielle à des quantités impliquant des dérivées d'ordre supérieur.

La généralisation du Théorème 2 est également limitée par l'apparition d'un bruit dans l'équation lente: le principe de la preuve n'est pas modifié, tandis qu'un terme supplémentaire apparaît dans les générateurs infinitésimaux $L_2$ et $\overline{L}$, dépendant de la dérivée seconde des fonctions tests:

$$L_2\psi(x,y) = < Ax + F(x,y), D_x\psi(x,y) > + \frac{1}{2}\mathrm{Tr}\big(\sigma(x,y)\sigma(x,y)^T D^2_{xx}\psi(x,y)\big),$$
$$\overline{L}\psi(x) = < Ax + \overline{F}(x), D_x\psi(x) > + \frac{1}{2}\mathrm{Tr}\big(\overline{\sigma}(x)\overline{\sigma}(x)^T D^2_{xx}\psi(x,y)\big).$$

La difficulté réside dans l'obtention de bornes indépendantes de la dimension: les termes de trace ne se compensent pas tous (même dans le cas d'un bruit additif) et la convergence des séries associées n'est pas assurée sans condition supplémentaire sur $\sigma$. Par exemple, le cas d'un bruit additif blanc, avec $\sigma(x,y) = Id$, ne semble pas pouvoir être traité, tandis que la preuve proposée ici se généralise facilement lorsque le bruit est supposé très régulier en espace.

Le traitement de la situation où l'équation lente est bruitée doit donc encore faire l'objet de recherches.

0.1.3. **Mesure invariante et schéma d'Euler pour EDPS.**

Cette section présente le contenu du Chapitre 2.

La méthode la plus simple d'approximation de solutions d'équations d'évolution, qu'elles soient déterministes ou stochastiques, en dimension finie ou infinie, est la méthode d'Euler explicite. On s'intéresse dans cette section à un schéma d'Euler semi-implicite appliqué à des EDPS; après avoir rappelé les principaux résultats de convergence de cette méthode, on étudie les propriétés asymptotiques de l'approximation. A pas de temps fixé, lorsque le nombre d'itérations tend vers l'infini, la loi du processus discrétisé approche l'unique mesure de probabilités invariante de l'équation continue; la convergence associée est d'ordre 1/2 par rapport au pas de temps. Ce résultat est fondamental pour expliquer l'efficacité de la méthode multi-échelle par rapport à une méthode directe.

Le cadre est le même que pour l'étude du principe de moyennisation: on considère une EDPS dans un espace de Hilbert $H$ du type

$$\begin{aligned} dY(t,y) &= \big(BY(t,y) + G(Y(t,y))\big)dt + dW(t), \\ Y(0,y) &= y. \end{aligned} \tag{0.18}$$

Il s'agit de l'équation rapide avec composante figée (0.3) du système (0.2), où pour simplifier les notations la composante lente $x$ n'est pas mentionnée. Les hypothèses de régularité des opérateurs $B$ et $G$ sont les mêmes que dans la partie concernant le principe de moyennisation. $W$ est un processus de Wiener cylindrique sur $H$.

Sous l'hypothèse de dissipativité faible (WD), le processus admet une unique mesure de probabilité invariante notée $\mu$. Remarquons qu'il n'est pas nécessaire de supposer ici que $G$ dérive d'un potentiel $U$, et l'expression de la mesure invariante $\mu$ en fonction de $U$ et $\nu$ (la mesure gaussienne invariante lorsque $G = 0$) est inutile - on n'a pas besoin de connaître le comportement de la mesure invariante en fonction du paramètre $x$. La convergence exponentielle à l'équilibre reste valable.

Pour discrétiser en temps l'EDPS (0.18), on introduit un pas de temps $\tau > 0$, et des approximations $Y_k(\tau, y)$ de $Y(k\tau, y)$ pour tout $k \in \mathbb{N}$ à l'aide de l'algorithme suivant:

$$\begin{aligned} Y_{k+1}(\tau,y) &= Y_k(\tau,y) + \tau B Y_{k+1}(\tau,y) + \tau G(Y_k(\tau,y)) + \sqrt{\tau}\chi_{k+1}, \\ Y_0(\tau,y) &= y, \end{aligned}$$

avec $\chi_{k+1} = \frac{1}{\sqrt{\tau}}\big(W((k+1)\tau) - W(k\tau)\big)$: le bruit $\sqrt{\tau}\chi_{k+1}$ représente l'accroissement du processus de Wiener entre les deux instants $k\tau$ et $(k+1)\tau$.

Il s'agit d'un schéma semi-implicite: la partie nonlinéaire est discrétisée de manière explicite, tandis que la partie linéaire l'est de manière implicite. En introduisant l'opérateur $R_\tau = (I - \tau B)^{-1}$, on dispose en fait d'une formule explicite

$$Y_{k+1} = R_\tau Y_k + \tau R_\tau G(Y_k) + \sqrt{\tau} R_\tau \chi_{k+1}. \tag{0.19}$$

Grâce aux conditions imposées à l'opérateur $B$ on peut montrer que $R_\tau$ est un opérateur de Hilbert-Schmidt; par conséquent le bruit $\sqrt{\tau}R_\tau\chi_{k+1}$ est bien à valeurs dans $H$. Ainsi $Y_n$ est bien défini pour tout $n \in \mathbb{N}$ et est à valeurs dans $H$.

Rappelons que l'approximation à un instant donné est d'ordre 1/4 au sens fort - voir [57] - et 1/2 au sens faible - voir [18]: pour tout $\tau_0 > 0$, pour tout $T > 0$, pour toute condition initiale $y \in H$, pour toute fonction test $\phi \in \mathcal{C}_b^2(H)$ et pour tout $0 < r < 1/4$, il existe une constante $C > 0$ telle que pour tout pas de temps $0 < \tau \leq \tau_0$ et tout indice $m \in \mathbb{N}$ tel que $m\tau \leq T$, on a

$$\begin{aligned} \mathbb{E}\big|Y_m(\tau,y) - Y(m\tau,y)\big| &\leq C\tau^{1/4-r}, \\ \big|\mathbb{E}[\phi(Y_m(\tau,y))] - \mathbb{E}[\phi(Y(m\tau,y))]\big| &\leq C\tau^{1/2-r}. \end{aligned}$$

La constante $C$ dépend a priori du temps final $T > 0$. Rappelons que lorsque l'équation n'est pas bruitée, l'approximation est d'ordre 1 - voir par exemple [14].

Par ailleurs, on sait que

$$\mathbb{E}[\phi(Y(m\tau,y))] \to \int_H \phi(z)\mu(dz)$$

lorsque $m$ tend vers l'infini, à vitesse exponentielle. Afin d'obtenir un résultat d'approximation de la mesure $\mu$, il est donc nécessaire d'obtenir une estimation de l'erreur indépendante du temps final; cela est possible pour l'estimation d'erreur faible. Précisément, on démontre le Théorème suivant, sous l'hypothèse de dissipativité faible (WD):

**Théorème 3.** *Pour tout $0 < \kappa < 1/2$, $\tau_0 > 0$ et pour toute fonction test $\phi \in \mathcal{C}_b^2$, il existe une constante $C > 0$ telle que pour tout $m \geq 2$, toute condition initiale $y \in H$ et tout pas de temps $0 < \tau \leq \tau_0$ on a*

$$\left|\mathbb{E}[\phi(Y(m\tau, y))] - \mathbb{E}[\phi(Y_m(\tau, y))]\right| \leq C(1 + |y|^3)(((m-1)\tau)^{-1/2+\kappa} + 1)\tau^{1/2-\kappa}.$$

Dans l'estimation, la fraction $\frac{1}{(m-1)\tau}$ disparaît lorsque $m$ tend vers l'infini à $\tau$ fixé; son apparition est due à des problèmes de régularité spécifiques aux EDPS.

Le comportement de $\mathbb{E}[\phi(Y(m\tau, y))]$ lorsque $m$ tend vers l'infini dépend de l'hypothèse de dissipativité. En général, l'existence d'une mesure invariante pour le processus à temps discret est valable, tandis que l'unicité n'est assurée a priori que sous l'hypothèse forte (SD). On obtient donc le résultat suivant:

**Corollaire 1.** *Pour tout $0 < \kappa < 1/2$, $\tau_0 > 0$ et pour toute fonction test $\phi \in \mathcal{C}_b^2$, il existe une constante $c > 0$ et une constante $C > 0$ telles que pour tout pas de temps $0 < \tau \leq \tau_0$, toute condition initiale $y \in H$ et tout indice $m \geq 1$*

$$\left|\mathbb{E}[\phi(Y_m(\tau, y))] - \int_H \phi d\overline{\mu}\right| \leq C(1 + |y|^3)\left(\frac{1}{m^{1/2-\kappa}} + \tau^{1/2-\kappa}\right) + C(1 + |y|^2)e^{-cm\tau}.$$

*De plus, si $\mu^\tau$ est une mesure de probabilité invariante et ergodique de $(Y_m(\tau, .))_{m\in\mathbb{N}}$, on a*

$$\left|\int_H \phi d\mu^\tau - \int_H \phi d\overline{\mu}\right| \leq C\tau^{1/2-\kappa};$$

Sous (SD), la probabilité invariante $\mu^\tau$ est unique; sous (WD), il est possible que plusieurs mesures invariantes ergodiques existent, auquel cas l'estimation fournit une estimation de l'écart entre deux de ces mesures. Pour des résultats concernant les mesures invariantes de schémas numériques, on peut regarder par exemple [40] et [51].

Alors que le cas des EDS a été traité par exemple dans [61], à notre connaissance le Théorème 3 est le premier résultat de ce type pour des EDPS.

Pour prouver le théorème 3 - dans le Chapitre 2 - on suit l'approche de [18] pour analyser l'erreur faible dans le cas des EDPS, et on obtient en outre le contrôle de chaque quantité par rapport au temps final. On décompose $\mathbb{E}[\phi(Y(m\tau, y))] - \mathbb{E}[\phi(Y_m(\tau, y))]$ en suivant une méthode classique - utilisée notamment dans [61]: on utilise la solution $u$ de l'équation de Kolmogorov associée au processus à temps continu $Y$. Plus précisément, on commence par utiliser une approximation de Galerkin, afin de se ramener à des quantités appartenant à des sous-espaces de dimension finie de $H$, qui sont construits à partir des éléments propres de l'opérateur $B$.

En définissant ensuite $u(t, y) = \mathbb{E}\phi(Y(t, y))$, on sait que $u$ est solution de l'équation de Kolmogorov

$$\frac{\partial u}{\partial t}(t, y) = Lu(t, y) = \frac{1}{2}\mathrm{Tr}\left(D^2u(t, y)\right) + < By + G(y), Du(t, y) > .$$

Grâce à la propriété de Markov et à la formule d'Itô, l'erreur dépend d'un contrôle des dérivées de $u$. Celui-ci dépend de deux contraintes: d'une part, on souhaite démontrer des estimations indépendantes du temps final $T = m\tau$; d'autre part, il faut que les bornes obtenues permettent un passage à la limite lorsque la dimension de l'approximation de Galerkin augmente.

Les difficultés liées à cette dernière exigence sont principalement les mêmes que dans la preuve de l'ordre de convergence faible avec un temps final $T$ fixé, dans [18]; la seule exigence supplémentaire est de prouver que l'estimation est valable avec des constantes indépendantes de $T$. Notons qu'on ne considère que le cas d'un bruit additif, et que cela engendre déjà de nombreux points délicats. Nous traitons en outre le cas des opérateurs nonlinéaires de type Nemytskii.

Rappelons deux idées essentielles de la preuve réutilisées ici. D'abord, le contrôle des dérivées première et seconde de $u$ dans les normes induites par la norme de l'espace de Hilbert $H$ est insuffisant; pour obtenir l'ordre 1/2 plutôt que l'ordre fort 1/4, on montre un effet de régularisation spatiale sur ces dérivées, avec des singularités en temps petit; précisément, on montre que les quantités considérées appartiennent aux

domaines des puissances fractionnaires de $-B$. Ensuite, le contrôle de certaines quantités est assuré par une transformation de l'expression utilisant une formule d'intégration par parties, issue du calcul de Malliavin.

On adapte ces deux idées, en surveillant la dépendance par rapport au temps des quantités. Concernant la première idée présentée ci-dessus, on obtient d'abord une borne de décroissance exponentielle sur $u$, dont la preuve est basée sur une estimation de couplage; l'effet régularisant de l'équation de Kolmogorov et la propriété de Markov sont alors utilisés pour obtenir des bornes sur les dérivées de $u$. Concernant la deuxième idée, une astuce est introduite pour contourner le problème suivant: la dérivée de Malliavin du processus $Y$ apparaît, et on ne peut pas la borner indépendamment du temps sous l'hypothèse de dissipativité faible (WD). On observe en fait qu'on peut séparer l'intervalle d'étude en deux parties, chacune étant concernée par un seul problème (convergence en dimension versus convergence en temps long): d'un côté, on effectue l'intégration par parties sur un intervalle de longueur majorée par 1 pour assurer une borne sur l'intégrale indépendante de la dimension; sur le reste de l'intervalle, on utilise directement les résultats à propos des dérivées de $u$, pour contrôler l'intégrale définie sur un domaine de longueur tendant vers l'infini avec le temps.

Remarquons que si l'hypothèse de stricte dissipativité (SD) est satisfaite, l'obtention de contrôle exponentiels peut être simplifiée à divers endroits de la preuve.

### 0.1.4. **La méthode multi-échelle.**

Cette section présente le contenu du Chapitre 3.

0.1.4.1. *Intérêt d'une méthode non directe.* Avant de présenter et d'analyser la méthode multi-échelle, justifions son introduction en analysant le comportement d'une méthode directe: pour un paramètre $\epsilon > 0$ fixé, la solution $(X^\epsilon, Y^\epsilon)$ du système (0.2) peut être approchée à l'aide d'une méthode d'Euler semi-implicite de pas de temps $\Delta t > 0$, appliquée à chaque composante: pour tout entier $n$ tel que $(n+1)\Delta t \leq T$,

$$\begin{aligned} \hat{X}_{n+1}^{\Delta t,\epsilon} &= S_{\Delta t}\hat{X}_n^{\Delta t,\epsilon} + \Delta t S_{\Delta t} F(\hat{X}_n^{\Delta t,\epsilon}, \hat{Y}_n^{\Delta t,\epsilon}), \\ \hat{Y}_{n+1}^{\Delta t,\epsilon} &= R_{\frac{\Delta t}{\epsilon}}\hat{Y}_n^{\Delta t,\epsilon} + \frac{\Delta t}{\epsilon} R_{\frac{\Delta t}{\epsilon}} G(\hat{X}_n^{\Delta t,\epsilon}, \hat{Y}_n^{\Delta t,\epsilon}) + \sqrt{\frac{\Delta t}{\epsilon}} R_{\frac{\Delta t}{\epsilon}} \hat{\chi}_{n+1}^{\Delta t,\epsilon}, \end{aligned} \tag{0.20}$$

avec conditions initiales $\hat{X}_0^{\Delta t,\epsilon} = x$ et $\hat{Y}_0^{\Delta t,\epsilon} = y$.

Les opérateurs linéaires $S_{\Delta t}$ et $R_{\frac{\Delta t}{\epsilon}}$ sont construits de telle sorte que la définition du schéma (0.20) corresponde à un schéma d'Euler implicite par rapport à sa partie linéaire, et explicite par rapport à sa partie nonlinéaire:

$$\begin{aligned} S_{\Delta t} &= (I - \Delta t A)^{-1}, \\ R_{\frac{\Delta t}{\epsilon}} &= (I - \frac{\Delta t}{\epsilon} B)^{-1}. \end{aligned}$$

Le bruit $\hat{\chi}_{n+1}^{\Delta t,\epsilon}$ est défini à partir des accroissements du processus de Wiener cylindrique $W$ apparaissant dans (0.2): pour tout entier $n$ tel que $(n+1)\Delta t \leq T$

$$\hat{\chi}_{n+1}^{\Delta t,\epsilon} = \frac{W((n+1)\frac{\Delta t}{\epsilon}) - W(n\frac{\Delta t}{\epsilon})}{\sqrt{\frac{\Delta t}{\epsilon}}}.$$

La méthode est bien définie dans $H$, puisque sous les hypothèses considérées pour l'opérateur $B$, $R_\tau \in \mathcal{L}(H,H)$ est un opérateur de Hilbert-Schmidt, pour tout $\tau > 0$.

Les résultats de convergence de la méthode d'Euler pour les EDP déterministes et stochastiques entraînent une estimation du type suivant pour l'erreur forte, d'après [57]:

$$\mathbb{E}|X^\epsilon(n\Delta t) - \hat{X}_n^{\Delta t,\epsilon}| + \mathbb{E}|Y^\epsilon(n\Delta t) - \hat{Y}_n^{\Delta t,\epsilon}| \leq C_\epsilon \left( \Delta t^r + \left(\frac{\Delta t}{\epsilon}\right)^{1/4-r} \right). \tag{0.21}$$

La constante $C_\epsilon$ dans (0.21) dépend a priori de $\epsilon$ via l'instant final $t/\epsilon$.

On contrôle ainsi l'erreur pour chaque composante. Remarquons que l'interdépendance entre les deux équations lente et rapide a pour conséquence que l'erreur forte sur la composante lente dépend du pas de temps effectif pour l'équation rapide $\tau = \frac{\Delta t}{\epsilon}$, avec l'ordre $1/4 - r$ correspondant à la discrétisation d'une EDPS, bien qu'a priori l'évolution a lieu à l'échelle lente, selon une équation déterministe.

De plus, lorsque le paramètre $\epsilon$ tend vers 0, l'estimation de l'erreur se dégrade, à moins d'utiliser un pas de temps $\Delta t$ qui soit petit devant $\epsilon$.

Par ailleurs, on a vu que la composante lente $X^\epsilon$ converge vers la solution $\overline{X}$ de l'équation moyennée, en loi pour les processus, et avec un ordre de convergence pour un instant fixé, au sens fort et au sens faible. On va construire une méthode numérique inspirée par ce principe de moyennisation, et basée sur le constat suivant: l'approximation de la solution $\overline{X}$ par un schéma d'Euler est d'ordre $1-r$ par rapport au pas de temps, pour tout $r \in (0,1)$. La difficulté restante concerne l'approximation du coefficient moyenné $\overline{F}$: un schéma construit à partir de l'équation rapide va fournir une valeur convenable.

0.1.4.2. *Construction de la méthode.* La méthode est construite selon le principe des Méthodes Multi-échelles Hétérogènes, qui s'applique à différents types de problèmes, décrits notamment dans l'article [24]. Dans le cas d'un système d'EDS, le principe du schéma est exposé dans [64], et l'analyse est effectuée dans [25]: on adapte leur construction dans le cas des EDPS, et on s'inspire de leurs preuves, en modifiant aussi parfois certains arguments.

Supposons dans un premier temps que le coefficient moyenné $\overline{F}$ est connu. On construit alors une méthode numérique, grâce aux deux étapes suivantes: on commence par approcher la composante lente par la solution de l'équation moyennée, puis on approche cette dernière à l'aide d'une méthode d'Euler. Plus précisément, étant donné un pas de temps $\Delta t > 0$ et un instant final $T > 0$, on définit pour tout entier $n$ tel que $(n+1)\Delta t \leq T$

$$\overline{X}_{n+1} = S_{\Delta t}\overline{X}_n + \Delta t S_{\Delta t}\overline{F}(\overline{X}_n),$$

avec la condition initiale $\overline{X}_0 = x$.

En décomposant l'erreur sous la forme

$$X^\epsilon(n\Delta t) - \overline{X}_n = \left(X^\epsilon(n\Delta t) - \overline{X}(n\Delta t)\right) + \left(\overline{X}(n\Delta t) - \overline{X}_n\right),$$

on voit qu'on obtient les estimées d'erreur forte et faible suivantes: pour tout instant final $T > 0$, tout $r \in (0,1/2)$, toute conditions initiales $x \in D(-A)^\theta$ pour un $\theta > 0$, $y \in H$, tout $\epsilon_0 > 0$, tout $\Delta t_0$, et toute fonction test $\phi \in \mathcal{C}_b^2(H)$, on a pour tout pas de temps $0 < \Delta t < \Delta t_0$, tout paramètre $0 < \epsilon < \epsilon_0$, et tout entier $n$ tel que $n\Delta t \leq T$

$$\mathbb{E}|X^\epsilon(n\Delta t) - \overline{X}_n| \leq C(\epsilon^{1/2-r} + \Delta t^{1-r}), |\mathbb{E}\phi(X^\epsilon(n\Delta t)) - \mathbb{E}\phi(\overline{X}_n)| \leq C(\epsilon^{1/2} + \Delta t^{1-r}). \tag{0.22}$$

Pour obtenir une méthode générale, il reste à définir un moyen d'approximation de $\overline{F}(x)$, pour tout $x \in H$, grâce aux identités

$$\begin{aligned}\overline{F}(x) &= \int_H F(x,y)\mu^x(dy) \\ &= \lim_{t\to+\infty} \mathbb{E}F(x, Y_x(t,y)) \\ &= \lim_{t\to+\infty} \frac{1}{t}\int_0^t \mathbb{E}F(x, Y_x(s,y))ds,\end{aligned}$$

où $Y_x(.y)$ est la solution de l'équation rapide avec composante lente figée (0.3); par ergodicité, les convergences sont valables pour toute condition initiale $y \in H$. La meilleure estimation de convergence est obtenue en utilisant la dernière expression. En fait, on va utiliser les résultats d'approximation de la mesure invariante par une méthode numérique présentés dans la section précédente. En définissant un schéma pour l'équation rapide, de pas de temps $\delta t > 0$, on va obtenir un terme supplémentaire d'erreur de l'ordre de $(\frac{\delta t}{\epsilon})^{1/2-r}$.

On va utiliser un schéma d'Euler pour chaque composante; en toute généralité, on peut utiliser deux méthodes distinctes. La composante lente est approchée par un "macro-schéma", avec macro-pas de temps $\Delta t$; le "micro-schéma", avec micro-pas de temps $\delta t$, est utilisé comme un procédé auxiliaire nécessaire au fonctionnement du macro-schéma. Une telle méthode est alors appelée Hétérogène, puisqu'elle traite différemment les deux échelles de temps.

On introduit donc un micro-pas de temps $\delta t > 0$, et des entiers non nuls $N, n_T, M$.

Pour chaque entier $n$ tel que $(n+1)\Delta t \leq T$, on pose

$$X_{n+1} = S_{\Delta t}X_n + \Delta t S_{\Delta t}\tilde{F}_n, \tag{0.23}$$

avec la condition initiale $X_0 = x$.

On veut que $\tilde{F}_n$ soit une approximation de $\overline{F}(X_n)$: on définit donc

$$\tilde{F}_n = \frac{1}{MN}\sum_{j=1}^{M}\sum_{m=n_T}^{n_T+N-1} F(X_n, Y_{n,m,j}), \tag{0.24}$$

où les quantités $Y_{n,m,j}$ sont définies à l'aide du micro-schéma: pour tout entier $m \geq 0$

$$Y_{n,m+1,j} = R_{\frac{\delta t}{\epsilon}} Y_{n,m,j} + \frac{\delta t}{\epsilon} R_{\frac{\delta t}{\epsilon}} G(X_n, Y_{n,m,j}) + \sqrt{\frac{\delta t}{\epsilon}} R_{\frac{\delta t}{\epsilon}} \zeta_{n,m+1,j}, \tag{0.25}$$

avec

$$\zeta_{n,m+1,j} = \frac{W^{(n,j)}_{(m+1)\delta t} - W^{(n,j)}_{m\delta t}}{\sqrt{\delta t}},$$

pour des processus de Wiener cylindriques sur $H$ indépendants $(W^{(n,j)})_{1\leq j\leq M, 0\leq n\leq n_T+N-1}$. Pour bien définir la méthode, il faut initialiser le micro-schéma, pour chaque macro-étape d'indice $n$. On fait le choix suivant:

$$Y_{0,0,j} = y \text{ pour tout } 1 \leq j \leq M; Y_{n,0,j} = Y_{n-1,n_T+N-1,j}, \text{ pour tout } n \geq 1 \text{ et tout } 1 \leq j \leq M.$$

A chaque macro-étape, on fait fonctionner le micro-schéma $M$ fois indépendamment, partant pour chaque réalisation de la dernière valeur calculée correspondante, durant $n_0 := n_T + N - 1$. Pour calculer alors la valeur de $\tilde{F}_n$, on moyennise sur les différentes réalisations et sur les $N$ dernières valeurs, ne tenant pas compte des $n_T$ premières.

L'erreur se décompose sous la forme

$$X^\epsilon(n\Delta t) - X_n = \left(X^\epsilon(n\Delta t) - \overline{X}_n\right) + \left(\overline{X}_n - X_n\right).$$

La première partie est contrôlée d'après (0.22); la deuxième partie dépend d'un contrôle de $\tilde{F}_n - \overline{F}(\overline{X}_n)$, qui dépend de trois sources d'erreur:

- l'erreur due au schéma de discrétisation en lui-même, qui dépend du pas de temps utilisé. Plus précisément, le résultat d'approximation de la mesure invariante par un schéma numérique est utilisé.
- l'approximation de l'espérance à l'aide d'une moyenne empirique sur $M$ réalisations, selon une méthode de Monte-Carlo;
- l'approximation du coefficient moyenné par une valeur en un temps fixé, selon un théorème ergodique, via le réglage des paramètres $n_T$ et $N$.

On obtient les résultats de convergence forte et faible suivants:

**Théorème 4** (Convergence forte)**.** *On suppose que l'hypothèse de dissipativité forte* (SD) *est satisfaite. Considérons une condition initiale $x, y \in H$.*

*Pour tous $0 < r < 1/2$, $0 < \kappa < 1/2$, $T > 0$, $\epsilon_0 > 0$, $\Delta_0 > 0$, $\tau_0 > 0$, il existe une constante $C > 0$ telle que pour tous $0 < \epsilon \leq \epsilon_0$, $0 < \Delta t \leq \Delta_0$, $\delta t > 0$ avec $\tau = \frac{\delta t}{\epsilon} \leq \tau_0$, et tout entier $n$ tel que $n \geq 1$ et $n\Delta t \leq T$, on a*

$$\begin{aligned}\mathbb{E}\left|X^\epsilon(n\Delta t) - X_n\right| \leq\ & C\left(\epsilon^{1/2-r} + \frac{1}{n} + \Delta t^{1-r}\right) \\ & + C\left((\frac{\delta t}{\epsilon})^{1/2-\kappa} + \frac{1}{\sqrt{N\frac{\delta t}{\epsilon}+1}} e^{-cn_T\frac{\delta t}{\epsilon}}\right) \\ & + C\frac{\sqrt{\Delta t}}{\sqrt{M(N\frac{\delta t}{\epsilon}+1)}}.\end{aligned}$$

**Théorème 5** (Convergence faible)**.** *On suppose que l'hypothèse de dissipativité faible* (WD) *est satisfaite. Considérons une condition initiale telle que $x \in D((-A)^\theta)$ et $y \in D((-B)^\beta)$, avec $\theta, \beta \in ]0, 1]$.*

*Soit $\Phi : H \to \mathbb{R}$ une fonction test de classe $\mathcal{C}^2$, bornée et à dérivées bornées.*

*Pour tous* $0 < r < 1/2$, $0 < \kappa < 1/2$, $T > 0$, $\epsilon_0 > 0$, $\Delta_0 > 0$, $\tau_0 > 0$, *il existe une constante* $C > 0$ *telle que pour tous* $0 < \epsilon \leq \epsilon_0$, $0 < \Delta t \leq \Delta_0$, $\delta t > 0$ *avec* $\tau = \frac{\delta t}{\epsilon} \leq \tau_0$, *et tout entier* $n$ *tel que* $n \geq 1$ *et* $n\Delta t \leq T$, *on a*

$$\left|\mathbb{E}\Phi(X^\epsilon(n\Delta t)) - \mathbb{E}\Phi(X_n)\right| \leq C\left(\epsilon^{1-r} + \frac{1}{n} + \Delta t^{1-r}\right) + C\left((\frac{\delta t}{\epsilon})^{1/2-\kappa}(1 + \frac{1}{((n_T-1)\frac{\delta t}{\epsilon})^{1/2-\kappa}}) + \frac{1}{N\frac{\delta t}{\epsilon}+1}e^{-cn_T\frac{\delta t}{\epsilon}}\right).$$

Dans ces estimations d'erreur, on peut clairement identifier les différentes sources d'erreur et le rôle de chacun des paramètres introduits. De plus, on voit qu'en choisissant convenablement les paramètres on peut rendre l'erreur aussi petite que voulu. On observe également que le paraètre $\epsilon$ n'intervient que sous la forme $\tau = \frac{\delta t}{\epsilon}$.

Une possibilité de comparaison entre le schéma HMM et une méthode directe consiste à définir un critère de coût. La définition du coût et l'obtention de paramètres tels que le schéma HMM est plus efficace sont détaillées dans le chapitre 3.

Les preuves des deux théorèmes 4 et 5 sont des transpositions de celles des résultats de convergence dans le principe de moyennisation correspondants: on décompose l'erreur forte de la même manière, globalement; pour étudier l'erreur faible, on utilise aussi la solution d'une équation de Kolmogorov en temps discret, construite à partir du schéma numérique, avec une condition initiale bien choisie, différente de la fonction test $\phi$ du Théorème.

Notons quelques différences avec la situation des EDS dans [25]. Premièrement, il faut noter que les ordres de convergence des méthodes d'Euler sont différents. De plus, le choix d'initialisation du processus numérique rapide n'est pas pleinement utilisé, comme il l'est dans [25]: à première vue, nos résultats sont moins bons, néanmoins nous avons l'avantage de présenter des preuves simplifiées. De plus, l'amélioration des estimées serait marginale, à cause de l'apparition du paramètre de régularité $\eta$ des coefficients de Nemytskii: l'analyse des coûts du schéma HMM et d'un schéma direct ne serait pas modifiée.

On voit que le nombre de réalisations $M$ n'intervient pas dans l'estimation du Théorème 5: la technique de preuve utilisée ne repose que sur des procédés d'analyse de l'erreur faible, sans besoin de l'erreur forte.

Notons que nous avons seulement mené une analyse de l'erreur induite par un schéma de discrétisation temporelle du sytème (0.2), en considérant les différences d'échelles de temps. Pour des applications pratiques, il faudrait ajouter une étape de discrétisation spatiale: son étude théorique n'a pas été effectuée jusqu'ici.

Finissons cette partie de l'introduction mettant en perspective la méthode numérique proposée dans cette thèse par rapport à d'autres travaux.

La construction de la méthode numérique se situe dans un contexte plus général, correspondant au formalisme "equation-free", présenté dans [42]. Il s'agit détudier des systèmes complexes, tels qu'un comportement macroscopique émerge à partir d'interactions à une échelle microscopique; plus particulièrement, la description macroscopique est théoriquement connue, mais inconnue sous une forme directement utilisable. Ici, on dispose de l'équation moyennée, mais la non-connaissance du coefficient $\overline{F}$ oblige à se servir de manière indirecte du modèle complet: on calcule $\tilde{F}_n$ à partir des $Y_{n,m,j}$ en étant guidé par le principe de moyennisation, et le choix des paramètres $M, N, n_T$ assure l'efficacité de la méthode.

On notera aussi l'utilisation dans [46] d'un algorithme de type pararéel micro-macro, pour l'approximation de solutions d'équations différentielles ordinaires avec deux échelles de temps. De manière itérative, une première méthode numérique est utilisée pour proposer une solution à l'échelle macroscopique, puis une correction est apportée sur chaque pas de temps, parallèlement, grâce à une méthode tenant compte du comportement microscopique.

## 0.2. Méthode semi-lagrangienne hybride

On présente maintenant le contenu des deux derniers chapitres de la thèse, qui est largement indépendant de la première partie constituée par les trois premiers chapitres et résumée ci-dessus.

0.2.1. **Description de la methode.** Le cadre semi-lagrangien est utilisé pour approximer la solution $u$ d'un problème d'évolution décrit par une EDP (ici parabolique), tel qu'il existe une formule reliant la valeur de $u$ à un instant $t_2$ et à une position $x$, et la valeur de $u$ à un instant $t_1 < t_2$ et à une position $\mathcal{X}(t_2, t_1, x)$. Cette dernière quantité est déterminée à partir des coefficients de l'EDP, par une équation des courbes caractéristiques. On produit ainsi une représentation Lagrangienne de la quantité Eulérienne $u$. Par exemple, on considère des équations pouvant contenir un terme de transport, ou un terme de diffusion - auquel cas la formule de représentation est probabiliste, impliquant l'espérance par rapport aux courbes caractéristiques devenues variables aléatoires. C'est ce cas d'équations avec diffusion qu'on va étudier plus précisément: le calcul exact d'espérance étant en général impossible, une méthode de Monte-Carlo est alors employée, rendant la méthode numérique hybride. Remarquons que le traitement de certaines équations non-linéaires est possible à l'aide de méthodes semi-implicites.

Pour clarifier la présentation, considérons le cas d'une EDP associée à un processus de diffusion $X$:

$$\begin{aligned} \frac{\partial u(t,x)}{\partial t} &= \frac{1}{2}\Delta u(t,x) + c(x).\nabla u(t,x), \text{ pour tout } 0 < t \leq T \text{ et tout } x \in \mathbb{R}^d \\ u(0,x) &= u_0(x), \text{ pour tout } x \in \mathbb{R}^d, \end{aligned} \tag{0.26}$$

telle que la condition initiale $u_0$ est supposée régulière. La fonction $c : \mathbb{R}^d \to \mathbb{R}$ est aussi supposée régulière.

Dans un premier temps, pour simplifier on considère que l'équation est posée dans le domaine $D = \mathbb{R}^d$, évitant ainsi le traitement de conditions au bord: une présentation plus générale est donnée dans le Chapitre 5.

Une formule de représentation probabiliste donne une expression de la solution $u$: pour tout $t \geq 0$ et tout $x \in \mathbb{R}^d$ on a

$$u(t,x) = \mathbb{E}u_0(X_t^x),$$

où $(X_t^x)_{t\geq 0}$ est solution de l'équation différentielle stochastique

$$dX_t^x = c(X_t^x)dt + dB_t, \quad X_0^x = x, \tag{0.27}$$

avec $(B_t)_{t\geq 0}$ un mouvement brownien $d$-dimensionnel standard.

Le générateur infinitésimal vérifie $\mathcal{L} = c.\nabla + \frac{1}{2}\Delta$, de telle sorte que $\frac{\partial u}{\partial t} = \mathcal{L}u$.

En général, pour un instant $t$ et une position initiale $x$ données, la loi de la variable aléatoire $X_t^x$ n'est pas connue, et on ne peut pas calculer explicitement l'espérance. La procédure classique pour traiter le problème repose sur la méthode de Monte-Carlo.

Rappelons que pour calculer une valeur approchée de l'espérance $\mathbb{E}Y$ d'une variable aléatoire $Y$, on simule un échantillon de la loi de $Y$ $(Y_1, \ldots, Y_N)$ de taille $N$ - i.e. $N$ réalisations indépendantes de $Y$ - puis on définit la moyenne empirique

$$\overline{Y} = \frac{1}{N}\sum_{m=1}^{N} Y_m.$$

D'après la Loi des Grands Nombres, si $Y$ est intégrable, alors presque sûrement lorsque $N \to \infty$ on a la convergence $\overline{Y} \to \mathbb{E}Y$; de plus si $Y$ admet un moment d'ordre 2, on a un contrôle de l'erreur:

$$\left(\mathbb{E}|\overline{Y} - \mathbb{E}Y|^2\right)^{1/2} \leq \frac{\sqrt{\mathrm{Var}(Y)}}{\sqrt{N}}.$$

Si on suppose qu'on est capable de générer un $N$-échantillon $(X_t^{x,m})_{1\leq m\leq N}$ de la loi de $X_t^x$, on peut donc approcher $u(t,x)$ par

$$\frac{1}{N}\sum_{m=1}^{N} u_0(X_t^{x,m}). \tag{0.28}$$

Notons qu'en général la variance des variables aléatoires $u_0(X_t^{x,m})$ est de taille $t$.

Si on a besoin d'une connaissance globale de la solution à un instant $t$, l'opération précédente doit *a priori* être répétée pour différentes valeurs de $x$: par exemple $x$ peut varier sur une grille $(x_j = j\delta x)_{j\in\mathbb{N},j\in\mathbb{N}}$ de taille $\delta x > 0$.

Toutefois une difficulté apparaît: on ne dispose pas d'une méthode de simulation exacte pour construire l'échantillon $(X_t^{x,m})_{1\leq m\leq N}$ dans (0.28). Pour définir une approximation, on peut utiliser un schéma numérique, par exempe de type Euler: si $\tau > 0$ désigne le pas de temps de la méthode, on définit récursivement

$$\begin{aligned} X_0(x) &= x, \\ X_{n+1}(x) &= X_n(x) + \tau c(X_n(x)) + (B_{(n+1)\tau} - B_{n\tau}), \text{ for } n \geq 0. \end{aligned} \tag{0.29}$$

L'erreur est d'ordre 1 par rapport à $\tau$, dans un sens faible: si $\varphi : \mathbb{R}^d \to \mathbb{R}$ est une fonction lisse, il existe une constante $C > 0$ telle que pour tout $n \geq 0$ vérifiant $n\tau \leq t$ on a

$$|\mathbb{E}\varphi(X_{n\tau}^x) - \mathbb{E}\varphi(X_n(x))| \leq C\tau.$$

Bien que l'utilisation combinée d'un schéma numérique et d'une méthode de Monte-Carlo soit possible, on constate que même lorsque $\delta t$ et $\delta x$ sont choisis assez petits, les variances des variables aléatoires simulées restent de taille $t$. De plus, il semble que chaque calcul de la solution en un nouveau point nécessite de reprendre la procédure d'approximation au départ. On propose au contraire une méthode qui fournit intrinsèquement une connaissance globale de la solution, avec de plus un contrôle des variances par rapport aux paramètres de discrétisation $\delta t$ et $\delta x$.

La construction du schéma repose sur un généralisation de la formule de représentation probabiliste de la solution de l'EDP: elle permet de relier les valeurs de la solutions entre deux instants $t_1 < t_2$, en utilisant la propriété de Markov du processus associé: on a pour tout $x \in \mathbb{R}^d$

$$u(t_2, x) = \mathbb{E}u(t_1, X_{t_2-t_1}^x), \tag{0.30}$$

Etant alors donnés un pas de temps $\delta t > 0$ et un pas d'espace $\delta x > 0$, (0.30) fournit un moyen pour approcher $u((n+1)\delta t, j\delta x)$ en fonction de la solution $u(n\delta t, .)$ à l'instant $n\delta t$. Pour obtenir une approximation de la loi de $X_{\delta t}^x$ on utilise un pas de temps dans le schéma d'Euler (0.29) avec $\tau = \delta t$.

Il reste à définir une méthode d'interpolation, afin de retrouver une fonction définie sur l'adhérence du domaine $\overline{D}$ à partir de valeurs données aux points de la grille $(j\delta x)_j$. Formellement, l'opérateur $\mathcal{J}$ transforme un vecteur $w = (w_j)_j$ en une fonction $\mathcal{J}(w) : \overline{D} \to \mathbb{R}$, avec $\mathcal{J}(w)(j\delta x) = w_j$.

On peut maintenant décrire entièrement la relation de récurrence définissant le schéma. On choisit la condition initiale $u^0 = (u_0(k\delta x))_k$. Alors étant donné $u^n = (u_k^n)_k$, on approche

$$u((n+1)\delta t, j\delta x) = \mathbb{E}u(n\delta t, X_{\delta t}^{j\delta x})$$

avec

$$\mathbb{E}u(n\delta t, j\delta x + \delta t c(j\delta x) + B_{\delta t}) = \mathbb{E}u(n\delta t, j\delta x + \delta t c(j\delta x) + \sqrt{\delta t}\mathcal{N}),$$

grâce à une itération du schéma d'Euler, où $\mathcal{N}$ désigne une variable de loi normale centrée réduite.

L'approximation suivante est donnée par l'interpolation, avec

$$v_j^{n+1} := \mathbb{E}\big(\mathcal{J}(u^n)\big)(j\delta x + \delta t c(j\delta x) + \sqrt{\delta t}\mathcal{N});$$

enfin la méthode de Monte-Carlo fournit la définition

$$u_j^{n+1} := \frac{1}{N}\sum_{m=1}^{N} \big(\mathcal{J}(u^n)\big)(j\delta x + \delta t c(j\delta x) + \sqrt{\delta t}\mathcal{N}^{n,m,j}),$$

où pour $n$ et $j$ fixés les variables aléatoires $(\mathcal{N}^{n,m,j})_{1\leq m\leq M}$ sont des variables Gaussiennes standard indépendantes. On requiert aussi l'indépendance de ces variables par rapport aux paramètres $n$ et $j$.

Selon le principe des méthodes semi-lagrangiennes, on a utilisé des courbes caractéristiques aléatoires sur des intervalles de temps de taille $\delta t$, ainsi qu'une méthode d'interpolation, pour relier la solution entre deux instants. L'ajout d'une approximation de type Monte-Carlo imposée par le caractère probabiliste de la représentation justifie l'utilisation de la terminologie "hybride".

Notons qu'on définit ainsi des méthodes explicites qui ne reposent pas sur une condition de type Courant-Friedrichs-Lewy (CFL), mais plutôt sur une condition anti-CFL issue du caractère semi-lagrangien.

0.2.2. **Analyse de l'erreur de type Monte-Carlo.** L'étude théorique menée dans le Chapitre 4 concerne le cas simple l'équation de la chaleur - avec $c = 0$ dans (0.26) - en dimension 1, avec conditions au bord périodiques sur le domaine $(0, 1)$:

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{1}{2}\frac{\partial^2 u}{\partial x^2}, \text{ pour } t > 0, x \in (0,1), \\ u(0,.) &= u_0, \\ \text{avec conditions au bord périodiques : } &\forall t \geq 0, u(t,1) = u(t,0). \end{aligned} \tag{0.31}$$

Même s'il s'agit d'une situation extrêmement simple, on va voir qu'elle permet d'obtenir, par le calcul et avec une preuve non triviale, une borne améliorée sur l'erreur de type Monte-Carlo.

L'opérateur $\frac{1}{2}\frac{\partial^2}{\partial x^2}$ est le générateur infinitésimal du Mouvement Brownien $(B_t)_{t\geq 0}$; la formule de représentation associée donne pour tout $t \geq 0$ et $x \in [0, 1]$

$$u(t, x) = \mathbb{E}u_0(x + B_t),$$

avec une interprétation modulo 1 de la quantité $x + B_t$.

La méthode semi-lagrangienne se construit grâce à la formule suivante, avec $t_1 < t_2$ et $x \in [0, 1]$:

$$u(t_2, x) = \mathbb{E}u(t_1, x + B_{t_2-t_1}).$$

En utilisant le schéma d'Euler et une interpolation linéaire, les quantités $u^n$ et $v^n$ prennent l'expression suivante:

$$\begin{aligned} u_j^{n+1} &= \frac{1}{N}\sum_{m=1}^{N}\sum_{k;0\leq k\delta x<1} v_k^n \phi_k(j\delta x + \sqrt{\delta t}\mathcal{N}^{n,m,j}), \\ v_j^{n+1} &= \sum_{k;0\leq k\delta x<1} v_k^n \mathbb{E}\phi_k(j\delta x + \sqrt{\delta t}\mathcal{N}) \end{aligned}$$

où les variables aléatoires $\mathcal{N}^{n,m,j}$ sont indépendantes et de même loi Gaussienne standard. La condition initiale est telle que $u_j^0 = v_j^0 = u_0(j\delta x)$. Les fonctions de base $\phi_k$ sont définies dans le chapitre 4.

L'indépendance par rapport au paramètre $m$ est intrinsèque à la méthode de Monte-Carlo appliquée à chaque calcul d'espérance. Par rapport à l'indice de temps $n$, l'indépendance est aussi naturelle: les accroissements du mouvement brownien sont indépendants. Enfin, l'indépendance par rapport à l'indice d'espace $j$ est a priori non nécessaire; mais on constate qu'elle a de nombreux effets: d'une part, on rompt une certaine symétrie du problème en injectant un bruit blanc en espace; on rend aussi les solutions d'autant plus irrégulières que le pas d'espace $\delta x$ est petit; néanmoins ce bruit a pour effet spectaculaire de fournir une estimation améliorée de l'erreur Monte-Carlo, où on observe essentiellement que la variance globale dans une norme bien choisie est de taille $\delta t$, au lieu d'être de taille 1 pour une méthode classique.

Plus généralement, au lieu de considérer des variables gaussiennes indépendantes par rapport à l'indice spatial $j$, on pourrait comme dans [41] introduire une matrice de corrélation $K$, et chercher à minimiser la variance par rapport au choix de $K$. En faisant ici le choix de la matrice identité pour $K$, on considère un bruit blanc en espace, et l'analyse de l'erreur proposée montre une estimation non triviale de la variance, due à un effet de moyennisation. Il est naturel de se demander si un autre choix de $K$ peut aussi réduire la variance; néanmoins on se contente dans ce travail du choix $K = I$.

On étudie l'erreur en norme $l^2$ discrète, définie par

$$\|u\|_{l^2}^2 = \delta x \sum_{k\in\mathbb{N};0\leq k\delta x<1} |u_k|^2.$$

L'erreur dépend de la condition initiale à travers sa semi-norme $h^1$ discrète, telle que

$$|u|_{h^1}^2 = \delta x \sum_{k\in\mathbb{N};0\leq k\delta x<1} \frac{|u_{k+1} - u_k|^2}{\delta x^2}.$$

Ces deux normes constituent de bonnes approximations des norme et semi-norme correspondantes pour des fonctions définies sur l'intervalle $(0, 1)$, respectivement sur $L^2(0, 1)$ et $H^1(0, 1)$.

Le problème étant linéaire, le schéma peut être naturellement écrit matriciellement, sous la forme

$$u^{n+1} = P^{(n)} u^n,$$
$$v^{n+1} = Q v^n;$$

l'essentiel de la preuve réside dans la compréhension du comportement des matrices $P^{(n)}$ et $Q$ par rapport aux normes et semi-normes.

Le résultat est le suivant:

**Théorème 6.** *Pour tout $p \in \mathbb{N}$ et tout temps final $T > 0$, il existe une constante $C_p > 0$, telle que pour tous $\delta t > 0$, $\delta x > 0$ et$N \in \mathbb{N}^*$ on a*

$$\begin{aligned} \mathbb{E}\|u^n - v^n\|_{l^2}^2 &= \delta x \sum_{j;0\leq j\delta x<1} \mathbb{E}|u_j^n - v_j^n|^2 \\ &\leq C_p |u^0|_{h^1}^2 (1 + \frac{\delta x^2}{\delta t}) \left(1 + \frac{\delta x}{\delta t} + \frac{\delta x^2}{\delta t^2}(1 + |\log(\delta t)|)\right)^p \left(\frac{\delta t}{N} + \frac{1}{N^{p+1}}\right). \end{aligned} \tag{0.32}$$

Le facteur le plus intéressant est le dernier: la puissance $p$ pouvant être arbitraire, le terme dominant est le premier. Essentiellement la variance globale est de taille $\frac{\delta t}{N}$, alors qu'on s'attendrait plus classiquement à la borner par $\frac{1}{N}$. A notre connaissance, il s'agit d'un résultat original dans le contexte des méthodes de Monte-Carlo.

L'analyse des autres facteurs est essentielle: on constate qu'une condition de type anti-CFL est nécessaire, et qu'elle est compatible avec la condition imposée par le cadre semi-lagrangien: on suppose qu'il existe une constante $c > 0$ telle qu'on choisit les paramètres $\delta t$ et $\delta x$ sous la contrainte

$$\frac{\delta x}{\delta t}\sqrt{|\log(\delta t)|} \leq c. \tag{0.33}$$

Pour prouver le Théorème 6, il faut comprendre comment s'accumule l'erreur au cours du temps. Tout d'abord, il faut analyser la variance de l'erreur sur un pas de temps: on montre qu'elle est de taille $\frac{\delta t + \delta x^2}{N}$, et qu'elle a deux sources, respectivement la variance associée à une diffusion sur un intervalle de temps de longueur $\delta t$, et l'erreur d'interpolation. Notons qu'elle dépend de la semi-norme $h^1$ de la solution à l'instant précédent. En sommant ces erreurs, on obtient (0.32) pour $p = 0$, avec déjà une condition de type anti-CFL.

L'amélioration de l'estimation demande une analyse plus précise des matrices $P^{(k)}$. Notons que l'écriture matricielle du schéma est bien adaptée et qu'on peut décomposer l'erreur en norme $l^2$. La remarque fondamentale est la suivante: l'indépendance au niveau spatial du bruit à chaque étape simplifie le contrôle de l'erreur en faisant apparaître un terme diagonal d'une puissance de $Q$, essentiellement de taille $\delta x$, au lieu de sa norme matricielle induite, uniquement bornée par 1. Néanmoins, on ne peut pas utiliser directement cette remarque: le contrôle de la variance sur un pas de temps demanderait en effet de contrôler $\mathbb{E}|P^{(k-1)} \dots P^{(0)} u^0|_{h^1}^2$. Or sur un pas de temps, on peut seulement prouver que pour tout vecteur $u^0$

$$\mathbb{E}|P^{(0)} u^0|_{h^1}^2 \leq C(1 + \frac{\delta t}{N \delta x^2})|u^0|_{h^1}^2.$$

La propriété d'indépendance en espace n'a donc pas que des avantages: plus le pas d'espace $\delta x$ diminue, plus les solutions sont irrégulières. Cette propriété se comprend en analysant les différences de comportement des matrices $Q$ et $P^{(k)}$.

Ces matrices partagent la propriété d'être stochastiques - leurs coefficients sont positifs, et leur somme est égale à 1 sur chaque ligne. Cette propriété a pour conséquence la décroissance suivante de la norme $l^2$ des solutions: pour tout $n \geq 0$, avec $0 \leq (n+1)\delta t \leq T$

$$\mathbb{E}\|u^{n+1}\|_{l^2}^2 \leq \mathbb{E}\|u^n\|_{l^2}^2,$$
$$\|v^{n+1}\|_{l^2} \leq \|v^n\|_{l^2}.$$

On constate que la matrice $Q$ a une propriété supplémentaire: elle est symétrique, ce qui rend le choix de la norme euclidienne normalisée $\|\,.\,\|_{l^2}$ bien adapté. De plus, on obtient, par un argument de type intégration par parties, que la décroissance se généralise à la semi-norme $h^1$:

$$|v^{n+1}|_{h^1} \leq |v^n|_{h^1}.$$

Comme on l'a vu, cette décroissance ne s'étend pas au matrices $P^{(k)}$. On propose l'explication suivante: la connaissance du rayon spectral d'une matrice ne suffit pas pour déterminer une estimation d'une norme induite. L'identité $\rho(A) = \sqrt{|||A^*A|||_2}$ repose en effet sur une diagonalisation en base orthonormale de la matrice autoadjointe $A^*A$: même si on sait que les valeurs propres d'une matrice $P^{(k)}$ sont de module inférieur à 1 - par propriété de stochasticité - on n'a pas d'informations sur le comportement de vecteurs propres par rapport à la semi-norme $h^1$, tandis que la matrice $Q$ est symétrique pour la norme $l^2$ et la semi-norme $h^1$!

Notons que le système déquations définissant le schéma forme un système dynamique aléatoire:

$$u^n = P^{(n-1)} \dots P^{(0)} u^0.$$

Pour comprendre le comportement de ce système, on peut s'intéresser à ses exposants de Lyapunov - définis presque sûrement, grâce au théorème ergodique multiplicatif d'Oseledets - voir [2]. La propriété vérifiée par chaque réalisation des matrices $P^{(k)}$ d'être une matrice stochastique entraîne que ces exposants sont négatifs. Or ces exposants ne dépendent pas du choix de norme équivalente: en définissant à partir de la semi-norme $|\,.\,|_{h^1}$ une norme telle que

$$\|u\|_{h^1}^2 = \|u\|_{l^2}^2 + |u|_{h^1}^2,$$

on obtient une information sur le comportement du système par rapport à cette norme, qui est toutefois insuffisante. En effet, une constante dépendant de $\delta x$ intervient lorsqu'on change de norme; de plus, le problème n'est pas lié au comportement en temps long du système.

Pour résoudre les problèmes et tirer profit des propriétés énoncées, on raffine le développement de l'erreur, afin d'utiliser un argument de type bootstrap: d'une part, on obtient un terme d'erreur de taille $\frac{\delta t}{N}$; d'autre part, on voit que l'erreur obtenue à l'étape précédente est multipliée par $\frac{1}{N}$; en itérant l'argument, on obtient (0.32).

0.2.3. **Extensions de la méthode.** Dans le Chapitre 5, on généralise l'approche hybride semi-lagrangienne du Chapitre 4, à de nombreuses situations plus générales et plus complexes: dimension supérieure, traitement de conditions au bord, présence de termes non-linéaires. Ce chapitre a un caractère prospectif: on propose des algorithmes et des simulations associées, sans aucune analyse théorique. Les simulations obtenues montrent que la méthode est adaptable et donne des résultats raisonnables.

On y propose également une expérience confirmant le Théorème 6, ainsi qu'une expérience suggérant que la même convergence doit avoir lieu pour des conditions au bord de Dirichlet plutôt que périodiques.

Premièrement, on peut approcher les solutions d'équations en dimension spatiale $d$ supérieure à 2: par exemple si on considère l'équation de la chaleur avec conditions au bord périodiques

$$\frac{\partial u}{\partial t} = \frac{1}{2}\Delta u, \text{ pour } t > 0, x \in (0,1)^d,$$
$$u(0,.) = u_0,$$
avec conditions au bord périodiques,

on dispose d'une formule de représentation probabiliste de la solution mettant en jeu un mouvement Brownien $d$-dimensionnel $(B_t^1, \dots, B_t^d)_{t\geq 0}$. Il est facile de généraliser les algorithmes. Notons que la méthode est explicite; lorsque la dimension $d$ augmente, on évite ainsi l'inversion de matrices de grande taille.

On montre ensuite que différents types de conditions au bord sont imposables: on peut les choisir de type Dirichlet ou Neumann. Dans le premier cas, il faut tuer la diffusion Brownienne lorsqu'elle atteint le bord; dans le deuxième cas, il faut la réfléchir. Le cadre semi-lagrangien impose de réaliser cette modification à chaque étape de temps. Numériquement, pour bien capter le comportement au bord des solutions, on effectue un traitement spécial selon la position par rapport au bord: on raffine l'intervalle de temps, et dans le cas des conditions de type Dirichlet on utilise en outre un test de type acceptation-rejet construit à partir de la loi exacte de sortie du domaine d'un pont Brownien, suivant [32].

Etant donnée une fonction $f$, on montre qu'on peut approcher les solutions d'équations de Poisson $\mathcal{L}u + f$, vues comme solutions stationnaires de $\frac{\partial u}{\partial t} = \mathcal{L}u + f$.

On fournit des simulations numériques pour l'équation de la chaleur en dimension 1 et 2, avec conditions au bord de type périodiques, Dirichlet et Neumann. On choisit différentes conditions initiales $u_0$ et fonctions

$f$, de telle manière qu'on dispose d'expressions explicites de la solution à tout instant et de la solution stationnaire, en utilisant les modes de Fourier adaptés.

Enfin la méthode s'applique à des équations plus compliquées du type

$$\frac{\partial u}{\partial t} = \nu \mathcal{L} u + f(u),$$

avec un paramètre de viscosité $\nu > 0$. L'approche est généralisée en utilisant des schémas de discrétisation temporelle semi-implicites.

Lorsque $\nu$ tend vers 0, la méthode peut être utiliser afin de capter des couches limites de la solution, comme le montrent des exemples proposés en dimension 1 et 2.

Les exemples de simulations donnés dans le Chapitre 5 concernent les équations de Burgers et de Navier-Stokes (en dimension 2) avec viscosité $\nu > 0$. Ces EDPs partagent l'apparition de la nonlinéarité quadratique $f(u) = u.\nabla u$, modélisant le phénomène de convection dans un fluide. Entre deux étapes de temps, on discrétise ce terme selon une formule du type $u^n.\nabla u^{n+1}$; en approchant le terme visqueux de manière implicite, on dispose sur chaque intervalle numérique d'une formule de représentation probabiliste - avec une diffusion dont le terme de dérive dépend de la solution calculée à l'instant précédent.

Notons aussi que la méthode a déjà été appliquée dans [27] pour obtenir des simulations, par exemple pour des solutions d'équations de type Ginzburg-Landau, ou de Fisher-KPP.

Dans le cas des équations de Burgers

$$\frac{\partial u}{\partial t} + (u.\nabla)u - \nu \Delta u = f,$$

avec une condition initiale et des conditions au bord, on applique la méthode avec différentes conditions initiales et en modifiant la viscosité $\nu$. Les simulations montrent qu'on parvient à capter les chocs dans la solution lorsque $\nu$ diminue.

Finalement, on propose quelques simulations de solutions des équations de Navier-Stokes incompressibles, en dimension 2, obtenues par une méthode de projection entièrement hybride. Rappelons le principe d'une telle méthode - voir [33]: on souhaite résoudre le problème dans le domaine $D$, avec conditions au bord de Dirichlet homogènes

$$\begin{aligned} \frac{\partial u}{\partial t} + (u.\nabla)u - \nu \Delta u + \nabla p = f, \text{ pour tout } (t,x) \in (0,T] \times D, \\ \operatorname{div}(u) = 0, \\ u(0,.) = u_0, \\ u_{|\partial D} = 0. \end{aligned} \tag{0.34}$$

Le gradient de pression $\nabla p$ s'interprète comme un terme de force correcteur, présent pour assurer la condition d'incompressibilité du fluide $\operatorname{div}(u) = 0$; c'est la combinaison de la nonlinéarité quadratique et de ce terme correcteur qui fait toute la difficulté de l'étude de ces équations.

Les méthodes de projection sont issues de la remarque suivante: un champ de vitesse (assez régulier) se décompose en la somme d'un champ solénoidal - i.e. de divergence nulle - et du gradient d'une fonction. A chaque étape de temps, on commence alors par calculer un premier champ de vitesse sans la contrainte d'incompressibilité, puis on effectue une correction. La première étape correspond à une évolution selon une EDP du type Burgers, pour laquelle on sait qu'une méthode semi-lagrangienne hybride s'applique et fournit une vitesse $v$; la deuxième étape consiste en la résolution d'une équation de type Poisson, de la forme

$$\Delta q = \operatorname{div}(v),$$

telle que le champ de vitesse recherché est $u = v - \nabla q$. La difficulté théorique et numérique repose dans le choix des conditions au bord. Puisque la vitesse doit vérifier des conditions de type Dirichlet homogènes, le choix le plus naturel est d'imposer des conditions de type Neumann homogènes; des méthodes plus sophistiquées peuvent rendre mieux compte du problème, mais celle proposée ici présente l'avantage de la simplicité, et de pouvoir être mise en oeuvre à l'aide d'une méthode semi-lagrangienne hybride - en interprétant $q$ comme solution stationnaire d'une équation d'évolution parabolique.

On obtient des simulations numériques dans le cas test classique de la cavité entraînée, avec différentes valeurs de la viscosité; on peut les comparer avec des simulations obtenues par des méthodes déterministes - par exemple dans [59]. On observe la formation attendue des tourbillons.

# Chapitre 1

# Strong and weak orders in averaging for SPDEs


Résumé

Dans ce chapitre, on s'intéresse à l'ordre de convergence dans le principe de moyennisation pour un système d'EDP stochastiques du type

$$\frac{\partial x^\epsilon(t,\xi)}{\partial t} = \frac{\partial^2 x^\epsilon(t,\xi)}{\partial \xi^2} + f(\xi, x^\epsilon(t,\xi), y^\epsilon(t,\xi)),$$
$$\frac{\partial y^\epsilon(t,\xi)}{\partial t} = \frac{1}{\epsilon}\frac{\partial^2 y^\epsilon(t,\xi)}{\partial \xi^2} + \frac{1}{\epsilon} g(\xi, x^\epsilon(t,\xi), y^\epsilon(t,\xi)) + \frac{1}{\sqrt{\epsilon}}\frac{\partial \omega(t,\xi)}{\partial t}.$$

On introduit notamment l'ensemble des hypothèses de travail.

On montre que sous une hypothèse de dissipation forte de l'équation rapide la convergence est d'ordre 1/2 au sens fort (lorsqu'on compare les trajectoires de la composante lente et du processus moyenné) et qu'en imposant une hypothèse de dissipation plus faible on observe que la convergence au sens faible est d'ordre 1 (lorsqu'on compare les lois).



Ce travail a fait l'objet d'une publication sous le titre
**Strong and weak orders in averaging for SPDEs**
dans la revue *Stochastic Processes and their Applications* (122(7), 2553–2593, 2012).

## 1.1. Introduction

In this Chapter, we consider a randomly-perturbed system of reaction-diffusion equations that can be written as

$$\begin{aligned} \frac{\partial x^\epsilon(t,\xi)}{\partial t} &= \frac{\partial^2 x^\epsilon(t,\xi)}{\partial \xi^2} + f(\xi, x^\epsilon(t,\xi), y^\epsilon(t,\xi)), \\ \frac{\partial y^\epsilon(t,\xi)}{\partial t} &= \frac{1}{\epsilon}\frac{\partial^2 y^\epsilon(t,\xi)}{\partial \xi^2} + \frac{1}{\epsilon} g(\xi, x^\epsilon(t,\xi), y^\epsilon(t,\xi)) + \frac{1}{\sqrt{\epsilon}}\frac{\partial \omega(t,\xi)}{\partial t}, \end{aligned} \tag{1.1}$$

for $t \geq 0, \xi \in (0,1)$, with initial conditions $x^\epsilon(0,\xi) = x(\xi)$ and $y^\epsilon(0,\xi) = y(\xi)$, and Dirichlet boundary conditions $x^\epsilon(t,0) = x^\epsilon(t,1) = 0$ and $y^\epsilon(t,0) = y^\epsilon(t,1) = 0$. The stochastic perturbation $\frac{\partial \omega(t,\xi)}{\partial t}$ is a space-time white noise and $\epsilon > 0$ is a small parameter.

Such a system presents a specific structure: while the variations of the first component *a priori* depend on the slow time $t$, the second component evolves with respect to the fast time $\frac{t}{\epsilon}$. These two natural time scales are coupled through the nonlinear terms in the two equations.

In this setting, the main idea of the *averaging principle*, see for instance [37], is to study the behaviour of the system when $\epsilon$ tends to 0 by exhibiting a limit equation - the so-called averaged equation - for the slow component $x^\epsilon$, and to prove the convergence of $x^\epsilon$ towards the solution of this averaged equation. Here, we show two approximation results - see Theorems 1.1 and 1.2 - and give explicit order of convergence with respect to $\epsilon$.

The averaged equation comes from the asymptotic behaviour of the fast equation. Heuristically, when $t > 0$ and $\epsilon \to 0$, the fast time $\frac{t}{\epsilon}$ goes to $+\infty$, so that we expect the solution of the fast equation to be quickly close to a stochastic equilibrium (and this is the case under the dissipativity assumptions made in this paper), and that we can replace $y^\epsilon(t,\xi)$ in the slow equation with some stationary - in the stochastic sense - process, which leads to the definition of averaged coefficients in the slow equation.

To give precise results, it is convenient to look at the equations in an abstract setting, where system (1.1) can be rewritten

$$\begin{aligned} dX^\epsilon(t) &= \big(AX^\epsilon(t) + F(X^\epsilon(t), Y^\epsilon(t))\big)dt \\ dY^\epsilon(t) &= \frac{1}{\epsilon}\big(BY^\epsilon(t) + G(X^\epsilon(t), Y^\epsilon(t))\big)dt + \frac{1}{\sqrt{\epsilon}} dW(t), \end{aligned} \tag{1.2}$$

with initial conditions given by $X^\epsilon(0) = x \in H$, $Y^\epsilon(0) = y \in H$, where $H$ is the Hilbert space $L^2(0,1)$, and $W$ is a cylindrical Wiener process on $H$ - see Section 1.2.1.2. In the case of system (1.1), the definitions of $A$ and $B$ are given in Example 1.4, and the definitions of $F$ and $G$ as Nemytskii operators are given in the second part of Example 1.12. Nevertheless the abstract setting allows for more general equations, and in the sequel we only work with system (1.2).

When $\epsilon$ tends to 0, the slow component $X^\epsilon$ is approximated by the process $\overline{X}$, which follows the deterministic evolution equation

$$d\overline{X}(t) = \big(A\overline{X}(t) + \overline{F}(\overline{X}(t))\big)dt, \tag{1.3}$$

with the initial condition $\overline{X}(0) = x$, where the nonlinear coefficient $\overline{F}$ is obtained via an averaging procedure - explained in detail in Section 1.2.2.

In this article, we analyse the error between $X^\epsilon(t)$ and $\overline{X}(t)$, with two different criteria. We focus on the order of convergence, i.e. we bound the error by $C\epsilon^\Lambda$, where $C$ is a constant and $\Lambda$ is the order, which gives an idea of the speed of convergence of $X^\epsilon(t)$ towards $\overline{X}(t)$. As a result, we control the error made when $X^\epsilon(t)$ is

approximated by $\overline{X}(t)$. For instance, the order of convergence is crucial for the analysis of numerical schemes used to approximate the slow component $X^\epsilon$. In Chapter 3, we extend a numerical scheme for SDEs analysed in [25] for systems of SPDEs satisfying the same structure assumptions as system (1.1). This scheme - called the Heterogeneous Multiscale Method - is deeply based on the averaging principle: instead of computing $X^\epsilon$, we approximate $\overline{X}$ - and we can control the error we make. Moreover, the nonlinear averaged coefficient $\overline{F}$ is never explicitly calculated in the scheme, but only approximated by using numerical approximations of the values of the fast component at large times. The theorems we prove here allow to analyse the convergence of such a scheme with the same kind of criteria, and without knowing the order of convergence in the averaging principle it would not be possible to control the error made in the numerical approximation.

The two main theorems give bounds on the error between $X^\epsilon$ and $\overline{X}$; they need different dissipativity conditions (SD) and (WD) which determine how the fast equation converges to its equilibrium, as explained below.

First, when Assumption 1.10 holds, the error can be estimated in a strong sense, where trajectories of the processes are compared at a given time $t$:

**Theorem 1.1** (Strong-order)**.** *Assume* (SD). *For any* $0 < r < 1/2$, $T > 0$, $x \in H$, $y \in H$, *there exists* $C = C(T, r, x, y) > 0$ - *depending also on the constants of the problem - such that for any* $\epsilon > 0$ *and* $0 \leq t \leq T$

$$\mathbb{E}\big|X^\epsilon(t) - \overline{X}(t)\big|_H \leq C\epsilon^{1/2-r}. \tag{1.4}$$

The error can also be estimated in a weak sense, where we are interested in the distance between the laws of the processes at a given time $t$; then only Assumption 1.11 is necessary, since we only need consequences of dissipativity at the level of the transition semi-group, instead of trajectories:

**Theorem 1.2** (Weak-order)**.** *Assume* (WD). *For any* $0 < r < 1$, $T > 0$, $0 < \theta \leq 1$, $x \in D(-A)^\theta$, $y \in H$, $\phi \in \mathcal{C}_b^2(H)$, *there exists* $C > 0$, *depending on* $r$, $T$, $\phi$, $|x|_{(-A)^\theta}$, $|y|$ *and the constants of the problem, such that for any* $\epsilon > 0$ *and* $t \leq T$

$$\big|\mathbb{E}[\phi(X^\epsilon(t))] - \mathbb{E}[\phi(\overline{X}(t))]\big| \leq C\epsilon^{1-r}. \tag{1.5}$$

The domains $D(-A)^\theta$ are usually the classical Sobolev spaces $H^{2\theta}$ with respect to the eigenbasis of $A$ - see Definition 1.5. We remark that for the first theorem no regularity is needed for the initial condition - i.e. we can take $\theta = 0$ - while we require $\theta > 0$ for the second one; this is explained in the proof of Theorem 1.2. We need to take a small parameter $r > 0$, which can be as small as possible, but different from 0. This is an effect of the infinite dimensional setting.

As a consequence, we can say that the strong order in averaging is $1/2$, while the weak order is 1. It is a general fact that the weak order is greater than the strong order - since test functions $\phi$ in the Theorem are Lipschitz continuous - but it is worth proving that there is a gap; this fact was known for SDEs - see [37], [38] - but had not been proved yet for SPDEs.

The strong convergence Theorem 1.1 is proved when the fast equation satisfies a strict dissipativity assumption: for any $x \in H$, the function $G(x, .)$ is Lipschitz continuous, with constant $L_g$ - independent of $x$ - satisfying the following condition:

$$L_g < \mu,$$

where $\mu$ is the smallest eigenvalue of the linear operator $-B$. Thanks to this assumption, we can easily analyse the asymptotic behaviour of the fast equation with frozen slow component; we can identify a unique invariant probability measure - depending on $x$ - and show some exponential convergence to equilibrium. More precisely, we control in a strong sense the difference between two solutions of this fast equation starting from different initial conditions, and driven by the same noise $W$: under the previous assumption, the ergodicity comes from properties of the deterministic equation only.

The weak convergence Theorem 1.2 needs a weaker dissipativity assumption (1.11), which yields the same ergodicity properties - unique invariant probability measure, exponential convergence to equilibrium - but with different arguments: the asymptotic behaviour of the transition semi-group can be analysed, thanks to the non-degeneracy of the noise - leading to a Strong Feller Property. A coupling method - adapted from the study of Markov processes, like in [44] or [50] - implies that the laws - instead of trajectories - of the fast process issued from two different initial conditions are exponentially closed. We refer to Section 1.2.2 for a

precise result, and to [19] for a detailed proof. It seems that for the first time an averaging result is obtained for SPDEs under a weak dissipativity condition.

Notice that we have assumed that the slow equation has no white noise term $dw(t)$; as a consequence, the averaged equation is a deterministic parabolic partial differential equation. Considering a more general situation with some additive noise terms in the slow equation, independent of the noise in the fast equation, we could still prove in a similar way a strong order result, the only changes being time regularity of solutions. We would obtain order $1/5$, which can also be compared with the order $1/3$ obtained for SDEs. But if we introduce noise in the slow equation, the method we used to prove the weak order theorem becomes more complicated, and we have not extended the result to this situation so far.

In the case of stochastic differential equations, averaging results are already well-known - see for instance [30], [37]. Convergence in law or in probability of $X^\epsilon$ to $\overline{X}$ in the space $\mathcal{C}([0,T],H)$ can be shown by different techniques: by using a Hasminskii technique based on a subdivision of the interval $[0,T]$, see [25], and [48]; a Poisson equation, see [56]; the method of perturbed test functions and of a martingale problem approach, see [29]; or an asymptotic expansion of the solutions of Kolmogorov equations, see [25] and [38].

As far as stochastic partial differential equations are concerned, in [10] both the Hasminskii technique and a martingale problem approach are used; in [12] a modified Poisson equation is the essential tool. Then convergence in law or in probability of $X^\epsilon$ to $\overline{X}$ in the space $\mathcal{C}([0,T],H)$ - the space of continuous functions from $[0,T]$ to $H$ - is proved; but order of convergence was never given.

Our proof of Theorem 1.1 relies on the Hasminskii technique already known for SDEs: we introduce an auxiliary process for which the slow component of the fast variable is frozen on small intervals of a subdivision. We use Hölder regularity of order $1-r$ in time of the slow component, for which we do not need $\theta > 0$.

To prove Theorem 1.2, we adapt the method of finding an expansion with respect to $\epsilon$ of the solutions of the Kolmogorov equations related to our system. This seems to be the first time that such a method is used to prove an averaging result for SPDEs. New technical difficulties due to infinite dimension arise: we use non bounded linear operators and non smooth nonlinear coefficients, and the Kolmogorov equations are more difficult to use. For these reasons, we use a reduction to finite dimension technique, keeping in mind that bounds must be independent of dimension, so that precise estimates are needed for each term appearing in the expansion. We interpret the necessity of $\theta > 0$ with a singularity which needs to be integrable.

In Section 1.2, we set the notations and give some results on the fast equation, allowing to define the averaged equation; we also precise the assumptions needed to prove the Theorems. Then in Section 1.3, we prove the strong-order result. In Section 1.4, we give the details of the method for proving the weak-order result. Finally in Section 1.5 and in the Appendix, we prove all the necessary estimates.

## 1.2. Preliminaries

### 1.2.1. **Assumptions and notations.**

1.2.1.1. *Test functions.* To study weak convergence, we use test functions $\phi$ in the space $\mathcal{C}_b^2(H,\mathbb{R})$ of functions from $H$ to $\mathbb{R}$ that are twice continuously differentiable, with first and second order bounded derivatives.

In the sequel, we often identify the first derivative $D\phi(x) \in \mathcal{L}(H,\mathbb{R})$ with the gradient in $H$, and the second derivative $D^2\phi(x)$ with a linear operator on $H$, via the formulae:

$$< D\phi(x), h >= D\phi(x).h \text{ for every } h \in H$$
$$< D^2\phi(x).h, k >= D^2\phi(x).(h,k) \text{ for every } h,k \in H.$$

1.2.1.2. *Stochastic integration in Hilbert spaces.* In this section, we recall the definition of the cylindrical Wiener process and of stochastic integral on a separable Hilbert space $H$ (its norm is denoted by $|.|_H$ or just $|.|$). For more details, see [15].

We first fix a filtered probability space $(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\geq 0},\mathbb{P})$. A cylindrical Wiener process on $H$ is defined with two elements:

- a complete orthonormal system of $H$, denoted by $(q_i)_{i\in I}$, where $I$ is a subset of $\mathbb{N}$;
- a family $(\beta_i)_{i\in I}$ of independent real Wiener processes with respect to the filtration $((\mathcal{F}_t)_{t\geq 0})$:

$$W(t) = \sum_{i \in I} \beta_i(t) q_i. \tag{1.6}$$

When $I$ is a finite set, we recover the usual definition of Wiener processes in the finite dimensional space $\mathbb{R}^{|I|}$. However the subject here is the study of some Stochastic Partial Differential Equations, so that in the sequel the underlying Hilbert space $H$ is infinite dimensional; for instance when $H = L^2(0,1)$, an example of complete orthonormal system is $(q_k) = (\sqrt{2}\sin(k\pi.))_{k\geq 1}$ - see Example 1.4.

A fundamental remark is that the series in (1.6) does not converge in $H$; but if a linear operator $\Psi : H \to K$ is Hilbert-Schmidt, then $\Psi W(t)$ converges in $L^2(\Omega, H)$ for any $t \geq 0$.

We recall that a linear operator $\Psi : H \to K$ is said to be Hilbert-Schmidt when

$$|\Psi|^2_{\mathcal{L}_2(H,K)} := \sum_{k=0}^{+\infty} |\Psi(q_k)|^2_K < +\infty,$$

where the definition is independent of the choice of the orthonormal basis $(q_k)$ of $H$. The space of Hilbert-Schmidt operators from $H$ to $K$ is denoted $\mathcal{L}_2(H,K)$; endowed with the norm $|.|_{\mathcal{L}_2(H,K)}$ it is an Hilbert space.

The stochastic integral $\int_0^t \Psi(s)dW(s)$ is defined in $K$ for predictable processes $\Psi$ with values in $\mathcal{L}_2(H,K)$ such that $\int_0^t |\Psi(s)|^2_{\mathcal{L}_2(H,K)} ds < +\infty$ a.s; moreover when $\Psi \in L^2(\Omega \times [0,t]; \mathcal{L}_2(H,K))$, the following two properties hold:

$$\mathbb{E}|\int_0^t \Psi(s)dW(s)|^2_K = \mathbb{E}\int_0^t |\Psi(s)|^2_{\mathcal{L}_2(H,K)} ds, \text{ (Itô isometry)},$$

$$\mathbb{E}\int_0^t \Psi(s)dW(s) = 0.$$

A generalization of Itô formula also holds - see [15].

For instance, if $v = \sum_{k\in\mathbb{N}} v_k q_k \in H$, we can define

$$< W(t), v >= \int_0^t < v, dW(s) >= \sum_{k\in\mathbb{N}} \beta_k(t) v_k;$$

we then have the following space-time white noise property

$$\mathbb{E} < W(t), v_1 >< W(s), v_2 >= t \wedge s < v_1, v_2 > .$$

Therefore to be able to integrate a process with respect to $W$ requires some strong properties on the integrand; in our SPDE setting, the Hilbert-Schmidt properties follow from the assumptions made on the linear coefficients of the equations.

1.2.1.3. *Assumptions on the linear operators.* We have to specify some properties of the linear operators $A$ and $B$ coming into the definition of system (1.2); we assume that the linear parts are of parabolic type, with space variable $\xi \in (0,1)$.

We assume that $A$ and $B$ are unbounded linear operators, with domains $D(A)$ and $D(B)$, which satisfy the following assumptions:

**Assumptions 1.3.** (1) *We assume that $(e_k)_{k\in\mathbb{N}}$ and $(f_k)_{k\in\mathbb{N}}$ are orthonormal basis of $H$, and $(\lambda_k)_{k\in\mathbb{N}}$ and $(\mu_k)_{k\in\mathbb{N}}$ are non-decreasing sequences of real positive numbers such that:*

$$Ae_k = -\lambda_k e_k \text{ for all } k \in \mathbb{N}$$
$$Bf_k = -\mu_k f_k \text{ for all } k \in \mathbb{N}.$$

*We use the notations $\lambda := \lambda_0 > 0$ and $\mu := \mu_0 > 0$ for the smallest eigenvalues of $A$ and $B$.*

(2) *The sequences $(\lambda_k)$ and $(\mu_k)$ go to $+\infty$; moreover we have some control of the behaviour of $(\mu_k)$ given by:*

$$\sum_{k=0}^{+\infty} \frac{1}{\mu_k^\alpha} < +\infty \Leftrightarrow \alpha > 1/2.$$

**Example 1.4.** *$A = B = \frac{d^2}{dx^2}$, with domain $H^2(0,1) \cap H_0^1(0,1) \in L^2(0,1)$ - homogeneous Dirichlet boundary conditions: in that case $\lambda_k = \mu_k = \pi^2(k+1)^2$, and $e_k(\xi) = f_k(\xi) = \sqrt{2}\sin((k+1)\pi\xi)$ - see* [8].

In the abstract setting, powers of $-A$ and $-B$, with their domains can be easily defined:

**Definition 1.5.** *For $a, b \in [0,1]$, we define the operators $(-A)^a$ and $(-B)^b$ by*

$$(-A)^a x = \sum_{k=0}^{\infty} \lambda_k^a x_k e_k \in H,$$
$$(-B)^b y = \sum_{k=0}^{\infty} \mu_k^b y_k f_k \in H,$$

*with domains*

$$D(-A)^a = \left\{ x = \sum_{k=0}^{+\infty} x_k e_k \in H; |x|^2_{(-A)^a} := \sum_{k=0}^{+\infty} (\lambda_k)^{2a} |x_k|^2 < +\infty \right\};$$
$$D(-B)^b = \left\{ y = \sum_{k=0}^{+\infty} y_k f_k \in H, |y|^2_{(-B)^b} := \sum_{k=0}^{+\infty} (\mu_k)^{2b} |y_k|^2 < +\infty \right\}.$$

The domains $D(-A)^a$ are related to Sobolev spaces $H^{2a}(0,1)$: therefore when $x$ belongs to a space $D(-A)^a$, the exponent $a$ represents some regularity of the function $x$.

The semi-groups $(e^{tA})_{t\geq 0}$ and $(e^{tB})_{t\geq 0}$ can be defined by the Hille-Yosida Theorem (see [8]). We use the following spectral formulae: if $x = \sum_{k=0}^{+\infty} x_k e_k \in H$ and $y = \sum_{k=0}^{+\infty} y_k f_k \in H$, then for any $t \geq 0$

$$e^{tA} x = \sum_{k=0}^{+\infty} e^{-\lambda_k t} x_k e_k \quad \text{and} \quad e^{tB} y = \sum_{k=0}^{+\infty} e^{-\mu_k t} y_k f_k.$$

For any $t \geq 0$, $e^{tA}$ and $e^{tB}$ are continuous linear operators in $H$, with respective operator norms $e^{-\lambda t}$ and $e^{-\mu t}$. The semi-group $(e^{tA})$ is used to define the solution $Z(t) = e^{tA} z$ of the linear Cauchy problem

$$\frac{dZ(t)}{dt} = AZ(t) \quad \text{with} \quad Z(0) = z.$$

To define solutions of more general PDEs of parabolic type, we use mild formulation, and Duhamel principle.

These semi-groups enjoy some smoothing properties that we often use in this work. Basically we need the following properties, which are easily proved using the above spectral properties. We write them for $A$, but they also hold with $B$.

**Proposition 1.6.** *Under Assumption 1.3, for any $\sigma \in [0,1]$, there exists $C_\sigma > 0$ such that we have:*

(1) *for any $t > 0$ and $x \in H$*
$$|e^{tA} x|_{(-A)^\sigma} \leq C_\sigma t^{-\sigma} e^{-\frac{\lambda}{2} t} |x|_H.$$

(2) *for any $0 < s < t$ and $x \in H$*
$$|e^{tA} x - e^{sA} x|_H \leq C_\sigma \frac{(t-s)^\sigma}{s^\sigma} e^{-\frac{\lambda}{2} s} |x|_H.$$

(3) *for any $0 < s < t$ and $x \in D(-A)^\sigma$*
$$|e^{tA} x - e^{sA} x|_H \leq C_\sigma (t-s)^\sigma e^{-\frac{\lambda}{2} s} |x|_{(-A)^\sigma}.$$

Under the previous assumptions on the linear coefficients, it is easy to show that the following stochastic integral is well-defined in $H$, for any $t \geq 0$:

$$W^B(t) = \int_0^t e^{(t-s)B} dW(s). \tag{1.7}$$

It is called a stochastic convolution, and it is the unique mild solution of

$$dZ(t) = BZ(t)dt + dW(t) \quad \text{with} \quad Z(0) = 0.$$

Under the second condition of Assumption 1.3, there exists $\delta > 0$ such that for any $t > 0$ we have $\int_0^t \frac{1}{s^\delta} |e^{sB}|^2_{\mathcal{L}_2(H)} ds < +\infty$; it can then be proved that $W^B$ has continuous trajectories - via the *factorization method*, see [15] - and that for any $1 \leq p < +\infty$ $\sup_{t \geq 0} \mathbb{E}|W^B(t)|^p_H < +\infty$.

1.2.1.4. *Assumptions on the nonlinear coefficients.* We now give the Assumptions on the nonlinear coefficients $F, G : H \times H \to H$. First, we need some regularity properties:

**Assumptions 1.7.** *We assume that there exists $0 \leq \eta < \frac{1}{2}$ and a constant $C$ such that the following directional derivatives are well-defined and controlled:*

- *For any $x, y \in H$ and $h \in H$, $|D_x F(x,y).h| \leq C|h|_H$ and $|D_y F(x,y).h| \leq C|h|_H$.*
- *For any $x, y \in H$, $h \in H$, $k \in D(-A)^\eta$, $|D^2_{xx} F(x,y).(h,k)| \leq C|h|_H |k|_{(-A)^\eta}$.*
- *For any $x, y \in H$, $h \in H$, $k \in D(-B)^\eta$, $|D^2_{yy} F(x,y).(h,k)| \leq C|h|_H |k|_{(-B)^\eta}$.*
- *For any $x, y \in H$, $h \in H$, $k \in D(-B)^\eta$, $|D^2_{xy} F(x,y).(h,k)| \leq C|h|_H |k|_{(-B)^\eta}$.*
- *For any $x, y \in H$, $h \in D(-A)^\eta$, $k \in H$, $|D^2_{xy} F(x,y).(h,k)| \leq C|h|_{(-A)^\eta} |k|_H$.*

*We moreover assume that $F$ is bounded.*

**Remark 1.8.** *We warn the reader that constants may vary from line to line during the proofs, and that in order to use lighter notations we usually forget to mention dependence on the parameters. We use the generic notation $C$ for such constants.*

We assume that the fast equation is a gradient system: for any $x$ the nonlinear coefficient $G(x,.)$ is the derivative of some potential $U$. We also assume regularity assumptions as for $F$.

**Assumptions 1.9.** *The function $G$ is defined through $G(x,y) = \nabla_y U(x,y)$, for some potential $U : H \times H \to \mathbb{R}$. Moreover we assume that $G$ is bounded, and that the regularity assumptions given in the Assumption 1.7 are also satisfied for $G$.*

Finally, we need to assume some dissipativity of the fast equation. Assumption 1.10 is necessary to prove Theorem 1.1, while Assumption 1.11 is weaker and is sufficient to prove Theorem 1.2.

**Assumptions 1.10** (Strict dissipativity)**.** *Let $L_g$ denote the Lipschitz constant of $G$ with respect to its second variable; then*

$$L_g < \mu, \tag{SD}$$

*where $\mu$ is defined in Assumption 1.3.*

**Assumptions 1.11** (Weak Dissipativity)**.** *There exist $c > 0$ and $C > 0$ such that for any $x \in H$ and $y \in D(B)$*

$$< By + G(x,y), y > \leq -c|y|^2 + C. \tag{WD}$$

Indeed, according to Assumptions 1.3 and 1.9, the weak dissipativity Assumption 1.11 is always satisfied, while strict dissipativity requires a condition on the Lipschitz constant of $G$.

**Example 1.12.** *We give some fundamental examples of nonlinearities for which the previous assumptions are satisfied:*

- *Functions $F, G : H \times H \to H$ of class $\mathcal{C}^2$, bounded and with bounded derivatives, such that $G(x,y) = \nabla_y U(x,y)$ and satisfying* (SD) *fit in the framework, with the choice $\eta = 0$.*
- *Functions $F$ and $G$ can be **Nemytskii** operators: let $f : (0,1) \times \mathbb{R}^2 \to \mathbb{R}$ be a bounded measurable function such that for almost every $\xi \in (0,1)$ $f(\xi,.)$ is twice continuously differentiable, bounded and with uniformly bounded derivatives. Then $F$ is defined for every $x, y \in H = L^2(0,1)$ by*

$$F(x,y)(\xi) = f(\xi, x(\xi), y(\xi)).$$

*For $G$, we assume that there exists a function $g$ with the same properties as $f$ above, such that $G(x,y)(\xi) = g(\xi, x(\xi), y(\xi))$. The strict dissipativity Assumption SD is then satisfied when*

$$\sup_{\xi \in (a,b), x \in \mathbb{R}, y \in \mathbb{R}} |\frac{\partial g}{\partial y}(\xi, x, y)| < \mu.$$

*The conditions in Assumption 1.7 are then satisfied for $F$ and $G$ as soon as there exists $\eta < 1/2$ such that $D(-A)^\eta$ and $D(-B)^\eta$ are continuously embedded into $L^\infty(0,1)$ - it is the case for $A$ and $B$ given in Example 1.4, with $\eta > 1/4$.*

We remark that under Assumption 1.9, if we define $U_0(x,y) = \int_0^1 < G(x,sy), y > ds$, we have $U_0(x,y) = U(x,y) - U(x,0)$; therefore $U_0$ is another potential for $G$, and is the only one such that for any $x \in H$ we have $U_0(x,0) = 0$. In the sequel, it is therefore not restrictive to assume $U = U_0$.

Now we can define solutions of system (1.2); under Assumptions 1.3, 1.7, 1.9, we notice that the nonlinearities $F$ and $G$ are Lipschitz continuous, and the following Proposition is classical - see [15]:

**Proposition 1.13.** *For every $\epsilon > 0$, $T > 0$, $x \in H$, $y \in H$, system* (1.2) *admits a unique mild solution $(X^\epsilon, Y^\epsilon) \in (L^2(\Omega, \mathcal{C}([0,T],H)))^2$:*

$$\begin{aligned} X^\epsilon(t) &= e^{tA}x + \int_0^t e^{(t-s)A}F(X^\epsilon(s), Y^\epsilon(s))ds \\ Y^\epsilon(t) &= e^{\frac{t}{\epsilon}B}y + \frac{1}{\epsilon}\int_0^t e^{\frac{(t-s)}{\epsilon}B}G(X^\epsilon(s), Y^\epsilon(s))ds + \frac{1}{\sqrt{\epsilon}}\int_0^t e^{\frac{(t-s)}{\epsilon}B}dW(s). \end{aligned} \tag{1.8}$$

In other words, system (1.2) is well-posed for any $\epsilon > 0$, on any finite time interval $[0,T]$.

Some properties - bounds on moments, space and time regularity, differentiability with respect to the parameters - of $X^\epsilon$ and $Y^\epsilon$ are given in the Appendix.

1.2.2. **Known results about the fast equation and the averaged equation.** In this section, we just recall without proof the main results on the fast equation with frozen slow component and on the averaged equation, defined below. Proofs can be found in [12] for the strict dissipative case, and the extension to the weakly dissipative situation relies on arguments explained below.

If $x \in H$, we define an equation on the fast variable where the slow variable is fixed and equal to $x$:

$$\begin{aligned} dY_x(t,y) &= (BY_x(t,y) + G(x, Y_x(t,y)))dt + dW(t), \\ Y_x(0,y) &= y. \end{aligned} \tag{1.9}$$

This equation admits a unique mild solution, defined on $[0, +\infty[$.

Since $Y^\epsilon$ is involved at time $t > 0$, heuristically we need to analyse the properties of $Y_x(\frac{t}{\epsilon}, y)$, with $\epsilon \to 0$, and by a change of time we need to understand the asymptotic behaviour of $Y_x(., y)$ when time goes to infinity.

Under the strict dissipativity Assumption 1.10, we obtain a contractivity of trajectories issued from different initial conditions and driven by the same noise:

**Proposition 1.14.** *With* (SD)*, for any $t \geq 0$, $x, y_1, y_2 \in H$ we have*

$$|Y_x(t,y_1) - Y_x(t,y_2)|_H \leq e^{-\frac{(\mu - L_g)}{2}t}|y_1 - y_2|_H.$$

Under the weak dissipativity Assumption 1.11, we obtain such an exponential convergence result for the laws instead of trajectories. The proof of this result is not straightforward, and can be found in [19].

**Proposition 1.15.** *With* (WD)*, there exist $c > 0$, $C > 0$ such that for any bounded test function $\phi$, any $t \geq 0$ and any $y_1, y_2 \in H$*

$$|\mathbb{E}\phi(Y(t,y_1)) - \mathbb{E}\phi(Y(t,y_2))| \leq C\|\phi\|_\infty(1 + |y_1|^2 + |y_2|^2)e^{-ct}. \tag{1.10}$$

The idea of coupling relies on the following formula: if $\nu_1$ and $\nu_2$ are two probability measures on a state space $S$, their total variation distance satisfies

$$d_{TV}(\nu_1, \nu_2) = \inf\left\{\mathbb{P}(X_1 \neq X_2)\right\},$$

which is an infimum over random variables $(X_1, X_2)$ defined on a same probability space, and such that $X_1 \sim \nu_1$ and $X_2 \sim \nu_2$.

The principle is to define a coupling $(Z_1(t,y_1,y_2), Z_2(t,y_1,y_2))_{t\geq 0}$ for the processes $(Y(t,y_1))_{t\geq 0}$ and $(Y(t,y_2))_{t\geq 0}$ such that the coupling time $\mathcal{T}$ of $Z_1$ and $Z_2$ - i.e. the first time the processes are equal - has an exponentially decreasing tail.

This technique was first used in the study of the asymptotic behaviour of Markov chains - see [7], [21], [47], [52] - and was later adapted for SDEs and more recently for SPDEs - see for instance [44], [50].

As a consequence, we can show that there exists a unique invariant probability measure associated with $Y_x$, and that the convergence to equilibrium is exponentially fast.

First, let $\nu = \mathcal{N}(0, (-B)^{-1}/2)$ be the centered Gaussian probability measure on $H$ with the covariance operator $(-B)^{-1}/2$ - which is positive and trace-class, thanks to Assumption 1.3.

Then $\mu^x$ defined by

$$\mu^x(dy) = \frac{1}{Z(x)} e^{2U(x,y)} \nu(dy), \tag{1.11}$$

where $Z(x) \in ]0, +\infty[$ is a normalization constant, is the unique probability invariant measure associated to $Y_x$. This expression comes from the gradient structure of equation (1.9), given in Assumption 1.9.

Second, under both dissipativity assumptions, the convergence to equilibrium is exponential in the following sense:

**Proposition 1.16.** *If we assume* (SD) *or* (WD)*, there exist constants $C, c > 0$ such that for any bounded function $\phi : H \to \mathbb{R}$ or $\phi : H \to H$, $t \geq 0$ and $x, y \in H$ we have*

$$|\mathbb{E}\phi(Y_x(t, y)) - \int_H \phi(z)\mu^x(dz)| \leq C\|\phi\|_\infty (1 + |y|_H^2) e^{-ct}.$$

Under the strict dissipativity Assumption 1.10, this is a consequence of Proposition 1.14 - see Theorem 3.5 and Remark 3.6 of [12]; under the weak dissipativity Assumption 1.11, Proposition 1.16 is a consequence of Proposition 1.15 and of the properties of the invariant measures $\mu^x$ - which have finite moments of any order, uniformly bounded with respect to $x$.

Now we define the averaged equation. First we define the averaged nonlinear coefficient $\overline{F}$:

**Definition 1.17.** *For any $x \in H$,*

$$\overline{F}(x) = \int_H F(x, y)\mu^x(dy). \tag{1.12}$$

Using Assumptions 1.7, 1.9 and the expression of $\mu^x$, we can easily prove the following properties on $\overline{F}$:

**Proposition 1.18.** *There exists $0 \leq \eta < 1$ and a constant $C$ such that the following directional derivatives of $\overline{F}$ are well-defined and controlled:*

- *For any $x \in H$, $h \in H$, $|D\overline{F}(x).h| \leq C|h|_H$.*
- *For any $x \in H$, $h \in H$, $k \in D(-A)^\eta$, $|D^2\overline{F}(x).(h, k)| \leq C|h|_H|k|_{(-A)^\eta}$.*

*Moreover, $\overline{F}$ is bounded and Lipschitz continuous.*

**Remark 1.19.** *Even when $F$ and $G$ are Nemytskii operators, $\overline{F}$ is not such an operator in general.*

Then the averaged equation - see (1.3) in the introduction - can be defined:

$$d\overline{X}(t) = (A\overline{X}(t) + \overline{F}(\overline{X}(t)))dt,$$

with initial condition $\overline{X}(0) = x \in H$. For any $T > 0$, this deterministic equation admits a unique mild solution $\overline{X} \in \mathcal{C}([0, T], H)$.

### 1.3. Proof of the strong-order result

The main idea - inspired by the work on SDEs of Khasminskii in [37] - of the proof of Theorem 1.1 is the construction of auxiliary processes $(\tilde{X}^\epsilon, \tilde{Y}^\epsilon)$ for any $\epsilon$, for which the analysis is simpler.

In this section, we assume that dissipativity of the fast equation is strict: we have SD.

Let $T > 0$, $x \in H$, $y \in H$ and $\epsilon > 0$ be fixed. We introduce the parameter

$$\delta = \delta(\epsilon) = \sqrt{\epsilon} \tag{1.13}$$

to define a subdivision of $[0, T]$. We also fix $r > 0$.

We define $\tilde{X}^\epsilon$ and $\tilde{Y}^\epsilon$ via a mild formulation: for any $0 \le t \le T$

$$\begin{aligned} \tilde{Y}^\epsilon(t) &= e^{\frac{t}{\epsilon}B}y + \frac{1}{\epsilon}\int_0^t e^{\frac{(t-s)}{\epsilon}B}G(X^\epsilon(\lfloor\frac{s}{\delta}\rfloor\delta), \tilde{Y}^\epsilon(s))ds + \frac{1}{\sqrt{\epsilon}}\int_0^t e^{\frac{(t-s)}{\epsilon}B}dW(s), \\ \tilde{X}^\epsilon(t) &= e^{tA}x + \int_0^t e^{(t-s)A}F(X^\epsilon(\lfloor\frac{s}{\delta}\rfloor\delta), \tilde{Y}^\epsilon(s))ds, \end{aligned} \tag{1.14}$$

where $\lfloor . \rfloor$ denotes the integer part function.

$\tilde{X}^\epsilon$ and $\tilde{Y}^\epsilon$ are continuous processes and they satisfy $\tilde{X}^\epsilon(0) = x = X^\epsilon(0)$ and $\tilde{Y}^\epsilon(0) = y = Y^\epsilon(0)$. Moreover, on any subinterval $[k\delta, (k+1)\delta]$, with $0 \le k \le N := \lfloor\frac{T}{\delta}\rfloor$, we have

$$\begin{aligned} d\tilde{X}^\epsilon(t) &= (A\tilde{X}^\epsilon(t) + F(X^\epsilon(k\delta), \tilde{Y}^\epsilon(t)))dt, \\ d\tilde{Y}^\epsilon(t) &= \frac{1}{\epsilon}(B\tilde{Y}^\epsilon(t) + G(X^\epsilon(k\delta), \tilde{Y}^\epsilon(t)))dt + \frac{1}{\sqrt{\epsilon}}dW(t). \end{aligned} \tag{1.15}$$

We remark that on such subintervals the fast component $\tilde{Y}^\epsilon$ does not depend on the slow component $\tilde{X}^\epsilon$, but only on the value of $X^\epsilon$ at the first point of the interval.

Since $F$ and $G$ are supposed to be bounded, we easily see that for any $t \ge 0$

$$\mathbb{E}|\tilde{X}^\epsilon(t)|_H^2 \le C(1 + |x|_H^2) \quad \text{and} \quad \mathbb{E}|\tilde{Y}^\epsilon(t)|_H^2 \le C(1 + |y|_H^2).$$

We show the following Lemmas:

**Lemma 1.20.** *There exists $C > 0$ such that for any $0 \le t \le T$ and any $\epsilon > 0$*

$$\mathbb{E}|X^\epsilon(t) - \tilde{X}^\epsilon(t)| \le C(\delta^{1-r} + \frac{\epsilon}{\delta}).$$

**Lemma 1.21.** *There exists $C > 0$ such that for any $0 \le t \le T$ and any $\epsilon > 0$*

$$\mathbb{E}|\tilde{X}^\epsilon(t) - \overline{X}(t)| \le C\left(\epsilon(1 + \delta^{-r})(1 + \frac{1}{1 - e^{-c\frac{\delta}{\epsilon}}})\right)^{1/2} + C(\delta^{1-r} + \frac{\epsilon}{\delta}).$$

The constant $C$ above depends on $r$, $T$, $x$, $y$, but not on $t$, $\epsilon$ or $\delta(\epsilon)$.

Lemma 1.20 explains why we can replace $\tilde{X}^\epsilon$ by $X^\epsilon$, and Lemma 1.21 gives an estimate of the distance between $\tilde{X}^\epsilon(t)$ and $\overline{X}(t)$. With the choice of $\delta(\epsilon) = \sqrt{\epsilon}$ given by (1.13), the proof of Theorem 1.1 is straightforward.

Proof of Lemma 1.20

- Estimate of $Y^\epsilon - \tilde{Y}^\epsilon$.

  We define $\rho^\epsilon(t) = Y^\epsilon(t) - \tilde{Y}^\epsilon(t)$ for any $0 \le t \le T$.

  We fix $k \ge 0$; then for any $t \in [k\delta, (k+1)\delta]$ we have

  $$d\rho^\epsilon(t) = \frac{1}{\epsilon}B\rho^\epsilon(t)dt + \frac{1}{\epsilon}(G(X^\epsilon(t), Y^\epsilon(t)) - G(X^\epsilon(k\delta), \tilde{Y}^\epsilon(t)))dt.$$

  Using a mild formulation and Gronwall Lemma, for any $t \in [k\delta, (k+1)\delta]$ we have

  $$\mathbb{E}|\rho^\epsilon(t)| \le e^{-\frac{\mu - L_g}{\epsilon}(t-k\delta)}\mathbb{E}|\rho^\epsilon(k\delta)| + \frac{C}{\epsilon}\int_{k\delta}^t e^{-\frac{\mu - L_g}{\epsilon}(t-s)}\mathbb{E}|X^\epsilon(s) - X^\epsilon(k\delta)|ds.$$

  Since $\mathbb{E}|Y^\epsilon(t)| \le C(1 + |y|)$ and $\mathbb{E}|\tilde{Y}^\epsilon(t)| \le C(1 + |y|)$, we have the same bound on $\rho^\epsilon$.

  We can integrate the previous inequality over the interval $t \in [k\delta, (k+1)\delta]$ and get

$$\begin{aligned} \int_{k\delta}^{(k+1)\delta}\mathbb{E}|\rho^\epsilon(t)|dt &\le C\int_{k\delta}^{(k+1)\delta} e^{-\frac{\mu - L_g}{\epsilon}(t-k\delta)}dt + \frac{C}{\epsilon}\int_{k\delta}^{(k+1)\delta}\int_{k\delta}^t e^{-\frac{\mu - L_g}{\epsilon}(t-s)}\mathbb{E}|X^\epsilon(s) - X^\epsilon(k\delta)|dsdt \\ &\le C\frac{\epsilon}{\mu - L_g} + C\int_{k\delta}^{(k+1)\delta}\mathbb{E}|X^\epsilon(s) - X^\epsilon(k\delta)|\int_s^{(k+1)\delta}\frac{1}{\epsilon}e^{-\frac{\mu - L_g}{\epsilon}(t-s)}dtds \\ &\le C\epsilon + C\int_{k\delta}^{(k+1)\delta}\mathbb{E}|X^\epsilon(s) - X^\epsilon(k\delta)|ds. \end{aligned}$$

  We recall that strict dissipativity $\mu - L_g > 0$ holds, thanks to Assumption 1.10.

It remains to take the sum over $k \in \{0, \dots, \lfloor \frac{t}{\delta} \rfloor\}$, where $t \leq T$; using Proposition 1.33 of the appendix, we then obtain

$$\int_0^t \mathbb{E}|\rho^\epsilon(s)|ds \leq C(r,T)(\frac{\epsilon}{\delta} + \delta^{1-r}).$$

- Estimate of $X^\epsilon - \tilde{X}^\epsilon$.
  We have for any $0 \leq t \leq T$

$$\begin{aligned}
|X^\epsilon(t) - \tilde{X}^\epsilon(t)| &= |\int_0^t e^{(t-s)A}(F(X^\epsilon(s), Y^\epsilon(s)) - F(X^\epsilon(\lfloor \frac{s}{\delta} \rfloor \delta), \tilde{Y}(s)))ds| \\
&\leq \int_0^t |F(X^\epsilon(s), Y^\epsilon(s)) - F(X^\epsilon(\lfloor \frac{s}{\delta} \rfloor \delta), \tilde{Y}(s))|ds \\
&\leq C \int_0^T (|X^\epsilon(s) - X^\epsilon(\lfloor \frac{s}{\delta} \rfloor \delta)| + |Y^\epsilon(s) - \tilde{Y}^\epsilon(s)|)ds.
\end{aligned}$$

Using the previous estimate and the regularity result from Proposition 1.33, we obtain for any $0 \leq t \leq T$

$$\mathbb{E}|X^\epsilon(t) - \tilde{X}^\epsilon(t)| \leq C(\delta^{1-r} + \frac{\epsilon}{\delta}). \tag{1.16}$$

□

Proof of Lemma 1.21 We introduce the following decomposition, for any $0 \leq t \leq T$:

$$\begin{aligned}
\tilde{X}^\epsilon(t) - \overline{X}(t) &= \int_0^t e^{(t-s)A}(F(X^\epsilon(\lfloor \frac{s}{\delta} \rfloor \delta), \tilde{Y}^\epsilon(s)) - \overline{F}(\overline{X}(s)))ds \\
&= \int_0^t e^{(t-s)A}(F(X^\epsilon(\lfloor \frac{s}{\delta} \rfloor \delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(\lfloor \frac{s}{\delta} \rfloor \delta)))ds \\
&+ \int_0^t e^{(t-s)A}(\overline{F}(X^\epsilon(\lfloor \frac{s}{\delta} \rfloor \delta)) - \overline{F}(X^\epsilon(s)))ds \\
&+ \int_0^t e^{(t-s)A}(\overline{F}(X^\epsilon(s)) - \overline{F}(\tilde{X}^\epsilon(s)))ds \\
&+ \int_0^t e^{(t-s)A}(\overline{F}(\tilde{X}^\epsilon(s)) - \overline{F}(\overline{X}(s)))ds \\
&= I_1(t) + I_2(t) + I_3(t) + I_4(t).
\end{aligned}$$

According to Proposition 1.18, $\overline{F}$ is Lipschitz continuous; thanks to Proposition 1.33 and Lemma 1.20, we then show that for any $0 \leq t \leq T$

$$\begin{aligned}
\mathbb{E}|I_2(t)| &\leq C \int_0^T \mathbb{E}|X^\epsilon(\lfloor \frac{s}{\delta} \rfloor \delta) - X^\epsilon(s)|ds \leq C\delta^{1-r} \\
\mathbb{E}|I_3(t)| &\leq C \int_0^T |X^\epsilon(s) - \tilde{X}^\epsilon(s)|ds \leq C(\delta^{1-r} + \frac{\epsilon}{\delta}) \\
\mathbb{E}|I_4(t)| &\leq CT \int_0^t \mathbb{E}|\tilde{X}^\epsilon(r) - \overline{X}(r)|ds.
\end{aligned}$$

The $I_4$ term is treated via the Gronwall Lemma.

It remains to focus on the $I_1$ term. It is fundamental to look at $\mathbb{E}|I_1(t)|^2$ and not only $\mathbb{E}|I_1(t)|$ in order to obtain the best estimate leading to order 1/2. For that, we use the subdivision of $[0,T]$ and we expand

the scalar product in $H$: we have for any $0 \le t \le T$

$$\begin{aligned}|I_1(t)|^2 &= |\int_0^t e^{(t-s)A}(F(X^\epsilon(\lfloor\frac{s}{\delta}\rfloor\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(\lfloor\frac{s}{\delta}\rfloor\delta)))ds|^2 \\ &= |\sum_{k=0}^{\lfloor\frac{t}{\delta}\rfloor}\int_{k\delta}^{(k+1)\delta\wedge t} e^{(t-s)A}(F(X^\epsilon(k\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(k\delta)))ds|^2 \\ &= A_1(t) + A_2(t),\end{aligned}$$

where

$$A_1(t) := \sum_{k=0}^{\lfloor\frac{t}{\delta}\rfloor}|\int_{k\delta}^{(k+1)\delta\wedge t} e^{(t-s)A}(F(X^\epsilon(k\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(k\delta)))ds|^2 \tag{1.17}$$

and

$$\begin{aligned}A_2(t) := 2\sum_{0\le i<j\le\lfloor\frac{t}{\delta}\rfloor}\langle &\int_{i\delta}^{(i+1)\delta\wedge t} e^{(t-s)A}(F(X^\epsilon(i\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(i\delta)))ds, \\ &\int_{j\delta}^{(j+1)\delta\wedge t} e^{(t-s)A}(F(X^\epsilon(j\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(j\delta)))ds\rangle.\end{aligned} \tag{1.18}$$

We claim that $\mathbb{E}A_1(t) \le C\epsilon$ and $\mathbb{E}A_2(t) \le C\epsilon(1+\delta^{-r})(1+\frac{1}{1-e^{-c\frac{\delta}{\epsilon}}})$, where $C = C(T, r, \theta, x, y)$. Using Gronwall Lemma, we get the result.

- We first prove the estimate on $\mathbb{E}A_1(t)$. We use conditional expectation with respect to $\mathcal{F}_s$. Then for any $0 \le k \le \lfloor\frac{t}{\delta}\rfloor$, using some symmetry for variables $s$ and $\sigma$,

$$\begin{aligned}&\mathbb{E}|\int_{k\delta}^{(k+1)\delta\wedge t} e^{(t-s)A}(F(X^\epsilon(k\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(k\delta)))ds|^2 \\ &= 2\mathbb{E}\int_{k\delta}^{(k+1)\delta\wedge t} ds\int_s^{(k+1)\delta\wedge t} d\sigma\langle e^{(t-s)A}(F(X^\epsilon(k\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(k\delta))), \\ &\qquad\qquad e^{(t-\sigma)A}(F(X^\epsilon(k\delta), \tilde{Y}^\epsilon(\sigma)) - \overline{F}(X^\epsilon(k\delta)))\rangle \\ &= 2\mathbb{E}\int_{k\delta}^{(k+1)\delta\wedge t} ds\int_s^{(k+1)\delta\wedge t} d\sigma\langle e^{(t-s)A}(F(X^\epsilon(k\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(k\delta))), \\ &\qquad\qquad e^{(t-\sigma)A}\mathbb{E}[F(X^\epsilon(k\delta), \tilde{Y}^\epsilon(\sigma)) - \overline{F}(X^\epsilon(k\delta))|\mathcal{F}_s]\rangle \\ &\le 2\int_{k\delta}^{(k+1)\delta}\int_s^{(k+1)\delta}\mathbb{E}\left(|F(X^\epsilon(k\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(k\delta))||\mathbb{E}[F(X^\epsilon(k\delta), \tilde{Y}^\epsilon(\sigma))|\mathcal{F}_s] - \overline{F}(X^\epsilon(k\delta))|\right).\end{aligned}$$

We now define the auxiliary function $\tilde{F}$. Propositions 1.43 and 1.44 - see the Appendix - give important properties of this function: we have some exponential control with respect to time $t$ of uniform and Lipschitz bounds with respect to $x$.

**Definition 1.22.** *For any $(x, y) \in H^2$ and $t \ge 0$*

$$\tilde{F}(x, y, t) = \mathbb{E}F(x, Y_x(t, y)) - \overline{F}(x). \tag{1.19}$$

Since $F$ is bounded, we can control the first factor in the integral; for the second factor, we use the definition of $\tilde{F}$, the Markov property and Proposition 1.43 to see that

$$\begin{aligned}&\mathbb{E}|\int_{k\delta}^{(k+1)\delta\wedge t} e^{(t-s)A}(F(X^\epsilon(k\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(k\delta)))ds|^2 \\ &\le C\int_{k\delta}^{(k+1)\delta} ds\int_s^{(k+1)\delta} d\sigma\mathbb{E}|\tilde{F}(X^\epsilon(k\delta), \tilde{Y}^\epsilon(s), \frac{\sigma-s}{\epsilon})| \\ &\le C\int_{k\delta}^{(k+1)\delta}\int_s^{(k+1)\delta} e^{-c\frac{\sigma-s}{\epsilon}}d\sigma ds \le C\delta\epsilon.\end{aligned}$$

Therefore we get $\mathbb{E}A_1(t) \leq C\epsilon$.

- Estimate of $\mathbb{E}A_2(t)$.

We have to introduce the following auxiliary processes, which generalize $\tilde{Y}^\epsilon$. $(Z_i^\epsilon(s))_{s\geq i\delta}$, where $i \in \{0, \dots, N\}$ is defined by:

$$\begin{aligned} dZ_i^\epsilon(s) &= \frac{1}{\epsilon}(BZ_i^\epsilon(s) + G(X^\epsilon(i\delta), Z_i^\epsilon(s)))ds + \frac{1}{\sqrt{\epsilon}}dW(s) \\ Z_i^\epsilon(i\delta) &= \tilde{Y}^\epsilon(i\delta). \end{aligned} \tag{1.20}$$

It is then clear that for $i\delta \leq s \leq (i+1)\delta$ we have $Z_i^\epsilon(s) = \tilde{Y}^\epsilon(s)$, and that $Z_k^\epsilon((k+1)\delta) = \tilde{Y}^\epsilon((k+1)\delta) = Z_{k+1}^\epsilon((k+1)\delta)$. Moreover the processes $(Z_i^\epsilon)$ are uniformly bounded with respect to $i$ and $\epsilon$.

It is then possible to rewrite the integrands appearing in the expression of $A_2$: when $i\delta \leq s \leq (i+1)\delta \leq j\delta \leq \tau \leq (j+1)\delta$,

$$\begin{aligned} &|\mathbb{E}\left\langle e^{(t-s)A}(F(X^\epsilon(i\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(i\delta))), e^{(t-\tau)A}(F(X^\epsilon(j\delta), \tilde{Y}^\epsilon(\tau)) - \overline{F}(X^\epsilon(j\delta)))\right\rangle| \\ &= |\mathbb{E}\left\langle e^{(t-s)A}(F(X^\epsilon(i\delta), \tilde{Y}^\epsilon(s)) - \overline{F}(X^\epsilon(i\delta))), e^{(t-\tau)A}\mathbb{E}[F(X^\epsilon(j\delta), \tilde{Y}^\epsilon(\tau)) - \overline{F}(X^\epsilon(j\delta))|\mathcal{F}_{(i+1)\delta}]\right\rangle| \\ &\leq C\mathbb{E}|\mathbb{E}[F(X^\epsilon(j\delta), \tilde{Y}^\epsilon(\tau)) - \overline{F}(X^\epsilon(j\delta))|\mathcal{F}_{(i+1)\delta}]|, \end{aligned}$$

since $F$ is bounded. We have $\tilde{Y}^\epsilon(\tau) = Z_j^\epsilon(\tau)$; however since $i < j$ we can use conditional expectation with respect to $\mathcal{F}_{(i+1)\delta}$ instead of $\mathcal{F}_{j\delta}$ in order to get a better estimate. We therefore propose the following decomposition

$$\begin{aligned} &\mathbb{E}|\mathbb{E}[F(X^\epsilon(j\delta), \tilde{Y}^\epsilon(\tau)) - \overline{F}(X^\epsilon(j\delta))|\mathcal{F}_{(i+1)\delta}]| \\ &\leq C\mathbb{E}|\mathbb{E}[F(X^\epsilon((i+1)\delta), Z_{i+1}^\epsilon(\tau)) - \overline{F}(X^\epsilon((i+1)\delta))|\mathcal{F}_{(i+1)\delta}]| \\ &+ C\mathbb{E}|\mathbb{E}\left[(F(X^\epsilon(j\delta), Z_j^\epsilon(\tau)) - \overline{F}(X^\epsilon(j\delta))) - (F(X^\epsilon((i+1)\delta), Z_{i+1}^\epsilon(\tau)) - \overline{F}(X^\epsilon((i+1)\delta)))|\mathcal{F}_{(i+1)\delta}\right]| \\ &=: B_1 + B_2. \end{aligned}$$

(1) First, using Markov property we have for any $j\delta \leq \tau \leq (j+1)\delta$

$$\begin{aligned} B_1 &= C\mathbb{E}|\mathbb{E}\left[F(X^\epsilon((i+1)\delta), Z_{i+1}^\epsilon(\tau)) - \overline{F}(X^\epsilon((i+1)\delta))|\mathcal{F}_{(i+1)\delta}\right]| \\ &= C\mathbb{E}|\mathbb{E}\left[\tilde{F}(X^\epsilon((i+1)\delta), Z_{(i+1)}^\epsilon((i+1)\delta), \frac{\tau-(i+1)\delta}{\epsilon})|\mathcal{F}_{(i+1)\delta}\right]| \\ &\leq Ce^{-c\frac{\tau-(i+1)\delta}{\epsilon}}, \end{aligned}$$

thanks to Proposition 1.43.

(2) $B_2$ can be rewritten using a telescoping sum, and some conditional expectation

$$\begin{aligned} B_2 = \mathbb{E}|\mathbb{E}\Big(\sum_{k=i+1}^{j-1} \mathbb{E}\Big[&[F(X^\epsilon((k+1)\delta), Z_{k+1}^\epsilon(\tau)) - \overline{F}(X^\epsilon((k+1)\delta))] \\ &- [F(X^\epsilon(k\delta), Z_k^\epsilon(\tau)) - \overline{F}(X^\epsilon(k\delta))]|\mathcal{F}_{k\delta}\Big]|\mathcal{F}_{(i+1)\delta}\Big)| \\ \leq \sum_{k=i+1}^{j-1} \mathbb{E}|\mathbb{E}\Big[&[F(X^\epsilon((k+1)\delta), Z_{k+1}^\epsilon(\tau)) - \overline{F}(X^\epsilon((k+1)\delta))] \\ &- [F(X^\epsilon(k\delta), Z_k^\epsilon(\tau)) - \overline{F}(X^\epsilon(k\delta))]|\mathcal{F}_{k\delta}\Big]|. \end{aligned}$$

Thanks to Markov property, we obtain

$$\begin{aligned} &\mathbb{E}\Big[F(X^\epsilon((k+1)\delta), Z_{k+1}^\epsilon(\tau)) - \overline{F}(X^\epsilon((k+1)\delta))|\mathcal{F}_{k\delta}\Big] \\ &\qquad = \mathbb{E}\Big[\tilde{F}(X^\epsilon((k+1)\delta), \tilde{Y}^\epsilon((k+1)\delta), \tau-(k+1)\delta, \epsilon)|\mathcal{F}_{k\delta}\Big] \end{aligned}$$

and

$$\mathbb{E}\Big[F(X^\epsilon(k\delta), Z_k^\epsilon(\tau)) - \overline{F}(X^\epsilon(k\delta))|\mathcal{F}_{k\delta}\Big] = \mathbb{E}\Big[\tilde{F}(X^\epsilon(k\delta), \tilde{Y}^\epsilon((k+1)\delta), \tau-(k+1)\delta, \epsilon)|\mathcal{F}_{k\delta}\Big].$$

Using the exponential decrease in time of the Lipschitz constant of $\tilde{F}$ with respect to $x$ given by Proposition 1.44 in the appendix, we have

$$\begin{aligned}
B_2 \leq & \sum_{k=i+1}^{j-1} \mathbb{E}|\mathbb{E}\Big[\tilde{F}(X^\epsilon((k+1)\delta), \tilde{Y}^\epsilon((k+1)\delta), \tau-(k+1)\delta, \epsilon)|\mathcal{F}_{k\delta}\Big] \\
& \qquad - \mathbb{E}\Big[\tilde{F}(X^\epsilon(k\delta), \tilde{Y}^\epsilon((k+1)\delta), \tau-(k+1)\delta, \epsilon)|\mathcal{F}_{k\delta}\Big]| \\
\leq & C\sum_{k=i+1}^{j-1} e^{-c\frac{\tau-(k+1)\delta}{\epsilon}} \mathbb{E}|X^\epsilon((k+1)\delta) - X^\epsilon(k\delta)|(1+\frac{\epsilon^\eta}{(\tau-(k+1)\delta)^\eta}) \\
\leq & C(1+\frac{\epsilon^\eta}{(\tau-j\delta)^\eta}) \sum_{k=i+1}^{j-1} e^{-c\frac{\tau-(k+1)\delta}{\epsilon}} \frac{\delta^{1-r}}{(k\delta)^{1-r}} \\
\leq & C(1+\frac{\epsilon^\eta}{(\tau-j\delta)^\eta}) \frac{\delta^{1-r}}{((i+1)\delta)^{1-r}} \frac{e^{-c\frac{\tau-j\delta}{\epsilon}}}{1-e^{-c\frac{\delta}{\epsilon}}},
\end{aligned}$$

by using the regularity proved in Proposition 1.33.

(3) We are now able to conclude, by using the estimates on $B_1$ and $B_2$: we have for any $0 \leq t \leq T$

$$\begin{aligned}
\mathbb{E}A_2(t) \leq & C \sum_{0\leq i<j\leq \lfloor \frac{t}{\delta} \rfloor} \int_{i\delta}^{(i+1)\delta\wedge t} ds \int_{j\delta}^{(j+1)\delta\wedge t} d\tau e^{-c\frac{\tau-(i+1)\delta}{\epsilon}} \\
& + C \sum_{0\leq i<j\leq \lfloor \frac{t}{\delta} \rfloor} \int_{i\delta}^{(i+1)\delta\wedge t} ds \int_{j\delta}^{(j+1)\delta\wedge t} d\tau \frac{\delta^{1-r}}{((i+1)\delta)^{1-r}} \frac{e^{-c\frac{\tau-j\delta}{\epsilon}}}{1-e^{-c\frac{\delta}{\epsilon}}}(1+\frac{\epsilon^\eta}{(\tau-j\delta)^\eta}) \\
\leq & C \sum_{0\leq i<j\leq \lfloor \frac{T}{\delta} \rfloor} \epsilon\delta e^{-c\frac{(j-i-1)\delta}{\epsilon}} + C\delta^{1-r}\epsilon \frac{1}{1-e^{-c\frac{\delta}{\epsilon}}} \sum_{0\leq i<j\leq \lfloor \frac{T}{\delta} \rfloor} \frac{\delta^{1-r}}{((i+1)\delta)^{1-r}} \\
\leq & C(T)\epsilon(1+\delta^{-r})(1+\frac{1}{1-e^{-c\frac{\delta}{\epsilon}}}).
\end{aligned}$$

□

## 1.4. Proof of the weak-order result

The proof of Theorem 1.2 relies on an expansion of the solution of the Kolmogorov equation associated with the stochastic system (1.2) with respect to the small parameter $\epsilon$; the zero-order term corresponds to the averaged equation, and we control the first-order term to get the result.

When working with SDEs, this strategy can be entirely followed; nevertheless in the case of SPDEs, the Kolmogorov equations involve the unbounded operators $A$ and $B$, and quantities like $\mathrm{Tr}(D^2_{yy}u)$, and this leads to technical problems - see [9] or [12].

The idea of the proof is to reduce the infinite dimensional problem to a finite dimensional one by Galerkin approximation; we apply the method in this finite dimensional setting, and we prove bounds that are uniform with respect to the dimension. We also show that taking the limit when dimension goes to infinity is meaningful and gives the desired result.

The key element in the construction of the expansion mentioned above is given in Lemma 1.25 below: a Poisson equation can be solved under ergodicity conditions on the fast equation. We notice that Assumption 1.11 is sufficient, since we can analyse the problem through the asymptotic properties of the transition semi-group of the fast equation, instead of trajectories.

First, we explain how we reduce the problem to a finite dimensional one - see Section 1.4.1; then we explain the method in this setting and show which expressions must be controlled - see Section 1.4.2; finally we prove the estimates.

1.4.1. **Reduction to a finite dimensional problem.** We use Galerkin approximations based on the orthonormal basis $(e_k)$ and $(f_k)$ of $H$, given by Assumption 1.3. We define the subspaces

$$H_N^{(1)} = \text{span}\,\{e_k; 0 \le k \le N-1\} \quad \text{and} \quad H_N^{(2)} = \text{span}\,\{f_k; 0 \le k \le N-1\}\,.$$

We denote by $P_N^{(1)} \in \mathcal{L}(H)$ - resp. $P_N^{(2)}$ - the orthogonal projection of $H$ onto $H_N^{(1)}$ - resp. $H_N^{(2)}$.

Then we define for $x \in H_N^{(1)}$ and $y \in H_N^{(2)}$

$$\begin{aligned} F_N(x,y) &= P_N^{(1)}(F(x,y)) \\ G_N(x,y) &= P_N^{(2)}(G(x,y)) \\ U_N(x,y) &= U(x,y). \end{aligned}$$

On $H_N^{(1)} \times H_N^{(2)}$ coefficients $F_N$ and $G_N$ are of class $\mathcal{C}^2$. The function $U_N$ is of class $\mathcal{C}^3$ with respect to $y$ and of class $\mathcal{C}^2$ with respect to $x$, and we have $D_y U_N(x,y) = G_N(x,y)$ for any $(x,y) \in H_N^{(1)} \times H_N^{(2)}$.

Moreover we have bounds on $F_N$, $G_N$, $U_N$ and on their derivatives which are uniform with respect to $N$ and satisfy bound like in Assumption 1.7. In particular we still have the weak dissipativity condition with $G$ replaced by $G_N$.

We can define the following approximation of system (1.2)

$$\begin{aligned} dX_N^\epsilon(t) &= (AX_N^\epsilon(t) + F_N(X_N^\epsilon(t), Y_N^\epsilon(t)))dt \\ dY_N^\epsilon(t) &= \frac{1}{\epsilon}(BY_N^\epsilon(t) + G_N(X_N^\epsilon(t), Y_N^\epsilon(t)))dt + \frac{1}{\sqrt{\epsilon}} dW_N(t), \end{aligned} \tag{1.21}$$

with initial conditions $X_N^\epsilon(0) = P_N^{(1)} x \in H_N^{(1)}$, $Y_N^\epsilon(0) = P_N^{(2)} y \in H_N^{(2)}$.

Since $H_N^{(1)}$ (resp. $H_N^{(2)}$) is a stable subspace of $A$ (resp. $B$), this system is well-posed in $H_N^{(1)} \times H_N^{(2)}$.

We have by definition $W_N(t) = P_N^{(2)} W(t)$; on $H_N^{(2)}$ it is a $N$-dimensional Brownian motion.

Below we explain that system (1.21) defines a good approximation of the initial problem (1.2). Moreover, we can check that the structure of the problem remains the same, with bounds independent of the dimension.

First we describe the ergodic properties, and in particular the relations between the invariant measures associated with the associated fast equations with frozen slow component. In Section 1.2.2, we have defined $\nu = \mathcal{N}(0, \frac{(-B)^{-1}}{2})$ and $\mu^x$ - see (1.11); we can do the same for the finite dimensional fast equation with frozen slow component $x \in H$ - and not only for $x \in H_N^{(1)}$:

$$\begin{aligned} dY_{x,N}(t,y) &= (BY_{x,N}(t,y) + P_N^{(2)} G(x, P_N^{(2)} Y_{x,N}(t,y)))dt + dW_N(t) \\ Y_{x,N}(0,y) &= P_N^{(2)} y. \end{aligned} \tag{1.22}$$

Let $\nu_N$ be the unique centered Gaussian probability measure on $H_N^{(2)}$ having for covariance operator the induced matrix from $\frac{(B)^{-1}}{2}$ on the subspace $H_N^{(2)}$. We can then build $\mu_N^x$ the unique - since we have weak dissipativity - invariant probability measure associated to (1.22): we naturally extend the definition of $U_N$ to $H \times H_N^{(2)}$, by $U_N(x,y) = U(x,y)$, and we define

$$\mu_N^x(dy) = \frac{1}{Z_N(x)} e^{2U_N(x,y)} \nu_N(dy), \tag{1.23}$$

where $Z_N(x) \in ]0, +\infty[$ is a normalization constant.

As $\nu_N$ is the image measure of $\nu$ by $P_N^{(2)}$, we can use the following change of variables formula for suitable test functions $\Phi$:

$$\begin{aligned} \int_{H_N^{(2)}} \Phi(y) \mu_N^x(dy) &= \frac{1}{Z_N(x)} \int_{H_N^{(2)}} \Phi(y) e^{2U(x,y)} \nu_N(dy) \\ &= \frac{1}{Z_N(x)} \int_H \Phi(P_N^{(2)} y) e^{2U(x, P_N^{(2)} y)} \nu(dy). \end{aligned}$$

This formula leads to convergence properties: first when $N \to +\infty$ we have

$$Z_N(x) \to Z(x) \text{ for any } x \in H.$$

Moreover if $\Phi : H \to H$ is a continuous function such that for any $y \in H$, $|\Phi(y)| \leq c(1+|y|)$, then

$$\int_{H_N^{(2)}} \Phi(y)\mu_N^x(dy) \to \int_H \Phi(y)\mu^x(dy).$$

For any $x \in H$ we define the averaged coefficient associated with the finite dimensional problem (1.22)

$$\overline{F_N}(x) = \int_{H_N^{(2)}} P_N^{(1)} F(x,y)\mu_N^x(dy), \tag{1.24}$$

and the new averaged equation

$$d\overline{X_N}(t) = (A\overline{X_N}(t) + \overline{F_N}(\overline{X_N}(t)))dt, \tag{1.25}$$

with initial condition $\overline{X_N}(0) = P_N^{(1)}x \in H_N^{(1)}$.

We notice that $\overline{F_N}$ is bounded, and of class $\mathcal{C}^2$ with bounded derivatives; moreover it satisfies the properties described in Proposition 1.18, with constants independent of $N$.

**Remark 1.23.** *Making the Galerkin projection and then averaging the coefficient with $\mu_N^x$ is not the same as averaging the coefficient with $\mu^x$ and then making the Galerkin projection. As a consequence $\overline{X_N}$ is not naturally defined by a Galerkin approximation from $\overline{X}$.*

The following Lemma gives the convergence of the finite dimensional approximations to the initial problem:

**Lemma 1.24.** (1) *For any fixed $\epsilon > 0$, $t \geq 0$, and any $x \in H$, $y \in H$, we have when $N \to +\infty$*

$$\mathbb{E}|X^\epsilon(t) - X_N^\epsilon(t)|^2 + \mathbb{E}|Y^\epsilon(t) - Y_N^\epsilon(t)|^2 \to 0.$$

(2) *For any $t \geq 0$, $x \in H$, we have when $N \to +\infty$*

$$|\overline{X}(t) - \overline{X_N}(t)| \to 0.$$

It remains to define an approximated test function: for any $x \in H_N^{(1)}$, we define $\phi_N(x) = \phi(x)$.

We can now show how the initial problem can be reduced to a finite dimensional one - provided we are able to give estimates uniform to the dimension: for any initial conditions $x, y \in H$ and $0 \leq t \leq T$,

$$\begin{aligned}\mathbb{E}[\phi(X^\epsilon(t))] - \mathbb{E}[\phi(\overline{X}(t))] &= \mathbb{E}[\phi(X^\epsilon(t))] - \mathbb{E}[\phi(X_N^\epsilon(t))] \\ &+ \mathbb{E}[\phi_N(X_N^\epsilon(t))] - \mathbb{E}[\phi_N(\overline{X_N}(t))] \\ &+ \mathbb{E}[\phi(\overline{X_N}(t))] - \mathbb{E}[\phi(\overline{X}(t))].\end{aligned}$$

The first and the last terms converge to 0 when $N \to +\infty$, according to the above Lemma 1.24; we later control the central term with an expression independent of $N$. Taking the limit as dimension goes to infinity then gives the result. So we have to control

$$\mathbb{E}[\phi_N(X_N^\epsilon(t))] - \mathbb{E}[\phi_N(\overline{X_N}(t))]. \tag{1.26}$$

From now on, we only work with the approximations to obtain estimates, but in order to simplify the notations, we forget the index $N$ - since bounds are uniform with respect to $N$. The spaces $H_N^{(i)}$ are denoted by $H^{(i)}$ in the next sections.

1.4.2. **The asymptotic expansion.** We define the following differential operators: for functions of class $\mathcal{C}^2$ $\psi : H^{(1)} \times H^{(2)} \to \mathbb{R}$, for any $(x,y) \in H^{(1)} \times H^{(2)}$

$$\begin{aligned}L_1\psi(x,y) &=< By + G(x,y), D_y\psi(x,y) > + \frac{1}{2}\mathrm{Tr}(D_{yy}^2\psi(x,y)) \\ L_2\psi(x,y) &=< Ax + F(x,y), D_x\psi(x,y) > \\ L^\epsilon &= \frac{1}{\epsilon}L_1 + L_2.\end{aligned}$$

We also define for $\psi : H^{(1)} \to \mathbb{R}$ of class $\mathcal{C}^1$ $\overline{L}\psi(x) =< Ax + \overline{F}(x), D_x\psi(x) >$.

We define the following functions $u^\epsilon$ and $\overline{u}$: for $x \in H^{(1)}$, $y \in H^{(2)}$ and $t \geq 0$

$$\begin{aligned}u^\epsilon(t,x,y) &= \mathbb{E}[\phi(X^\epsilon(t,x,y))] \\ \overline{u}(t,x) &= \phi(\overline{X}(t,x)),\end{aligned} \tag{1.27}$$

where we have mentioned explicitly the dependence on the initial conditions $x, y$ in $X^\epsilon$ and $\overline{X}$.

Since the test function $\phi$ is of class $\mathcal{C}_b^2$, $u^\epsilon$ and $\overline{u}$ are of class $\mathcal{C}^1$ with respect to $t$ and of class $\mathcal{C}_b^2$ with respect to $x, y$; we also know that $u^\epsilon$ and $\overline{u}$ are solutions of the following Kolmogorov equations:

$$\begin{aligned} \frac{\partial u^\epsilon}{\partial t}(t,x) &= L^\epsilon u^\epsilon(t,x) \\ u^\epsilon(0,x) &= \phi(x); \end{aligned} \tag{1.28}$$

$$\begin{aligned} \frac{\partial \overline{u}}{\partial t}(t,x,y) &= \overline{L}\overline{u}(t,x,y) \\ \overline{u}(0,x,y) &= \phi(x). \end{aligned} \tag{1.29}$$

We remark that the second equation is a linear transport equation with no diffusion term. We then rewrite the expression we want to study (see (1.26)):

$$\mathbb{E}[\phi(X^\epsilon(T,x,y))] - \mathbb{E}[\phi(\overline{X}(T,x))] = u^\epsilon(T,x,y) - \overline{u}(T,x). \tag{1.30}$$

Our strategy is to look for an expansion of $u^\epsilon$ with respect to the small parameter $\epsilon$:

$$u^\epsilon = u_0 + \epsilon u_1 + v^\epsilon, \tag{1.31}$$

where $v^\epsilon$ is a residual term, while $u_0$ and $u_1$ are smooth and are constructed below.

The identification with respect to the powers of $\epsilon$ gives the following equations:

$$L_1 u_0 = 0, \tag{1.32}$$

$$\frac{\partial u_0}{\partial t} = L_1 u_1 + L_2 u_0. \tag{1.33}$$

The operator $L_1$ satisfies the following property on the solutions of Poisson equations:

**Lemma 1.25.** *We fix $x \in H$.*

- *If $\Psi$ is a bounded continuous function such that $\int_H \Psi(y)\mu^x(dy) = 0$, then if $\Phi$ is a function of class $\mathcal{C}^2$ satisfying $L_1\Phi = -\Psi$ then for any $y \in H$ we have*
$$\Phi(y) = \int_H \Phi\mu^x + \int_0^{+\infty} \mathbb{E}[\Psi(Y_x(s,y))]ds.$$
- *Moreover if $\Psi$ is of class $\mathcal{C}_b^2$, then $\Phi$ defined by*
$$\Phi(y) = \int_0^{+\infty} \mathbb{E}[\Psi(Y_x(s,y))]ds$$
*is of class $\mathcal{C}^2$, satisfies $L_1\Phi = -\Psi$, and there exists a constant $C$ - independent on $N$ - such that for any $y \in H$ we have*
$$|\Phi(y)| \leq C(1+|y|^2)\|\Psi\|_\infty.$$

Proof The first part of the Lemma is an easy consequence of Itô formula and of equation (1.10), after integration with respect to $y_2$ under $\mu^x$. To prove the second part of the lemma, we first see that for any fixed $s \in \mathbb{R}^+$ the function $y \mapsto \mathbb{E}[\Psi(Y_x(s,y))] =: v_x(s,y)$ is of class $\mathcal{C}^2$. To be able to exchange integration in $s$ and derivation with respect to $y$, we need to prove an estimate of the first and the second derivatives which is integrable with respect to $s$. The derivatives of $Y_x(s,y)$ with respect to the initial condition satisfy the following equations - to simplify notations we do not write dependence in $x$ in those derivatives:

$$\begin{aligned} \frac{d\tilde{\eta}^{h,y}(s)}{ds} &= B\tilde{\eta}^{h,y}(s) + D_y G(x, Y_x(s,y)).\tilde{\eta}^{h,y}(s,y) \\ \tilde{\eta}^{h,y}(0) &= h, \end{aligned}$$

and

$$\begin{aligned} \frac{d\tilde{\zeta}^{h,k,y}(s)}{ds} &= B\tilde{\zeta}^{h,k,y}(s) + D_y G(x, Y_x(s,y)).\tilde{\zeta}^{h,k,y}(s) + D^2_{yy} G(x, Y_x(s,y)).(\tilde{\eta}^{h,y}(s), \tilde{\eta}^{k,y}(s)) \\ \tilde{\zeta}^{h,k,y}(0) &= 0. \end{aligned}$$

Without any further dissipativity assumption than (WD), we only get bounds on finite time intervals like $[0,1]$: there exists $C>0$ such that for any $y,h,k\in H$ and $0\le s\le 1$

$$|D_y v_x(s,y).h|\le C|h|$$
$$|D^2_{yy} v_x(s,y).(h,k)|\le C|h||k|.$$

However, using the estimate (1.10) and a Bismut-Elworthy-Li formula, we can indeed prove some exponential convergence with respect to $s$ of the derivatives $D_y v_x(s,y)$ and $D^2_{yy}v_x(s,y)$.

Let $\Psi^0$ be a function such that $|\Psi^0(y)|\le C(\Psi^0)(1+|y|^2)$ for any $y\in H$. If we define $v^0_x(s,y):=\mathbb{E}\Psi^0(Y_x(s,y))$ for any $y\in H$, $h,k\in H$, we get for the first order derivative

$$\begin{aligned} D_y v^0_x(s,y).h &= \frac{1}{s}\mathbb{E}[\int_0^s <\tilde\eta^{h,y}(\sigma), dW(\sigma)>\Psi^0(Y_x(s,y))] \\ &= \frac{2}{s}\mathbb{E}[\int_0^{s/2} <\tilde\eta^{h,y}(\sigma), dW(\sigma)> v^0_x(s/2, Y_x(s/2,y))], \end{aligned} \tag{1.34}$$

with the observation that $v^0_x(s,y)=\mathbb{E}v^0_x(s/2,Y_x(s/2,y))$ thanks to the Markov property; the second order derivative satisfies

$$\begin{aligned} D^2_{yy}v^0_x(s,y).(h,k) &= \frac{2}{s}\mathbb{E}[\int_0^{s/2}\tilde\zeta^{h,k,y}(\sigma), dW(\sigma)> v^0_x(s/2,Y(s/2,y))] \\ &+\frac{2}{s}\mathbb{E}[\int_0^{s/2} <\tilde\eta^{h,y}(\sigma), dW(\sigma)> D_y v^0_x(s/2,Y(s/2)).\tilde\eta^{k,y}(s/2)]. \end{aligned} \tag{1.35}$$

Since $Y_x$ can be controlled in a $L^2$ norm according to Proposition 1.31 in the appendix, we see that there exists $C>0$ such that for any $0<s\le 1$, $y\in H$, $h,k\in H$

$$\begin{aligned} |D_y v^0_x(s,y).h| &\le \frac{C}{\sqrt{s}}C(\Psi^0)(1+|y|^2)|h|, \\ |D^2_{yy}v^0_x(s,y).(h,k)|| &\le \frac{C}{s}C(\Psi^0)(1+|y|^2)|h||k|. \end{aligned} \tag{1.36}$$

Now when $s\ge 1$ the Markov property implies that $v_x(s,y)=\mathbb{E}v_x(s-1,Y_x(1,y))$, and choosing $y_1=y$ and by integrating with respect to $\mu^x(dy_2)$ in (1.10) we have

$$|v_x(s-1,y)|\le Ce^{-c(s-1)}(1+|y|^2).$$

By (1.36) at time 1, we obtain for $s\ge 1$

$$|D_y v_x(s,y).h|\le Ce^{-c(s-1)}(1+|y|^2)|h|$$
$$|D^2_{yy}v_x(s,y).(h,k)|\le Ce^{-c(s-1)}(1+|y|^2)|h||k|.$$

Moreover we have a uniform control when $0\le s\le 1$, so that with a change of constants we get the result.

□

As a consequence we see from (1.32) that $u_0$ is independent of $y$; we then write $u_0(t,x,y)=u_0(t,x)$. We also choose the initial condition $u_0(0,x)=\phi(x)$. The second equation (1.33) then yields

$$\begin{aligned} \frac{\partial u_0}{\partial t}(t,x) &= \int_{H^{(2)}}\frac{\partial u_0}{\partial t}(t,x)\mu^x(dy) \\ &= \int_{H^{(2)}} L_1u_1(t,x,y)\mu^x(dy)+\int_{H^{(2)}}L_2u_0(t,x)\mu^x(dy) \\ &= < Au_0(t,x)+\int_{H^{(2)}}F(x,y)\mu^x(dy), D_xu_0(t,x)> \\ &= \overline{L}u_0(t,x). \end{aligned}$$

$u_0$ and $\overline{u}$ are solutions of the same evolution equation, with the same initial condition; we can then conclude that $u_0=\overline{u}$.

Then the second equation can be transformed: $\overline{L}u_0 = L_1u_1 + L_2u_0$. We then obtain an equation on $u_1$:

$$L_1u_1(t,x,y) =< \overline{F}(x) - F(x,y), D_xu_0(t,x) >=: -\chi(t,x,y), \tag{1.37}$$

where $\chi$ is of class $\mathcal{C}_b^2$ with respect to $y$, and satisfies for any $t \geq 0$ and $x \in H^{(1)}$

$$\int_{H^{(2)}} \chi(t,x,y)\mu^x(dy) = 0.$$

Thanks to Lemma 1.25 above, we thus obtain the following solution to equation (1.37)

$$u_1(t,x,y) = \int_0^{+\infty} \mathbb{E}[\chi(t,x,Y_x(s,y))]ds. \tag{1.38}$$

Moreover we are able to show regularity of $u_1$ with respect to $t$ and $x, y$.

The remainder $v^\epsilon = u^\epsilon - u_0 - \epsilon u_1$ satisfies

$$(\partial_t - \frac{1}{\epsilon}L_1 - L_2)v^\epsilon = \epsilon(L_2u_1 - \frac{\partial u_1}{\partial t}). \tag{1.39}$$

Due to non-integrability in 0 of some bounds below, we introduce a parameter $\rho(\epsilon) = \epsilon^{1/\theta} \leq \epsilon$ (since $0 < \theta \leq 1$); it satisfies $\rho(\epsilon) \to 0$ when $\epsilon \to 0$.

Using a variation of constant formula, we obtain

$$\begin{aligned} v^\epsilon(T,x,y) = \mathbb{E}[v^\epsilon(\rho(\epsilon), X^\epsilon(T-\rho(\epsilon),x,y), Y^\epsilon(T-\rho(\epsilon),x,y))] \\ + \epsilon\mathbb{E}[\int_{\rho(\epsilon)}^T (L_2u_1 - \frac{\partial u_1}{\partial t})(t, X^\epsilon(T-t,x,y), Y^\epsilon(T-t,x,y))dt] \end{aligned}$$

By (1.31), and since $u_0 = \overline{u}$, we then have

$$\begin{aligned} u^\epsilon(T,x,y) - \overline{u}(T,x,y) &= \epsilon u_1(T,x,y) \\ &+ \mathbb{E}[v^\epsilon(\rho(\epsilon), X^\epsilon(T-\rho(\epsilon),x,y), Y^\epsilon(T-\rho(\epsilon),x,y))] \\ &+ \epsilon\mathbb{E}[\int_{\rho(\epsilon)}^T (L_2u_1 - \frac{\partial u_1}{\partial t})(t, X^\epsilon(T-t,x,y), Y^\epsilon(T-t,x,y))dt]. \end{aligned} \tag{1.40}$$

The following estimates are proved below:

**Lemma 1.26.** *There exists a constant $C$ such that for any $0 < t \leq T$, $x, y \in H$,*

$$\begin{aligned} |u_1(t,x,y)| &\leq C(1+|x|+|y|) \\ |\frac{\partial u_1}{\partial t}(t,x,y)| &\leq C(1+\frac{1}{t})(1+|x|+|y|)^2 \\ |L_2u_1(t,x,y)| &\leq C(1+|x|+|y|)(1+|Ax|). \end{aligned}$$

Using estimates on $X^\epsilon$ and $Y^\epsilon$ proved in the appendix - see Propositions 1.31 and 1.35 - the first and the last expressions of (1.40) are bounded by

$$C\epsilon(1+|x|+|y|) + C\epsilon^{1-r/2}(1+|\log(\rho(\epsilon))|)(1+|x|_{(-A)^\theta}+|y|)^2,$$

which is dominated by $C\epsilon^{1-r}$ with the choice of $\rho(\epsilon)$ given above.

We notice that the Assumption $\theta > 0$ is essential to control the part involving $|Ax|$.

We now explain how the central term of (1.40) is controlled; for that we estimate for any $x, y \in H$

$$\begin{aligned} v^\epsilon(\rho(\epsilon),x,y) &= u^\epsilon(\rho(\epsilon),x,y) - u_0(\rho(\epsilon),x) - \epsilon u_1(\rho(\epsilon),x,y) \\ &= -\epsilon u_1(\rho(\epsilon),x,y) \\ &+ [u^\epsilon(\rho(\epsilon),x,y) - u^\epsilon(0,x,y)] - [u_0(\rho(\epsilon),x) - u_0(0,x)], \end{aligned} \tag{1.41}$$

since the initial condition $\phi$ is the same for $u^\epsilon$ and $\overline{u}$.

Using Lemma 1.26, the first term above is easily controlled by $C\epsilon(1+|x|+|y|)$. We now use another method to control the two other terms.

First, we use the definition (1.27) of $\overline{u} = u_0$ to write

$$\begin{aligned}|u_0(\rho(\epsilon),x) - u_0(0,x)| &= |\int_0^{\rho(\epsilon)} \frac{\partial}{\partial t} u_0(t,x)dt| \\ &= |\int_0^{\rho(\epsilon)} \frac{\partial}{\partial t}\phi(\overline{X}(t,x))dt| \\ &= |\int_0^{\rho(\epsilon)} D\phi(\overline{X}(t,x)).\frac{d}{dt}\overline{X}(t,x)dt| \\ &\leq C\int_0^{\rho(\epsilon)} |\frac{d}{dt}\overline{X}(t,x)|dt.\end{aligned}$$

By definition of $\overline{X}$ (see (1.3)), and using Proposition 1.38, we get for any $t > 0$

$$|\frac{d}{dt}\overline{X}(t,x)| \leq C(1 + t^{\theta-1})(1 + |x|_{(-A)^\theta}).$$

As a consequence, since $\theta > 0$, we get

$$|u_0(\rho(\epsilon),x) - u_0(0,x)| \leq C(\rho(\epsilon) + \frac{\rho(\epsilon)^\theta}{\theta})(1 + |x|_{(-A)^\theta}).$$

The other expression is controlled in the same way; it is important to notice that the assumption that $\phi$ only depends on the slow variable $x$ is fundamental in this estimate.

$$\begin{aligned}|u^\epsilon(\rho(\epsilon),x,y) - u^\epsilon(0,x,y)| &= |\mathbb{E}\phi(X^\epsilon(\rho(\epsilon),x,y)) - \phi(x)| \\ &\leq C\mathbb{E}|X^\epsilon(\rho(\epsilon),x,y) - x| \\ &\leq C\int_0^{\rho(\epsilon)} \mathbb{E}|AX^\epsilon(t,x,y) + F(X^\epsilon(t,x,y), Y^\epsilon(t,x,y))|.\end{aligned}$$

We now use the estimate on $\mathbb{E}|AX^\epsilon(t,x,y)|$ of Proposition 1.35 in the appendix and the boundedness of $F$, and we obtain

$$|u^\epsilon(\rho(\epsilon),x,y) - u^\epsilon(0,x,y)| \leq C(\rho(\epsilon) + \frac{\rho(\epsilon)^\theta}{\theta} + \epsilon^{-r/2}\rho(\epsilon))(1 + |x|_{(-A)^\theta} + |y|).$$

Then by (1.41), and using $\rho(\epsilon) = \epsilon^{1/\theta} \leq \epsilon$, we get

$$|v^\epsilon(\rho(\epsilon),x,y)| \leq C\epsilon^{1-r/2}(1 + |x|_{(-A)^\theta} + |y|).$$

Then thanks to (1.40) and to Proposition 1.32, we get for $\epsilon \leq 1$

$$|u^\epsilon(T,x,y) - \overline{u}(T,x,y)| \leq C\epsilon^{1-r}.$$

As explained at the end of Section 1.4.1, we have indeed proved a bound on (1.26). It is now enough to notice that the above constant $C$ is independent of dimension $N$, and to let $N$ go to $+\infty$, and Theorem 1.2 follows.

## 1.5. Proof of Lemma 1.26

We use results gathered in the appendix 1.6 and 1.7.

1.5.1. **Estimate of $u_1$.** Since $u_1$ is defined by (1.38), by using Lemma 1.25 we have

$$|u_1(t,x,y)| \leq C(1 + |y|^2)\|y \mapsto \chi(t,x,y)\|_\infty.$$

According to (1.37), we indeed have for any $y \in H^{(2)}$

$$\chi(t,x,y) =< F(x,y) - \overline{F}(x), D_x u_0(t,x) >,$$

and therefore we just have to bound $|D_x u_0(t,x)|$, thanks to the following lemma:

**Lemma 1.27.** *For any $T \in ]0,+\infty[$, there exists $C_0 > 0$ such that for any $0 \leq t \leq T$ and $x \in H^{(1)}$*

$$|D_x u_0(t,x)|_H \leq C_T \sup_{z\in H} |D\phi(z)|_H.$$

Proof $u_0$ is the solution of the equation

$$\frac{\partial u_0}{\partial t}(t,x) =< Ax + \overline{F}(x), D_x u_0(t,x) > \tag{1.42}$$
$$u_0(0,x) = \phi(x).$$

We have a representation formula $u_0(t,x) = \phi(\overline{X}(t,x))$, where $\overline{X}$ is solution of (1.3).
We can differentiate (1.3) with respect to the initial condition $x$, and we have for any $h \in H^{(1)}$

$$D_x u_0(t,x).h = D\phi(\overline{X}(t,x)).\eta^h(t,x),$$

where $\eta^h(t,x)$ is the derivative of $\overline{X}$ with respect to $x$ in direction $h$, and is solution of the variational equation (1.47) in the appendix.

Using Proposition 1.39, we get $|D_x u_0(t,x).h| \leq C_T \sup_{z\in H} |D\phi(z)||h|$, and taking the supremum over $h$ gives the result.

□

Therefore, we obtain the first estimate of Lemma 1.26.

1.5.2. **Estimate of $\frac{\partial u_1}{\partial t}$.** First we check that

$$\int_{H^{(2)}} \frac{\partial \chi}{\partial t}(t,x,z)\mu^x(dz) = \frac{\partial}{\partial t}\int_{H^{(2)}} \chi(t,x,z)\mu^x(dz) = 0.$$

By definition (1.38) of $u_1$, it is easy to show that we can differentiate with respect to $t$, and that

$$\frac{\partial u_1}{\partial t}(t,x,y) = \int_0^{+\infty} \mathbb{E}[\frac{\partial \chi}{\partial t}(t,x,Y_x(s,y))]ds. \tag{1.43}$$

We then obtain

$$|\frac{\partial u_1}{\partial t}(t,x,y)| \leq C(1+|y|^2)\|y \mapsto \frac{\partial \chi}{\partial t}(t,x,y)\|_\infty.$$

Since by (1.37) we have

$$\frac{\partial \chi}{\partial t}(t,x,y) =< F(x,y) - \overline{F}(x), \frac{\partial}{\partial t} D_x u_0(t,x) >,$$

we just need to control $|\frac{\partial}{\partial t} D_x u_0(t,x)|$:

**Lemma 1.28.** *For any $T > 0$, there exists $C_T > 0$ such that for any $0 < t \leq T$, $x \in H^{(1)}$ and $h \in H^{(1)}$ we have*

$$|\frac{\partial}{\partial t} D_x u_0(t,x).h| \leq C(1+|x|_H)(1+t^{-1})|h|.$$

Proof For any $h \in H^{(1)}$, we have

$$\frac{\partial}{\partial t}(D_x u_0(t,x).h) = D^2\phi(\overline{X}(t,x))(\eta^h(t,x), \frac{d}{dt}\overline{X}(t,x))$$
$$+ D\phi(\overline{X}(t,x)).\frac{d}{dt}\eta^h(t,x).$$

(1) Thanks to Proposition 1.36, we have $|\eta^h(t,x)| \leq C(1+|x|)$ for any $t \geq 0$.
Moreover $\frac{d}{dt}\overline{X}(t,x) = A\overline{X}(t,x) + \overline{F}(\overline{X}(t,x))$.
On the one hand, $\overline{F}$ is bounded; on the other hand, thanks to Proposition 1.38 we have

$$|A\overline{X}(t,x)|_H \leq C_\theta(1+t^{-1})(1+|x|_H).$$

Therefore

$$|\frac{d}{dt}\overline{X}(t,x)|_H \leq C(1+t^{-1})(1+|x|_H).$$

(2) It remains to control

$$|\frac{d}{dt}\eta^h(t,x)| = |A\eta^h(t,x) + D\overline{F}(\overline{X}(t,x)).\eta^h(t,x)|.$$

Since $\overline{F}$ is Lipschitz continuous, and using Proposition 1.36, we get an estimate of the second term.

Moreover Proposition 1.41 gives

$$|A\eta^h(t,x)| \leq C(t^{-1}+1)(1+|x|_H)|h|.$$

Therefore

$$|\frac{d}{dt}\eta^h(t,x)| \leq C(t^{-1}+1)(1+|x|_H)|h|.$$

(3) We then have for any $h \in H^{(1)}$

$$|\frac{\partial}{\partial t}(D_x u_0(t,x).h)| \leq C(t^{-1}+1)(1+|x|_H)|h|.$$

□

We then obtain the second estimate of Lemma 1.26.

1.5.3. **Estimate of $L_2u_1$.** To prove Lemma 1.26, it remains to control the part involving $L_2u_1$.

By definition of $L_2$, we have

$$L_2u_1(t,x,y) =< Ax + F(x,y), D_x u_1(t,x,y) > . \tag{1.44}$$

Therefore we have to estimate $|D_x u_1(t,x,y)|$. We explain how $D_x u_1(t,x,y).h$ can be calculated for any $h \in H^{(1)}$.

Recall that $u_1$ defined by (1.38) satisfies

$$L_1u_1(t,x,y) =< \overline{F}(x) - F(x,y), D_x u_0(t,x) >= -\chi(t,x,y),$$

where we explicitly write the dependence of the operator $L_1$ in the two variables $x$ and $y$.

We fix $t \geq 0$, $x \in H^{(1)}$, $y \in H^{(2)}$, and $h \in H^{(1)}$. Then for any $\xi \neq 0$ we have

$$\begin{aligned} L_1(x,y)\frac{u_1(t,x+\xi h,y) - u_1(t,x,y)}{\xi} &= -\frac{\chi(t,x+\xi h,y) - \chi(t,s,y)}{\xi} \\ &\quad - < \frac{G(x+\xi h,y) - G(x,y)}{\xi}, D_y u_1(t,x+\xi h,y) > \\ &=: -\Gamma(t,x,y,h,\xi), \end{aligned}$$

where $\Gamma$ is regular with respect to $y$; therefore by using Lemma 1.25 we get

$$\begin{aligned} \frac{u_1(t,x+\xi h,y) - u_1(t,x,y)}{\xi} - \int_{H^{(2)}} \frac{u_1(t,x+\xi h,y) - u_1(t,x,y)}{\xi}\mu^x(dy) \\ = \int_0^{+\infty} \mathbb{E}[\Gamma(t,x,Y_x(s,y),h,\xi)]ds. \end{aligned}$$

We want to take the limit when $\xi \to 0$, in order to prove that we can differentiate, and to obtain an expression that we are able to control.

First, we notice that for any $t,x,y$ we have $\int_{H^{(2)}} u_1(t,x,y)\mu^x(dy) = 0$; so we can write that

$$\int_{H^{(2)}} \frac{u_1(t,x+\xi h,y) - u_1(t,x,y)}{\xi}\mu^x(dy) = -\int_{H^{(2)}} u_1(t,x+\xi h,y)\frac{V(x+\xi h,y) - V(x,y)}{\xi}\nu(dy),$$

where $V(x,y) := \frac{1}{Z(x)}e^{2U(x,y)}$ (so that we have $\mu^x(dy) = V(x,y)\nu(dy)$).

When $\xi \to 0$, we obtain

$$\begin{aligned} \int_{H^{(2)}} \frac{u_1(t,x+\xi h,y) - u_1(t,x,y)}{\xi}\mu^x(dy) &\to \int_{H^{(2)}} u_1(t,x,y)D_xV(x,y).h\nu(dy) \\ &= \int_{H^{(2)}} u_1(t,x,y)H(x,y).hV(x,y)\nu(dy), \end{aligned}$$

where $H(x,y) = 2D_xU(x,y) - 2\int_{H^{(2)}} D_xU(x,z)\mu^x(dz)$.

Moreover $|\int_{H^{(2)}} u_1(t,x,y)H(x,y).hV(x,y)\nu(dy)| \leq C(1+|x|)|h|$.
Second, we look at the part involving $\Gamma$: we notice that when $\xi \to 0$,

$$\Gamma(t,x,y,h,\xi) \to \Theta(t,x,y).h,$$

where

$$\Theta(t,x,y).h = D_x\chi(t,x,y).h + < D_xG(x,y).h, D_yu_1(t,x,y) > . \tag{1.45}$$

Below, we prove the following estimate on the function $\Theta$:

**Lemma 1.29.** *There exists a constant $C$ such that for any $x \in H^{(1)}$, $t \geq 0$, $h \in H^{(1)}$ we have for any $y \in H^{(2)}$*

$$|\Theta(t,x,y).h| \leq C(1+|y|^2)|h|.$$

We notice that for any $t,x,\xi,h$ we have by definition of $\Gamma$ $\int_{H^{(2)}} \Gamma(t,x,y,\xi,h)\mu^x(dy) = 0$; then using the bound of the previous Lemma and the dominated convergence Theorem we obtain $\int_{H^{(2)}} \Theta(t,x,y).h\mu^x(dy) = 0$ for any $x \in H^{(1)}$, $t \geq 0$, $h \in H^{(1)}$. Using this result, Proposition 1.15 - with integration with respect to $\mu^x(dy_2)$ - and the estimate in the previous Lemma, we then see that $u_1$ can be differentiated with respect to $x$, and that the following formula holds:

$$D_xu_1(t,x,y).h = \int_{H^{(2)}} u_1(t,x,y)H(x,y).hV(x,y)\nu(dy) + \int_0^{+\infty} \mathbb{E}[\Theta(t,x,Y_x(s,y)).h]ds; \tag{1.46}$$

According to Lemma 1.29, we do not know whether $\Theta$ is a bounded function, but we only know that it has quadratic growth. However, the result of Proposition 1.15 can easily be extended to such function.
Now we obtain that

$$|D_xu_1(t,x,y).h| \leq C(1+|y|^2)|h|$$

and therefore - see (1.44):

$$|L_2u_1(t,x,y)| \leq C(1+|y|^2)(1+|Ax|),$$

which is the third estimate of Lemma 1.26.
It remains to prove Lemma 1.29.
We fix $t \geq 0$, $x \in H^{(1)}$, $h \in H^{(1)}$, and $y, y' \in H^{(2)}$.

- On the one hand, $\chi$ being defined by (1.37), we have

$$D_x\chi(t,x,y).h = < D_xF(x,y).h, D_xu_0(t,x) > + D^2_{xx}u_0(t,x).(h,F(x,y)).$$

Using the boundedness of the first derivative of $F$, and Lemma 1.27, we easily have

$$| < D_xF(x,y).h, D_xu_0(t,x) > | \leq C|h|.$$

The other part can be controlled thanks to the following Lemma:

**Lemma 1.30.** *For any $0 \leq t \leq T$, $x \in H^{(1)}$, $h,k \in H^{(1)}$, we have*

$$|D^2_{xx}u_0(t,x).(h,k)| \leq C(T,\phi)|h|_H|k|_H.$$

Proof We have

$$u_0(t,x) = \phi(\overline{X}(t,x))$$
$$D_xu_0(t,x).h = D\phi(\overline{X}(t,x)).(D_x\overline{X}(t,x).h)$$

and

$$D^2_{xx}u_0(t,x).(h,k) = D^2\phi(\overline{X}(t,x))(D_x\overline{X}(t,x).h, D_x\overline{X}(t,x).k) + D\phi(\overline{X}(t,x)).(D^2_{xx}\overline{X}(t,x).(h,k)).$$

Using Proposition 1.39, we control $\eta^h(t,x) = D_x\overline{X}(t,x).h$; moreover we notice that the second derivative $\xi^{h,k}(t,x) := D^2_{xx}\overline{X}(t,x).(h,k)$ satisifies equation (1.48); using Proposition 1.42, we get the result.

□

Therefore $|D_x\chi(t,x,y).h| \leq C|h|$.

- On the other hand,

$$| < D_xG(x,y).h, D_yu_1(t,x,y) > | \leq C|h||D_yu_1(t,x,y)|,$$

But we have proved in Lemma 1.25 how to control the derivatives of $u_1$ with respect to $y$: we obtain

$$|D_yu_1(t,x,y).h| \leq C(1+|y|^2)|h|.$$

Therefore we have

$$| < D_xG(x,y).h, D_yu_1(t,x,y) > | \leq C(1+|y|^2)|h|,$$

and now the result is easily obtained:

$$|\Theta(t,x,y).h| \leq C(1+|y|^2)|h|.$$

## 1.6. Appendix:Properties of $(X^\epsilon, Y^\epsilon)$

The results of this section only require Assumptions 1.3, 1.7 and 1.9; in particular no dissipativity is assumed.

The first important property is the control of moments of any order:

**Proposition 1.31.** *For any $1 \leq p < +\infty$, there exists $c_p > 0$ such that for any $(x,y) \in H^2$, $t \geq 0$ and $\epsilon > 0$*

$$\mathbb{E}[|X^\epsilon(t)|_H^p] \leq c_p(1+e^{-\lambda t}|x|_H^p) \quad \text{and} \quad \mathbb{E}[|Y^\epsilon(t)|_H^p] \leq c_p(1+e^{-\mu t}|y|_H^p).$$

We can also give bounds on the moments with respect to $|\,.\,|_{(-A)^a}$ and $|\,.\,|_{(-B)^b}$ norms, for $0 < a < 1$ and $0 < b < 1/4$ (the case $a = 1$ is treated in Proposition 1.35 below).

**Proposition 1.32.** *For any $p \geq 1$, $a \in (0,1)$, $b \in (0,1/4)$, there exists $C_{p,a,b} > 0$ such that for any $x \in D(-A)^a$ and $y \in D(-B)^b$, we have:*

$$\mathbb{E}|X^\epsilon(t,x,y)|_{(-A)^a}^p \leq C_p(1+|x|_{(-A)^a}^p) \quad \text{and} \quad \mathbb{E}|Y^\epsilon(t,x,y)|_{(-B)^b}^p \leq C_p(1+|y|_{(-B)^b}^p).$$

We now give some regularity estimates of $X^\epsilon$ and $Y^\epsilon$ in the time variable. We do not assume any regularity assumption on $x$ or $y$; as a consequence, we obtain singularities at the origin, which are integrable.

**Proposition 1.33.** *For any $0 < r < 1$, there exists $C_r > 0$ such that for any $x, y \in H$, for any $0 < s \leq t$ and $\epsilon > 0$ we have*

$$\left(\mathbb{E}|X^\epsilon(t) - X^\epsilon(s)|_H^2\right)^{1/2} \leq C_r|t-s|^{1-r}(1+\frac{1}{s^{1-r}})(1+|x|_H).$$

Proof If we fix $0 < s \leq t$, $x, y \in H$, we have

$$\begin{aligned} X^\epsilon(t) - X^\epsilon(s) &= e^{tA}x - e^{sA}x \\ &\quad + \int_0^t e^{(t-\sigma)A}F(X^\epsilon(\sigma), Y^\epsilon(\sigma))d\sigma - \int_0^s e^{(s-\sigma)A}F(X^\epsilon(\sigma), Y^\epsilon(\sigma))d\sigma. \end{aligned}$$

For the first term, if $x = \sum_{k=0}^{+\infty} x_k e_k$, we can use Proposition 1.6 to get

$$|e^{tA}x - e^{sA}x|_H \leq C_r\frac{(t-s)^{1-r}}{s^{1-r}}|x|_H.$$

For the second term, we use the following decomposition:

$$\begin{aligned} \int_0^t e^{(t-\sigma)A}F(X^\epsilon(\sigma), Y^\epsilon(\sigma))d\sigma - \int_0^s e^{(s-\sigma)A}F(X^\epsilon(\sigma), Y^\epsilon(\sigma))d\sigma \\ = \int_s^t e^{(t-\sigma)A}F(X^\epsilon(\sigma), Y^\epsilon(\sigma))d\sigma \\ + \int_0^s (e^{(t-\sigma)A} - e^{(s-\sigma)A})F(X^\epsilon(\sigma), Y^\epsilon(\sigma))d\sigma. \end{aligned}$$

First, by the Cauchy-Schwarz inequality, we have

$$\begin{aligned}\mathbb{E}|\int_s^t e^{(t-\sigma)A}F(X^\epsilon(\sigma),Y^\epsilon(\sigma))d\sigma|_H^2 &\leq (t-s)\mathbb{E}\int_s^t |e^{(t-\sigma)A}F(X^\epsilon(\sigma),Y^\epsilon(\sigma))|_H d\sigma\\ &\leq C(t-s)^2,\end{aligned}$$

since $F$ is assumed to be bounded.

Second, we use the second inequality of Proposition 1.6 to control the last expression:

$$\begin{aligned}\mathbb{E}|\int_0^s e^{(s-\sigma)A}(e^{(t-s)A}-I)F(X^\epsilon(\sigma),Y^\epsilon(\sigma))d\sigma|_H^2&\\ &\leq \mathbb{E}[\int_0^s |\left(e^{(t-\sigma)A}-e^{(s-\sigma)A}\right)F(X^\epsilon(\sigma),Y^\epsilon(\sigma))d\sigma|_H]^2\\ &\leq C_r^2(t-s)^{2(1-r)}\mathbb{E}(\int_0^s \frac{e^{-\frac{\lambda}{2}(s-\sigma)}}{(s-\sigma)^{1-r}}|F(X^\epsilon(\sigma),Y^\epsilon(\sigma))|_H d\sigma)^2\\ &\leq C_r^2(t-s)^{2(1-r)},\end{aligned}$$

since $\int_0^{+\infty}\frac{e^{-\frac{\lambda}{2}s}}{s^{1-r}}ds<+\infty$.

□

**Proposition 1.34.** *For any $0<r<1/4$, there exists a constant $C_r$ such that if $x,y\in H$, then for any $0<s<t$ and $\epsilon>0$*

$$\mathbb{E}|Y^\epsilon(t)-Y^\epsilon(s)|^2\leq C(1+|x|_H^2+|y|_H^2)[\left(\frac{t-s}{s}\right)^{2r}+\left(\frac{t-s}{\epsilon}\right)^{2r}].$$

Proof

- For any $0<s<t$,

$$\begin{aligned}Y^\epsilon(t)-Y^\epsilon(s)=&(e^{\frac{t}{\epsilon}B}-e^{\frac{s}{\epsilon}B})y\\ &+\frac{1}{\epsilon}\int_s^t e^{\frac{(t-\sigma)}{\epsilon}B}G(X^\epsilon(\sigma),Y^\epsilon(\sigma))d\sigma\\ &+\frac{1}{\epsilon}\int_0^s \left(e^{\frac{(t-\sigma)}{\epsilon}B}-e^{\frac{(s-\sigma)}{\epsilon}B}\right)G(X^\epsilon(\sigma),Y^\epsilon(\sigma))d\sigma\\ &+W^{\epsilon,B}(t)-W^{\epsilon,B}(s),\end{aligned}$$

  where $W^{\epsilon,B}(r)=\frac{1}{\sqrt{\epsilon}}\int_0^r e^{\frac{(r-\sigma)}{\epsilon}B}dW(\sigma)$. We remark that only the last expression can not bounded almost surely (since we assume that $G$ is bounded).
- For the first term, using the second inequality of Proposition 1.6, we have for any $0<s<t$

$$|(e^{\frac{t}{\epsilon}B}-e^{\frac{s}{\epsilon}B})y|_H\leq C_r(\frac{t-s}{s})^r|y|_H.$$

- For the second term, we have for any $0<s<t$

$$\begin{aligned}|\frac{1}{\epsilon}\int_s^t e^{\frac{(t-\sigma)}{\epsilon}B}G(X^\epsilon(\sigma),Y^\epsilon(\sigma))d\sigma|&\leq \frac{1}{\epsilon}\int_s^t |e^{\frac{(t-\sigma)}{\epsilon}B}|_{\mathcal{L}(H)}\|G\|_\infty d\sigma\\ &\leq \frac{C}{\epsilon}\int_s^t e^{-\mu(t-\sigma)/\epsilon}d\sigma\\ &\leq C\int_0^{(t-s)/\epsilon}e^{-\mu\sigma}d\sigma\\ &\leq C\frac{(t-s)^r}{\epsilon^r}.\end{aligned}$$

- For the third term, we use the second estimate of Proposition 1.6, and we have for any $0 < s < t$

$$\begin{aligned}|\frac{1}{\epsilon}\int_0^s \left(e^{\frac{(t-\sigma)}{\epsilon}B} - e^{\frac{(s-\sigma)}{\epsilon}B}\right) G(X^\epsilon(\sigma), Y^\epsilon(\sigma))d\sigma| &\leq \frac{1}{\epsilon}\int_0^s |e^{\frac{(t-\sigma)}{\epsilon}B} - e^{\frac{(s-\sigma)}{\epsilon}B}|_{\mathcal{L}(H)}\|G\|_\infty d\sigma \\ &\leq \frac{C_r}{\epsilon}\int_0^s \frac{(t-s)^r}{(s-\sigma)^r} e^{-\frac{\mu(s-\sigma)}{2\epsilon}} d\sigma \\ &\leq C_r \frac{(t-s)^r}{\epsilon^r}\int_0^{+\infty} \frac{1}{\sigma^r} e^{-\frac{\mu\sigma}{2}} d\sigma.\end{aligned}$$

- For the fourth term, we have for any $0 < s < t$,

$$\begin{aligned}\mathbb{E}|W^{\epsilon,B}(t) - W^{\epsilon,B}(s)|^2 &= \mathbb{E}|\frac{1}{\sqrt{\epsilon}}\int_s^t e^{(t-\sigma)B/\epsilon} dW(\sigma) + \frac{1}{\sqrt{\epsilon}}\int_0^s (e^{(t-\sigma)B/\epsilon} - e^{(s-\sigma)B/\epsilon}) dW(\sigma)|^2 \\ &= \mathbb{E}|\frac{1}{\sqrt{\epsilon}}\int_s^t e^{(t-\sigma)B/\epsilon} dW(\sigma)|^2 \\ &+ \mathbb{E}|\frac{1}{\sqrt{\epsilon}}\int_0^s (e^{(t-\sigma)B/\epsilon} - e^{(s-\sigma)B/\epsilon}) dW(\sigma)|^2 \\ &= \frac{1}{\epsilon}\int_s^t |e^{(t-\sigma)B/\epsilon}|^2_{\mathcal{L}_2(H)} d\sigma \\ &+ \frac{1}{\epsilon}\int_0^t |e^{(t-\sigma)B/\epsilon} - e^{(s-\sigma)B/\epsilon}|^2_{\mathcal{L}_2(H)} d\sigma.\end{aligned}$$

  On the one hand,

$$\begin{aligned}\frac{1}{\epsilon}\int_s^t |e^{(t-\sigma)B/\epsilon}|^2_{\mathcal{L}_2(H)} d\sigma &= \frac{1}{\epsilon}\int_s^t \sum_{k=0}^{+\infty} e^{-2(t-\sigma)\mu_k/\epsilon} d\sigma \\ &= \sum_{k=0}^{+\infty}\int_0^{(t-s)/\epsilon} e^{-2\sigma\mu_k} d\sigma \\ &= \sum_{k=0}^{+\infty} \frac{1}{2\mu_k}(1 - e^{-2\mu_k(t-s)/\epsilon}) \\ &\leq C_\zeta \sum_{k=0}^{+\infty} \frac{\mu_k^{2r}}{\mu_k}\left(\frac{t-s}{\epsilon}\right)^{2r},\end{aligned}$$

  and we know (by Assumption 1.3) that the above sum is finite if and only if $r < 1/4$; on the other hand,

$$\begin{aligned}\frac{1}{\epsilon}\int_0^s |e^{(t-\sigma)B/\epsilon} - e^{(s-\sigma)B/\epsilon}|^2_{\mathcal{L}_2(H)} d\sigma &= \frac{1}{\epsilon}\int_0^s \sum_{k=0}^{+\infty} e^{-2(s-\sigma)\mu_k/\epsilon}(1 - e^{-(t-s)\mu_k/\epsilon})^2 d\sigma \\ &\leq C_\zeta \sum_{k=0}^{+\infty}\left(\frac{t-s}{\epsilon}\right)^{2r} \frac{\mu_k^{2r}}{\mu_k}(1 - e^{-2s\mu_k}).\end{aligned}$$

□

Finally the following Proposition gives a control for $AX^\epsilon$. We assume $\theta > 0$, even if the proof is valid for $\theta = 0$.

**Proposition 1.35.** *For any $0 < r < 1$, there exists a constant $C_r$ such that if $x \in D((-A)^\theta)$ and $y \in H$, then for any $t > 0$ and $\epsilon > 0$*

$$(\mathbb{E}[|AX^\epsilon(t)|^2])^{1/2} \leq C_r(1 + t^{\theta-1})|x|_{(-A)^\theta} + C_r(1 + \epsilon^{-\frac{r}{2}})(1 + |x|_H + |y|_H).$$

Proof We remark that in Lemma 4.4 of [12] $\mathbb{E}|AX^\epsilon(t)|$ is controlled, but the same approach gives the result for $\mathbb{E}|AX^\epsilon(t)|^2$. We have

$$\begin{aligned} X^\epsilon(t) &= e^{tA}x + \int_0^t e^{(t-s)A}F(X^\epsilon(s), Y^\epsilon(s))ds \\ &= e^{tA}x + \int_0^t e^{(t-s)A}F(X^\epsilon(t), Y^\epsilon(t))ds \\ &+ \int_0^t e^{(t-s)A}\left(F(X^\epsilon(s), Y^\epsilon(s)) - F(X^\epsilon(t), Y^\epsilon(t))\right)ds. \end{aligned}$$

For the first term, we have for any $t > 0$

$$|Ae^{tA}x| \leq Ct^{\theta-1}|x|_{(-A)^\theta}.$$

For the second term, we have

$$|A\int_0^t e^{(t-s)A}F(X^\epsilon(t), Y^\epsilon(t))ds| = |(e^{tA} - I)F(X^\epsilon(t), Y^\epsilon(t))| \leq C.$$

For the third term, we have

$$\begin{aligned} &|A\int_0^t e^{(t-s)A}\left(F(X^\epsilon(s), Y^\epsilon(s)) - F(X^\epsilon(t), Y^\epsilon(t))\right)ds| \\ &\qquad \leq \int_0^t \frac{Ce^{-\frac{\lambda}{2}(t-s)}}{t-s}\left(|X^\epsilon(s) - X^\epsilon(t)| + |Y^\epsilon(s) - Y^\epsilon(t)|\right)ds. \end{aligned}$$

Using Minkowski inequality, we get

$$\begin{aligned} \mathbb{E}\left(\int_0^t \frac{Ce^{-\frac{\lambda}{2}(t-s)}}{t-s}|X^\epsilon(s) - X^\epsilon(t)|ds\right)^2 &\leq \left(\int_0^t \frac{Ce^{-\frac{\lambda}{2}(t-s)}}{t-s}(\mathbb{E}|X^\epsilon(t) - X^\epsilon(s)|^2)^{1/2}ds\right)^2 \\ \mathbb{E}\left(\int_0^t \frac{Ce^{-\frac{\lambda}{2}(t-s)}}{t-s}|Y^\epsilon(s) - Y^\epsilon(t)|ds\right)^2 &\leq \left(\int_0^t \frac{Ce^{-\frac{\lambda}{2}(t-s)}}{t-s}(\mathbb{E}|Y^\epsilon(t) - Y^\epsilon(s)|^2)^{1/2}ds\right)^2. \end{aligned}$$

Using Propositions 1.33 and 1.34, we obtain a regularity result which gives convergent integrals. It is then easy to conclude.

□

## 1.7. Appendix: Properties of $\overline{X}$

Again the results of this section only require Assumptions 1.3, 1.7 and 1.9; in particular no dissipativity is assumed.

Recall that $\overline{X}(t,x)$ is defined via (1.3).

**Proposition 1.36.** *There exists $C > 0$ such that for any $x \in H$ and any $t \geq 0$*

$$|\overline{X}(t,x)| \leq C(1 + e^{-\lambda t}|x|).$$

Proof We use the mild representation formula: for any $t \geq 0$ and $x \in H$,

$$\begin{aligned} |\overline{X}(t,x)| &= |e^{tA}x + \int_0^t e^{(t-s)A}\overline{F}(\overline{X}(s,x))ds| \\ &\leq e^{-\lambda t}x + \int_0^t e^{-\lambda(t-s)}|\overline{F}(\overline{X}(s,x))|ds \\ &\leq C(1 + e^{-\lambda t}|x|), \end{aligned}$$

since $\overline{F}$ is bounded.

□

**Proposition 1.37.** *For any $0 < r < 1$ and $0 < \theta \leq 1$, there exists $C_r > 0$ such that for any $x \in H$, for any $0 < s \leq t$, we have*

$$|\overline{X}(t,x) - \overline{X}(s,x)| \leq C_r |t-s|^{1-r}(1 + \frac{1}{s^{1-r}})(1 + |x|_H).$$

Proof

- If $0 \leq s < t \leq T$, we can write

$$\begin{aligned}\overline{X}(t,x) - \overline{X}(s,x) = (e^{tA} - e^{sA})x \\ + \int_s^t e^{(t-\sigma)A}\overline{F}(\overline{X}(\sigma,x))d\sigma \\ + \int_0^s (e^{(t-\sigma)A} - e^{(s-\sigma)A})\overline{F}(\overline{X}(\sigma,x))d\sigma.\end{aligned}$$

- For the first term, it is easy to see that $|(e^{tA} - e^{sA})x|_H \leq C_r|t-s|^{1-r}(1 + \frac{1}{s^{1-r}})|x|_H$.
- For the second term, since $\overline{F}$ is bounded we have $|\int_s^t e^{(t-\sigma)A}\overline{F}(\overline{X}(\sigma,x))d\sigma|_H \leq C(t-s)$.
- For the third term, we have

$$\begin{aligned}|\int_0^s (e^{(t-\sigma)A} - e^{(s-\sigma)A})\overline{F}(\overline{X}(\sigma,x))d\sigma|_H &\leq C_r \int_0^s \frac{e^{-\frac{\lambda}{2}(s-\sigma)}}{(s-\sigma)^{1-r}}(t-s)^{1-r}|\overline{F}(\overline{X}(\sigma,x))|_H d\sigma \\ &\leq C_r(t-s)^{1-r},\end{aligned}$$

□

**Proposition 1.38.** *For any $0 < \theta \leq 1$, there exists $C(\theta) > 0$ such that if $x \in D(-A)^\theta$, then for any $t > 0$*

$$|A\overline{X}(t,x)|_H \leq C_\theta(1 + t^{\theta-1})(1 + |x|_{(-A)^\theta}).$$

Proof We first write that for any $t \geq 0$

$$\begin{aligned}\overline{X}(t,x) &= e^{tA}x + \int_0^t e^{(t-s)A}\overline{F}(\overline{X}(s,x))ds \\ &= e^{tA}x + \int_0^t e^{(t-s)A}\overline{F}(\overline{X}(t,x))ds \\ &+ \int_0^t e^{(t-s)A}(\overline{F}(\overline{X}(s,x)) - \overline{F}(\overline{X}(t,x)))ds.\end{aligned}$$

We have $|Ae^{tA}x|_H \leq C|x|_{(-A)^\theta}t^{\theta-1}$.
For the second term, we have

$$\begin{aligned}|A\int_0^t e^{(t-s)A}\overline{F}(\overline{X}(t,x))ds|_H &= |(e^{tA} - I)\overline{F}(\overline{X}(t,x))|_H \\ &\leq |\overline{F}(\overline{X}(t,x))|_H \\ &\leq C.\end{aligned}$$

The third term can be controlled by

$$|A\int_0^t e^{(t-s)A}(\overline{F}(\overline{X}(s,x)) - \overline{F}(\overline{X}(t,x)))ds| \leq C\int_0^t \frac{e^{-c(t-s)}}{t-s}|\overline{X}(t,x) - \overline{X}(s,x)|_H ds.$$

In order to get a convergent integral, we use the regularity result of $\overline{X}$ proved in Proposition 1.37; therefore we obtain the result.

□

The next three Propositions deal with $\eta^h(t,x)$ the derivative of $\overline{X}(t,x)$ with respect to $x$ in direction $h \in H$, at time $t$: it is the solution of

$$\begin{aligned}\frac{d\eta^h(t,x)}{dt} &= A\eta^h(t,x) + D\overline{F}(\overline{X}(t,x)).\eta^h(t,x) \\ \eta^h(0,x) &= h.\end{aligned} \tag{1.47}$$

Notice that we have to consider a finite horizon $T > 0$.

**Proposition 1.39.** *For any $T > 0$, there exists $C_T > 0$ such that for any $x \in H$, $h \in H$ and $0 < t \leq T$*

$$|\eta^h(t,x)| \leq C_T|h|$$
$$|\eta^h(t,x)|_{(-A)^\eta} \leq C_T(1 + \frac{1}{t^\eta})|h|.$$

Proof We use that $A$ is a negative operator to prove say that for any $t \geq 0$

$$\begin{aligned}\frac{1}{2}\frac{d|\eta^h(t,x)|^2}{dt} &=< A\eta^h(t,x), \eta^h(t,x) > + < D\overline{F}(\overline{X}(t,x)).\eta^h(t,x), \eta^h(t,x) > \\ &\leq [\overline{F}]_{\mathrm{Lip}}|\eta^h(t,x)|^2 \\ &\leq C|\eta^h(t,x)|^2.\end{aligned}$$

Gronwall Lemma then yields the first estimate. The second one is proved by using the mild formulation for $\eta^h(t,x)$:

$$\eta^h(t,x) = e^{tA}h + \int_0^t e^{(t-s)A}D\overline{F}(\overline{X}(s,x)).\eta^h(s,x)ds;$$

thanks to the previous estimate, the integral is bounded by a constant, while $|e^{tA}h|_{(-A)^\eta} \leq \frac{C}{t^\eta}|h|_H$ (see Proposition 1.6).

□

**Proposition 1.40.** *For any $T > 0$, $0 < r < 1$, there exists $C_{T,r} > 0$ such that for any $x \in H$, $h \in H$ and $0 < s \leq t \leq T$*

$$|\eta^h(t,x) - \eta^h(s,x)| \leq C_{T,r}(t-s)^{1-r}(1 + \frac{1}{s^{1-r}})|h|.$$

Proof

- For $0 \leq s < t \leq T$ we can write that

$$\begin{aligned}\eta^h(t,x) - \eta^h(s,x) = (e^{tA} - e^{sA})h &+ \int_s^t e^{(t-\sigma)A}D\overline{F}(\overline{X}(\sigma,x)).\eta^h(\sigma,x)d\sigma \\ &+ \int_0^s (e^{(t-\sigma)A} - e^{(s-\sigma)A})D\overline{F}(\overline{X}(\sigma,x)).\eta^h(\sigma,x)d\sigma.\end{aligned}$$

- For the first term, we can see that $|(e^{tA} - e^{sA})h| \leq C_r\frac{(t-s)^{1-r}}{s^{1-r}}|h|$.
- For the second term, we simply have

$$|\int_s^t e^{(t-\sigma)A}D\overline{F}(\overline{X}(\sigma,x)).\eta^h(\sigma,x)d\sigma| \leq C(t-s)|h|_H.$$

- For the third term,

$$\begin{aligned}&|\int_0^s (e^{(t-\sigma)A} - e^{(s-\sigma)A})D\overline{F}(\overline{X}(\sigma,x)).\eta^h(\sigma,x)d\sigma| \\ &\qquad \leq C_\delta \int_0^s \frac{(t-s)^{1-r}}{(s-\sigma)^{1-r}}|D\overline{F}(\overline{X}(\sigma,x)).\eta^h(\sigma,x)|_H d\sigma \\ &\qquad \leq C_{r,T}(t-s)^{1-r}|h|_H.\end{aligned}$$

□

**Proposition 1.41.** *For any $T > 0$, there exists $C_T$ such that for any $x \in H$, $h \in H$ and $0 < t \leq T$*

$$|A\eta^h(t,x)| \leq C_T(t^{-1} + 1)(1 + |x|)|h|.$$

Proof For any $t \geq 0$,

$$\begin{aligned}\eta^h(t,x) &= e^{tA}h + \int_0^t e^{(t-s)A} D\overline{F}(\overline{X}(s,x)).\eta^h(s,x)ds \\ &= e^{tA}h + \int_0^t e^{(t-s)A} D\overline{F}(\overline{X}(t,x)).\eta^h(t,x)ds \\ &+ \int_0^t e^{(t-s)A}(D\overline{F}(\overline{X}(s,x)).\eta^h(s,x) - D\overline{F}(\overline{X}(t,x)).\eta^h(t,x))ds.\end{aligned}$$

For the first term, we have $|Ae^{tA}h|_H \leq Ct^{-1}|h|$.
For the second term,

$$\begin{aligned}|A\int_0^t e^{(t-s)A} D\overline{F}(\overline{X}(t,x)).\eta^h(t,x)ds|_H &= |(e^{tA}-I)D\overline{F}(\overline{X}(t,x)).\eta^h(t,x)|_H \\ &\leq 2|D\overline{F}(\overline{X}(t,x)).\eta^h(t,x)|_H \\ &\leq C|\eta^h(t,x)|_H \\ &\leq C|h|_H.\end{aligned}$$

For the third term, we have

$$\begin{aligned}&|A\int_0^t e^{(t-s)A}(D\overline{F}(\overline{X}(s,x)).\eta^h(s,x) - D\overline{F}(\overline{X}(t,x)).\eta^h(t,x))ds|_H \\ &\qquad\leq \int_0^t \frac{C}{t-s}|D\overline{F}(\overline{X}(t,x)).\eta^h(t,x) - D\overline{F}(\overline{X}(s,x)).\eta^h(s,x)|_H ds.\end{aligned}$$

To get a convergent integral, we need to show some regularity property.
For any $0 \leq s < t \leq T$,

$$\begin{aligned}D\overline{F}(\overline{X}(t,x)).\eta^h(t,x) - D\overline{F}(\overline{X}(s,x)).\eta^h(s,x) &= [D\overline{F}(\overline{X}(t,x)) - D\overline{F}(\overline{X}(s,x))].\eta^h(t,x) \\ &+ D\overline{F}(\overline{X}(s,x)).(\eta^h(t,x) - \eta^h(s,x)).\end{aligned}$$

On the one hand, using Proposition 1.18 on the regularity of $\overline{F}$, we have

$$\begin{aligned}|[D\overline{F}(\overline{X}(t,x)) - D\overline{F}(\overline{X}(s,x))].\eta^h(t,x)| &\leq C|\overline{X}(t,x) - \overline{X}(s,x)||\eta^h(t,x)|_{(-A)^\eta} \\ &\leq C(1+|x|)(t-s)^r(1+\frac{1}{s^r})(1+\frac{1}{s^\eta})|h|_H,\end{aligned}$$

thanks to Propositions 1.37 and 1.39. Here $r$ must satisfy $r > 0$ and $\eta + r < 1$.
On the other hand, using Proposition 1.40,

$$\begin{aligned}|D\overline{F}(\overline{X}(s,x)).(\eta^h(t,x) - \eta^h(s,x))|_H &\leq C|\eta^h(t,x) - \eta^h(s,x)|_H \\ &\leq C|h||t-s|^{1-r}(1+\frac{1}{s^{1-r}}).\end{aligned}$$

By integration, we then obtain the result.

□

Finally we focus on $\xi^{h,k}(t,x)$ the second derivative of $\overline{X}(t,x)$ with respect to $x$ in directions $h,k \in H$, at time $t$: it is solution of

$$\begin{aligned}\frac{d\xi^{h,k}(t,x)}{dt} &= A\xi^{h,k}(t,x) + D_x\overline{F}(\overline{X}(t,x)).(\xi^{h,k}(t,x)) \\ &+ D^2_{xx}\overline{F}(\overline{X}(t,x)).(\eta^h(t,x),\eta^k(t,x)).\end{aligned} \tag{1.48}$$

**Proposition 1.42.** *For any $T > 0$, there exists $C_T > 0$ such that for any $x \in H$, $h,k \in H$ and $0 \leq t \leq T$*

$$|\xi^{h,k}(t,x)| \leq C_T|h||k|.$$

Proof We have - since $A$ is negative, and using the estimates of Proposition 1.18:

$$\begin{aligned}\frac{1}{2}\frac{d|\xi^{h,k}(t,x)|^2}{dt} &\leq |D_x\overline{F}(\overline{X}(t,x))||\xi^{h,k}(t,x)|^2\\ &+ C|\eta^h(t,x)||\eta^k(t,x)|_{(-A)^\eta}|\xi^{h,k}(t,x)|\\ &\leq C|\xi^{h,k}(t,x)|^2 + C|\eta^h(t,x)|^2|\eta^k(t,x)|^2_{(-A)^\eta},\end{aligned}$$

Using Proposition 1.39, the Assumption $\eta < \frac{1}{2}$, and the Gronwall Lemma, we get the result.

□

## 1.8. Appendix: Properties of the auxiliary function $\tilde{F}$

The results of this section are used only for the proof of the strong convergence Theorem 1.1. Here we need the strict dissipativity Assumption 1.10.

In the proof of Lemma 1.21, we need to use an auxiliary function $\tilde{F}$ - see definition 1.22.

Thanks to Proposition 1.16, we get:

**Proposition 1.43.** *There exists $c > 0$, $C > 0$ such that for any $(x, y) \in H^2$ and $t \geq 0$,*

$$|\tilde{F}(x,y,t)| \leq Ce^{-ct}(1 + |x|_H + |y|_H).$$

We also need the following estimate on the Lipschitz constant of $\tilde{F}$ with respect to $x$, which depends on the regularity assumptions made on $F$ and $G$ - see Assumptions 1.7 and 1.9:

**Proposition 1.44.** *There exists $c > 0$, $C > 0$, such that for any $x_1, x_2, y \in H$ and $t \geq 0$*

$$|\tilde{F}(x_1,y,t) - \tilde{F}(x_2,y,t)| \leq C(1 + |y|)e^{-ct}(1 + \frac{1}{t^\eta})|x_1 - x_2|.$$

Proof For any $t_0 > 0$, we define the following function:

$$\tilde{F}_{t_0}(x,y,t) = \hat{F}(x,y,t) - \hat{F}(x,y,t+t_0),$$

where $\hat{F}(x,y,t) := \mathbb{E}F(x, Y_x(t,y))$.

We claim that it satisfies the following properties:

- $\tilde{F}_{t_0}(x,y,t) \to \tilde{F}(x,y,t)$ when $t_0 \to +\infty$.
- For any $t_0$, for any $x, y, t$ and any $h$, $\tilde{F}_{t_0}$ is differentiable with respect to $x$ at $(x,y,t)$ and in direction $h \in H$.
- We have $|D_x\tilde{F}_{t_0}(x,y,t).h| \leq Ce^{-ct}(1 + \frac{1}{t^\eta})(1 + |y|)|h|$, $C$ being independent of $t_0$.

The first two ones are obvious, thanks to regularity properties of $F$; moreover as soon as we have the third property, the proof of the Proposition can be finished as follows: if we fix $x_1, x_2, y, t, h$, then for any $t_0 > 0$

$$|\tilde{F}_{t_0}(x_1,y,t) - \tilde{F}_{t_0}(x_2,y,t)| \leq Ce^{-ct}(1 + \frac{1}{t^\eta})(1 + |y|)|x_1 - x_2|.$$

Letting $t_0 \to +\infty$, we get

$$|\tilde{F}(x_1,y,t) - \tilde{F}(x_2,y,t)| \leq Ce^{-ct}(1 + \frac{1}{t^\eta})(1 + |y|)|x_1 - x_2|.$$

It remains to estimate $|D_x\tilde{F}_{t_0}(x,y,t).h|$ for any $h \in H$.

First we notice that thanks to the Markov property we have

$$\begin{aligned}\tilde{F}_{t_0}(x,y,t) &= \hat{F}(x,y,t) - \hat{F}(x,y,t+t_0)\\ &= \hat{F}(x,y,t) - \mathbb{E}F(x, Y_x(t+t_0,y))\\ &= \hat{F}(x,y,t) - \mathbb{E}\hat{F}(x, Y_x(t_0,y),t).\end{aligned}$$

Therefore we have for any $h$

$$\begin{aligned}D_x\tilde{F}_{t_0}(x,y,t).h &= D_x\hat{F}(x,y,t).h - \mathbb{E}D_x\left(\hat{F}(x,Y_x(t_0,y),t)\right).h\\ &= D_x\hat{F}(x,y,t).h - \mathbb{E}D_x\hat{F}(x,Y_x(t_0,y),t).h - \mathbb{E}D_y\hat{F}(x,Y_x(t_0,y),t).(D_xY_x(t_0,y).h).\end{aligned}$$

Then we see that we have to analyse

$$D_x\hat{F}(x,y,t).h - D_x\hat{F}(x,z,t).h$$

and

$$D_y\hat{F}(x,y,t).$$

- For any $y, z \in H$, we have

$$\begin{aligned}|\hat{F}(x,y,t) - \hat{F}(x,z,t)| &= |\mathbb{E}F(x,Y_x(t,y)) - \mathbb{E}F(x,Y_x(t,z))| \\ &\leq C\mathbb{E}|Y_t^x(y) - Y_t^x(z)| \\ &\leq Ce^{-ct}|y-z|,\end{aligned}$$

  and we deduce that $|D_y\hat{F}(x,y,t).k| \leq Ce^{-ct}|k|$.
- Moreover we know that $U_t^{x,h}(y) = D_xY_x(t,y).h$ is solution of

$$\begin{aligned}dU_t^{x,h}(y) &= \left(BU_t^{x,h}(y) + D_xG(x,Y_x(t,y)).h + D_yG(x,Y_x(t,y)).U_t^{x,h}(y)\right)dt \\ U_0^{x,h}(y) &= 0.\end{aligned}$$

  We deduce the following property: $|U_t^{x,h}(y)| \leq C|h|$ a.s.
  As a consequence $|\mathbb{E}D_y\hat{F}(x,Y_x(t_0,y),t).(D_xY_x(t_0,y).h)| \leq Ce^{-ct}|h|$.
- Now we take $x, y, z, t, h$, and we compute

$$\begin{aligned}D_x\hat{F}(x,y,t).h - D_x\hat{F}(x,z,t).h &= \mathbb{E}\left(D_xF(x,Y_x(t,y)).h - D_xF(x,Y_x(t,z)).h\right) \\ &+ \mathbb{E}\left(D_yF(x,Y_x(t,y)).U_t^{x,h}(y) - D_yF(x,Y_x(t,z)).U_t^{x,h}(z)\right) \\ &= \mathbb{E}\left(D_xF(x,Y_x(t,y)).h - D_xF(x,Y_x(t,z)).h\right) \\ &+ \mathbb{E}\left([D_yF(x,Y_x(t,y)) - D_yF(x,Y_x(t,z))].U_t^{x,h}(y)\right) \\ &+ \mathbb{E}\left(D_yF(x,Y_x(t,z)).(U_t^{x,h}(y) - U_t^{x,h}(z))\right)\end{aligned}$$

  First, we have

$$\begin{aligned}|\mathbb{E}\left(D_xF(x,Y_x(t,y)).h - D_xF(x,Y_x(t,z)).h\right)| &\leq \mathbb{E}|D_xF(x,Y_x(t,y)).h - D_xF(x,Y_x(t,z)).h| \\ &\leq C|h|_H\mathbb{E}|Y_x(t,y) - Y_x(t,z)|_{(-B)^\eta} \\ &\leq C|h|_He^{-ct}(1+\frac{1}{t^\eta})|y-z|_H,\end{aligned}$$

  using the regularity assumptions on $F$ (see (1.7)), and using the following estimate:

$$\mathbb{E}|Y_x(t,y) - Y_x(t,z)|_{(-B)^\eta} \leq Ce^{-ct}(1+\frac{1}{t^\eta})|y-z|_H,$$

  for some $c > 0$.
  Second,

$$\begin{aligned}|\mathbb{E}([D_yF(x,Y_x(t,y)) - D_yF(x,Y_x(t,z))].U_t^{x,h}(y))| &\\ &\leq \mathbb{E}|[D_yF(x,Y_x(t,y)) - D_yF(x,Y_x(t,z))].U_t^{x,h}(y)| \\ &\leq C\mathbb{E}|U_t^{x,h}(y)|_H|Y_x(t,y) - Y_x(t,z)|_{(-B)^\eta} \\ &\leq Ce^{-ct}(1+\frac{1}{t^\eta})|h|_H|y-z|_H.\end{aligned}$$

  Third,

$$\begin{aligned}|\mathbb{E}\left(D_yF(x,Y_x(t,z)).(U_t^{x,h}(y) - U_t^{x,h}(z))\right)| &\leq \mathbb{E}|D_yF(x,Y_x(t,z)).(U_t^{x,h}(y) - U_t^{x,h}(z))| \\ &\leq C\mathbb{E}|U_t^{x,h}(y) - U_t^{x,h}(z)|_H;\end{aligned}$$

It remains to look at $|U_t^{x,h}(y) - U_t^{x,h}(z)|_H$; we indeed have

$$|U_t^{x,h}(y) - U_t^{x,h}(z)|^2 \leq C|h|_H^2|y-z|_H^2 e^{-c_0 t},$$

where $c_0 > 0$.

We use these inequalities with $z := Y_x(t_0, y)$; recalling that for any $t_0$

$$\mathbb{E}|Y_x(t_0, y)| \leq C(1 + |y|),$$

we get

$$|D_x \tilde{F}_{t_0}(x, y, t).h| \leq C(1 + |y|)e^{-ct}(1 + \frac{1}{t^\eta})|h|.$$

□

# Chapitre 2

# Utilisation d'un schéma numérique de type Euler pour approcher la loi invariante d'une EDPS


Résumé

On s'intéresse ici à la comparaison des lois à un instant $t$ de la solution d'une EDPS et de son approximation obtenue grâce à un schéma numérique de type Euler. Sous une hypothèse de dissipation faible, on montre que l'erreur faible associée est d'ordre 1/2 par rapport au pas de temps, uniformément par rapport au temps $t$. On en déduit un résultat d'approximation de la loi invariante de l'EDPS considéée.

Ce résultat est essentiel dans la construction du schéma numérique multi-échelle présenté dans le chapitre suivant: on peut approcher la loi invariante de l'équation rapide à la base du principe de moyennisation du chapitre 1.



Ce travail a fait l'objet d'une prépublication sous le titre

**Approximation of the invariant measure with an Euler scheme for Stochastic PDEs driven by Space-Time White Noise**.

# Chapitre 2. Approximation of the invariant measure with an Euler scheme for Stochastic PDEs driven by Space-Time White Noise

## 2.1. Introduction

In this Chapter, we are interested in the discretization in time of the following stochastic reaction-diffusion equation

$$\frac{\partial y(t,\xi)}{\partial t} = \frac{\partial^2 y(t,\xi)}{\partial \xi^2} + g(\xi, y(t,\xi)) + \frac{\partial \omega(t,\xi)}{\partial t}, \tag{2.1}$$

for $t \geq 0, \xi \in (0,1)$, with the initial condition $y(0,\xi) = y(\xi)$, and homogeneous Dirichlet boundary conditions $y(t,0) = y(t,1) = 0$. The stochastic perturbation $\frac{\partial \omega(t,\xi)}{\partial t}$ is a space-time white noise: the rigorous interpretation of (2.1) is given by an abstract evolution equation (2.2) - in the sense of [15] - in the Hilbert space $H = L^2(0,1)$, driven by a Wiener cylindrical process - see Section 2.2.3.

The numerical approximation of stochastic equations has been extensively studied during the last thirty years. First we recall that one can look at strong approximation results - when the trajectories of the continuous and the discrete-time processes are compared - or at weak approximation results - when the laws at a fixed time are compared. The simplest method in the case of SDEs is the Euler-Maruyama scheme; it is built as a straightforward extension of the well-known explicit Euler method for ODEs, which is of order 1: if we consider in $\mathbb{R}^d$ a SDE with regular coefficients

$$dX_t = f(X_t)dt + \sigma(X_t)dB_t, X_0 = x,$$

its numerical approximation is defined for a given step-size $\Delta t$ by

$$X_0 = x,$$
$$X_{n+1} = X_n + f(X_n)\Delta t + \sigma(X_n)(B_{t_{n+1}} - B_{t_n}).$$

Due to the regularity properties of the Brownian Motion, this scheme is in general only of order $1/2$ in the strong sense - i.e. for $n\Delta t \leq T$ we have $\mathbb{E}|X_{n\Delta t} - X_n|^2 \leq C(T)\Delta t$ - while it is of order 1 in the weak sense - i.e. for test functions $\phi$ of class $\mathcal{C}^3$, with bounded derivatives, we have a bound $|\mathbb{E}\phi(X_{n\Delta t}) - \mathbb{E}\phi(X_n)| \leq C(T,\phi)\Delta t$. The idea for proving that weak order is 1 and not $1/2$ is to consider the Kolmogorov equation associated with the process $X_t$, which is satified by the function $(t,x) \mapsto \mathbb{E}\phi(X(t,x))$ - see [60], [62]. The books [43] and [54] - see also [53] - contain various numerical schemes - like the well-known Milstein scheme, some implicit schemes, and methods based on stochastic Taylor expansions - with their order of convergence in both strong and weak senses.

Numerical methods for SPDEs like (2.1) need discretization both in time and in space. For example, time discretization leads to explicit or implicit methods, while discretization in space can be done with finite difference or finite element methods. Basically, the result for space-time white noise driven equations is convergence with strong order $1/2$ in space and only $1/4$ in time - under some Courant-Friedrichs-Lewy conditions when necessary: see [17], [34], [35], [36], [39], [65].

If we look at the abstract formulation of (2.1) in the Hilbert space $H$, results have also been proved for the time discretization using semi-implicit Euler schemes: the strong order of convergence is $1/4$ - see [57] - while the weak order of convergence is $1/2$ - see [18]. Moreover in [20], the authors have studied weak convergence in the case of linear equations when using a finite element method in space. We follow here the framework of [18]: we consider the stochastic evolution equation in $H$ and we use a time discretization with a semi-implicit Euler scheme, with no discretization in space.

In this work, we are interested in the behaviour of the weak convergence estimates when the final time $T$ goes to infinity: can we replace constants $C(T,\phi)$ by a constant $C(\phi)$ independent from $T$? Passing to

the limit, we thus ask the more general following question: can we use a numerical scheme to approximate the invariant probability measure of the continuous time process - which is assumed to be unique, ergodic and with exponential convergence to equilibrium? The SDE case has been studied with various methods: in [61], the weak error analysis is made by showing that the time derivatives of the solution of the Kolmogorov equation are exponentially decreasing in time; in [40], some general conditions are given for the ergodicity of the numerical scheme, thanks to the theory of geometric ergodicity of Markov Chains. In [51], the approximation result is shown thanks to the use of a Poisson equation.

In the case of SPDEs, general ergodicity results have only been obtained recently - see Section 2.4.3 - and the problem of approximation of invariant measures by numerical schemes has not been studied yet.

We prove the following result:

**Theorem 2.1.** *For any $0 < \kappa < 1/2$, $\tau_0 > 0$ and for any $\mathcal{C}_b^2$ function $\phi$, there exists a constant $C > 0$ such that for any $m \geq 2$, $y \in H$ and $0 < \tau \leq \tau_0$*

$$\left|\mathbb{E}[\phi(Y(m\tau, y))] - \mathbb{E}[\phi(Y_m(\tau, y))]\right| \leq C(1 + |y|^3)(((m-1)\tau)^{-1/2+\kappa} + 1)\tau^{1/2-\kappa}.$$

The continuous process $Y$ is defined by Equation (2.2) below, and the numerical approximation $(Y_m(\tau, y))$ with time step $\tau$ and initial condition $y$ is defined by (2.8).

We notice that the right-hand side of the previous estimate contains a singularity when $m = 1$, which is due to a lack of regularity of the infinite-dimensional processes - details are given in Sub-Section 2.6.1 below. However, if we look at the error at a fixed time $T = m\tau$, if $\tau$ is small enough we just need to change the constant $C = C(T)$; moreover since we are interested in the behaviour when $m$ goes to infinity, this term plays no role.

With the assumptions precised below, we show in Section 2.4.3 that the SPDE admits a unique invariant probability measure $\overline{\mu}$, which is ergodic and strongly mixing, with exponential convergence to equilibrium; nevertheless we can in general only show the existence, not the uniqueness, of invariant measures for the numerical approximation, and we can prove the following result:

**Corollary 2.2.** *For any $0 < \kappa < 1/2$, $\tau_0 > 0$ and for any $\mathcal{C}_b^2$ function $\phi$, there exists constants $c > 0$, $C > 0$ such that for any $0 < \tau \leq \tau_0$, any initial condition $y \in H$ and any $m \geq 1$*

$$\left|\mathbb{E}[\phi(Y_m(\tau, y))] - \int_H \phi d\overline{\mu}\right| \leq C(1 + |y|^3)\left(\frac{1}{m^{1/2-\kappa}} + \tau^{1/2-\kappa}\right) + C(1 + |y|^2)e^{-cm\tau}.$$

*Moreover, if $\mu^\tau$ is an ergodic invariant probability measure of the numerical scheme $(Y_m(\tau, .))_{m\in\mathbb{N}}$, we have*

$$\left|\int_H \phi d\mu^\tau - \int_H \phi d\overline{\mu}\right| \leq C\tau^{1/2-\kappa};$$

*all the ergodic invariant measures of the numerical approximation are then close to the unique invariant probability measure of the continuous process.*

Up to our knowledge, this is the first result of this kind for SPDEs.

The key point, like in [61], for obtaining bounds independent from the time $T = m\tau$, is to prove that derivatives of the solution $u$ of the underlying Kolomogorov equation - mentioned above - decrease exponentially in time: this is done in Section 2.5.2, using a coupling method.

Moreover, an essential tool in [18] is the use of Malliavin calculus and of an integration by parts formula in order to transform a stochastic integral - which is not regular enough in space in the infinite dimensional setting - into an expression which can be controlled; in Lemma 2.17, we see that the involved Malliavin derivatives may increase exponentially fast with respect to time. The solution is to separate the lack of regularity problem and this badly controlled growth with a decomposition of the interval into two parts: in the first one we need the integration by parts formula, and we can control the Malliavin derivatives, while in the other one we can directly give an appropriate bound. We also provide an improvement with respect to [18]: we can consider a more general nonlinear coefficient $G$ - like a Nemytskii operator, see Example 2.9.

We also notice that some more general equations, with additive noise which is white in time but colored in space, can be studied with our method, with suitable assumptions on the coefficients. For a very smooth noise, the numerical analysis on finite time is easier to treat, but then ergodic properties and long-time behaviour require different techniques; this will be treated elsewhere.

The case of some equations with multiplicative noise which satisfy the Strong Feller Property is also covered by our technique of proof, but with some additional difficulties - see for instance the restrictive condition on the diffusion coefficient in [18]. Since all the necessary ideas are contained here and in [18], we only focus on the additive noise case.

The paper is organized as follows: in Section 2.2, we precise the assumptions made on the coefficients of the equations, and we define the numerical method in Section 2.3. In Section 2.4, we give some bounds on the solutions of the continuous and discrete equations, and we study their asymptotic behaviour: existence of invariant measures, and uniqueness for the continuous equation, following from Proposition 2.19. In Section 2.5, we explain the proof of Theorem 2.1, and we give a proof of Corollary 2.2; more precisely, in Section 2.5.2 we prove the exponential decreasing in time of the derivatives of the function $u$. Eventually Section 2.6 contains the proofs of the remaining estimates

## 2.2. Notations and assumptions

Let $H$ be a separable Hilbert space, with norm denoted by $|.|_H$ or simply $|.|$. We consider equations of the form

$$\begin{aligned} dY(t,y) &= \big(BY(t,y) + G(Y(t,y))\big)dt + dW(t) \\ Y(0,y) &= y. \end{aligned} \tag{2.2}$$

In the next paragraphs, we explain the assumptions on the linear operator $B$ and on the nonlinear coefficient $G$; we also recall how the cylindrical Wiener process $W$ is defined, and how we can construct solutions of this equation.

2.2.1. **Test functions.** To quantify the weak approximation, we use test functions $\phi$ in the space $\mathcal{C}_b^2(H,\mathbb{R})$ of functions from $H$ to $\mathbb{R}$ that are twice continuously differentiable, bounded, with first and second order bounded derivatives.

**Remark 2.3.** *In the sequel, we often identify the first derivative $D\phi(x) \in \mathcal{L}(H,\mathbb{R})$ with the gradient in the Hilbert space $H$, and the second derivative $D^2\phi(x)$ with a linear operator on $H$, via the formulas:*

$$< D\phi(x), h >= D\phi(x).h \text{ for every } h \in H$$
$$< D^2\phi(x).h, k >= D^2\phi(x).(h,k) \text{ for every } h,k \in H.$$

In the sequel, we use the following notations:

$$\|\Phi\|_\infty = \sup_{x\in H} |\Phi(x)|_H$$
$$\|\Phi\|_1 = \sup_{x\in H} |D\Phi(x)|_H$$
$$\|\Phi\|_2 = \sup_{x\in H} |D^2\Phi(x)|_{\mathcal{L}(H)}.$$

2.2.2. **Assumptions on the coefficients.**

2.2.2.1. *The linear operator.* We denote by $\mathbb{N} = \{0,1,2,\ldots\}$ the set of nonnegative integers.

We suppose that the following properties are satisfied:

**Assumptions 2.4.** (1) *We assume that there exists a complete orthonormal system of elements of $H$ denoted by $(f_k)_{k\in\mathbb{N}}$, and a non-decreasing sequence of real positive numbers $(\mu_k)_{k\in\mathbb{N}}$ such that:*

$$Bf_k = -\mu_k f_k \text{ for all } k \in \mathbb{N}.$$

(2) *The sequence $(\mu_k)$ goes to $+\infty$ and*

$$\sum_{k=0}^{+\infty} \frac{1}{\mu_k^\alpha} < +\infty \Leftrightarrow \alpha > 1/2.$$

The smallest eigenvalue of $-B$ is then $\mu_0$.

**Example 2.5.** *The equation* (2.1) *enters in this framework: we can choose* $B = \frac{d^2}{dx^2}$*, with the domain* $H^2(0,1) \cap H_0^1(0,1) \subset L^2(0,1)$ *- corresponding to homogeneous Dirichlet boundary conditions. In this case for any* $k \in \mathbb{N}$ $\mu_k = \pi^2(k+1)^2$*, and* $f_k(\xi) = \sqrt{2}\sin((k+1)\pi\xi)$ *- see* [8].

For a $N \in \{1, 2, \ldots\}$, we denote by $H_N$ the subspace of $H$ spanned by $f_0, \ldots, f_{N-1}$, and by $P_N$ the orthogonal projection of $H$ onto $H_N$.

The domain $D(B)$ of $B$ is equal to $D(B) = \left\{ y = \sum_{k=0}^{+\infty} y_k f_k \in H, \sum_{k=0}^{+\infty} (\mu_k)^2 |y_k|^2 < +\infty \right\}$. We can more generally define fractional powers of $-B$, for $b \in [0,1]$:

$$(-B)^b y = \sum_{k=0}^{\infty} \mu_k^b y_k f_k \in H,$$

with the domains

$$D(-B)^b = \left\{ y = \sum_{k=0}^{+\infty} y_k f_k \in H, \quad |y|_b^2 := \sum_{k=0}^{+\infty} (\mu_k)^{2b} |y_k|^2 < +\infty \right\}.$$

When $b \in [0,1]$, we can also define the spaces $D(-B)^{-b}$ and operators $(-B)^{-b}$, with norm denoted by $|.|_{-b}$; when $y = \sum_{k=0}^{+\infty} y_k f_k \in H$, we have $(-B)^{-b} y = \sum_{k=0}^{+\infty} \mu_k^{-b} y_k f_k$ and $|y|_{-b}^2 := \sum_{k=0}^{+\infty} (\mu_k)^{-2b} |y_k|^2$.

The semi-group $(e^{tB})_{t \geq 0}$ can be defined by the Hille-Yosida Theorem - see [8]. We use the following spectral formula: if $y = \sum_{k=0}^{+\infty} y_k f_k \in H$, then for any $t \geq 0$

$$e^{tB} y = \sum_{k=0}^{+\infty} e^{-\mu_k t} y_k f_k.$$

For any $t \geq 0$, $e^{tB}$ is a continuous linear operator in $H$, with operator norm $e^{-\mu_0 t}$. The semi-group $(e^{tB})$ is used to define the solution $Z(t) = e^{tB} z$ of the linear Cauchy problem

$$\frac{dZ(t)}{dt} = BZ(t) \quad \text{with} \quad Z(0) = z.$$

To define solutions of more general PDEs of parabolic type, we use mild formulation, and Duhamel principle.

This semi-group enjoys some smoothing properties that we often use in this work. Basically we need the following properties, which are easily proved using the above spectral properties.

**Proposition 2.6.** *Under Assumption 2.4, for any* $\sigma \in [0,1]$*, there exists* $C_\sigma > 0$ *such that we have:*

(1) *for any* $t > 0$ *and* $y \in H$*,*

$$|e^{tB} y|_\sigma \leq C_\sigma t^{-\sigma} e^{-\frac{\mu_0}{2} t} |y|_H.$$

(2) *for any* $0 < s < t$ *and* $y \in H$*,*

$$|e^{tB} y - e^{sB} y|_H \leq C_\sigma \frac{(t-s)^\sigma}{s^\sigma} e^{-\frac{\mu_0}{2} s} |y|_H.$$

(3) *for any* $0 < s < t$ *and* $y \in D(-A)^\sigma$*,*

$$|e^{tB} y - e^{sB} y|_H \leq C_\sigma (t-s)^\sigma e^{-\frac{\mu_0}{2} s} |y|_\sigma.$$

2.2.2.2. *The nonlinear operator.* The nonlinear operator $G$ is assumed to satisfy some general assumptions. In Example 2.9, we give the two main kind of operators that can be used in our framework.

**Assumptions 2.7.** *The function* $G : H \to H$ *is assumed to be bounded and Lipschitz continuous. We denote by* $L_G$ *the Lipschitz constant of* $G$

*We also define for each* $N \geq 1$ *a function* $G_N : H_N \to H_N$*, with* $G_N(y) = P_N G(y)$ *for any* $y \in H_N$*. We assume that each* $G_N$ *is twice differentiable, and that we have the following bounds on the derivatives, uniformly with respect to* $N$*:*

- *There exists a constant* $C_1$ *such that for any* $N \geq 1$*,* $y \in H_N$ *and* $h \in H_N$

$$|DG_N(y).h|_H \leq C_1 |h|_H.$$

- *There exists $\eta \in [0,1)$ and a constant $C_2$ such that for any $N \geq 1$, $y \in H_N$ and any $h, k \in H_N$ we have*
$$|(-B)^{-\eta} D^2 G_N(y).(h,k)| \leq C_2 |h|_H |k|_H.$$
- *Moreover, there exists a constant $C_3$ such that for any $N \geq 1$, $y \in H_N$ and any $h, k \in H_N$*
$$|D^2 G_N(y).(h,k)| \leq C_3 |h|_{(-B)^\eta} |k|_H.$$

Since $G$ is bounded, the following property is easily satisfied:

**Proposition 2.8** (Dissipativity)**.** *There exist $c > 0$ and $C > 0$ such that for any $y \in D(B)$*

$$< By + G(y), y > \leq -c|y|^2 + C. \tag{2.3}$$

We remark that we have uniform control with respect to the dimension $N$ of the bounds on $G_N$ and on its derivatives, and that (2.3) is also satisfied for $G_N$, with constants $c$ and $C$ independent from $N$.

**Example 2.9.** *We give some fundamental examples of nonlinearities for which the previous assumptions are satisfied:*

- *A function $G : H \to H$ of class $\mathcal{C}^2$, bounded and with bounded derivatives, fits in the framework, with the choice $\eta = 0$.*
- *The function $G$ can be a **Nemytskii** operator: let $g : (0,1) \times \mathbb{R} \to \mathbb{R}$ be a measurable, bounded, function such that for almost every $\xi \in (0,1)$ $g(\xi, .)$ is twice continuously differentiable, with uniformly bounded derivatives. Then $G(y)$ is defined for every $y \in H = L^2(0,1)$ by*
$$G(y)(\xi) = g(\xi, y(\xi)).$$
*In general, such functions are not Fréchet differentiable, but only Gâteaux differentiable, with the following expressions:*
$$[DG(y).h](\xi) = \frac{\partial g}{\partial y}(\xi, y(\xi)) h(\xi)$$
$$[D^2 G(y).(h,k)](\xi) = \frac{\partial^2 g}{\partial y^2}(\xi, y(\xi)) h(\xi) k(\xi).$$
*If $h$ and $k$ are only $L^2$ functions, $D^2G(y).(h,k)$ may only be $L^1$; however if $h$ or $k$ is $L^\infty$, it is $L^2$. The conditions in Assumption 2.7 are then satisfied as soon as there exists $\eta < 1$ such that $D(-B)^\eta$ is continuously embedded into $L^\infty(0,1)$ - it is the case for $B$ given in Example 2.5, with $\eta > 1/4$. Then the finite dimensional spaces $H_N$ are subspaces of $L^\infty$, and differentiability can be shown.*

2.2.3. **The cylindrical Wiener process and stochastic integration in $H$.** In this section, we recall the definition of the cylindrical Wiener process and of stochastic integral on a separable Hilbert space $H$ with norm $|.|_H$. For more details, see [15].

We first fix a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. A cylindrical Wiener process on $H$ is defined with two elements:

- a complete orthonormal system of $H$, denoted by $(q_i)_{i\in I}$, where $I$ is a subset of $\mathbb{N}$;
- a family $(\beta_i)_{i\in I}$ of independent real Wiener processes with respect to the filtration $((\mathcal{F}_t)_{t\geq 0})$;

then $W$ is defined by

$$W(t) = \sum_{i\in I} \beta_i(t) q_i. \tag{2.4}$$

When $I$ is a finite set, we recover the usual definition of Wiener processes in the finite dimensional space $\mathbb{R}^{|I|}$. However the subject here is the study of some Stochastic Partial Differential Equations, so that in the sequel the underlying Hilbert space $H$ is infinite dimensional; for instance when $H = L^2(0,1)$, an example of complete orthonormal system is $(q_k) = (\sqrt{2}\sin(k\pi.))_{k\geq 1}$ - see Example 2.5.

A fundamental remark is that the series in (2.4) does not converge in $H$; but if a linear operator $\Psi : H \to K$ is Hilbert-Schmidt, then $\Psi W(t)$ converges in $L^2(\Omega, H)$ for any $t \geq 0$.

We recall that a bounded linear operator $\Psi : H \to K$ is said to be Hilbert-Schmidt when

$$|\Psi|^2_{\mathcal{L}_2(H,K)} := \sum_{k=0}^{+\infty} |\Psi(q_k)|^2_K < +\infty,$$

where the definition is independent of the choice of the orthonormal basis $(q_k)$ of $H$. The space of Hilbert-Schmidt operators from $H$ to $K$ is denoted $\mathcal{L}_2(H,K)$; endowed with the norm $|.|_{\mathcal{L}_2(H,K)}$ it is an Hilbert space.

The stochastic integral $\int_0^t \Psi(s)dW(s)$ is defined in $K$ for predictible processes $\Psi$ with values in $\mathcal{L}_2(H,K)$ such that $\int_0^t |\Psi(s)|^2_{\mathcal{L}_2(H,K)}ds < +\infty$ a.s; moreover when $\Psi \in L^2(\Omega \times [0,t]; \mathcal{L}_2(H,K))$, the following two properties hold:

$$\mathbb{E}|\int_0^t \Psi(s)dW(s)|^2_K = \mathbb{E}\int_0^t |\Psi(s)|^2_{\mathcal{L}_2(H,K)}ds \text{ (Itô isometry)},$$
$$\mathbb{E}\int_0^t \Psi(s)dW(s) = 0.$$

A generalization of Itô formula also holds - see [15].

For instance, if $v = \sum_{k\in\mathbb{N}} v_k q_k \in H$, we can define

$$< W(t), v >= \int_0^t < v, dW(s) >= \sum_{k\in\mathbb{N}} \beta_k(t)v_k;$$

we then have the following space-time white noise property

$$\mathbb{E} < W(t), v_1 >< W(s), v_2 >= t \wedge s < v_1, v_2 > .$$

Therefore to be able to integrate a process with respect to $W$ requires some strong properties on the integrand; in our SPDE setting, the Hilbert-Schmidt properties follow from the assumptions made on the linear coefficients of the equations.

Thanks to Assumption 2.4, it is easy to show that the following stochastic integral is well-defined in $H$, for any $t \geq 0$:

$$W^B(t) = \int_0^t e^{(t-s)B}dW(s). \tag{2.5}$$

It is called a stochastic convolution, and it is the unique mild solution of

$$dZ(t) = BZ(t)dt + dW(t) \quad \text{with} \quad Z(0) = 0.$$

Under the second condition of Assumption 2.4, there exists $\delta > 0$ such that for any $t > 0$ we have $\int_0^t \frac{1}{s^\delta}|e^{sB}|^2_{\mathcal{L}_2(H)}ds < +\infty$; it can then be proved that $W^B$ has continuous trajectories - via the *factorization method*, see [15] - and that for any $1 \leq p < +\infty$

$$\mathbb{E}\sup_{t\geq 0}|W^B(t)|^p_H < +\infty. \tag{2.6}$$

We can now define solutions to Equation (2.2), thanks to the assumptions made on the coefficients: the following result is classical - see [15]:

**Proposition 2.10.** *For every $T > 0$, $y \in H$, the equation (2.2) admits a unique mild solution $Y \in L^2(\Omega, \mathcal{C}([0,T],H))$:*

$$Y(t) = e^{tB}y + \int_0^t e^{(t-s)B}G(Y(s))ds + \int_0^t e^{(t-s)B}dW(s). \tag{2.7}$$

### 2.3. Definition of the numerical scheme

We now define the numerical approximation of $Y$: denoting by $\tau$ the time step, we have

$$Y_{k+1}(\tau,y) = Y_k(\tau,y) + \tau BY_{k+1}(\tau,y) + \tau G(Y_k(\tau,y)) + \sqrt{\tau}\chi_{k+1}$$
$$Y_0(\tau,y) = y,$$

where $\chi_{k+1} = \frac{1}{\sqrt{\tau}}(W((k+1)\tau) - W(k\tau))$.

To simplify the equations, we omit the dependence of $Y_k$ on the time-step $\tau$ and on the initial condition $y$.

This expression does not make sense in $H$. Defining $R_\tau = (I - \tau B)^{-1}$, this last equation can be replaced by

$$Y_{k+1} = R_\tau Y_k + \tau R_\tau G(Y_k) + \sqrt{\tau} R_\tau \chi_{k+1}, \tag{2.8}$$

which is valid, since $R_\tau$ is a Hilbert-Schmidt operator on $H$.

**Remark 2.11.** *Later, we often use the following expression for $Y_k$:*

$$Y_k = R_\tau^k y + \tau \sum_{l=0}^{k-1} R_\tau^{k-l} G(Y_l) + \sqrt{\tau} \sum_{l=0}^{k-1} R_\tau^{k-l} \chi_{l+1}. \tag{2.9}$$

*The following expression is also useful:*

$$\sqrt{\tau} \sum_{l=0}^{k-1} R_\tau^{k-l} \chi_{l+1} = \int_0^{t_k} R_\tau^{k-l_s} dW(s), \tag{2.10}$$

*where $l_s = \lfloor \frac{s}{\tau} \rfloor$ - with the notation $\lfloor . \rfloor$ for the integer part.*

We need the following technical estimate:

**Lemma 2.12.** *For any $0 \leq \kappa \leq 1$ and $j \geq 1$,*

$$|(-B)^{1-\kappa} R_\tau^j|_{\mathcal{L}(H)} \leq c \frac{1}{(j\tau)^{1-\kappa}} \frac{1}{(1+\mu_0\tau)^{j\kappa}}.$$

Proof For any $z \in H$,

$$\begin{aligned}
|(-B)^{1-\kappa} R_\tau^j z|_H^2 &= \sum_{i=0}^{+\infty} \mu_i^{2(1-\kappa)} \frac{1}{(1+\mu_i\tau)^{2j}} |z_i|^2 \\
&= \frac{1}{(j\tau)^{2(1-\kappa)}} \sum_{i=0}^{+\infty} |z_i|^2 \mu_i^{2(1-\kappa)} (j\tau)^{2(1-\kappa)} \frac{1}{(1+\mu_i\tau)^{2j(1-\kappa)}} \frac{1}{(1+\mu_i\tau)^{2j\kappa}} \\
&\leq \frac{1}{(j\tau)^{2(1-\kappa)}} \sum_{i=0}^{+\infty} \left( \frac{\mu_i j\tau}{1+\mu_i j\tau} \right)^{2(1-\kappa)} \frac{1}{(1+\mu_0\tau)^{2j\kappa}} |z_i|^2 \\
&\leq c|z|_H^2 \frac{1}{(j\tau)^{2(1-\kappa)}} \frac{1}{(1+\mu_0\tau)^{2j\kappa}}. \quad \square
\end{aligned}$$

## 2.4. Preliminary results

We warn the reader that constants may vary from line to line during the proofs, and that in order to use lighter notations we usually forget to mention dependence on the parameters. We use the generic notation $C$ for such constants.

We fix the time step $\tau$, as well as $m \in \mathbb{N}$; we then introduce the notation $T = m\tau$. We also define $t_k = k\tau$. $\kappa > 0$ is a parameter, which is be supposed to be small enough. We also control $\tau$: for some $\tau_0 > 0$, $\tau \leq \tau_0$.

2.4.1. **Galerkin approximation.** The first step of the proof is to consider finite dimensional approximations of the $H$-valued processes $(Y(t))_{t\in\mathbb{R}^+}$ and $(Y_k)_{k\in\mathbb{N}}$: if we fix $N \geq 1$, we define $(Y^{(N)}(t))_{t\in\mathbb{R}^+}$ and $(Y_k^{(N)})_{k\in\mathbb{N}}$ by the equations

$$dY^{(N)}(t) = BY^{(N)}(t)dt + G_N(Y^{(N)}(t))dt + dW^{(N)}(t)$$

and

$$Y_{k+1}^{(N)} = Y_k^{(N)} + \tau BY_{k+1}^{(N)} + \tau G_N(Y_k^{(N)}) + \sqrt{\tau} P_N \chi_{k+1},$$

with the initial conditions $Y_{t=0}^{(N)} = Y_{k=0}^{(N)} = P_N y$.

The projection $P_N$ and the nonlinear coefficient $G_N$ have been defined above. $W^{(N)} = P_N W$ is a $N$-dimensional Wiener process on the subspace $H_N$. We remark that the above equations are well-defined on $H_N$ - which is a stable subspace of $B$.

The important and not difficult to prove result is the following: for any fixed $t \in \mathbb{R}^+$ and $k \in \mathbb{N}$, when $N \to +\infty$ we have

$$\mathbb{E}|Y(t) - Y^{(N)}(t)|^2 \to 0 \quad \text{and} \quad \mathbb{E}|Y_k - Y_k^{(N)}|^2 \to 0.$$

We need test functions $\Phi_N$ adapted to the finite-dimensional approximation: for any $N \geq 1$, by restriction we define $\Phi_N(y) = \Phi(y)$ for any $y \in H_N$; we obtain the following decomposition

$$\begin{aligned}\mathbb{E}\Phi(Y(m\tau)) - \mathbb{E}\Phi(Y_m) &= \mathbb{E}\Phi(Y(m\tau)) - \mathbb{E}\Phi(Y^{(N)}(m\tau)) \\ &+ \mathbb{E}\Phi_N(Y^{(N)}(m\tau)) - \mathbb{E}\Phi_N(Y_m^{(N)}) \\ &+ \mathbb{E}\Phi(Y_m^{(N)}) - \mathbb{E}\Phi(Y_m);\end{aligned}$$

the first and the third terms converge to 0 when $N \to +\infty$. In the sequel, we prove an estimate of the second term, which is uniform with respect to dimension $N$; letting $N \to +\infty$ then yields an estimate on the left hand side.

Hence we work with the finite dimensional approximation, but we omit the parameter $N$. The constants appearing below are independent of $N$.

In Section 2.4.2, we prove some estimates on $Y(t)$ and $Y_m$, and in Section 2.4.3 we focus on the asymptotic behaviour of the processes.

2.4.2. **Some useful estimates.** Bounds on moments of $Y_t$ and $Y_k$ can be proved, uniformly with respect to time.

**Lemma 2.13.** *For any $p \geq 1$, there exists a constant $C_p > 0$ such that for every $t \geq 0$ and $y \in H$*

$$\mathbb{E}|Y(t,y)|^p \leq C_p(1 + |y|^p).$$

Proof If we define $Z(t) = Y(t) - W^B(t)$, we have $Z(0) = Y(0) = y$, and

$$\frac{dZ(t)}{dt} = BZ(t) + G(Y(t)),$$

and by Proposition 2.8

$$\begin{aligned}\frac{1}{2}\frac{d|Z(t)|^2}{dt} &=< BZ(t) + G(Y(t)), Z(t) > \\ &=< BZ(t) + G(Z(t)), Z(t) > + < G(Y(t)) - G(Z(t)), Z(t) > \\ &\leq -c|Z(t)|^2 + C + \|G\|_\infty |Z(t)| \\ &\leq -c'|Z(t)|^2 + C',\end{aligned}$$

for some new constants $c', C'$.

Then almost surely we have for any $t \geq 0$

$$|Z(t)| \leq C(1 + |y|).$$

Thanks to (2.6), the conclusion easily follows. □

**Lemma 2.14.** *For any $p \geq 1$, $\tau_0 > 0$, there exists a constant $C > 0$ such that for every $0 < \tau \leq \tau_0$, $k \in \mathbb{N}$ and $y \in H$*

$$\mathbb{E}|Y_k|^p \leq C(1 + |y|^p).$$

Proof As in the proof of Lemma 2.13 above, we introduce $Z_m = Y_m - w_m$, where the process $(w_m)$ is the numerical approximation of $W^B$ with the numerical scheme (2.8) - with $G = 0$; it is defined by

$$w_{m+1} = R_\tau w_m + \sqrt{\tau} R_\tau \chi_{m+1}.$$

Using Theorem 3.2 of [57], giving the strong order 1/4 for the numerical scheme - when the initial condition is 0, with no nonlinear coefficient, with a constant diffusion term and under the assumptions made here - we obtain the following estimate: for any $p \geq 1$, $\tau_0 > 0$ and $0 < r < 1/2$ there exists $C > 0$ such that for any $0 < \tau \leq \tau_0$ and $m \geq 0$

$$\mathbb{E}|w_m - W^B(m\tau)|^{2p} \leq C\tau^{(1/2-r)p}. \tag{2.11}$$

Thanks to (2.6) and (2.11), we get that for any $\tau_0 > 0$, there exists $C > 0$ such that for $0 < \tau \leq \tau_0$ and $m \geq 0$

$$\mathbb{E}|w_m|^2 \leq C. \tag{2.12}$$

Now $Z_m$ defined above satisfies $Z_0 = Y_0 = y$ and

$$Z_{m+1} = R_\tau Z_m + \tau R_\tau G(Y_m);$$

since $|R_\tau|_{\mathcal{L}(H)} \leq \frac{1}{1+\mu_0\tau}$, we obtain the almost sure estimates

$$|Z_{m+1}| \leq \frac{1}{1+\mu_0\tau}|Z_m| + C\tau$$

and

$$|Z_m| \leq C(1+|y|).$$

Thanks to (2.12), we therefore obtain the result. □

We now introduce the following process: for $0 \leq k \leq m-1$ and $t_k \leq t \leq t_{k+1}$

$$\tilde{Y}(t) = Y_k + \int_{t_k}^t [B_\tau Y_k + R_\tau G(Y_k)]ds + \int_{t_k}^t R_\tau dW(s), \tag{2.13}$$

where $B_\tau = BR_\tau$. The process $\tilde{Y}$ is a natural interpolation of the numerical solution $(Y_k)$ defined by (2.8): $\tilde{Y}(t_k) = Y_k$.

Thanks to Lemma 2.14, we get

**Lemma 2.15.** *For any $p \geq 1$, $\tau_0 > 0$, there exists $C > 0$ such that for any $0 < \tau \leq \tau_0$, $t \geq 0$ and $y \in H$*

$$\mathbb{E}|\tilde{Y}(t)|^p \leq C(1+|y|^p).$$

In the next Lemma, we give a control on Malliavin derivatives of $Y_k$ used in the proof. For an introduction to Malliavin calculus, see [55], [58]. The notations here are the same as in [18], where the following useful integration by parts formula is given - see Lemma 2.1 therein:

**Lemma 2.16.** *For any $F \in \mathbb{D}^{1,2}(H)$, $u \in \mathcal{C}_b^2(H)$ and $\Psi \in L^2(\Omega \times [0,T], \mathcal{L}_2(H))$ an adapted process,*

$$\mathbb{E}[Du(F).\int_0^T \Psi(s)dW(s)] = \mathbb{E}[\int_0^T Tr(\Psi(s)^* D^2u(F)D_sF)ds], \tag{2.14}$$

*where $D_sF : h \in H \mapsto D_s^hF \in H$ stands for the Malliavin derivative of $F$, and $\mathbb{D}^{1,2}(H)$ is the set of $H$-valued random variables $F = \sum_{i\in\mathbb{N}} F_i f_i$, with $F_i \in \mathbb{D}^{1,2}$ the domain of the Malliavin derivative for $\mathbb{R}$-valued random variables for any $i$, and $\sum_{i\in\mathbb{N}} \int_0^T \mathbb{E}|D_sF_i|^2 ds < +\infty$.*

Without any further dissipativity assumption, we are not able to give a uniform control with respect to time of the Malliavin derivative of $\tilde{Y}$. In the proof below, we take care of this problem by using these derivatives only at times $t_k = k\tau$ and $s$ such that $t_{k-l_s} \leq 1$.

**Lemma 2.17.** *For any $0 \leq \beta < 1$ and $\tau_0 > 0$, there exists a constant $C > 0$ such that for every $h \in H$, $k \geq 1$, $0 < \tau \leq \tau_0$ and $s \in [0, t_k]$*

$$|D_s^h Y_k|_\beta \leq C(1+L_G\tau)^{k-l_s}(1 + \frac{1}{(1+\mu_0\tau)^{(1-\beta)(k-l_s)}t_{k-l_s}^\beta})|h|.$$

Proof For any $k \geq 1$, $h \in H$ and $s \in [0, t_k]$, using the chain rule for Malliavin calculus and expressions (2.9) and (2.10), we have

$$D_s^h Y_k = R_\tau^{k-l_s}h + \tau \sum_{l=l_s+1}^{k-1} R_\tau^{k-l} DG(Y_l).D_s^h Y_l.$$

Indeed, recall that $l_s$ denotes the integer part of $\frac{s}{\tau}$, so that when $l \leq l_s$ we have $D_s^h Y_l = 0$.

As a consequence, for $k \geq l_s + 1$

$$|D_s^h Y_k| \leq (1+L_G\tau)^{k-l_s}|h|.$$

Now using Lemma 2.12, we have

$$|(-B)^\beta D_s^h Y_k| \leq \frac{1}{(1+\mu_0\tau)^{(1-\beta)(k-l_s)} t_{k-l_s}^\beta}|h| + L_G\tau \sum_{l=l_s+1}^{k-1} \frac{(1+L_G\tau)^{l-l_s}}{(1+\mu_0\tau)^{(1-\beta)(k-l)} t_{k-l}^\beta}|h|.$$

To conclude, we see that

$$\tau \sum_{l=l_s+1}^{k-1} \frac{1}{(1+\mu_0\tau)^{(1-\beta)(k-l)} t_{k-l}^\beta} \leq C \int_0^{+\infty} t^{-\beta} \frac{1}{(1+\mu_0\tau)^{(1-\beta)\frac{t}{\tau}}} dt \leq C < +\infty,$$

when $0 < \tau \leq \tau_0$. □

2.4.3. **Asymptotic behaviour of the processes.** The results of this section are obtained for the initial $H$-valued processes, and for their finite dimensional approximations.

First, we focus on the existence of invariant measures for the continuous and discrete time processes. We use the well-known Krylov-Bogoliubov criterion - see [16]. Tightness comes from two facts: $D(-B)^\gamma$ is compactly embedded in $H$ when $\gamma > 0$, and when $\gamma < 1/4$ we can control moments:

**Lemma 2.18.** *For any $0 < \gamma < 1/4$, $\tau > 0$ and any $y \in H$, there exists $C(\gamma, \tau, y), C(\gamma, y) > 0$ such that for any $m \geq 1$ and $t \geq 1$*

$$\mathbb{E}|Y_m(\tau, y)|_\gamma^2 \leq C(\gamma, \tau, y) \quad \text{and} \quad \mathbb{E}|Y(t, y)|_\gamma^2 \leq C(\gamma, y)$$

Uniqueness of the invariant probability measure for the continuous time process $(Y(t))_{t\in\mathbb{R}^+}$ can be deduced from the well-known Doob Theorem - see [16]. Indeed, since in equation (2.2) noise is additive and non-degenerate, the Strong Feller property - see also Lemma 2.25 below - and irreducibility can be easily proved. In the proof of the main Theorem 2.1, we also need speed of convergence, and thanks to a coupling argument we get the following exponential convergence result - for a proof see Section 6.1 in [19]:

**Proposition 2.19.** *There exist $c > 0$, $C > 0$ such that for any bounded test function $\phi$, any $t \geq 0$ and any $y_1, y_2 \in H$*

$$|\mathbb{E}\phi(Y(t, y_1)) - \mathbb{E}\phi(Y(t, y_2))| \leq C\|\phi\|_\infty (1 + |y_1|^2 + |y_2|^2) e^{-ct}. \tag{2.15}$$

The idea of coupling relies on the following formula: if $\nu_1$ and $\nu_2$ are two probability measures on a state space $S$, their total variation distance satisfies

$$d_{TV}(\nu_1, \nu_2) = \inf \{\mathbb{P}(X_1 \neq X_2)\},$$

which is an infimum over random variables $(X_1, X_2)$ defined on a same probability space, and such that $X_1 \sim \nu_1$ and $X_2 \sim \nu_2$.

Roughly speaking, the principle is to define a coupling $(Z_1(t, y_1, y_2), Z_2(t, y_1, y_2))_{t\geq 0}$ for the processes $(Y(t, y_1)_{t\geq 0}$ and $Y((t, y_2))_{t\geq 0}$ such that the coupling time $\mathcal{T}$ of $Z_1$ and $Z_2$ - i.e. the first time the processes are equal - has an exponentially decreasing tail.

This technique was first used in the study of the asymptotic behaviour of Markov chains - see [7], [21], [47], [52] - and was later adapted for SDEs and more recently for SPDEs - see for instance [44], [50].

In fact, uniqueness of an invariant probability measure $\overline{\mu}$ is an easy consequence of this Proposition, and moreover we get for any $y \in H$ and any $t \geq 0$

$$|\mathbb{E}\phi(Y(t, y)) - \int_H \phi d\overline{\mu}| \leq C\|\phi\|_\infty (1 + |y|^2) e^{-ct}. \tag{2.16}$$

In general, we do not know whether uniqueness also holds for the numerical approximation $(Y_k(\tau, .))_{k\in\mathbb{N}}$.

**Remark 2.20.** *A sufficient condition for the uniqueness of the invariant probability measure of the discrete time process $(Y_k)_{k\in\mathbb{N}}$ is the strict dissipativity assumption*

$$L_G < \mu_0,$$

*where we recall that $L_G$ denotes the Lipschitz constant of $G$.*

*Then trajectories of the processes $(Y_t)_{t\in\mathbb{R}^+}$ and $(Y_k)_{k\in\mathbb{N}}$ issued from different initial conditions $y_1$ and $y_2$ and driven by the same noise process are exponentially close when time increases: for any $\tau_0 > 0$, there exists $c > 0$ such that for any $0 < \tau \leq \tau_0$, $k \geq 0$ and $t \geq 0$ we have almost surely*

$$|Y(t, y_1) - Y(t, y_2)| \leq e^{-(\mu_0 - L_G)t}|y_1 - y_2|$$
$$|Y_k(\tau, y_1) - Y_k(\tau, y_2)| \leq e^{-ck\tau}|y_1 - y_2|.$$

*Proof of uniqueness is then straightforward - and we do not need Proposition 2.19 above.*

## 2.5. Presentation of the proof of the weak approximation result

The proof of Theorem 2.1 is very technical, so for pedagogy we first introduce the decomposition of the error, and identify the term which we control later in Section 2.6. Some crucial estimates on the derivatives of the semi-group with respect to the initial conditions - regularization, long-time behaviour - are proved below in Sub-Section 2.5.2.

2.5.1. **Strategy.** We define

$$u(t, y) = \mathbb{E}[\phi(Y(t, y))], \tag{2.17}$$

which is solution of a finite dimensional Kolmogorov equation associated with the finite dimensional approximation of (2.2):

$$\frac{\mathrm{d}u}{\mathrm{d}t}(t, y) = Lu(t, y) = \frac{1}{2}\mathrm{Tr}\left(D^2u(t, y)\right) + < By + G(y), Du(t, y) > .$$

As explained in the introduction, this is one of the essential tools in the proof of the weak approximation result.

The weak error at time $T = m\tau$ can be decomposed with a telescoping sum - where to simplify the dependence of the numerical approximation in $\tau$ and $y$ is not written, so that $Y_m = Y_m(\tau, y)$:

$$\begin{aligned}\mathbb{E}[\phi(Y(T, y))] - \mathbb{E}[\phi(Y_m)] &= u(T, y) - \mathbb{E}[u(0, Y_m)] \\ &= \sum_{k=0}^{m-1} (\mathbb{E}[u(T - t_k, Y_k)] - \mathbb{E}[u(T - t_{k+1}, Y_{k+1})]) \\ &= u(T, y) - \mathbb{E}[u(T - \tau, Y_1(\tau, y))] + \sum_{k=1}^{m-1}(a_k + b_k + c_k),\end{aligned} \tag{2.18}$$

where for $1 \leq k \leq m - 1$

$$\begin{aligned}a_k &= \mathbb{E}\int_{t_k}^{t_{k+1}} < B\tilde{Y}(t) - B_\tau Y_k, Du(T - t, \tilde{Y}(t)) > dt, \\ b_k &= \mathbb{E}\int_{t_k}^{t_{k+1}} < G(\tilde{Y}(t)) - R_\tau G(Y_k), Du(T - t, \tilde{Y}(t)) > dt, \\ c_k &= \frac{1}{2}\mathbb{E}\int_{t_k}^{t_{k+1}} \mathrm{Tr}((I - R_\tau R_\tau^*)D^2u(T - t, \tilde{Y}(t)))dt.\end{aligned} \tag{2.19}$$

This follows from the use of the Kolmogorov equation and of the Itô formula. We recall that $\tilde{Y}$ is defined in (2.13).

2.5.2. **Bounds on the derivatives of the transition semi-group.** By (2.17), $u(t, y) = \mathbb{E}[\phi(Y(t, y))]$; since $\phi$ is of class $\mathcal{C}^2$, bounded and with bounded derivatives, we are able to prove that with respect to $y$ the function $u$ is twice differentiable, and that the derivatives can be calculated in the following way:

- For any $h \in H$, we have

$$Du(t, y).h = \mathbb{E}[D\phi(Y(t, y)).\eta^{h,y}(t)], \tag{2.20}$$

where $\eta^{h,y}$ is the solution of

$$\frac{d\eta^{h,y}(t)}{dt} = B\eta^{h,y}(t) + DG(Y(t,y)).\eta^{h,y}(t),$$
$$\eta^{h,y}(0) = h.$$

- For any $h, k \in H$, we have

$$D^2u(t,y).(h,k) = \mathbb{E}[D^2\phi(Y(t,y)).(\eta^{h,y}(t), \eta^{k,y}(t)) + D\phi(Y(t,y)).\zeta^{h,k,y}(t)], \tag{2.21}$$

where $\zeta^{h,k,y}$ is the solution of

$$\frac{d\zeta^{h,k,y}(t)}{dt} = B\zeta^{h,k,y}(t) + DG(Y(t,y)).\zeta^{h,k,y}(t) + D^2G(Y(t,y)).(\eta^{h,y}(t), \eta^{k,y}(t)),$$
$$\zeta^{h,k,y}(0) = 0.$$

In [18], the key point for obtaining the weak order 1/2 is to control the derivatives $|Du(t,y)|_\beta$ and $|(-B)^\beta D^2u(t,y)(-B)^\gamma|_{\mathcal{L}(H)}$, with $\beta < 1/2$ and $\gamma < 1/2$ - with the identification of Remark 2.3. Moreover, to obtain a long-time weak estimate we need to prove some exponential decreasing of such quantities when time $t$ goes to infinity. The two Propositions below are the essential results we thus need.

**Proposition 2.21.** *There exists a constant $\tilde{\mu} > 0$ such that for any $0 \leq \beta < 1/2$, for any $t > 0$ and $y \in H$*

$$|Du(t,y)|_\beta \leq C_\beta(1 + \frac{1}{t^\beta})e^{-\tilde{\mu}t}(1 + |y|^2). \tag{2.22}$$

**Proposition 2.22.** *There exists a constant $\tilde{\mu} > 0$ such that for any $0 \leq \beta, \gamma < 1/2$, for any $t > 0$ and $y \in H$*

$$|(-B)^\beta D^2u(t,y)(-B)^\gamma|_{\mathcal{L}(H)} \leq C_{\beta,\gamma}(1 + \frac{1}{t^\eta} + \frac{1}{t^{\beta+\gamma}})e^{-\tilde{\mu}t}(1 + |y|^2). \tag{2.23}$$

The singularity $t^{-\eta}$ in (2.23) is a consequence of the regularity properties satisfied by $G$. Since in general during the proof of Theorem 2.1, we need $\beta + \gamma$ to be close to 1, and therefore greater than $\eta$, only the second singularity $t^{-\beta-\gamma}$ plays a role.

The proofs require several steps. First in Lemma 2.23 below we prove estimates for a finite horizon and general $0 \leq \beta, \gamma < 1/2$; then in Lemma 2.25 we study the long-time behaviour in the particular case $\beta = \gamma = 0$; we finally conclude with the proofs of Propositions 2.21 and 2.22.

First, we prove estimates of these quantities for $0 < t \leq 1$ - see Lemmas 4.4 and 4.5 in [18], with a difference coming from the assumptions made on the nonlinear coefficient $G$:

**Lemma 2.23.** *For any $0 \leq \beta < 1/2$, $0 \leq \gamma < 1/2$, there exist constants $C_\beta$ and $C_{\beta,\gamma}$ such that for any $y \in H$ and any $0 < t \leq 1$*

$$|Du(t,y)|_\beta \leq \frac{C_\beta}{t^\beta}$$
$$|(-B)^\beta D^2u(t,y)(-B)^\gamma|_{\mathcal{L}(H)} \leq C_{\beta,\gamma}(\frac{1}{t^\eta} + \frac{1}{t^{\beta+\gamma}}).$$

**Remark 2.24.** *If we take another time interval $]0, T_{max}]$ instead of $]0, 1]$, the constants $C_\beta$ and $C_{\beta,\gamma}$ are a priori exponentially increasing in $T_{max}$.*

Proof Owing to (2.20) and (2.21), we only need to prove the following almost sure estimates, for some constants $C_\beta$ and $C_{\beta,\gamma}$ - which may vary from line to line below: for any $0 < t \leq 1$

$$|\eta^{h,y}(t)| \leq \frac{C_\beta}{t^\beta}|h|_{-\beta}$$
$$|\zeta^{h,k,y}(t)| \leq \frac{C_{\beta,\gamma}}{t^\eta}|h|_{-\beta}|k|_{-\gamma}, \tag{2.24}$$

where the parameter $\eta$ is defined in Assumption 2.7.

We use mild formulations, and the regularization properties of the semi-group given in Proposition 2.6:

$$|\eta^{h,y}(t)| = |e^{tB}h + \int_0^t e^{(t-s)B} DG(Y(s,y)).\eta^{h,y}(s)ds|$$
$$\leq \frac{C_\beta}{t^\beta}|h|_{-\beta} + C\int_0^t |\eta^{h,y}(s)|ds,$$

and by the Gronwall Lemma we get the result.

For the second-order derivative, we moreover use the properties of $G$ in Assumption 2.7 to get

$$\begin{aligned}|\zeta^{h,k,y}(t)| = |&\int_0^t e^{(t-s)B} DG(Y(s,y)).\zeta^{h,k,y}(s)ds \\ &+ \int_0^t e^{(t-s)B} D^2G(Y(s,y)).(\eta^{h,y}(s),\eta^{k,y}(s))ds| \\ &\leq C\int_0^t |\zeta^{h,k,y}(s)|ds + \int_0^t \frac{C_{\beta,\gamma}}{(t-s)^\eta}|\eta^{h,y}(s)||\eta^{k,y}(s)|ds \\ &\leq C\int_0^t |\zeta^{h,k,y}(s)|ds + C_{\beta,\gamma}|h|_{-\beta}|k|_{-\gamma}t^{1-\eta-\beta-\gamma}\int_0^1 \frac{1}{(1-s)^\eta s^{\beta+\gamma}}ds.\end{aligned}$$

To conclude, it remains to use the Gronwall Lemma, since for any $0 < t \leq 1$ we get $t^{1-\eta-\beta-\gamma} \leq t^{-\eta}$, thanks to the assumption $\beta + \gamma < 1$. □

Thanks to the dissipativity property expressed in Proposition 2.8, we can prove the result in the case $\beta = \gamma = 0$. We notice that the proof would be straightforward under a strict dissipativity assumption - since then $\eta^{h,y}(t)$ and $\zeta^{h,k,y}(t)$ would decrease exponentially in $t$; in the general case $\eta^{h,y}(t)$ and $\zeta^{h,k,y}(t)$ are exponentially increasing in time so that we can not work directly. Here the result comes from the estimate (2.15) of Proposition 2.19.

**Lemma 2.25.** *There exist constants $C$ and $c > 0$ such that for any $t \geq 0$ and any $y \in H$*

$$|Du(t,y)| \leq Ce^{-ct}(1+|y|^2) \quad \text{and} \quad |D^2u(t,y)|_{\mathcal{L}(H)} \leq Ce^{-ct}(1+\frac{1}{t^\eta})(1+|y|^2). \tag{2.25}$$

Proof The Bismut-Elworthy-Li formula states that if $\Phi : H \to \mathbb{R}$ is a function of class $\mathcal{C}^2$ with bounded derivatives and with at most quadratic growth - i.e. there exists $M(\Phi) > 0$ such that for any $y \in H$ we have $|\Phi(y)| \leq M(\Phi)(1+|y|^2)$ - then we can calculate the first and the second order derivatives of $(t,y) \mapsto v(t,y) := \mathbb{E}\Phi(Y(t,y))$ with respect to $y$. First, we have for any $y \in H$ and $h \in H$

$$\begin{aligned}Dv(t,y).h &= \frac{1}{t}\mathbb{E}[\int_0^t < \eta^{h,y}(s), dW(s) > \Phi(Y(t,y))] \\ &= \frac{2}{t}\mathbb{E}[\int_0^{t/2} < \eta^{h,y}(s), dW(s) > v(t/2, Y(t/2,y))];\end{aligned} \tag{2.26}$$

the second equality is a consequence of the identity $v(t,y) = \mathbb{E}v(t/2, Y(t/2,y))$ obtained with the Markov property, and of the first equality applied with the function $v(t/2,.)$.

Using the second formula of (2.26), we obtain a formula for the second order derivative: for any $y \in H$ and $h, k \in H$,

$$\begin{aligned}D^2v(t,y).(h,k) = &\frac{2}{t}\mathbb{E}[\int_0^{t/2} < \zeta^{h,k,y}(s), dW(s) > v(t/2, Y(T/2,y))] \\ &+ \frac{2}{t}\mathbb{E}[\int_0^{t/2} < \eta^{h,y}(s), dW(s) > Dv(t/2, Y(t/2)).\eta^{k,y}(t/2)].\end{aligned} \tag{2.27}$$

We then see, using Lemmas 2.13 and 2.23 - with $\beta = \gamma = 0$ - that there exists $C > 0$ such that for any $0 < t \leq 1$, $y \in H$, $h, k \in H$

$$\begin{aligned}|Dv(t,y).h| &\leq \frac{C}{\sqrt{t}}M(\Phi)(1+|y|^2)|h|, \\ |D^2v(t,y).(h,k)| &\leq \frac{C}{t}M(\Phi)(1+|y|^2)|h||k|.\end{aligned} \tag{2.28}$$

Now when $t \geq 1$ the Markov property implies that $u(t,y) = \mathbb{E}u(t-1, Y(1,y))$, and by (2.16) we have

$$|u(t-1,y) - \int_H \phi d\overline{\mu}| \leq Ce^{-c(t-1)}(1+|y|^2).$$

If we choose $\Phi_t(y) = u(t-1,y) - \int_H \phi d\overline{\mu}$, we have $u(t,y) = \mathbb{E}\Phi_t(Y(1,y)) + \int_H \phi d\overline{\mu}$, with $M(\Phi_t) \leq Ce^{-c(t-1)}$. With (2.28) at time 1, we obtain for $t \geq 1$

$$|Du(t,y).h| \leq Ce^{-c(t-1)}(1+|y|^2)|h|$$
$$|D^2u(t,y).(h,k)| \leq Ce^{-c(t-1)}(1+|y|^2)|h||k|.$$

Moreover by Lemma 2.23 we have a control when $0 \leq t \leq 1$, so that with a change of constants we get the result. □

We can finally prove the Propositions 2.21 and 2.22. The key tool is the Markov property of the process $Y$ which yields the following formula: for any $t \geq 1$

$$u(t,y) = \mathbb{E}[u(t-1, Y_1(y))]. \tag{2.29}$$

To get the exponential decreasing, we use Lemma 2.25 at time $t-1$ when $t \geq 1$, while $|h|_{-\beta}$ appears from $\eta_{h,y}(1)$, and with estimates coming from Lemma 2.23.

Proof of Propositions 2.21 and 2.22 Using equation (2.29) and Lemma 2.25, for any $t \geq 1$ we have

$$|Du(t,y).h| \leq Ce^{-c(t-1)}\mathbb{E}[(1+|Y(1,y)|^2)|\eta^{h,y}(1)|] \leq Ce^{-c(t-1)}(1+|y|^2)|h|_{-\beta},$$

where the last estimate comes from Lemmas 2.13 and 2.23. Combining this estimate with the result of Lemma 2.23, which gives an estimate for $t \leq 1$, we obtain (2.22). For the second order derivatives, Lemma 2.23 gives an estimate for $t \leq 1$, and for $t \geq 1$ we use (2.29) to see that

$$\begin{aligned} D^2u(t,y).(h,k) &= \mathbb{E}[D^2[u(t-1,Y(1,y))].(h,k)] \\ &= \mathbb{E}D^2u(t-1,Y(1,y)).(\eta^{h,y}(1), \eta^{k,y}(1)) + \mathbb{E}Du(t-1,Y(1,y)).\zeta^{h,k,y}(1). \end{aligned}$$

Using Lemma 2.25, we get an exponential decreasing; thanks to Lemma 2.13 and to the estimates in the proof of Lemma 2.23 at time 1, we obtain

$$|D^2u(t,y).(h,k)| \leq Ce^{-c(t-1)}(1+|y|^2)|h|_{-\beta}|k|_{-\gamma}.$$

Then (2.23) easily follows. □

2.5.3. **Proof of Corollary 2.2.** The first estimate is a simple consequence of the Theorem, and of the exponential convergence to equilibrium of the continuous-time process - see (2.16). We then get

$$|\mathbb{E}[\phi(Y_m(\tau,y))] - \int_H \phi d\overline{\mu}| \leq C(1+|y|^3)(\frac{1}{m^{1/2-\kappa}} + \tau^{1/2-\kappa}) + C(1+|y|^2)e^{-cm\tau}.$$

If $\mu^\tau$ is an ergodic invariant probability measure of $(Y_m(\tau,.))_m$, then since $\phi$ is bounded for $\mu^\tau$-almost any $y \in H$ we have by the ergodic Theorem the following convergence when $M \to +\infty$:

$$\frac{1}{M}\sum_{m=1}^{M} \mathbb{E}[\phi(Y_m(\tau,y))] \to \int_H \phi(y)\mu^\tau(dy).$$

To conclude, it remains to choose a initial condition $y$ in this non-empty set, and to use Cesaro Lemma on the right-hand side of the estimate.

We notice that if $\mu^\tau$ is an invariant probability measure, not necessarily ergodic, having a finite moment of order 3, then it is enough to integrate the inequality above with respect to $\mu^\tau$.

## 2.6. Proof of the estimates

We need to control the terms given in (2.19), according to the decomposition (2.18). We recall that constants $C$ must be independent from the dimension $N$ and the final time $T = m\tau$.

2.6.1. **Estimate of $u(T,x) - \mathbb{E}[u(T-\tau, Y_1)]$.** The Markov property gives

$$u(T,y) = \mathbb{E}[\phi(Y(T,y))] = \mathbb{E}[u(T-\tau, Y(\tau,y))].$$

If $0 < \kappa < 1/2$, using Lemma 2.14 and Proposition 2.21 we get

$$|u(T,y) - \mathbb{E}[u(T-\tau,Y_1)]| \leq C(1+(T-\tau)^{-1/2+\kappa})e^{-\tilde{\mu}(T-\tau)}(\mathbb{E}|Y(\tau,y)-Y_1|^2_{-1/2+\kappa})^{1/2}(1+|y|^2).$$

We can write

$$Y(\tau,y) - Y_1 = (e^{\tau B} - R_\tau)y + \int_0^\tau e^{(\tau-s)B}G(Y(s,y))ds - \tau R_\tau G(y)$$
$$+ \int_0^\tau e^{(\tau-s)B}dW(s) - \sqrt{\tau}R_\tau\chi_1.$$

We use the following properties to estimate the first line in this equality:

$$|(-B)^{-1/2+\kappa}(e^{\tau B} - R_\tau)|_{\mathcal{L}(H)} \leq c\tau^{1/2-\kappa},$$
$$|e^{sB}|_{\mathcal{L}(H)} \leq 1 \text{ for } s \geq 0,$$
$$|R_\tau|_{\mathcal{L}(H)} \leq 1,$$
$$|(-B)^{-1/2+\kappa}.| \leq c|.|;$$

therefore the first line in the last expression is almost surely bounded by $C(\tau^{1/2-\kappa}+\tau)(1+|y|)$.

For the second line, we have

$$\begin{aligned}\mathbb{E}|(-B)^{-1/2+\kappa}\int_0^\tau e^{(\tau-s)B}dW(s)|^2 &= \mathbb{E}\int_0^\tau |(-B)^{-1/2+\kappa}e^{(\tau-s)B}|^2_{\mathcal{L}_2(H)}ds \\ &\leq \tau|(-B)^{-1/2+\kappa}|^2_{\mathcal{L}_2(H)} \\ &\leq c\tau;\end{aligned}$$

the last term is controlled in the same way: $\mathbb{E}|(-B)^{-1/2+\kappa}\sqrt{\tau}R_\tau\chi_1|^2 \leq c\tau$. Therefore we have

$$|u(T,x) - \mathbb{E}[u(T-\tau,Y_1)]| \leq C(1+|y|^3)(1+(T-\tau)^{-1/2+\kappa})e^{-\tilde{\mu}(T-\tau)}\tau^{1/2-\kappa}. \tag{2.30}$$

We thus understand that to obtain weak order 1/2 requires to be careful in the estimate. Here we used Lemma 2.14 instead of Lemma 2.25; otherwise looking at $\mathbb{E}|Y(\tau,y)-Y_1|^2$ would have not been sufficient. The control of the other terms must be done in the same spirit.

2.6.2. **Estimate of $a_k$.** We have

$$\begin{aligned}a_k &= \mathbb{E}\int_{t_k}^{t_{k+1}} < B\tilde{Y}(t) - B_\tau Y_k, Du(T-t,\tilde{Y}(t)) > dt \\ &= \mathbb{E}\int_{t_k}^{t_{k+1}} < (B-B_\tau)Y_k, Du(T-t,\tilde{Y}(t)) > dt \\ &+ \mathbb{E}\int_{t_k}^{t_{k+1}} < B(\tilde{Y}(t)-Y_k), Du(T-t,\tilde{Y}(t)) > dt \\ &:= a_k^1 + a_k^2.\end{aligned}$$

2.6.2.1. *Estimate of $a_k^1$.* We use the equality $B_\tau - B = \tau R_\tau B^2$. We also decompose $a_k^1$ using expression (2.9):

$$a_k^{1,1} = -\tau\mathbb{E}\int_{t_k}^{t_{k+1}} < R_\tau B^2 R_\tau^k y, Du(T-t,\tilde{Y}(t)) > dt$$
$$a_k^{1,2} = -\tau\mathbb{E}\int_{t_k}^{t_{k+1}} < R_\tau B^2 \tau\sum_{l=0}^{k-1} R_\tau^{k-l}G(Y_l), Du(T-t,\tilde{Y}(t)) > dt$$
$$a_k^{1,3} = -\tau\mathbb{E}\int_{t_k}^{t_{k+1}} < R_\tau B^2\sqrt{\tau}\sum_{l=0}^{k-1} R_\tau^{k-l}\chi_{l+1}, Du(T-t,\tilde{Y}(t)) > dt;$$

then $a_k^1 = a_k^{1,1} + a_k^{1,2} + a_k^{1,3}$.

(1) **Estimate of $a_k^{1,1}$**

The idea is to "share" $B^2$ between different factors - thanks to regularization properties of the semi-group $(R_\tau^k)_{k\in\mathbb{N}}$ and to Lemma 2.14 - in order to increase the order of convergence.

$$|a_k^{1,1}| \leq \tau\mathbb{E}\int_{t_k}^{t_{k+1}} |R_\tau(-B)^{1/2+2\kappa}|_{\mathcal{L}(H)}|(-B)^{1-\kappa}R_\tau^k|_{\mathcal{L}(H)}|y|_H|(-B)^{1/2-\kappa}Du(T-t,\tilde{Y}(t))|dt$$
$$\leq C|y|\tau\tau^{-1/2-2\kappa}t_k^{-1+\kappa}\int_{t_k}^{t_{k+1}}(1+(T-t)^{-1/2+\kappa})e^{-\tilde{\mu}(T-t)}\mathbb{E}(1+|\tilde{Y}(t)|^2)dt,$$

thanks to Lemma 2.12 and Proposition 2.21.

By taking expectation, thanks to Lemma 2.15 we have

$$\sum_{k=1}^{m-1}|a_k^{1,1}| \leq C|y|\tau^{1/2-2\kappa}\sum_{k=1}^{m-1}\frac{1}{t_k^{1-\kappa}}\int_{t_k}^{t_{k+1}}(1+\frac{1}{(T-t)^{1/2-\kappa}})e^{-\tilde{\mu}(T-t)}(1+|y|^2)dt$$
$$\leq C_\kappa(1+|y|^3)\tau^{1/2-2\kappa}\int_0^T\frac{1}{t^{1-\kappa}(T-t)^{1/2-\kappa}}dt$$
$$\leq T^{-(1/2-2\kappa)}C_\kappa(1+|y|^3)\tau^{1/2-2\kappa}\int_0^1\frac{1}{s^{1-\kappa}(1-s)^{1/2-\kappa}}ds,$$

and we obtain

$$\sum_{k=1}^{m-1}|a_k^{1,1}| \leq C(1+|y|^3)T^{-(1/2-2\kappa)}\tau^{1/2-2\kappa}. \tag{2.31}$$

(2) **Estimate of $a_k^{1,2}$**

First we write

$$|a_k^{1,2}| \leq C\tau\mathbb{E}\int_{t_k}^{t_{k+1}}|R_\tau(-B)^{1/2+2\kappa}|_{\mathcal{L}(H)}|\tau(-B)^{1-\kappa}\sum_{l=0}^{k-1}R_\tau^{k-l}G(Y_l)||(-B)^{1/2-\kappa}Du(T-t,\tilde{Y}(t))|dt.$$

Using Lemma 2.12, we can prove the following useful inequality: for $\tau \leq \tau_0$ and any $k \geq 1$

$$\tau\sum_{l=1}^{k}\frac{1}{(l\tau)^{1-\kappa}}\frac{1}{(1+\mu_0\tau)^{l\kappa}} \leq C_\kappa. \tag{2.32}$$

Indeed,

$$\tau\sum_{l=1}^{k}\frac{1}{(l\tau)^{1-\kappa}}\frac{1}{(1+\mu_0\tau)^{l\kappa}} \leq C\int_0^{t_k}\frac{1}{t^{1-\kappa}}\frac{1}{(1+\mu_0\tau)^{\kappa\frac{t}{\tau}}}dt$$
$$\leq \int_0^\infty\frac{1}{t^{1-\kappa}}e^{-t\frac{\kappa}{\tau}\log(1+\mu_0\tau)}dt$$
$$\leq \int_0^\infty\frac{1}{s^{1-\kappa}}e^{-s}ds\left(\frac{\tau}{\kappa\log(1+\mu_0\tau)}\right)^\kappa$$
$$\leq C_\kappa.$$

Since $G$ is supposed to be bounded, the estimate (2.32) yields

$$|\tau(-B)^{1-\kappa}\sum_{l=0}^{k-1}R_\tau^{k-l}G(Y_l)| \leq C\|G\|_\infty\tau\sum_{l=1}^{k}\frac{1}{(l\tau)^{1-\kappa}}\frac{1}{(1+\mu_0\tau)^{l\kappa}} \leq C_\kappa.$$

With Lemma 2.15 and Proposition 2.22, we can now write

$$|a_k^{1,2}| \leq C(1+|y|^2)\tau^{1/2-2\kappa}\int_{t_k}^{t_{k+1}}(1+\frac{1}{(T-t)^{1/2-\kappa}})e^{-\tilde{\mu}(T-t)}dt,$$

and we get

$$\sum_{k=1}^{m-1} |a_k^{1,2}| \leq C(1+|y|^2)\tau^{1/2-2\kappa}. \tag{2.33}$$

(3) **Estimate of $a_k^{1,3}$**

The analysis of this term is more complicated. We recall that since noise is white in space, for any $t>0$, $\mathbb{E}|(-B)^\gamma W^B(t)|^2 < +\infty$ if and only if $\gamma < 1/4$; as a consequence, the strategy used above to control $a_k^{1,1}$ can not be used directly - otherwise we could only obtain order $1/4$.

In [18], an integration by parts formula is used to deal with the lack of regularity of the stochastic integral appearing in the definition of $a_k^{1,3}$. An additionnal difficulty arises in our situation because the estimate given in Lemma 2.17 is not uniform with respect to time. Instead, we remark that the two problems - lack of regularity and bad time dependence - do not occur at the same time; therefore a decomposition of the interval $[0,t_k]$ into $[0,t_k-1]$ and $[t_k-1,t_k]$ - for $k$ large enough - can help to treat the problems separately.

Let us explain more concretely the situation at the continuous time level: we have to treat - at the finite dimensional approximation level, but with a bound independent from dimension - an expression involving $B^2\int_0^t e^{(t-s)B}dW(s)$. The idea to get rid of this expression is to use an integration by parts formula on the whole interval. We can also see that we can do this integration by parts only on a subinterval of size independent from $t$: indeed if $t\geq 1$

$$\int_0^t e^{(t-s)B}dW(s) = \int_0^{t-1} e^{(t-s)B}dW(s) + \int_{t-1}^t e^{(t-s)B}dW(s).$$

The first term is equal to $e^B\int_0^{t-1} e^{(t-1-s)B}dW(s)$, so that thanks to regularization properties of the semi-group $(e^{tB})$ - see Proposition 2.6 - multiplication by $B^2$ is possible - in other words we do not require an integration by parts to get a bound independent from the dimension; to treat the second term, the lack of regularity remains but can still be treated by the integration by parts, with the advantage of involving a smaller interval, where at the discrete time level below we can use a uniform control of all quantities, thanks to Lemma 2.17. Then the same idea of "sharing" $B^2$ can be used again to get order $1/2$.

Let us now explain develop this program for the discrete time situation: by using (2.10), we make the decomposition

$$\begin{aligned}
a_k^{1,3} &= -\tau\mathbb{E}\int_{t_k}^{t_{k+1}} < R_\tau B^2\sqrt{\tau}\sum_{l=0}^{k-1} R_\tau^{k-l}\chi_{l+1}, Du(T-t,\tilde{Y}(t))>dt \\
&= -\tau\mathbb{E}\int_{t_k}^{t_{k+1}} < \int_0^{t_k} R_\tau B^2 R_\tau^{k-l_s}dW(s), Du(T-t,\tilde{Y}(t))>dt \\
&= -\tau\mathbb{E}\int_{t_k}^{t_{k+1}} < \int_0^{(t_k-1)\vee 0} R_\tau B^2 R_\tau^{k-l_s}dW(s), Du(T-t,\tilde{Y}(t))>dt \\
&- \tau\mathbb{E}\int_{t_k}^{t_{k+1}} < \int_{(t_k-1)\vee 0}^{t_k} R_\tau B^2 R_\tau^{k-l_s}dW(s), Du(T-t,\tilde{Y}(t))>dt.
\end{aligned}$$

For the first term - which is equal to 0 when $t_k<1$ - we use the Cauchy-Schwarz inequality and we directly get

$$\begin{aligned}
|\mathbb{E} < \int_0^{(t_k-1)\vee 0} & R_\tau B^2 R_\tau^{k-l_s}dW(s), Du(T-t,\tilde{Y}(t))>| \\
&\leq (\mathbb{E}|\int_0^{(t_k-1)\vee 0} R_\tau B^2 R_\tau^{k-l_s}dW(s)|^2)^{1/2}(\mathbb{E}|Du(T-t,\tilde{Y}(t))|^2)^{1/2} \\
&\leq C(1+|y|^2)e^{-c(T-t)},
\end{aligned}$$

thanks to Lemmas 2.25 and 2.15, and to the following inequality - we remark that in the integral below $t_{k-l_s} \geq 1$:

$$\begin{aligned}
\mathbb{E}|\int_0^{(t_k-1)\vee 0} R_\tau B^2 R_\tau^{k-l_s} dW(s)|^2 &= \int_0^{(t_k-1)\vee 0} |R_\tau B^2 R_\tau^{k-l_s}|^2_{\mathcal{L}_2(H)} ds \\
&= \int_0^{(t_k-1)\vee 0} \mathrm{Tr}(R_\tau^2 B^4 R_\tau^{2(k-l_s)}) ds \\
&\leq \int_0^{(t_k-1)\vee 0} |(R_\tau^2 B^{4+1/2+\kappa} R_\tau^{2(k-l_s)}|_{\mathcal{L}(H)} ds \mathrm{Tr}((-B)^{-1/2-\kappa}) \\
&\leq C \int_0^{(t_k-1)\vee 0} \frac{1}{(1+\mu_0\tau)^{k-l_s} t_{k-l_s}^{4+1/2+\kappa}} ds \\
&\leq C \int_0^{(t_k-1)\vee 0} \frac{1}{(1+\mu_0\tau)^{k-l_s}} ds \\
&\leq C \int_0^{+\infty} \frac{1}{(1+\mu_0\tau)^{s/\tau}} ds \\
&\leq C,
\end{aligned}$$

when $\tau \leq \tau_0$. Then

$$|\tau \mathbb{E} \int_{t_k}^{t_{k+1}} < \int_0^{(t_k-1)\vee 0} R_\tau B^2 R_\tau^{k-l_s} dW(s), Du(T-t, \tilde{Y}(t)) > dt| \leq C \int_{t_k}^{t_{k+1}} e^{-\tilde{\mu}(T-t)} dt (1+|y|^2)\tau.$$

For the second term, we use the integration by parts formula of Lemma 2.16 to get

$$\begin{aligned}
\tau \mathbb{E} \int_{t_k}^{t_{k+1}} &< \int_{(t_k-1)\vee 0}^{t_k} R_\tau B^2 R_\tau^{k-l_s} dW(s), Du(T-t, \tilde{Y}(t)) > dt \\
&= -\tau \mathbb{E} \int_{t_k}^{t_{k+1}} \int_{(t_k-1)\vee 0}^{t_k} \mathrm{Tr}\left( R_\tau^{k-l_s} B^2 R_\tau D^2 u(T-t, \tilde{Y}(t)) D_s \tilde{Y}(t) \right) ds dt.
\end{aligned}$$

If $h \in H$, $(t_k-1)\vee 0 \leq s \leq t_k \leq t < t_{k+1}$, with (2.13) we see that

$$\begin{aligned}
D_s^h \tilde{Y}(t) &= D_s^h Y_k + \int_{t_k}^t (B_\tau D_s^h Y_k + R_\tau G(Y_k) D_s^h Y_k) d\lambda \\
&\quad + R_\tau D_s^h (W(t) - W(t_k)) \\
&= D_s^h Y_k + (t-t_k)(B_\tau D_s^h Y_k + R_\tau G(Y_k) D_s^h Y_k).
\end{aligned}$$

Therefore, since $|\tau B_\tau|_{\mathcal{L}(H)} \leq C$

$$|D_s^h \tilde{Y}(t)|_\beta \leq c |D_s^h Y_k|_\beta,$$

and taking supremum over $h$ with $|h| \leq 1$ we get

$$|(-B)^\beta D_s \tilde{Y}(t)|_{\mathcal{L}(H)} \leq c |(-B)^\beta D_s Y_k|_{\mathcal{L}(H)}. \tag{2.34}$$

The last quantity is estimated thanks to Lemma 2.17:

$$|D_s^h Y_k|_\beta \leq C(1+L_G\tau)^{k-l_s} (1 + \frac{1}{(1+\mu_0\tau)^{(1-\beta)(k-l_s)} t_{k-l_s}^\beta}) |h|.$$

When $\tau \leq \tau_0$ and $(t_k-1)\vee 0 \leq s \leq t_k \leq t < t_{k+1}$, we see that $(1+L_G\tau)^{k-l_s}$ is bounded by a constant.

We can then control the second term of $a_k^{1,3}$ with

$$\tau\mathbb{E}\int_{t_k}^{t_{k+1}}\int_{(t_k-1)\vee 0}^{t_k}|R_\tau(-B)^{1/2+2\kappa}|_{\mathcal{L}(H)}|(-B)^{1-3\frac{\kappa}{2}}R_\tau^{k-l_s}|_{\mathcal{L}(H)}\mathrm{Tr}((-B)^{-1/2-\frac{\kappa}{2}})$$
$$\times|(-B)^{1/2-\kappa/2}D^2u(T-t,\tilde{Y}(t))(-B)^{1/2-\kappa/2}|_{\mathcal{L}(H)}|(-B)^{\kappa}D_s\tilde{Y}(t)|_{\mathcal{L}(H)}dsdt$$
$$\leq C\tau^{1/2-2\kappa}\int_{t_k}^{t_{k+1}}\int_{(t_k-1)\vee 0}^{t_k}t_{k-l_s}^{-1+3\frac{\kappa}{2}}\frac{1}{(1+\mu_0\tau)^{(k-l_s)3\frac{\kappa}{2}}}(1+t_{k-l_s}^{-\kappa}\frac{1}{(1+\mu_0\tau)^{(k-l_s)(1-\kappa)}})ds$$
$$\times(1+\frac{1}{(T-t)^\eta}+\frac{1}{(T-t)^{1-\kappa}})e^{-\tilde{\mu}(T-t)}(1+|y|^2)dt,$$

using Proposition 2.22 and Lemmas 2.17 and 2.12.

On the one hand, we have

$$\int_{(t_k-1)\vee 0}^{t_k}t_{k-l_s}^{-1+3\frac{\kappa}{2}}\frac{1}{(1+\mu_0\tau)^{(k-l_s)3\frac{\kappa}{2}}}ds\leq\int_0^{t_k}\frac{1}{s^{1-3\frac{\kappa}{2}}}\frac{1}{(1+\mu_0\tau)^{3\frac{\kappa}{2}s/\tau}}ds\leq C<+\infty,$$

for $\tau\leq\tau_0$, thanks to (2.32).

On the other hand,

$$\sum_{k=1}^{m-1}\int_{t_k}^{t_{k+1}}(1+\frac{1}{(T-t)^\eta}+\frac{1}{(T-t)^{1-\kappa}})e^{-\tilde{\mu}(T-t)}dt\leq\int_0^{+\infty}(1+\frac{1}{t^\eta}+\frac{1}{t^{1-\kappa}})e^{-\tilde{\mu}t}dt<+\infty.$$

Therefore

$$\sum_{k=1}^{m-1}|a_k^{1,3}|\leq C(1+|y|^2)\tau^{1/2-2\kappa}. \tag{2.35}$$

2.6.2.2. *Estimate of* $a_k^2$. We decompose $a_k^2$ using the definition of $\tilde{Y}$ - see (2.13):

$$a_k^{2,1}=\mathbb{E}\int_{t_k}^{t_{k+1}}(t-t_k)<BB_\tau Y_k,Du(T-t,\tilde{Y}(t))>dt$$
$$a_k^{2,2}=\mathbb{E}\int_{t_k}^{t_{k+1}}(t-t_k)<BR_\tau G(Y_k),Du(T-t,\tilde{Y}(t))>dt$$
$$a_k^{2,3}=\mathbb{E}\int_{t_k}^{t_{k+1}}<\int_{t_k}^{t}BR_\tau dW(s),Du(T-t,\tilde{Y}(t))>dt;$$

then $a_k^2=a_k^{2,1}+a_k^{2,2}+a_k^{2,3}$.

(1) **Estimate of** $a_k^{2,1}$

Since $BB_\tau=R_\tau B^2$, $a_k^{2,1}$ is bounded by the same expression as $a_k^1$: by (2.31), (2.33), (2.35) we have

$$\sum_{k=1}^{m-1}|a_k^{2,1}|\leq C(1+|y|^3)(1+T^{-(1/2-2\kappa)})\tau^{1/2-2\kappa}. \tag{2.36}$$

(2) **Estimate of** $a_k^{2,2}$

We have

$$|a_k^{2,2}|\leq\tau\mathbb{E}\int_{t_k}^{t_{k+1}}|(-B)^{1/2+\kappa}R_\tau|_{\mathcal{L}(H)}|G(Y_k)||(-B)^{1/2-\kappa}Du(T-t,\tilde{Y}(t))|dt$$
$$\leq\|G\|_\infty\tau^{1/2-\kappa}\int_{t_k}^{t_{k+1}}(1+\frac{1}{(T-t)^{1/2-\kappa}})e^{-\tilde{\mu}(T-t)}dt.$$

We then have

$$\sum_{k=1}^{m-1}|a_k^{2,2}|\leq C\tau^{1/2-\kappa}. \tag{2.37}$$

(3) **Estimate of $a_k^{2,3}$**

We again use the integration by parts formula to rewrite $a_k^{2,3}$:

$$\begin{aligned} a_k^{2,3} &= \mathbb{E}\int_{t_k}^{t_{k+1}} < \int_{t_k}^{t} BR_\tau dW(s), Du(T-t,\tilde{Y}(t)) > dt \\ &= \mathbb{E}\int_{t_k}^{t_{k+1}} \int_{t_k}^{t} \mathrm{Tr}(R_\tau B D^2 u(T-t,\tilde{Y}(t)) D_s \tilde{Y}(t)) ds dt. \end{aligned}$$

From (2.13), for $t_k \le s \le t \le t_{k+1}$ we have $D_s^h \tilde{Y}(t) = R_\tau h$; as a consequence, we do not need to use the same trick as in the control of $a_k^{1,3}$.

Then we have

$$\begin{aligned} |a_k^{2,3}| &\le \mathbb{E}\int_{t_k}^{t_{k+1}} (t-t_k)\mathrm{Tr}(R_\tau B D^2 u(T-t,\tilde{Y}(t)) R_\tau) dt \\ &\le c\tau \int_{t_k}^{t_{k+1}} |R_\tau (-B)^{1/2+\kappa/2}|_{\mathcal{L}(H)} \mathrm{Tr}((-B)^{-1/2-\kappa/2}) |(-B)^{\kappa} R_\tau|_{\mathcal{L}(H)} \\ &\quad |(-B)^{1/2-\kappa/2} D^2 u(T-t,\tilde{Y}(t)) (-B)^{1/2-\kappa/2}|_{\mathcal{L}(H)} dt \\ &\le c(1+|y|^2)\tau^{1/2-3\kappa/2} \int_{t_k}^{t_{k+1}} (1 + \frac{1}{(T-t)^{\eta}} + \frac{1}{(T-t)^{1-\kappa}}) e^{-\tilde{\mu}(T-t)} dt. \end{aligned}$$

Therefore

$$\sum_{k=1}^{m-1} |a_k^{2,3}| \le C(1+|y|^2)\tau^{1/2-3\kappa/2}. \tag{2.38}$$

With the previous estimates on $a^1$ and $a^2$, we get

$$\sum_{k=1}^{m-1} |a_k| \le C(1+|y|^3)(1+T^{-(1/2-2\kappa)})\tau^{1/2-2\kappa}. \tag{2.39}$$

2.6.3. **Estimate of $b_k$.** We have

$$\begin{aligned} b_k &= \mathbb{E}\int_{t_k}^{t_{k+1}} < G(\tilde{Y}(t)) - R_\tau G(Y_k), Du(T-t,\tilde{Y}(t)) > dt \\ &= \mathbb{E}\int_{t_k}^{t_{k+1}} < (I-R_\tau) G(Y_k), Du(T-t,\tilde{Y}(t)) > dt \\ &+ \mathbb{E}\int_{t_k}^{t_{k+1}} < G(\tilde{Y}(t)) - G(Y_k), Du(T-t,\tilde{Y}(t)) > dt \\ &:= b_k^1 + b_k^2. \end{aligned}$$

2.6.3.1. *Estimate of $b_k^1$.* This term is easy to treat: we have

$$\begin{aligned} |b_k^1| &\le \mathbb{E}\int_{t_k}^{t_{k+1}} |(-B)^{-1/2+\kappa}(I-R_\tau)|_{\mathcal{L}(H)} |G(Y_k)| |(-B)^{1/2-\kappa} Du(T-t,\tilde{Y}(t))| dt \\ &\le C\tau^{1/2-\kappa} \int_{t_k}^{t_{k+1}} (1 + \frac{1}{(T-t)^{1/2-\kappa}}) e^{-\tilde{\mu}(T-t)} dt, \end{aligned}$$

where we have used Proposition 2.21, and the following inequality for $0 \le \beta \le 1$:

$$|(-B)^{-\beta}(I-R_\tau)|_{\mathcal{L}(H)} \le C_\beta \tau^{\beta}. \tag{2.40}$$

Then we see that

$$\sum_{k=1}^{m-1} |b_k^1| \le C\tau^{1/2-\kappa}. \tag{2.41}$$

2.6.3.2. *Estimate of* $b_k^2$. To estimate $|b_k^2|$, we write the scalar product in coordinates with respect to the orthonormal basis $(f_i)$, and then we expand the terms thanks to the Itô formula.

If we note $G_i =< G, f_i >$ and $\partial_i =< D., f_i >$, we have

$$< G(\tilde{Y}(t)) - G(Y_k), Du(T-t, \tilde{Y}(t)) >= \sum_i (G_i(\tilde{Y}(t)) - G_i(Y_k))\partial_i u(T-t, \tilde{Y}(t)).$$

The above sum is finite, because we work with finite dimensional approximations.

Itô formula gives for $t_k \leq t < t_{k+1}$

$$\begin{aligned} G_i(\tilde{Y}(t)) - G_i(Y_k) = &\frac{1}{2}\int_{t_k}^t \mathrm{Tr}(R_\tau R_\tau^* D^2 G_i(\tilde{Y}(s)))ds \\ &+ \int_{t_k}^t < B_\tau Y_k, DG_i(\tilde{Y}(s)) > ds \\ &+ \int_{t_k}^t < R_\tau G(Y_k), DG_i(\tilde{Y}(s)) > ds \\ &+ \int_{t_k}^t < DG_i(\tilde{Y}(s)), R_\tau dW(s) > . \end{aligned}$$

We naturally define $b_k^{2,j}$, for $j \in \{1,2,3,4\}$, and we now control each term.

(1) **Estimate of** $b_k^{2,1}$

By definition, we have

$$b_k^{2,1} = \int_{t_k}^{t_{k+1}} \mathbb{E}\frac{1}{2}\int_{t_k}^t \sum_i \mathrm{Tr}(R_\tau R_\tau^* D^2 G_i(\tilde{Y}(s)))ds\partial_i u(T-t, \tilde{Y}(t))dt.$$

Using the orthonormal basis $(f_k)_k$ given by assumption 2.4, and recalling that the sums are finite, we can calculate:

$$\begin{aligned} \sum_i \mathrm{Tr}(R_\tau R_\tau^* D^2 G_i(\tilde{Y}(s)))\partial_i u(T-t, \tilde{Y}(t)) &= \sum_i \mathrm{Tr}(D^2 G_i(\tilde{Y}(s))R_\tau R_\tau^*)\partial_i u(T-t, \tilde{Y}(t)) \\ &= \sum_i \sum_j < D^2 G_i(\tilde{Y}(s))\frac{1}{(1+\mu_j\tau)^2} f_j, f_j > \partial_i u(T-t, \tilde{Y}(t)) \\ &= \sum_i \sum_j \frac{1}{(1+\mu_j\tau)^2} D^2 G_i(\tilde{Y}(s)).(f_j, f_j)\partial_i u(T-t, \tilde{Y}(t)). \end{aligned}$$

Using the Cauchy-Schwarz inequality (where $j$ is fixed), we get

$$\begin{aligned} &|\sum_i D^2 G_i(\tilde{Y}(s)).(f_j, f_j)\partial_i u(T-t, \tilde{Y}(t))| \\ &\qquad \leq \left(\sum_i \frac{|D^2 G_i(\tilde{Y}(s)).(f_j, f_j)|^2}{\mu_i^{2\eta}}\right)^{1/2} \left(\sum_i \mu_i^{2\eta} |\partial_i u(T-t, \tilde{Y}(t))|^2\right)^{1/2}. \end{aligned}$$

The second factor of this expression is $|(-B)^\eta Du(T-t, \tilde{Y}(t))|_H$; we control it thanks to Proposition 2.21. The first factor is controlled thanks to Assumption 2.7:

$$\begin{aligned} \left(\sum_i \frac{|D^2 G_i(\tilde{Y}(s)).(f_j, f_j)|^2}{\mu_i^{2\eta}}\right)^{1/2} &= |(-B)^{-\eta} D^2 G(\tilde{Y}(s)).(f_j, f_j)| \\ &\leq C|f_j|_H |f_j|_H \leq C, \end{aligned}$$

since $(f_j)_j$ is an orthonormal system.

Therefore

$$
\begin{aligned}
&|\sum_i \mathrm{Tr}(R_\tau R_\tau^* D^2 G_i(\tilde{Y}(s)))\partial_i u(T-t,\tilde{Y}(t))| \\
&\qquad \le C(1+|y|^2)(1+\frac{1}{(T-t)^\eta})e^{-\tilde{\mu}(T-t)}\sum_{j=0}^{\infty}\frac{1}{(1+\mu_j\tau)^2} \\
&\qquad \le C(1+|y|^2)(1+\frac{1}{(T-t)^\eta})e^{-\tilde{\mu}(T-t)}\tau^{-1/2-\kappa}\sum_{j=0}^{\infty}\frac{(\mu_j\tau)^{1/2+\kappa}}{(1+\mu_j\tau)^2}\frac{1}{\mu_j^{1/2+\kappa}} \\
&\qquad \le C(1+|y|^2)(1+\frac{1}{(T-t)^\eta})e^{-\tilde{\mu}(T-t)}\tau^{-1/2-\kappa}.
\end{aligned}
$$

Then

$$
|b_k^{2,1}| \le C(1+|y|^2)\tau^{1/2-\kappa}\int_{t_k}^{t_{k+1}}(1+\frac{1}{(T-t)^\eta})e^{-\tilde{\mu}(T-t)}dt,
$$

and

$$
\sum_{k=1}^{m-1}|b_k^{2,1}| \le C(1+|y|^2)\tau^{1/2-\kappa}. \tag{2.42}
$$

(2) **Estimate of $b_k^{2,2}$**

Thanks to (2.9) and (2.10), we have

$$
\begin{aligned}
b_k^{2,2} &= \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\sum_i < B_\tau R_\tau^k y + B_\tau\tau\sum_{l=0}^{k-1}R_\tau^{k-l}G(Y_l), DG_i(\tilde{Y}(s)) > \partial_i u(T-t,\tilde{Y}(t))dsdt \\
&+ \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\sum_i < B_\tau\int_0^{t_k}R_\tau^{k-l_r}dW(r), DG_i(\tilde{Y}(s)) > \partial_i u(T-t,\tilde{Y}(t))dsdt \\
&:= b_k^{2,2,1} + b_k^{2,2,2}.
\end{aligned}
$$

*(i)* For the first term, recalling that $B_\tau = BR_\tau$ and that $G$ is bounded, we have

$$
\begin{aligned}
|b_k^{2,2,1}| &= |\mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t} < DG(\tilde{Y}(s)).(B_\tau R_\tau^k y + B_\tau\tau\sum_{l=0}^{k-1}R_\tau^{k-l}G(Y_l)), Du(T-t,\tilde{Y}(t)) > dsdt| \\
&\le \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}|(-B)^\kappa R_\tau|_{\mathcal{L}(H)}(|(-B)^{1-\kappa}R_\tau^k y| + \tau\sum_{l=0}^{k-1}|(-B)^{1-\kappa}R_\tau^{k-l}|_{\mathcal{L}(H)}|G(Y_l)|) \\
&\times |Du(T-t,\tilde{Y}(t))|dsdt \\
&\le C\tau^{1-\kappa}\int_{t_k}^{t_{k+1}}(1+|y|^2)e^{-\tilde{\mu}(T-t)}dt(t_k^{-1+\kappa}|y| + \tau\sum_{l=0}^{k-1}t_{k-l}^{(1-\kappa)}\frac{1}{(1+\mu_0\tau)^{(k-l)\kappa}}) \\
&\le C\tau^{1-\kappa}(1+|y|^3)\frac{1}{t_k^{1-\kappa}}\int_{t_k}^{t_{k+1}}e^{-\tilde{\mu}(T-t)}dt,
\end{aligned}
$$

if $\tau \le \tau_0$ - see (2.32).

Therefore

$$\begin{aligned}\sum_{k=1}^{m-1}|b_k^{2,2,1}| &\leq C\tau^{1-\kappa}(|y|+1)\sum_{k=1}^{m-1}\frac{1}{t_k^{1-\kappa}}\int_{t_k}^{t_{k+1}}e^{-\tilde{\mu}(T-t)}dt\\ &\leq C\tau^{1-\kappa}(1+|y|^3)\int_0^T\frac{1}{t^{1-\kappa}}e^{-\tilde{\mu}(T-t)}dt\\ &\leq C\tau^{1-\kappa}(1+|y|^3)\int_0^T\frac{1}{t^{1-\kappa}}\frac{C}{(T-t)^{1/2-\kappa}}dt\\ &\leq C\tau^{1-\kappa}(1+|y|^3)T^{-(1/2-2\kappa)}\int_0^1\frac{1}{s^{1-\kappa}}\frac{C}{(1-s)^{1/2-\kappa}}dt\\ &\leq C\tau^{1-\kappa}(1+|y|^3)T^{-(1/2-2\kappa)}.\end{aligned}$$

*(ii)* For the second term, we again use an integration by parts, after a decomposition of the time interval - as in the estimates for $a_k^{1,3}$. First,

$$\begin{aligned}&b_k^{2,2,2}=\mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\sum_i<B_\tau\int_0^{t_k}R_\tau^{k-l_r}dW(r),DG_i(\tilde{Y}(s))>\partial_i u(T-t,\tilde{Y}(t))dsdt\\ &=\mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\sum_i<B_\tau\int_0^{(t_k-1)\vee 0}R_\tau^{k-l_r}dW(r),DG_i(\tilde{Y}(s))>\partial_i u(T-t,\tilde{Y}(t))dsdt\\ &+\mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\sum_{i,j,m}<B_\tau\int_{(t_k-1)\vee 0}^{t_k}R_\tau^{k-l_r}f_m,f_j>d\beta_m(r)\partial_j G_i(\tilde{Y}(s))\partial_i u(T-t,\tilde{Y}(t))dsdt\\ &=:b_k^{2,2,2,1}+b_k^{2,2,2,2}.\end{aligned}$$

For $b_k^{2,,2,2,1}$, we can work directly and see that

$$\begin{aligned}|b_k^{2,2,2,1}|&\leq|\mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\sum_i<B_\tau\int_0^{(t_k-1)\vee 0}R_\tau^{k-l_r}dW(r),DG_i(\tilde{Y}(s))>\partial_i u(T-t,\tilde{Y}(t))dsdt|\\ &\leq\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\mathbb{E}|<DG(\tilde{Y}(s)).B_\tau\int_0^{(t_k-1)\vee 0}R_\tau^{k-l_r}dW(r),Du(T-t,\tilde{Y}(t))>|dsdt\\ &\leq\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}(\mathbb{E}|B_\tau\int_0^{(t_k-1)\vee 0}R_\tau^{k-l_r}dW(r)|^2)^{1/2}(\mathbb{E}|Du(T-t,\tilde{Y}(t))>|^2)^{1/2}dsdt\\ &\leq C\tau\int_{t_k}^{t_{k+1}}e^{-\tilde{\mu}(T-t)}(1+|y|^2),\end{aligned}$$

thanks to Lemmas 2.15, 2.25 and to the following estimate for $\tau\leq\tau_0$

$$\mathbb{E}|B_\tau\int_0^{(t_k-1)\vee 0}R_\tau^{k-l_r}dW(r)|^2\leq\mathbb{E}|B^2R_\tau\int_0^{(t_k-1)\vee 0}R_\tau^{k-l_r}dW(r)|^2\leq C,$$

thanks to the estimate proved to control $a_k^{1,3}$.

For $b_k^{2,2,2,2}$, we can write thanks to a Malliavin integration by parts and with the chain rule

$$b_k^{2,2,2,2} = \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\sum_{i,j,m} < B_\tau \int_{(t_k-1)\vee 0}^{t_k} R_\tau^{k-l_r} f_m, f_j > d\beta_m(r)\partial_j G_i(\tilde{Y}(s))\partial_i u(T-t,\tilde{Y}(t))dsdt$$

$$= \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\int_{(t_k-1)\vee 0}^{t_k}\sum_{i,j,m,n} < B_\tau R_\tau^{k-l_r} f_m, f_j > \partial^2_{j,n} G_i(\tilde{Y}(s)) < D_r^m \tilde{Y}(s), f_n > \partial_i u(T-t,\tilde{Y}(t))drdsdt$$

$$+ \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\int_{(t_k-1)\vee 0}^{t_k}\sum_{i,j,m,n} < B_\tau R_\tau^{k-l_r} f_m, f_j > \partial_j G_i(\tilde{Y}(s))\partial^2_{i,n} u(T-t,\tilde{Y}(t)) < D_r^m \tilde{Y}(t), f_n > drdsdt$$

$$= \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\int_{(t_k-1)\vee 0}^{t_k}\sum_{i,m} D^2 G_i(\tilde{Y}(s))(B_\tau R_\tau^{k-l_r} f_m, D_r^m \tilde{Y}(s))\partial_i u(T-t,\tilde{Y}(t))drdsdt$$

$$+ \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\int_{(t_k-1)\vee 0}^{t_k}\sum_{i,m} < \mathcal{B}_i(s,t) B_\tau R_\tau^{k-l_r} f_m, D_r^m \tilde{Y}(t) > drdsdt$$

$$= \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\int_{(t_k-1)\vee 0}^{t_k}\sum_{i} \mathrm{Tr}\left((D_r\tilde{Y}(s))^* D^2 G_i(\tilde{Y}(s)) B_\tau R_\tau^{k-l_r}\right)\partial_i u(T-t,\tilde{Y}(t))drdsdt$$

$$+ \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\int_{(t_k-1)\vee 0}^{t_k}\sum_{i} \mathrm{Tr}\left((D_r\tilde{Y}(t)^* \mathcal{B}_i(s,t) B_\tau R_\tau^{k-l_r}\right) drdsdt,$$

where we define a linear operator on $H$ by

$$< \mathcal{B}_i(s,t)h, k > = < DG_i(\tilde{Y}(s)), h > \sum_{n=0}^{+\infty} \partial^2_{i,n} u(T-t,\tilde{Y}(t)) < k, f_n >$$
$$= < DG_i(\tilde{Y}(s)), h > < D^2 u(T-t,\tilde{Y}(t)).f_i, k > .$$

We have $\sum_i < \mathcal{B}_i(s,t)h, k >= D^2 u(T-t,\tilde{Y}(t)).(DG(\tilde{Y}(s)).h, k)$, and

$$|\sum_i \mathcal{B}_i(s,t)|_{\mathcal{L}(H)} \leq |DG(\tilde{Y}(s))|_{\mathcal{L}(H)} |D^2 u(T-t,\tilde{Y}(t))|_{\mathcal{L}(H)};$$

so we can write, for $(t_k-1)\vee 0 \leq r \leq t_k$

$$|\sum_i \mathrm{Tr}\left((D_r\tilde{Y}(t)^* \mathcal{B}_i(s,t) B_\tau R_\tau^{k-l_r}\right)|$$
$$\leq |D_r\tilde{Y}(t)|_{\mathcal{L}(H)} |\sum_i \mathcal{B}_i(s,t)|_{\mathcal{L}(H)} |(-B)^{1-3\kappa/2} R_\tau^{k-l_r}|_{\mathcal{L}(H)} |R_\tau (-B)^{1/2+2\kappa}|_{\mathcal{L}(H)} \mathrm{Tr}((-B)^{-1/2-\kappa/2})$$
$$\leq C\tau^{-1/2-2\kappa} t_{k-l_r}^{-1+3\kappa/2} \frac{1}{(1+\mu_0\tau)^{(k-l_r)3\kappa/2}} e^{-\tilde{\mu}(T-t)},$$

using Proposition 2.22, Lemma 2.17 - since $(1+L_G\tau)^{k-l_r} \leq C$ - Lemma 2.12 and estimate (2.34).

The other term is a little more complicated, because we are not able to control $D^2 G(\tilde{Y}(s))$ in $H$. We proceed as in the estimate of $b_k^{2,1}$, and we directly calculate the trace.

$$|\sum_i \mathrm{Tr}\left((D_r\tilde{Y}(s))^* D^2 G_i(\tilde{Y}(s)) B_\tau R_\tau^{k-l_r}\right)\partial_i u(T-t,\tilde{Y}(t))|$$
$$\leq |D_r\tilde{Y}(s)|_{\mathcal{L}(H)} |\sum_i \mathrm{Tr}\left(D^2 G_i(\tilde{Y}(s)) B_\tau R_\tau^{k-l_r}\right)\partial_i u(T-t,\tilde{Y}(t))|$$
$$\leq |D_r\tilde{Y}(s)|_{\mathcal{L}(H)} \sum_{i,j} \frac{|D^2 G_i(\tilde{Y}(s)).(f_j,f_j)|}{\mu_i^\eta} \frac{\mu_j}{(1+\mu_j\tau)^{1+k-l_r}} \mu_i^\eta |\partial_i u(T-t,\tilde{Y}(t))|$$
$$\leq |D_r\tilde{Y}(s)|_{\mathcal{L}(H)} |(-B)^\eta Du(T-t,\tilde{Y}(t))|_H \sum_j |(-B)^{-\eta} D^2 G(\tilde{Y}(s)).(f_j,f_j)| \frac{\mu_j}{(1+\mu_j\tau)^{1+k-l_r}},$$

thanks to the Cauchy-Schwarz inequality.

By using the same analysis as in the estimation of $b_k^{2,1}$, we see that the above expression is bounded by

$$C|D_r\tilde{Y}(s)|_{\mathcal{L}(H)}|(-B)^\eta Du(T-t,\tilde{Y}(t))|_H \sum_j \frac{\mu_j}{(1+\mu_j\tau)^{1+k-l_r}};$$

but the last sum is equal to $\mathrm{Tr}(B_\tau R_\tau^{k-l_r})$, so that we see that indeed the two expressions in $b_k^{2,2,2}$ are bounded by the same expression.

Therefore

$$\begin{aligned}
&|b_k^{2,2,2,2}|\\
&\leq \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\int_{(t_k-1)\vee 0}^{t_k} C\tau^{-1/2-2\kappa}t_{k-l_r}^{-1+3\kappa/2}\frac{e^{-\tilde{\mu}(T-t)}}{(1+\mu_0\tau)^{(k-l_r)3\kappa/2}}(1+\frac{1}{(T-t)^\eta})(1+|y|^2)drdsdt\\
&\leq C(1+|y|^2)\tau^{1/2-2\kappa}\int_{t_k}^{t_{k+1}}(1+\frac{1}{(T-t)^\eta})e^{-\tilde{\mu}(T-t)}dt\int_0^{t_k} t_{k-l_r}^{-1+3\kappa/2}\frac{1}{(1+\mu_0\tau)^{(k-l_r)3\kappa/2}}dr\\
&\leq C(1+|y|^2)\tau^{1/2-2\kappa}\int_{t_k}^{t_{k+1}}(1+\frac{1}{(T-t)^\eta})e^{-\tilde{\mu}(T-t)}dt,
\end{aligned}$$

as already proved - see (2.32).

Now gathering estimates for $b_k^{2,2,2,1}$ and $b_k^{2,2,2,2}$, we obtain

$$\sum_{k=1}^{m-1}|b_k^{2,2,2}| \leq C(1+|y|^2)\tau^{1/2-2\kappa}. \tag{2.43}$$

(3) **Estimate of** $b_k^{2,3}$ We have

$$\begin{aligned}
b_k^{2,3} &= \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\sum_i < R_\tau G(Y_k), DG_i(\tilde{Y}(s)) > \partial_i u(T-t,\tilde{Y}(t))dsdt\\
&= \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t} < Du(T-t,\tilde{Y}(t)), DG(\tilde{Y}(s)).(R_\tau G(Y_k)) > dsdt.
\end{aligned}$$

Using that $G$ and $DG$ are bounded, we easily see that

$$|b_k^{2,3}| \leq C(1+|y|^2)\tau\int_{t_k}^{t_{k+1}} e^{-\tilde{\mu}(T-t)}dt,$$

and that

$$\sum_{k=1}^{m-1}|b_k^{2,3}| \leq C(1+|y|^2)\tau. \tag{2.44}$$

(4) **Estimate of** $b_k^{2,4}$

We use the integration by parts formula of Proposition 2.16 to get

$$\begin{aligned}
b_k^{2,4} &= \int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\sum_i < DG_i(\tilde{Y}(s)), R_\tau dW(s) > \partial_i u(T-t,\tilde{Y}(t))dt\\
&= \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\mathrm{Tr}\left((D_s\tilde{Y}(t))^* D^2u(T-t,\tilde{Y}(t))DG(\tilde{Y}(s))R_\tau\right)dsdt\\
&= \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t}\mathrm{Tr}\left(R_\tau D^2u(T-t,\tilde{Y}(t))DG(\tilde{Y}(s))R_\tau\right)dsdt,
\end{aligned}$$

using the identity $D_s^h\tilde{Y}(t) = R_\tau h$ when $t_k \leq s \leq t \leq t_{k+1}$, as in the estimate of $a_k^{2,3}$.

Now

$$|b_k^{2,4}| \leq \mathbb{E}\int_{t_k}^{t_{k+1}}\int_{t_k}^{t} |(R_\tau(-B)^{1/2+\kappa}|_{\mathcal{L}(H)}|DG(\tilde{Y}(s))|_{\mathcal{L}(H)}|R_\tau|_{\mathcal{L}(H)} \times |D^2u(T-t,\tilde{Y}(t))|_{\mathcal{L}(H)}\mathrm{Tr}((-B)^{-1/2-\kappa})dsdt \leq C(1+|y|^2)\tau^{1/2-\kappa}\int_{t_k}^{t_{k+1}}(1+\frac{1}{(T-t)^\eta})e^{-\tilde{\mu}(T-t)}dt,$$

and

$$\sum_{k=1}^{m-1}|b_k^{2,4}| \leq C(1+|y|^2)\tau^{1/2-\kappa}. \tag{2.45}$$

2.6.3.3. *Estimate of $b_k$: conclusion.* With (2.41), (2.42), (2.43), (2.44) and (2.45), we get

$$\sum_{k=1}^{m-1}|b_k| \leq C\tau^{1/2-2\kappa}. \tag{2.46}$$

2.6.4. **Estimate of $c_k$.** We have, using the symmetry of $R_\tau$,

$$\frac{1}{2}I-\frac{1}{2}R_\tau R_\tau^* = R_\tau(I-R_\tau)^* + \frac{1}{2}(I-R_\tau)(I-R_\tau)^*,$$

and

$$\begin{aligned} c_k &= \frac{1}{2}\mathbb{E}\int_{t_k}^{t_{k+1}}\mathrm{Tr}((I-R_\tau R_\tau^*)D^2u(T-t,\tilde{Y}(t)))dt \\ &= \frac{1}{2}\mathbb{E}\int_{t_k}^{t_{k+1}}\mathrm{Tr}((I-R_\tau)(I-R_\tau)^*D^2u(T-t,\tilde{Y}(t)))dt \\ &+ \mathbb{E}\int_{t_k}^{t_{k+1}}\mathrm{Tr}(R_\tau(I-R_\tau)^*D^2u(T-t,\tilde{Y}(t)))dt \\ &:= c_k^1 + c_k^2. \end{aligned}$$

2.6.4.1. *Estimate of $c_k^1$.* We have, using inequality (2.40)

$$\begin{aligned} |c_k^1| &\leq \frac{1}{2}\mathbb{E}\int_{t_k}^{t_{k+1}}\mathrm{Tr}((-B)^{-1/2+\kappa}(I-R_\tau)^2(-B)^{-1/2+\kappa}) \\ &\qquad \times |(-B)^{1/2-\kappa}D^2u(T-t,\tilde{Y}(t))(-B)^{1/2-\kappa}|_{\mathcal{L}(H)}dt \\ &\leq C(1+|y|^2)\int_{t_k}^{t_{k+1}}|(-B)^{-1/2+3\kappa}(I-R_\tau)|_{\mathcal{L}(H)}|I-R_\tau|_{\mathcal{L}(H)}\mathrm{Tr}((-B)^{-1/2-\kappa}) \\ &\qquad \times (1+\frac{1}{(T-t)^\eta}+\frac{1}{(T-t)^{1-\kappa}})e^{-\tilde{\mu}(T-t)}dt \\ &\leq C(1+|y|^2)\tau^{1/2-3\kappa}\int_{t_k}^{t_{k+1}}(1+\frac{1}{(T-t)^\eta}+\frac{1}{(T-t)^{1-\kappa}})e^{-\tilde{\mu}(T-t)}dt. \end{aligned}$$

Then

$$\sum_{k=1}^{m-1}|c_k^1| \leq C(1+|y|^2)\tau^{1/2-3\kappa}. \tag{2.47}$$

2.6.4.2. *Estimate of* $c_k^2$. We have, using inequality (2.40)

$$\begin{aligned}
|c_k^2| &\leq \mathbb{E}\int_{t_k}^{t_{k+1}} \mathrm{Tr}((-B)^{-1/2+\kappa}R_\tau(I-R_\tau)(-B)^{-1/2+\kappa}) \\
&\qquad \times |(-B)^{1/2-\kappa}D^2u(T-t,\tilde{Y}(t))(-B)^{1/2-\kappa}|_{\mathcal{L}(H)}dt \\
&\leq C(1+|y|^2)\int_{t_k}^{t_{k+1}} |(-B)^{-1/2+\kappa}(I-R_\tau)(-B)^{2\kappa}|_{\mathcal{L}(H)}\mathrm{Tr}((-B)^{-1/2-\kappa}) \\
&\qquad \times (1+\frac{1}{(T-t)^\eta}+\frac{1}{(T-t)^{1-\kappa}})e^{-\tilde{\mu}(T-t)}dt \\
&\leq C(1+|y|^2)\tau^{1/2-3\kappa}\int_{t_k}^{t_{k+1}}(1+\frac{1}{(T-t)^\eta}+\frac{1}{(T-t)^{1-\kappa}})e^{-\tilde{\mu}(T-t)}dt.
\end{aligned}$$

Then

$$\sum_{k=1}^{m-1}|c_k^2| \leq C(1+|y|^2)\tau^{1/2-3\kappa}. \tag{2.48}$$

2.6.4.3. *Estimate of* $c_k$*: conclusion.* With (2.47) and (2.48), we get

$$\sum_{k=1}^{m-1}|c_k| \leq C(1+|y|^2)\tau^{1/2-3\kappa}. \tag{2.49}$$

2.6.5. **Conclusion.** We put together estimates (2.39), (2.46), (2.49) and (2.30); then passing to the limit with respect to dimension, we get the result.

# Chapitre 3

# Schéma numérique multi-échelle de type HMM pour un système d'EDPS hautement oscillant


Résumé

On applique les résultats des chapitres précédents pour définir et analyser la discrétisation temporelle du système d'EDPS présenté dans le chapitre 1. Le principe de la méthode est de remplacer l'approximation de la composante lente par celle du processus moyenné (selon le chapitre 1), et d'utiliser en parallèlle un schéma numérique sur l'équation lente afin d'approcher un coefficient manquant dépendant de la loi invariante de l'équation rapide (selon le chapitre 2).

Sous les mêmes hypothèses de dissipation forte et faible du chapitre 1, on obtient des estimations de l'erreur, avec ordres de convergence aux sens respectivement fort et faible.



Ce travail a fait l'objet d'une prépublication sous le titre
**Analysis of a HMM time-discretization scheme for a system of Stochastic PDEs**.

**Chapitre 3. Analysis of a HMM time-discretization scheme for a system of Stochastic PDEs**

## 3.1. Introduction

In this Chapter, we are interested in the numerical approximation of a randomly-perturbed system of reaction-diffusion equations that can be written

$$\begin{aligned} \frac{\partial x^\epsilon(t,\xi)}{\partial t} &= \frac{\partial^2 x^\epsilon(t,\xi)}{\partial \xi^2} + f(\xi, x^\epsilon(t,\xi), y^\epsilon(t,\xi)), \\ \frac{\partial y^\epsilon(t,\xi)}{\partial t} &= \frac{1}{\epsilon}\frac{\partial^2 y^\epsilon(t,\xi)}{\partial \xi^2} + \frac{1}{\epsilon} g(\xi, x^\epsilon(t,\xi), y^\epsilon(t,\xi)) + \frac{1}{\sqrt{\epsilon}}\frac{\partial \omega(t,\xi)}{\partial t}, \end{aligned} \tag{3.1}$$

for $t \geq 0, \xi \in (0,1)$, with initial conditions $x^\epsilon(0,\xi) = x(\xi)$ and $y^\epsilon(0,\xi) = \xi$, and homogeneous Dirichlet boundary conditions $x^\epsilon(t,0) = x^\epsilon(t,1) = 0, y^\epsilon(t,0) = y^\epsilon(t,1) = 0$. The stochastic perturbation $\frac{\partial \omega(t,\xi)}{\partial t}$ is a space-time white noise and $\epsilon > 0$ is a small parameter.

In the first Chapter, we have proved that an Averaging Principle holds for such a system, and we have exhibited an order of convergence - with respect to $\epsilon$ - in a strong and in a weak sense: the slow component $x^\epsilon$ is approximated thanks to the solution of an averaged equation. In this Chapter, we analyze a numerical method of time discretization which reproduces this averaging effect at the discrete time level. More precisely, our aim is to build a numerical approximation of the slow component $x^\epsilon$, taking care of the stiffness induced by the time scale separation. The Heterogeneous Multiscale Method - HMM - procedure can be used, as it is done in [25] for SDEs of the same kind. First we recall the general principle of such a method, which has been developed in various contexts, both deterministic and stochastic - see the review article [24] and the references therein, as well as [22], [23], [64]. We can also mention the paper [1], which justifies the introduction of indirect methods for systems of SDEs with two time-scales.

Such systems of equations as (3.1) with two time-scales may be used for the study of problems in chemistry or in biology, and in finance, when two phenomena occur at different speeds. An historical example where averaging occurs concerns celestial mechanics. The book [56] is a good reference for multiscale systems of equations. Precise assumptions on the coefficients $f$ and $g$ are given in Section 3.3.

In system (3.1), the two components evolve at different time scales; $x^\epsilon$ is the slow component of the problem, while the fast component $y^\epsilon$ has fast variations in time. We are indeed interested in evaluating the slow component, which can be thought as the mathematical model for a phenomenon appearing at the natural time scale of the experiment, whereas the fast component can often be interpreted as data for the slow component, taking care of effects at a faster time scale. Instead of using a direct numerical method, which might require a very small time step size because of the fast component, we use a different solver for each time scale: a macrosolver and a microsolver. The macrosolver leads to the approximation of the slow component; it takes into account data from the evolution at the fast time scale. The microsolver is then a procedure for estimating the unknown necessary data, using the evolution at the microtime scale, which also depends on the evolution at the macrotime scale. We focus on the situation where both solvers are constructed thanks to a semi-implicit Euler scheme - while in [25] the analysis is more general: first, we obtain a simple method which already contains the fundamental technical difficulties of the Heterogeneous Multiscale Methods, and we can use the results of [5] about the approximation of the invariant measure of a SPDE thanks to a numerical scheme.

In Section 3.4.1, we state the two main Theorems of this article: we show a strong convergence result - Theorem 3.11 - as well as a weak convergence result - Theorem 3.12 - which are similar to the available results for SDEs. Compared to [25], we propose modified and simplified proofs leading to apparently weaker error estimates; we made this choice for various reasons. First, even if apparently we get weaker estimates, under an appropriate choice of the parameters the cost of the method remains of the same order. Second,

the generalization of the finite dimensional results would not yield the same bounds, due to the regularity assumptions we make on the nonlinear coefficients of the equations. Finally, we can extend the weak convergence result to the situation where the fast equation only satisfies a weak dissipativity assumption.

In the case of a linear fast equation - when $g$ is equal to 0 - it is well-known that the second equation in (3.1) is dissipative. In the general case, we make assumptions on $g$ so that this property is preserved for $y^\epsilon$ - see Assumptions 3.7 and 3.8 below. The fast equation with frozen slow component - defined by (3.16) in the abstract framework - then admits a unique invariant probability measure, which is ergodic and strongly mixing - with exponential convergence to equilibrium. Under the strict dissipativity condition 3.7, we can prove that the averaging principle holds in the strong and in the weak sense; moreover the "fast" numerical scheme has the same asymptotic behaviour as the continuous time equation. If we only assume weak dissipativity of Assumption 3.8, the averaging principle only holds in the weak sense, and we can not prove uniqueness of the invariant law of the fast numerical scheme. Nevertheless, in the general setting [5] gives an approximation result of the invariant law of the continuous time equation with the numerical method which is used to prove Theorem 3.12; the order of convergence is 1/2 with respect to the timestep size - the precise result is recalled in Theorem 3.24 and is studied in Chapter 2.

The Chapter is organized as follows. In Section 3.2, we give the definition of the numerical scheme. We then state the main assumptions made on the system of equations. In Section 3.4.1 we state the two main Theorems proved in this article, while in Section 3.4.2 we compare the efficiency of the HMM scheme with a direct one in order to justify the use of a new method. Before proving the theorems, we give some useful results on the numerical schemes. Finally the last two sections contain the proofs of the strong and weak convergence theorems.

## 3.2. Description of the numerical scheme

Instead of working directly with system (3.1), we work with abstract stochastic evolution equations in Hilbert spaces $H$:

$$\begin{aligned} dX^\epsilon(t) &= \big(AX^\epsilon(t) + F(X^\epsilon(t), Y^\epsilon(t))\big)dt \\ dY^\epsilon(t) &= \frac{1}{\epsilon}\big(BY^\epsilon(t) + G(X^\epsilon(t), Y^\epsilon(t))\big)dt + \frac{1}{\sqrt{\epsilon}}dW(t), \end{aligned} \tag{3.2}$$

with initial conditions $X^\epsilon(0) = x \in H$, $Y^\epsilon(0) = y \in H$.

To get system (3.2) from (3.1), we take $H = L^2(0,1)$; the linear operators $A$ and $B$ are Laplace operators with homogeneous Dirichlet boundary conditions and the nonlinearities $F$ and $G$ are Nemytskii operators - see Example 3.9. The process $(W(t))_{t\geq 0}$ is a cylindrical Wiener process on $H$. For precise assumptions on the coefficients, we refer to Section 3.3.

We recall the idea of the Averaging Principle - proved in the previous article [4]: when $\epsilon$ goes to 0, $X^\epsilon$ can be approximated by $\overline{X}$ defined by the averaged equation

$$\begin{aligned} \frac{d\overline{X}(t)}{dt} &= A\overline{X}(t) + \overline{F}(\overline{X}(t)) \\ \overline{X}(t=0) &= x \in H; \end{aligned} \tag{3.3}$$

the error is controlled in a strong sense by $C\epsilon^{1/2-r}$ and in a weak sense by $C\epsilon^{1-r}$, where $r > 0$ can be chosen as small as necessary, and where $C$ is a constant.

The averaged coefficient $\overline{F}$ - see (3.19) - satisfies

$$\overline{F}(X) = \int_H F(X,y)\mu^X(dy) = \lim_{s\to+\infty} \mathbb{E}[F(X, Y_X(s,y))],$$

where $\mu^X$ is the unique invariant probability measure of the fast process $Y_X$ with frozen slow component - more details are given in Section 3.5.1.

To apply the HMM strategy, we need to define a macrosolver and a microsolver. We denote by $\Delta t$ the macrotime step size, and by $\delta t$ the microtime step size. Let also $T > 0$ be a given final time.

The construction of the macrosolver is deeply based on the averaging principle: for $n\Delta t \leq T$ $X^\epsilon(n\Delta t)$ can be approximated by $\overline{X}(n\Delta t)$. If the averaged coefficient $\overline{F}$ was known, one could build an approximation

with a deterministic numerical scheme on the averaged equation; nevertheless in general it is not the case, and the idea is to calculate an approximation of this coefficient on-the-fly, by using the microsolver.

Therefore the macrosolver is defined in the following way: for any $0 \leq n \leq \lfloor \frac{T}{\Delta t} \rfloor := n_0$,

$$X_{n+1} = X_n + \Delta t A X_{n+1} + \Delta t \tilde{F}_n,$$

with the initial condition $X_0 = x$. $\tilde{F}_n$ has to be defined; before that, we notice that the above definition leads to a semi-implicit Euler scheme - we use implicitness on the linear part, but the nonlinear part is explicit. If we define a bounded linear operator on $H$ by $S_{\Delta t} = (I - \Delta t A)^{-1}$, we rather use the following explicit formula

$$X_{n+1} = S_{\Delta t} X_n + \Delta t S_{\Delta t} \tilde{F}_n. \tag{3.4}$$

We want $\tilde{F}_n$ to be an approximation of $\overline{F}(X_n)$. The role of the microsolver is to give an approximation of $Y_{X_n}$ - the fast process with frozen slow component $X_n$, when $n$ is fixed; moreover we compute a finite number $M$ of independent replicas of the process, in order to approximate theoretical expectations by discrete averages over different realizations of the random variables, in a Monte-Carlo spirit. Therefore the microsolver is defined in the following way: we fix a realization index $j \in \{1, \dots, M\}$, and a macrotime step $n$; then for any $m \geq 0$

$$Y_{n,m+1,j} = Y_{n,m,j} + \frac{\delta t}{\epsilon} B Y_{n,m+1,j} + \frac{\delta t}{\epsilon} G(X_n, Y_{n,m,j}) + \sqrt{\frac{\delta t}{\epsilon}} \zeta_{n,m+1,j}.$$

As above we can give an explicit formula

$$Y_{n,m+1,j} = R_{\frac{\delta t}{\epsilon}} Y_{n,m,j} + \frac{\delta t}{\epsilon} R_{\frac{\delta t}{\epsilon}} G(X_n, Y_{n,m,j}) + \sqrt{\frac{\delta t}{\epsilon}} R_{\frac{\delta t}{\epsilon}} \zeta_{n,m+1,j}, \tag{3.5}$$

where for any $\tau > 0$ $R_\tau = (I - \tau B)^{-1}$.

The noises $\zeta_{n,m,j}$ are defined by

$$\zeta_{n,m+1,j} = \frac{W^{(n,j)}_{(m+1)\delta t} - W^{(n,j)}_{m\delta t}}{\sqrt{\delta t}},$$

where $(W^{(n,j)})_{1 \leq j \leq M, 0 \leq n \leq n_0}$ are independent cylindrical Wiener processes on $H$. It is essential to use independent noises at each macrotime step. It is important to remark that this equation is well-posed in the Hilbert space $H$ since $R_\tau$ is a Hilbert-Schmidt operator from $H$ to $H$, under assumptions given in Section 3.3.

The missing definition can now be written: $\tilde{F}_n$ is given by

$$\tilde{F}_n = \frac{1}{MN} \sum_{j=1}^{M} \sum_{m=n_T}^{n_T+N-1} F(X_n, Y_{n,m,j}). \tag{3.6}$$

$n_T$ is the number of microtime steps that are not used in the evaluation of the average in $\tilde{F}_n$, while $N$ is the number of microtime steps that are then used for this evaluation. Each macrotime step then requires the computation of $m_0 = n_T + N - 1$ values of the microsolver.

At each macrotime step $Y_{n,m,j}$ must be initialized at time $m = 0$. In our proofs, this is not as important as in [25], but for definiteness we use the same method: we initialize with the last value computed during the previous macrotime step:

$$\begin{aligned} Y_{n+1,0,j} &= Y_{n,m_0,j} \text{ for } n \geq 0, \\ Y_{0,0,j} &= y. \end{aligned} \tag{3.7}$$

The aim of the analysis for HMM schemes is to prove that under an appropriate choice of the parameters $n_T, N, M$ of the scheme, we can bound the error by expressions of the following kind, where $r > 0$ is chosen as small as necessary: for $n = n_0 = \lfloor \frac{T}{\Delta t} \rfloor$, we have the strong error estimate

$$\mathbb{E}|X^\epsilon(n\Delta t) - X_n| \leq C \left( \epsilon^{1/2-r} + \Delta t^{1-r} + \left( \frac{\delta t}{\epsilon} \right)^{1/2-r} \right),$$

and if $\Phi$ is a test function of class $\mathcal{C}^2$, bounded and with bounded derivatives, we have the weak error estimate

$$|\mathbb{E}\Phi(X^\epsilon(n\Delta t)) - \mathbb{E}\Phi(X_n)| \leq C\left(\epsilon^{1-r} + \Delta t^{1-r} + \left(\frac{\delta t}{\epsilon}\right)^{1/2-r}\right).$$

The origin of the three error terms appears clearly in the proofs - see Sections 3.6 and 3.7: the first one is the averaging error, the second one is the error in a deterministic scheme with the macrotime step, and the third one is the weak error in a scheme for stochastic equations with the microtime step. We recall that in the SPDE case the strong order of the semi-implicit Euler scheme for the microsolver used here is 1/4, while the weak order is 1/2, while in the SDE situation the respective orders are 1/2 and 1. The macrosolver is deterministic, so that the order is 1. Precise results for any choice of $n_T, N, M$ are given in Theorems 3.11 and 3.12 below, while the choice of these parameters is explained in Section 3.4.2.

## 3.3. Assumptions

As mentioned above, System (3.2) satisfies an averaging principle, and strong and weak order of convergence can be given. The HMM method relies on that idea. The natural assumptions are basically the same as the hypothesis needed to prove those results, but must be strengthened sometimes.

3.3.1. **Assumptions on the linear operators.** First, we have to specify some properties of the linear operators $A$ and $B$ coming into the definition of system (3.2); we assume that the linear parts are of parabolic type, with space variable $\xi \in (0,1)$.

We assume that $A$ and $B$ are unbounded linear operators, with domains $D(A)$ and $D(B)$, which satisfy the following assumptions:

**Assumptions 3.1.** (1) *There exist complete orthonormal systems of $H$, $(e_k)_{k\in\mathbb{N}}$ and $(f_k)_{k\in\mathbb{N}}$, and $(\lambda_k)_{k\in\mathbb{N}}$ and $(\mu_k)_{k\in\mathbb{N}}$ non-decreasing sequences of real positive numbers such that:*

$$Ae_k = -\lambda_k e_k \text{ for all } k \in \mathbb{N}$$
$$Bf_k = -\mu_k f_k \text{ for all } k \in \mathbb{N}.$$

*We use the notations $\lambda := \lambda_0 > 0$ and $\mu := \mu_0 > 0$ for the smallest eigenvalues of $A$ and $B$.*

(2) *For every $k \in \mathbb{N}$, $f_k$ is Lipschitz continuous and bounded on $[0,1]$, with a uniform control with respect to $k$: there exists $C > 0$ such that for any $\xi_1, \xi_2 \in [0,1]$*

$$|f_k(\xi_1)| \leq C \quad \text{and} \quad |f_k(\xi_1) - f_k(\xi_2)| \leq C\sqrt{\mu_k}|\xi_1 - \xi_2|.$$

(3) *The sequences $(\lambda_k)$ and $(\mu_k)$ go to $+\infty$; moreover we have some control of the behaviour of $(\mu_k)$ given by:*

$$\sum_{k=0}^{+\infty} \frac{1}{\mu_k^\alpha} < +\infty \Leftrightarrow \alpha > 1/2.$$

In the abstract setting, powers of $-A$ and $-B$, with their domains can be easily defined:

**Definition 3.2.** *For $a, b \in [0,1]$, we define the operators $(-A)^a$ and $(-B)^b$ by*

$$(-A)^a x = \sum_{k=0}^{\infty} \lambda_k^a x_k e_k \in H, \quad \text{and} \quad (-B)^b y = \sum_{k=0}^{\infty} \mu_k^b y_k f_k \in H,$$

*with domains*

$$D(-A)^a = \left\{ x = \sum_{k=0}^{+\infty} x_k e_k \in H; |x|^2_{(-A)^a} := \sum_{k=0}^{+\infty} (\lambda_k)^{2a} |x_k|^2 < +\infty \right\};$$
$$D(-B)^b = \left\{ y = \sum_{k=0}^{+\infty} y_k f_k \in H, |y|^2_{(-B)^b} := \sum_{k=0}^{+\infty} (\mu_k)^{2b} |y_k|^2 < +\infty \right\}.$$

On $D(-A)^a$, the norm $|\,.\,|_{(-A)^a}$ and the Sobolev norm of $H^{2a}$ are equivalent: when $x$ belongs to a space $D(-A)^a$, the exponent $a$ represents some regularity of the function $x$.

When $a \geq 0$, we can also define a bounded linear operator $(-A)^{-a}$ in $H$ with

$$(-A)^{-a}x = \sum_{k=0}^{+\infty} \lambda_k^{-a} x_k \in H,$$

where $x = \sum_{k=0}^{+\infty} x_k e_k \in H$.

Under the previous assumptions on the linear coefficients, the following stochastic integral is well-defined in $H$, for any $t \geq 0$, when $(W(t))_{t\geq 0}$ is a cylindrical Wiener process on $H$ - see [15] for the definition and the properties of $W$:

$$W^B(t) = \int_0^t e^{(t-s)B} dW(s). \tag{3.8}$$

It is called a stochastic convolution, and it is the unique mild solution of

$$dZ(t) = BZ(t)dt + dW(t), \quad Z(0) = 0.$$

Under the second condition of Assumption 3.1, there exists $\delta > 0$ such that for any $t > 0$ we have $\int_0^t \frac{1}{s^\delta} |e^{sB}|^2_{\mathcal{L}_2(H)} ds < +\infty$; it can then be proved that $W^B$ has continuous trajectories - via the *factorization method*, see [15] - and that for any $1 \leq p < +\infty$, any $0 \leq \gamma < 1/4$, there exists a constant $C(p,\gamma) < +\infty$ such that for any $t \geq 0$

$$\mathbb{E}|W^B(t)|^p_{(-B)^\gamma} \leq C(p,\gamma). \tag{3.9}$$

3.3.2. **Assumptions on the nonlinear coefficients.** We now give the Assumptions on the nonlinear coefficients $F, G : H \times H \to H$. First, we need some regularity properties:

**Assumptions 3.3.** *We assume that there exists $0 \leq \eta < \frac{1}{2}$ and a constant $C$ such that the following directional derivatives are well-defined and controlled:*

- *For any $x, y \in H$ and $h \in H$, $|D_x F(x,y).h| \leq C|h|_H$ and $|D_y F(x,y).h| \leq C|h|_H$.*
- *For any $x, y \in H$, $h, k \in H$, if the right-hand side is finite we have*

$$|D^2_{\cdot\cdot} F(x,y).(h,k)|_H \leq C|h|_H |k|_{(-\mathcal{D})},$$

*where $D^2_{\cdot\cdot}$ stands for a second directional derivative with respect to either $x$ or $y$ and with $\mathcal{D} = (-A)^\eta$ or $\mathcal{D} = (-B)^\eta$ according to the situation.*

*We moreover assume that $F$ is bounded.*

We also need:

**Assumptions 3.4.** *For $\eta$ defined in the previous Assumption 3.3, we have for any $x, y \in H$ and $h, k \in H$*

$$|(-A)^{-\eta} D^2_{\cdot\cdot} F(x,y).(h,k)| \leq C|h|_H |k|_H.$$

We assume that the fast equation is a gradient system: for any $x$ the nonlinear coefficient $G(x,.)$ is the derivative of some potential $U$. We also assume regularity assumptions as for $F$.

**Assumptions 3.5.** *The function $G$ is defined through $G(x,y) = \nabla_y U(x,y)$, for some potential $U : H \times H \to \mathbb{R}$. Moreover we assume that $G$ is bounded, and that the regularity assumptions given in the Assumption 3.3 are also satisfied for $G$.*

This condition only plays a role when we get the expression of the invariant probability measure of the fast process with frozen slow component, given by (3.18). This expression is essential to study the regularity properties of the averaged coefficient with respect to the $x$ variable given in Proposition 3.18. Such results are required to obtain the orders of convergence in the averaging principle, given in [4]. We recall that the averaging principle still holds if this gradient assumption is removed - see [10] - but that the order of the convergence is then a priori unknown.

For $G$, we need a stronger hypothesis than for $F$ - in order to get Proposition 3.18. Assumption 3.4 becomes:

**Assumptions 3.6.** *We have for any $x, y \in H$, $h, k \in H$, $z \in L^\infty(0,1)$*

$$| < D_{xx}G(x,y).(h,k), z >_H | \leq C|h|_H|k|_H|z|_{L^\infty(0,1)}.$$

Finally, we need to assume some dissipativity of the fast equation. Assumption 3.7 is necessary to obtain strong convergence in the Averaging Principle, while Assumption 3.8 is weaker and can lead to the weak convergence.

**Assumptions 3.7** (Strict dissipativity)**.** *Let $L_g$ denote the Lipschitz constant of $G$ with respect to its second variable; then*

$$L_g < \mu, \tag{SD}$$

*where $\mu$ is defined in Assumption 3.1.*

**Assumptions 3.8** (Weak Dissipativity)**.** *There exist $c > 0$ and $C > 0$ such that for any $y \in D(B)$*

$$< By + G(y), y > \leq -c|y|^2 + C. \tag{WD}$$

The second Assumption is satisfied as soon as $G$ is bounded, while the first one requires some knowledge of the Lipschitz constant of $G$.

**Example 3.9.** *We give some fundamental examples of nonlinearities for which the previous assumptions are satisfied:*

- *Functions $F, G : H \times H \to H$ of class $\mathcal{C}^2$, bounded and with bounded derivatives, such that $G(x,y) = \nabla_y U(x,y)$ fit in the framework, with the choice $\eta = 0$.*
- *Functions $F$ and $G$ can be **Nemytskii** operators: let $f : (0,1) \times \mathbb{R}^2 \to \mathbb{R}$ be a bounded measurable function such that for almost every $\xi \in (0,1)$ $f(\xi, .)$ is twice continuously differentiable, bounded and bounded derivatives, uniformly with respect to $\xi$. Then $F$ is defined for every $x, y \in H = L^2(0,1)$ by*

  $$F(x,y)(\xi) = f(\xi, x(\xi), y(\xi)).$$

  *For $G$, we assume that there exists a function $g$ with the same properties as $f$ above, such that $G(x,y)(\xi) = g(\xi, x(\xi), y(\xi))$. The strict dissipativity Assumption* (SD) *is then satisfied when*

  $$\sup_{\xi \in (a,b), x \in \mathbb{R}, y \in \mathbb{R}} |\frac{\partial g}{\partial y}(\xi, x, y)| < \mu.$$

  *The conditions in Assumption 3.3 are then satisfied for $F$ and $G$ as soon as there exists $\eta < 1/2$ such that $D(-A)^\eta$ and $D(-B)^\eta$ are continuously embedded into $L^\infty(0,1)$.*

We then deduce that the system (3.2) is well-posed for any $\epsilon > 0$, on any finite time interval $[0,T]$. Under Assumptions 3.1, 3.3, 3.5, the nonlinearities $F$ and $G$ are Lipschitz continuous, and the following Proposition is classical - see [15]:

**Proposition 3.10.** *For every $\epsilon > 0$, $T > 0$, $x \in H$, $y \in H$, system* (3.2) *admits a unique mild solution $(X^\epsilon, Y^\epsilon) \in (L^2(\Omega, \mathcal{C}([0,T], H)))^2$: for any $t \in [0,T]$,*

$$\begin{aligned} X^\epsilon(t) &= e^{tA}x + \int_0^t e^{(t-s)A}F(X^\epsilon(s), Y^\epsilon(s))ds \\ Y^\epsilon(t) &= e^{\frac{t}{\epsilon}B}y + \frac{1}{\epsilon}\int_0^t e^{\frac{(t-s)}{\epsilon}B}G(X^\epsilon(s), Y^\epsilon(s))ds + \frac{1}{\sqrt{\epsilon}}\int_0^t e^{\frac{(t-s)}{\epsilon}B}dW(s). \end{aligned} \tag{3.10}$$

## 3.4. Convergence results

3.4.1. **Statement of the Theorems.** We can now state the main results: the numerical process $(X_n)$ defined by (3.4) approximates the slow component $X^\epsilon(n\Delta t)$ of system (3.2), with strong and weak error estimates given in the Theorems.

If the Strict Dissipativity Assumption 3.7 is satisfied, we can prove the following:

**Theorem 3.11** (Strong convergence)**.** *Assume that $x, y \in H$. With* (SD)*, for any $0 < r < 1/2$, $0 < \kappa < 1/2$, $T > 0$, $\epsilon_0 > 0$, $\Delta_0 > 0$, $\tau_0 > 0$, there exists $C > 0$ such that for any $0 < \epsilon \leq \epsilon_0$, $0 < \Delta t \leq \Delta_0$, $\delta t > 0$ such that $\tau = \frac{\delta t}{\epsilon} \leq \tau_0$ and $1 < n \leq n_0 = \lfloor \frac{T}{\Delta t} \rfloor$*

$$\begin{aligned} \mathbb{E}\big|X^\epsilon(n\Delta t) - X_n\big| \leq C\bigg(\epsilon^{1/2-r} + \frac{1}{n} + \Delta t^{1-r}\bigg) \\ + C\bigg((\frac{\delta t}{\epsilon})^{1/2-\kappa} + \frac{1}{\sqrt{N\frac{\delta t}{\epsilon}+1}} e^{-cn_T\frac{\delta t}{\epsilon}}\bigg) \\ + C\frac{\sqrt{\Delta t}}{\sqrt{M(N\frac{\delta t}{\epsilon}+1)}}. \end{aligned} \tag{3.11}$$

Under the more general Weak Dissipativity Assumption 3.8, we can prove the following:

**Theorem 3.12** (Weak convergence)**.** *Assume that $x \in D((-A)^\theta)$ and $y \in H$, for some $\theta \in ]0,1]$. Let $\Phi : H \to \mathbb{R}$ bounded, of class $\mathcal{C}^2$, with bounded first and second order derivatives. Then with the weak dissipativity assumption* (WD)*, for any $0 < r < 1$, $\kappa < 1/2$, $T > 0$, $\epsilon_0 > 0$, $\Delta_0 > 0$, $\tau_0 > 0$, there exists $C > 0$ such that for any $0 < \epsilon \leq \epsilon_0$, $0 < \Delta t \leq \Delta_0$, $\delta t > 0$ such that $\tau = \frac{\delta t}{\epsilon} \leq \tau_0$ and $1 < n \leq n_0 = \lfloor \frac{T}{\Delta t} \rfloor$*

$$\begin{aligned} \big|\mathbb{E}\Phi(X^\epsilon(n\Delta t)) - \mathbb{E}\Phi(X_n)\big| \leq C\bigg(\epsilon^{1-r} + \frac{1}{n} + \Delta t^{1-r}\bigg) \\ + C\bigg((\frac{\delta t}{\epsilon})^{1/2-\kappa}(1 + \frac{1}{((n_T-1)\frac{\delta t}{\epsilon})^{1/2-\kappa}}) + \frac{1}{N\frac{\delta t}{\epsilon}+1} e^{-cn_T\frac{\delta t}{\epsilon}}\bigg). \end{aligned} \tag{3.12}$$

**Remark 3.13.** *If we indeed assume* (SD) *in Theorem 3.12, the factor $(1 + \frac{1}{((n_T-1)\frac{\delta t}{\epsilon})^{1/2-\kappa}})$ can be replaced by 1. We also notice that this factor is absent in* [25]*, where only strictly dissipative situations are considered.*

If we look at the estimates of Theorems 3.11 and 3.12 at time $n = n_0$, the factor $\frac{1}{n}$ is of size $\Delta t = \mathrm{O}(\Delta t^{1-r})$. The parameters $r$ and $\kappa$ are positive, but can be chosen as small as necessary.

We can remark that in both Theorems we require no condition on the fast initial condition $y$, while some regularity of the slow initial condition $x$ is required in the analysis of the weak error: we assume $x \in D((-A)^\theta)$, with $\theta > 0$. The reason lies in the use of the estimate of the averaging error in the weak sense, given by Theorem 1.2 in [4].

The special structure in the error bounds is a consequence of the heterogeneity in the treatment of the components in the numerical scheme.

### 3.4.2. **Some comments on the convergence results.**

3.4.2.1. *Interpretation of the error terms.* The proofs rely on the following decomposition, which explains the origin of the different error terms:

$$X^\epsilon(n\Delta t) - X_n = X^\epsilon(n\Delta t) - \overline{X}(n\Delta t) + \overline{X}(n\Delta t) - \overline{X}_n + \overline{X}_n - X_n. \tag{3.13}$$

The numerical process $(\overline{X}_n)$ is defined in (3.26) below: it is the solution of the macrosolver with a known $\overline{F}$, while $(X_n)$ is solution of the macrosolver using $\tilde{F}_n$. The continuous processes $\overline{X}$ and $X^\epsilon$ are defined in (3.2) and (3.3).

The first term is bounded thanks to the averaging result, using strong and weak order of convergence results obtained in [4]. The second term is the error in a deterministic numerical scheme, for which convergence results are classical; we recall the estimate in Proposition 3.26. The third term is the difference between the two numerical approximations, and the main task is the control of this part: we show that an extension of the averaging effect holds at the discrete time level, where $\overline{X}_n$ plays the role of an averaged process for $X_n$.

When we look at the Theorems 3.11 and 3.12, we first remark that we obtain the same kind of bounds as in the finite dimensional case of [25]. However we notice some differences; they are due both to the infinite dimensional setting and to different proofs.

First, the weak order of the Euler scheme for SPDEs is only 1/2 - see [18] - while it is 1 for SDEs; we can remark that in (3.11) and (3.12) only the weak order of the microsolver appears, and this is one of the main

theoretical advantages of the method. For completeness, we recall that the strong order is 1/4 in the SPDE case - see [57] - and 1/2 for SDEs. In fact, at least in the strictly dissipative case, we are comparing the invariant measures of the continuous fast equation with the invariant measure of its numerical approximation - see Theorem 3.24 and Corollary 3.25, obtained from [5]. In the weakly dissipative case, we use the weak error estimates where the constants do not depend on the final time.

The proofs of Theorems 3.11 and 3.12 are inspired by the ones in [25], but are different. The strong error is analysed in a global way, as in the proof of the strong order Theorem in [4]. Moreover we do not need a counterpart of Lemma 2.6 in [25] - which gives an estimate of $\tilde{F}_n - \overline{F}(X_n)$ - and we thus think that our method is more natural. For the control of the weak error, we also introduce a new appropriate auxiliary function to control everything in the weak sense: as a consequence we observe that the number of independent realizations of the microsolver does not appear in (3.12).

However, we also present some simplified proofs, and a comparison of the results reveals the absence of a factor denoted by $R$ in [25]: the difference is due to the way we use the initialization of the microsolver at each macrotime step.

3.4.2.2. *Role of the initialization of the microsolver.* The principal effect of the definition (3.7) is the control of the moments of the fast numerical component $Y_{n,m,j}$, uniformly with respect to $n, m, j$: see Lemma 3.21. Precisely, two constraints are imposed on the variables $Y_{n,0,j}$: we require that for $n \geq 1$ the variable $Y_{n,0,j}$ is measurable with respect to the $\sigma$-field $\mathcal{G}_n$ defined below in (3.31) and that the estimate of Lemma 3.21 is satisfied.

However, at least at an intuitive level, the choice of (3.7) should improve the convergence results, thanks to a quicker relaxation to equilibrium of the microsolver at each macrotime step. In order to take into account this effect, we would need to develop more complex proofs of our results, without modifying the regimes for the parameters $m_0, n_T, M$ of the scheme.

Precisely, as in [25], some terms of the error bounds can be multiplied by a factor $R$: instead of $R = 1$, we could get

$$R = \frac{\Delta t^{1-\eta-r}}{1 - e^{-m_0 \frac{\delta t}{\epsilon}}},$$

where the value of $\eta$ is imposed by the regularity assumptions on the coefficients $F$ and $G$ and $r > 0$ is chosen as small as necessary. In the finite dimensional case, with $\eta = r = 0$, this factor can play a role: for instance, if high order numerical methods are used. Here, such a factor is useless in the analysis of the cost of the scheme with respect to the parameters, even if it seems that the bound is improved.

Therefore, to simplify the proofs, we only show how to obtain $R = 1$.

3.4.2.3. *Efficiency of HMM.* We now show how the Heterogeneous Multiscale Method used here is more efficient of a direct method when the parameter $\epsilon$ goes to 0.

The fundamental remark is that $\epsilon$ only appears in the expressions with $\tau = \frac{\delta t}{\epsilon}$, and that we never have $\frac{\Delta t}{\epsilon}$ in the scheme and in the convergence estimate. The parameter $\tau$ can therefore be interpreted as the effective time step for the fast numerical scheme defined with (3.5). The reason for this absence of $\epsilon$ relies in the construction of the scheme: instead of approximating $X^\epsilon$, we approximate $\overline{X}$ thanks to the averaging principle. As explained before, the microsolver is only used to get an approximation of the averaged coefficient $\overline{F}$; the evolution in the macrosolver uses the time step $\Delta t$ is linked with $\epsilon$ only through the definition of $\tilde{F}_n$, with only $\tau$.

What is then important is to check that if $\Delta t$ and $\tau$ are chosen small enough, and $M$, $N$ and $n_T$ are large enough, each error term due to the numerical scheme in the Theorems 3.11 and 3.12 goes to 0.

The choices must be uniform with respect to $\epsilon$. The error term due to the application of the averaging principle is treated apart, since it only depends on $\epsilon$.

To see this, we first choose $\Delta t$ and $\tau$.

Then we take $n_T$ such that $n_T \tau \to +\infty$; we observe that the corresponding error terms go to 0 exponentially fast, as a consequence of the dissipation in the evolution. It is better to choose $n_T$ before $N$ and $M$, since the exponential convergence is faster than the other convergences.

Apparently, the parameters $M$ and $N$ play no role: indeed in (3.11) the orders of convergence with respect to $\Delta t$ in the terms

$$\Delta t \quad \text{and} \quad \frac{\sqrt{\Delta t}}{\sqrt{M(N\tau+1)}}$$

are not the same: a better choice is to choose $M$ and $N$ such that $\frac{1}{MN\tau}$ and $\Delta t$ go to 0 at the same speed. Moreover, choosing $\Delta t$ after $M$, $N$ and $n_T$ also changes the analysis.

The question of an optimal joint choice of the parameters is related to the analysis of the cost of the scheme. Even if the multiscale method seems more efficient than a direct one when $\epsilon$ goes to 0, we have to check that for a fixed value of the parameter $\epsilon$ the approximation of the slow component at a fixed time $T$ by our method does not require too much computational time. Following the approach in [25], we consider as a unit of time the computation of one realization of a step of the microsolver - i.e. the computation of $Y_{n,m+1,j}$ knowing $Y_{n,m,j}$. We can consider that the computation of $\tilde{F}_n$ in (3.6) requires a negligible amount of time; we can then consider that one step of the macrosolver also requires 1 unit of time.

For each step of the macrosolver, $m_0.M$ steps of the microsolver are necessary; then we require $\frac{T}{\Delta t}$ macrosteps, so that we define the cost - with $T = 1$:

$$\text{cost(HMM)} = \frac{m_0.M}{\Delta t};$$

for a direct scheme with time step $\delta t$, the corresponding cost is defined by

$$\text{cost(direct)} = \frac{1}{\delta t} = \frac{1}{\epsilon}\frac{1}{\delta t/\epsilon}.$$

Working as in [25], we introduce a tolerance in the discretization error $tol$, and we choose the parameters in order to bound the error with

$$C\epsilon^{\Lambda} + tol,$$

with $\Lambda = 1/2 - r$ for the strong estimate (3.11) and $\Lambda = 1 - r$ for the weak estimate (3.12), with $r > 0$ chosen as small as necessary. In other words, we only focus on the difference between $\overline{X}(n\Delta t)$ and $X_n$, and consider that the averaging error is negligible. The costs are then functions of $tol$.

For simplicity, we take in the presentation $r = \kappa = 0$.

The time scale separation parameter $\epsilon$ is supposed to be very small, whereas we fix some tolerance $tol$ for the error in the numerical scheme: more precisely, we want the error to satisfy

$$\begin{aligned} \mathbb{E}|X^{\epsilon}(n\Delta t) - X_n| &\leq C(\epsilon^{1/2} + tol) \text{ (strong error)} \\ |\mathbb{E}\Phi(X^{\epsilon}(n\Delta t)) - \mathbb{E}\Phi(X_n)| &\leq C(\epsilon^{1} + tol) \text{ (weak error)}, \end{aligned} \tag{3.14}$$

in the regimes $\epsilon^{1/2} = \text{o}(tol)$ (strong error) and $\epsilon^1 = \text{o}(tol)$ (weak error). We want to show that we can choose the parameters of the scheme such that each term of the estimate of the Theorem 3.11 are of size $tol$, except the averaging one, and such that the cost of the scheme is lower than the cost of a direct scheme.

- We first notice that the choice for $\Delta t$ and $\frac{\delta t}{\epsilon}$ is easy:

$$\Delta t \approx tol \quad \text{and} \quad \frac{\delta t}{\epsilon} \approx tol^2, \tag{3.15}$$

  The choice of other parameters $N, M, n_T$ changes when we look at the strong or at the weak error.

  We can remark that the time step $\Delta t$ for the slow equation does not depend on $\epsilon$. Moreover, $\epsilon$ only appears in the scheme and in the error bounds through $\tau = \frac{\delta t}{\epsilon}$, which is interpreted as the effective time step in the fast equation.
- We first focus on the strong convergence case. For simplicity, we consider separately the simple cases with $M = 1$ or $N = 1$. We obtain

| | parameters | cost |
|---|---|---|
| $M = 1$ | $N \approx tol^{-3}$ | |
| | $n_T \approx tol^{-2}\log(tol^{-1})$ | $tol^{-4}$ |
| $N = 1$ | $M \approx tol^{-1}$ | |
| | $n_T \approx \log(tol^{-1})tol^{-2}$ | $tol^{-4}\log(tol^{-1})$ |

Up to a logarithmic factor, there is no difference between these situations.

We can now make a comparison with the cost coming from the use of a direct scheme with a single time step $\tilde{\delta t}$: since the strong order of Euler scheme for SPDEs is $1/4$, the error is at least of size $C(\frac{\tilde{\delta t}}{\epsilon})^{1/4}$ for some constant $C$. Therefore to have a bound of size $tol$, $\tilde{\delta t}$ much satisfy $\frac{\tilde{\delta t}}{\epsilon} \approx tol^{\frac{1}{1/4}}$. This leads to a cost

$$\frac{1}{\tilde{\delta t}} \approx \epsilon^{-1} tol^{-4};$$

we conclude that in this situation the HMM numerical method is better, since the ratio of the costs tends to 0 when $\epsilon \to 0$.

- We now focus on the weak error estimate. Here $M$ plays no role, so that $M = 1$ is the good choice! The time steps $\Delta t$ and $\delta t$ are still given by (3.15). It remains to look for parameters $N$ and $n_T$ such that

$$\frac{1}{N\frac{\delta t}{\epsilon} + 1} e^{-cn_T \frac{\delta t}{\epsilon}} \approx \Delta t.$$

Once again, we need to choose $n_T$ and $N$ such that $N\frac{\delta t}{\epsilon}$ or $n_T \frac{\delta t}{\epsilon}$ is large. Since exponential decrease is faster than polynomial decrease, the best choice is $e^{-cn_T \frac{\delta t}{\epsilon}} \approx \Delta t$, and therefore $n_T \approx tol^{-2} \log(tol^{-1})$, while $N = 1$.

We then see that $\frac{1}{((n_T-1)\frac{\delta t}{\epsilon})^{1/2}} \leq C$, so that the additional factor appearing when only weak dissipativity is satisfied plays no role.

The corresponding cost of the scheme is then of order $\frac{m_0}{\Delta t} \approx \log(tol^{-1}) tol^{-3}$.

We can again compare this cost with the cost coming from the use of a direct scheme: the weak order of Euler scheme for SPDEs is $1/2$; to obtain the second estimate of (3.14), we need a cost

$$\frac{1}{\tilde{\delta t}} \approx \epsilon^{-1} tol^{-2},$$

and again the HMM scheme is better than the direct scheme for such a range of parameters.

Finally, we have not treated here the discretization in space, since the main issue is the presence of $\epsilon$ in the evolution in time of the system of SPDEs. To complete the analysis, we need to take into account the cost of the discretization in space of one step of the microsolver, with an additional factor; then one realization of the microsolver would not require one unit of time, but a number of units depending on the size of the discretization. The same remark also holds for a direct method, and we see that the comparison of the costs of time discretization is sufficient to prove the better efficiency of the Heterogeneous Multiscale Method.

## 3.5. Preliminary results

3.5.1. **Known results about the fast equation and the averaged equation.** In this section, we just recall without proof the main results on the fast equation with frozen slow component and on the averaged equation, defined below. Proofs can be found in [12] for the strict dissipative case, and the extension to the weakly dissipative situation relies on arguments explained below.

If $x \in H$, we define an equation on the fast variable where the slow variable is fixed and equal to $x$:

$$\begin{aligned} dY_x(t,y) &= (BY_x(t,y) + G(x, Y_x(t,y)))dt + dW(t), \\ Y_x(0,y) &= y. \end{aligned} \tag{3.16}$$

This equation admits a unique mild solution, defined on $[0, +\infty[$.

Since $Y^\epsilon$ is involved at time $t > 0$, heuristically we need to analyse the properties of $Y_x(\frac{t}{\epsilon}, y)$, with $\epsilon \to 0$, and by a change of time we need to understand the asymptotic behaviour of $Y_x(., y)$ when time goes to infinity.

Under the strict dissipativity Assumption 3.7, we obtain a contractivity of trajectories issued from different initial conditions and driven by the same noise:

**Proposition 3.14.** *With* (SD)*, for any $t \geq 0$, $x, y_1, y_2 \in H$ we have*

$$|Y_x(t,y_1) - Y_x(t,y_2)|_H \leq e^{-\frac{(\mu - L_g)}{2} t} |y_1 - y_2|_H.$$

Under the weak dissipativity Assumption 3.8, we obtain such an exponential convergence result for the laws instead of trajectories. The proof of this result is not staightforward, and can be found in [19] - see also [5] for further references.

**Proposition 3.15.** *With* (WD)*, there exist $c > 0$, $C > 0$ such that for any bounded test function $\phi$, any $t \geq 0$ and any $y_1, y_2 \in H$*

$$|\mathbb{E}\phi(Y_x(t, y_1)) - \mathbb{E}\phi(Y_x(t, y_2))| \leq C\|\phi\|_\infty(1 + |y_1|^2 + |y_2|^2)e^{-ct}. \tag{3.17}$$

The proofs of Propositions 3.14 and 3.15 are only based on the respective dissipativity Assumption; the gradient structure of the equation is not necessary at this stage.

As a consequence, we can show that there exists a unique invariant probability measure $\mu^x$ associated with $Y_x$, and that the convergence to equilibrium is exponentially fast:

**Proposition 3.16.** *If we assume* (SD) *or* (WD)*, the fast process $Y_x$ with frozen slow component $x$ admits a unique invariant probability measure $\mu^x$, and there exist constants $C, c > 0$ such that for any bounded function $\phi : H \to \mathbb{R}$ or $\phi : H \to H$, $t \geq 0$ and $x, y \in H$ we have*

$$|\mathbb{E}\phi(Y_x(t, y)) - \int_H \phi(z)\mu^x(dz)| \leq C\|\phi\|_\infty(1 + |y|_H^2)e^{-ct}.$$

This result is not sufficient to obtain the orders of convergence in the averaging principle of [4], and to prove Theorems 3.11 and 3.12. Thanks to the existence of a potential $U$ such that $G = D_y U$, we are moreover able to give an explicit formula (3.18) for $\mu^x$.

Let $\nu = \mathcal{N}(0, (-B)^{-1}/2)$ be the centered Gaussian probability measure on $H$ with the covariance operator $(-B)^{-1}/2$ - which is positive and trace-class, thanks to Assumption 3.1. Then $\mu^x$ satisfies

$$\mu^x(dy) = \frac{1}{Z(x)} e^{2U(x,y)}\nu(dy), \tag{3.18}$$

where $Z(x) \in ]0, +\infty[$ is a normalization constant.

Now we define the averaged equation. First we define the averaged nonlinear coefficient $\overline{F}$:

**Definition 3.17.** *For any $x \in H$,*

$$\overline{F}(x) = \int_H F(x, y)\mu^x(dy). \tag{3.19}$$

Using Assumptions 3.3, 3.5 and the expression of $\mu^x$, it is easy to prove that $\overline{F}$ is bounded and Lipschitz continuous.

Under Assumptions 3.3, 3.5 and thanks to the expression (3.18) of $\mu^x$, we can easily prove the following properties on $\overline{F}$:

**Proposition 3.18.** *There exists $0 \leq \eta < 1$ and a constant $C$ such that the following directional derivatives of $\overline{F}$ are well-defined and controlled:*

- *For any $x \in H$, $h \in H$, $|D\overline{F}(x).h| \leq C|h|_H$.*
- *For any $x \in H$, $h \in H$, $k \in D(-A)^\eta$, $|D^2\overline{F}(x).(h, k)| \leq C|h|_H|k|_{(-A)^\eta}$.*
- *For any $x \in H$, $h, k \in H$, $|(-A)^{-\eta}D^2\overline{F}(x).(h, k)| \leq C|h|_H|k|_H$.*

*Moreover, $\overline{F}$ is bounded and Lipschitz continuous.*

The last estimate above is a consequence of Assumptions 3.4 and 3.6, and of the following fact: we have almost surely $W^B(t) \in L^\infty(0, 1)$ for any $t \geq 0$, and

$$\int_H |z|_{L^\infty(0,1)}\nu(dz) < +\infty. \tag{3.20}$$

We can notice that if $\eta > 1/4$ - which is the right condition in the case of linear Laplace operators and of nonlinear Nemytskii operators - we have $\int_H |z|_{(-A)^\eta}\nu(dz) = +\infty$: we thus need the restrictive condition in Assumption 3.6.

**Remark 3.19.** *Even when $F$ and $G$ are Nemytskii operators, $\overline{F}$ is not such an operator in general.*

Then the averaged equation (see (3.3) in the introduction) can be defined:

$$\frac{d\overline{X}(t)}{dt} = A\overline{X}(t) + \overline{F}(\overline{X}(t)),$$

with initial condition $\overline{X}(0) = x \in H$. For any $T > 0$, this deterministic equation admits a unique mild solution $\overline{X} \in \mathcal{C}([0,T], H)$.

3.5.2. **Estimates on the numerical solutions.** We can give uniform estimates on $X_n$ and $Y_{n,m,j}$, defined by 3.4 and 3.5:

**Lemma 3.20.** *There exists $C > 0$ such that we have $\mathbb{P}$-almost surely*

$$|X_n| \leq C(1 + |x|),$$

*for any $0 \leq n \leq n_0$.*

Proof The linear operator $S_{\Delta t}$ satifies $|S_{\Delta t}|_{\mathcal{L}(H)} \leq \frac{1}{1+\lambda\Delta t}$; moreover $F$ is bounded, so that by (3.6) we almost surely have for any $n \geq 0$ $|\tilde{F}_n| \leq \|F\|_\infty$. The end of the proof is then straightforward; we also notice that $C$ does not depend on the final time $T$. □

**Lemma 3.21.** *There exists $C > 0$ - which does not depend on $T > 0$, on $N$, on $n_T$ or on $M$ - such that for any $\Delta t > 0$, $\tau = \frac{\delta t}{\epsilon} > 0$, $0 \leq n \leq n_0$, $0 \leq m \leq m_0$ and $1 \leq j \leq M$, we have*

$$\mathbb{E}|Y_{n,m,j}|^2 \leq C(|y|^2 + 1).$$

Proof We introduce $\omega_{n,m,j}$ defined by the fast numerical scheme with no nonlinear coefficient - see (3.5) - with the notation $\tau = \frac{\delta t}{\epsilon}$: for any $0 \leq j \leq M$, $0 \leq n \leq n_0$ and $0 \leq m \leq m_0$

$$w_{n,m+1,j} = R_\tau w_{n,m,j} + \sqrt{\tau} R_\tau \zeta_{n,m+1,j}, \tag{3.21}$$

with the initial condition $w_{n,0,j} = 0$. It is a classical result that for any $\tau_0 > 0$, there exists a constant $C(\tau_0)$ such that for any $0 \leq \tau < \tau_0$, $0 \leq n \leq n_0$, $1 \leq j \leq M$ and $0 \leq m \leq m_0$ we have

$$\mathbb{E}|\omega_{n,m,j}|^2 \leq C(\tau_0).$$

Now for any $0 \leq m \leq m_0$ we define $D_{n,m,j} = Y_{n,m,j} - w_{n,m,j}$; it is enough to control $|D_{n,m,j}|^2$, $\mathbb{P}$-almost surely. By (3.5), we have the following expression: for any $0 \leq m \leq m_0$

$$D_{n,m+1,j} = R_\tau D_{n,m,j} + \tau R_\tau G(X_n, Y_{n,m,j}).$$

Since $G$ is bounded, and using the inequality $|R_\tau|_{\mathcal{L}(H)} \leq \frac{1}{1+\mu\tau}$, we get

$$|D_{n,m+1,j}| \leq \frac{1}{1+\mu\tau}|D_{n,m,j}| + C\tau.$$

Therefore we have for any $0 \leq m \leq m_0$

$$(1+\mu\tau)^m |D_{n,m,j}| \leq |D_{n,0,j}| + C[(1+\mu\tau)^m - 1].$$

But $D_{n,0,j} = D_{n-1,m_0,j}$. So we get

$$(1+\mu\tau)^{m_0} |D_{n+1,0,j}| \leq |D_{n,0,j}| + C[(1+\mu\tau)^{m_0} - 1].$$

Therefore

$$(1+\mu\tau)^{nm_0} |Z_{n,0,j}| \leq |D_{0,0,j}| + C(1+\mu\tau)^{nm_0} = |y| + C(1+\mu\tau)^{nm_0}.$$

As a consequence, we get for any $0 \leq n \leq n_0$ and any $0 \leq m \leq m_0$, $|D_{n,0,j}| \leq C(1 + |y|)$, and then $|Z_{n,m,j}| \leq 2C(1 + |y|)$. □

We can remark that Lemma 3.21 is a consequence of the choice of the initialization of the microsolver at each macrotime step (3.7), through the equality $D_{n,0,j} = D_{n-1,m_0,j}$ appearing at the end of the proof. However, other choices for $Y_{n,0,j}$ can lead to the same kind of estimate.

3.5.3. **Asymptotic behaviour of the "fast" numerical scheme.** At the continuous time level, the averaging principle proved in [4] comes from the asymptotic behaviour of the fast equation with frozen slow component (3.16), as described in Section 3.5.1. The underlying idea of the HMM method in our setting is to prove a similar averaging effect at the discrete time level: we therefore study the asymptotic behaviour of the fast numerical scheme which defines the microsolver, with frozen slow component - in other words we are looking at the evolution of the microsolver during one fixed macrotime step.

In Section 3.5.1, we have seen that under the weak dissipativity Assumption 3.8 the fast equation with frozen slow component admits a unique invariant probability measure $\mu^x$. At the discrete time level, this Assumption only yields the existence of invariant laws; to get a unique invariant law $\mu^{x,\tau}$, we need strict dissipativity (SD), and we obtain:

**Theorem 3.22.** *Under Assumption 3.7, for any $\tau > 0$ and any $x \in H$; the numerical scheme 3.23 admits a unique ergodic invariant probability measure $\mu^{x,\tau}$. Moreover, we have convergence to equilibrium in the following sense: for any $\tau_0 > 0$, there exist $c > 0$ and $C > 0$ such that for any $0 < \tau \leq \tau_0$, $x \in H$, $y \in H$, any Lipschitz continuous function $\phi$ from $H$ to $R$, and $m \geq 0$, we have*

$$|\mathbb{E}(\phi(Y_m^x(y))) - \int_H \phi(z)\mu^{x,\tau}(dz)| \leq C(1+|y|)[\phi]_{Lip}e^{-cm\tau}. \tag{3.22}$$

We recall the notation $\tau = \frac{\delta t}{\epsilon}$ for the effective time step; the noise is defined with a cylindrical Wiener process $\tilde{W}$: $\tilde{\zeta}_{m+1} = \frac{\tilde{W}_{(m+1)\tau} - \tilde{W}_{m\tau}}{\sqrt{\tau}}$. If we fix the slow component $x \in H$, we define

$$Y_{m+1}^x(y) = R_\tau Y_m^x(y) + \tau R_\tau G(x, Y_m^x(y)) + \sqrt{\tau} R_\tau \tilde{\zeta}_{m+1}, \tag{3.23}$$

with the initial condition $Y_0^x(y) = y$.

The proof of Theorem 3.22 is divided into 2 steps. First in Section 3.5.3.1 we focus on the existence, which is obtained under the weak dissipativity Assumption (WD) thanks to the Krylov-Bogoliubov criterion - see [16]. Second, in Section 3.5.3.2 we show that uniqueness holds if the strict dissipativity Assumption (SD) is satisfied.

In the case when (WD) is satisfied while (SD) is not true, it seems that there is no general uniqueness result at the discrete time level.

3.5.3.1. *Existence of an invariant law.* With Equation (3.23), we associate the transition semi-group $(P_m^{x,\tau})$: if $\phi$ is a bounded measurable function from $H$ to $\mathbb{R}$, if $y \in H$ and $m \geq 0$

$$P_m^{x,\tau}\phi(y) = \mathbb{E}[\phi(Y_m^x(y))]. \tag{3.24}$$

We also denote by $\nu_{m,y}^{x,\tau}$ the law of $Y_m^{x,\tau}(y)$: then

$$P_m^{x,\tau}\phi(y) = \mathbb{E}[\phi(Y_m^{x,\tau}(y))] = \int_H \phi(z)\nu_{m,y}^{x,\tau}(dz).$$

We notice that the semi-group $(P_m)$ satisfies the Feller property: if $\phi$ is bounded and continuous, then $P_m\phi \in$ is bounded and continuous.

The required tightness property for the use of the Krylov-Bogoliubov criterion is a consequence of the following estimate, which can be proved thanks to regularization properties of the semi-group $(R_\tau^m)_m$: for any $0 < \gamma < 1/4$, $\tau > 0$, there exists $C(\gamma, \tau) > 0$ such that for any $m \geq 1$ and $\tau \leq \tau_0$

$$\mathbb{E}|Y_m(y)|^2_{(-B)^\gamma} \leq C(\gamma, \tau).$$

Moreover if $0 < \gamma < 1/4$, the embedding of $D(-B)^\gamma$ in $H$ is compact.

We then see that for any $y \in H$ the family of probability measures $(\frac{1}{m}\sum_{k=1}^m \nu_{k,y}^{x,\tau})$ is tight.

3.5.3.2. *Uniqueness under strict dissipativity.* The key estimate to prove uniqueness is the following contractivity property, which holds thanks to Assumption 3.7:

**Proposition 3.23.** *For any $\tau_0 > 0$, there exists $c > 0$ such that for any $0 < \tau \leq \tau_0$, $m \geq 0$, $y_1, y_2 \in H$, $x \in H$ we have $\mathbb{P}$-almost surely*

$$|Y_m^x(y_1) - Y_m^x(y_2)| \leq e^{-cm\tau}|y_1 - y_2|. \tag{3.25}$$

Proof If we define $r_m = Y_m(y_1) - Y_m(y_2)$, then we have the equation

$$r_{m+1} = r_m + \tau B R_{m+1} + \tau(G(x, Y_m(y_1)) - G(x, Y_m(y_2))),$$
$$r_0 = y_1 - y_2.$$

If we take the scalar product in $H$ of this equation with $r_{m+1}$, we get

$$|r_{m+1}|^2 - < r_m, r_{m+1} > = \tau < B r_{m+1}, r_{m+1} >$$
$$+ \tau < G(x, Y_m(y_1)) - G(x, Y_m(y_2)), r_{m+1} >;$$

The left hand-side is equal to $\frac{1}{2}(|r_{m+1}|^2 - |r_m|^2) + \frac{1}{2}|r_{m+1} - r_m|^2$, and we get

$$\frac{1}{2}(|r_{m+1}|^2 - |r_m|^2) \leq -\tau|(-B)^{1/2} r_{m+1}|^2 + \tau L_g |r_m||r_{m+1}|$$
$$\leq -\mu\tau|r_{m+1}|^2 + \frac{1}{2}\tau L_g(|r_{m+1}|^2 + |r_m|^2).$$

Therefore we have $(1 + \tau(2\mu - L_g))|r_{m+1}|^2 \leq (1 + \tau L_g)|r_m|^2$. We remark that (SD) implies that for any $\tau_0 > 0$, there exists $c > 0$ such that if $\tau \leq \tau_0$ we have $\rho = \frac{1+\tau L_g}{1+\tau(2\mu - L_g)} \leq e^{-2c\tau}$; therefore

$$|r_m|^2 \leq \rho^m |y_1 - y_2|^2 \leq e^{-2cm\tau}|y_1 - y_2|^2.$$

□

As a consequence, there exists a unique ergodic invariant probability measure $\mu^{x,\tau}$, which is strongly mixing. Moreover one can prove that there exists $C > 0$ such that for any $\tau > 0$ and any $x \in H$ $\int_H |y| \mu^{x,\tau}(dy) \leq C$; we therefore get

$$|\mathbb{E}(\phi(Y_m^x(y))) - \int_H \phi(z)\mu^{x,\tau}(dz)| = |\mathbb{E}(\phi(Y_m^x(y))) - \int_H \mathbb{E}\phi(Y_m^x(z))\mu^{x,\tau}(dz)|$$
$$\leq \int_H \mathbb{E}|\phi(Y_m^x(y)) - \phi(Y_m^x(z)|\mu^{x,\tau}(dz)$$
$$\leq [\phi]_{\mathrm{Lip}} \int_H e^{-cm\tau}|y - z|\mu^{x,\tau}(dz)$$
$$\leq C(1 + |y|)[\phi]_{\mathrm{Lip}} e^{-cm\tau}.$$

3.5.3.3. *Approximation of the invariant law $\mu^x$ by the fast numerical scheme.* We recall that $\mu^x$ denotes the invariant law of the continuous time fast equation with frozen slow component (3.16). Thanks to the fast numerical scheme, we can get an approximation result, which is proved in Chapter 2 - see also [5]: with test functions of class $\mathcal{C}_b^2$, we can control the weak error for any time with a convergence of order $1/2$ with respect to the time step $\tau$. Moreover the estimate is easily seen to be independent from the slow component $x$.

We define for any $\Phi$ of class $\mathcal{C}_b^2$

$$\|\Phi\|_{(2)} = \sup_{y \in H} |\Phi(y)| + \sup_{y \in H, h \in H, |h|=1} |D_y\Phi(y).h| + \sup_{y \in H, h,k \in H, |h|=|k|=1} |D_{yy}^2\Phi(y).(h,k)|.$$

**Theorem 3.24.** *With the dissipativity condition* (WD)*, for any $0 < \kappa < 1/2$, for any $\tau_0 > 0$, there exists $C, c > 0$ such that for any $\Phi$ of class $\mathcal{C}_b^2$, for any $x, y \in H$, for any $0 < \tau \leq \tau_0$ and any integer $2 \leq m < +\infty$*

$$|\mathbb{E}[\Phi(Y_x(m\tau, y))] - \mathbb{E}[\Phi(Y_m^x(y))]| \leq C\|\Phi\|_{(2)}(1 + |y|^3)(((m-1)\tau)^{-1/2+\kappa} + 1)\tau^{1/2-\kappa}.$$

We can remark that this Theorem is proved without requiring the gradient structure of the fast equation with frozen slow component.

As explained in Section 3.5.3, the existence of invariant probability measures for the numerical scheme is true with only a weak dissipativity condition (WD), while uniqueness is *a priori* only satisfied when the strict dissipativity Assumption 3.7 holds; the unique invariant law is then denoted by $\mu^{x,\tau}$.

**Corollary 3.25.** *Under the assumptions of Theorem 3.24:*

(i) *we have for any $m \geq 2$*

$$|\int_H \Phi d\mu^x - \mathbb{E}[\Phi(Y_m^x)]| \leq C\|\Phi\|_{(2)}(1 + |y|^3)(((m-1)\tau)^{-1/2+\kappa} + 1)\tau^{1/2-\kappa} + CN(\Phi)(1 + |y|^2)e^{-cm\tau}.$$

(ii) *if moreover* (SD) *is satisfied,*

$$|\int_H \Phi d\mu^x - \int_H \Phi d\mu^{x,\tau}| \leq C\|\Phi\|_{(2)}\tau^{1/2-\kappa}.$$

The result *(i)* is sufficient for the proof of Theorem 3.12, while the result *(ii)* is necessary to obtain the strong convergence in Theorem 3.11.

We recall that in the case of Euler scheme for SDEs this kind of results holds with the order of convergence 1.

3.5.4. **Error in the deterministic scheme** (3.26)**.** We define a scheme based on the macrosolver, for theoretical purpose, in the situation when $\overline{F}$ is known:

$$\begin{aligned} \overline{X}_{n+1} &= S_{\Delta t}\overline{X}_n + \Delta t S_{\Delta t}\overline{F}(\overline{X}_n) \\ \overline{X}_0 &= x. \end{aligned} \tag{3.26}$$

We can look at the error between $\overline{X}_n$, defined by (3.26), and $\overline{X}(n\Delta t)$, defined by (3.3). Here quantities are deterministic, and the following result is classical - see [45], [14], or the details of the proofs in [57]:

**Proposition 3.26.** *For any $0 < r < 1$, $\Delta t_0 > 0$ and $T > 0$, there exists $C > 0$, such that for any $0 < \Delta t \leq \Delta t_0$ and $1 \leq n \leq \lfloor \frac{T}{\Delta t} \rfloor$*

$$|\overline{X}_n - \overline{X}(n\Delta t)| \leq \frac{C}{n} + C(1+|x|)\Delta t^{1-r}.$$

## 3.6. Proof of the strong convergence Theorem 3.11

The final time $T$ is fixed and we recall the notation $n_0 = \lfloor \frac{T}{\Delta t} \rfloor$.

To simplify notations, we do not always precise the range of summation in the expressions below: the indices $j, j_1, j_2$ belong to $\{1, \dots, M\}$, and $m, m_1, m_2$ belong to $\{n_T, \dots, n_T + N - 1 = m_0\}$.

We recall that according to the decomposition of the error (3.13), we have to control

$$\begin{aligned} \mathbb{E}|X^\epsilon(n\Delta t) - X_n| \leq\ & \mathbb{E}|X^\epsilon(n\Delta t) - \overline{X}(n\Delta t)| \\ & + |\overline{X}(n\Delta t) - \overline{X}_n| \\ & + \mathbb{E}|\overline{X}_n - X_n|. \end{aligned} \tag{3.27}$$

The first part is controlled thanks to the strong order Theorem of [4]: for any $0 < r < 1/2$, we have for any $0 \leq n \leq n_0$

$$\mathbb{E}|X^\epsilon_{n\Delta t} - \overline{X}_{n\Delta t}| \leq C\epsilon^{1/2-r}.$$

The second part is deterministic and is controlled thanks to Proposition 3.26:

$$|\overline{X}_n - \overline{X}(n\Delta t)| \leq \frac{C}{n} + C(1+|x|)\Delta t^{1-r},$$

where $C$ depends on $T, r, x, y$.

It remains to focus on the third part $e_n = \overline{X}_n - X_n$. Instead of analyzing the local error like in [25], we adopt a global point of view, and we follow the idea of the proof of Theorem 1.1 in [4]: for any $0 \leq n \leq n_0$

$$X_n - \overline{X}_n = S^n_{\Delta t}x + \Delta t \sum_{k=0}^{n-1} S^{n-k}_{\Delta t}\tilde{F}_k - S^n_{\Delta t}x - \Delta t \sum_{k=0}^{n-1} S^{n-k}_{\Delta t}\overline{F}(\overline{X}_k). \tag{3.28}$$

The averaged coefficient $\overline{F}$ is Lipschitz continuous, and $|S_{\Delta t}|_{\mathcal{L}(H)} \leq 1$; moreover we can define the averaged coefficient $\overline{F}^\tau$ with respect to the invariant measure $\mu^{x,\tau}$ of the fast numerical scheme - which is unique since we assume strict dissipativity (SD): for any $x \in H$

$$\overline{F}^\tau(x) = \int_H F(x,y)\mu^{x,\tau}(dy). \tag{3.29}$$

The error in (3.28) can then be decomposed in the following way - the idea of looking at the square of the norm in the second expression is an essential tool of the proof:

$$\begin{aligned}\mathbb{E}|X_n-\overline{X}_n| &\leq C\Delta t\sum_{k=0}^{n-1}\mathbb{E}|X_k-\overline{X}_k| \\ &+\left(\mathbb{E}|\Delta t\sum_{k=0}^{n-1}S_{\Delta t}^{n-k}\tilde{F}_k-S_{\Delta t}^{n-k}\overline{F}^{\tau}(X_k)|^2\right)^{1/2} \\ &+\Delta t\sum_{k=0}^{n-1}\mathbb{E}|S_{\Delta t}^{n-k}\overline{F}(X_k)-S_{\Delta t}^{n-k}\overline{F}^{\tau}(X_k)|.\end{aligned} \tag{3.30}$$

If we can control the two last terms by a certain quantity $Q$, by a discrete Gronwall Lemma we get for any $0\leq k\leq n_0$ $\mathbb{E}|X_n-\overline{X}_n|\leq e^{CT}Q$.

First, the third term in (3.30) is linked to the distance between the invariant measures $\mu^x$ and $\mu^{x,\tau}$ - since we assume strict dissipativity for this strong estimate - which is evaluated thanks to Theorem 3.24 and Corollary 3.25 for test functions of class $\mathcal{C}_b^2$. Since by regularization properties of the semi-groups we have $|(-A)^{\eta}S_{\Delta t}^{n-k}|_{\mathcal{L}(H)}\leq\frac{C}{((n-k)\Delta t)^{\eta}}$, we can apply Corollary 3.25 with the regular test function $S_{\Delta t}^{n-k}\overline{F}$, which thanks to Assumption 3.4 satisfies for some constant $C>0$ and for any $x,h,k\in H$

$$|(S_{\Delta t}^{n-k}\overline{F})(x)|_H\leq C, |D_x(S_{\Delta t}^{n-k}\overline{F})(x).h|_H\leq C|h|,$$

$$|D_{xx}^2(S_{\Delta t}^{n-k}\overline{F})(x).(h,k)|_H=|(-A)^{\eta}S_{\Delta t}^{n-k}(-A)^{-\eta}D_{xx}^2(S_{\Delta t}^{n-k}\overline{F})(x).(h,k)|\leq\frac{C}{((n-k)\Delta t)^{\eta}}|h||k|.$$

We thus obtain

$$|S_{\Delta t}^{n-k}\overline{F}(X_k)-S_{\Delta t}^{n-k}\overline{F}^{\tau}(X_k)|\leq\frac{C}{((n-k)\Delta t)^{\eta}}\tau^{1/2-\kappa},$$

and summing we get

$$\Delta t\sum_{k=0}^{n-1}\mathbb{E}|S_{\Delta t}^{n-k}\overline{F}(X_k)-S_{\Delta t}^{n-k}\overline{F}^{\tau}(X_k)|\leq C\tau^{1/2-\kappa}.$$

The control of the other term is more complicated; in order to get a precise estimate, we expand the square of the norm of the sum. We can then use some conditional expectations, which allow to use exponential convergence to equilibrium via Theorem 3.22. Therefore we obtain the following expansion and we treat separately each term:

$$\begin{aligned}\mathbb{E}|\Delta t\sum_{k=0}^{n-1}S_{\Delta t}^{n-k}\tilde{F}_k-S_{\Delta t}^{n-k}\overline{F}^{\tau}(X_k)|^2 &= \Delta t^2\sum_{k=0}^{n-1}\mathbb{E}|S_{\Delta t}^{n-k}(\tilde{F}_k-\overline{F}^{\tau}(X_k))|^2 \\ &+2\Delta t^2\sum_{0\leq k_1<k_2\leq n-1}\mathbb{E}<S_{\Delta t}^{n-k_1}(\tilde{F}_{k_1}-\overline{F}^{\tau}(X_{k_1})),S_{\Delta t}^{n-k_2}(\tilde{F}_{k_2}-\overline{F}^{\tau}(X_{k_2}))> \\ &=:\Sigma_1+\Sigma_2.\end{aligned}$$

*(i)* We first treat $\Sigma_1$.

We introduce the following notation: $\mathbb{E}_n$ denotes conditional expectation with respect to the $\sigma$-field

$$\mathcal{G}_n=\sigma(\zeta_{k,m,j},0\leq k\leq n-1,1\leq m\leq m_0,1\leq j\leq M). \tag{3.31}$$

We notice that $X_n$ is $\mathcal{G}_n$-measurable, but that $\tilde{F}_n$ is not.

From (3.6), for any $0\leq k\leq n-1$ we have

$$\tilde{F}_k=\frac{1}{MN}\sum_{j=1}^{M}\sum_{m=n_T}^{n_T+N-1}F(X_k,Y_{k,m,j});$$

therefore we can see that

$$\mathbb{E}|S_{\Delta t}^{n-k}(\tilde{F}_k - \overline{F}^\tau(X_k))|^2$$
$$= \frac{1}{M^2N^2}\sum_{j_1,j_2}\sum_{m_1,m_2}\mathbb{E}\mathbb{E}_k < S_{\Delta t}^{n-k}(F(X_k, Y_{k,m_1,j_1}) - \overline{F}^\tau(X_k)), S_{\Delta t}^{n-k}(F(X_k, Y_{k,m_2,j_2}) - \overline{F}^\tau(X_k)) >,$$

with the conditional expectation $\mathbb{E}_k$ with respect to $\mathcal{G}_k$ - see (3.31).

When $j_1 \neq j_2$, $\zeta_m^{(j_1)}$ and $\zeta_m^{(j_2)}$ are independent, so that if we treat differently the cases $j_1 = j_2$ and $j_1 \neq j_2$ in the above summation we obtain

$$M^2N^2\mathbb{E}|S_{\Delta t}^{n-k}(\tilde{F}_k - \overline{F}^\tau(X_k))|^2$$
$$= \sum_{j_1\neq j_2}\mathbb{E} < \sum_{m_1}\mathbb{E}_k S_{\Delta t}^{n-k}(F(X_k, Y_{k,m_1,j_1}) - \overline{F}^\tau(X_k)), \sum_{m_2}\mathbb{E}_k S_{\Delta t}^{n-k}(F(X_k, Y_{k,m_2,j_2}) - \overline{F}^\tau(X_k)) >$$
$$+ \sum_{j=1}^{M}\sum_{m_1,m_2}\mathbb{E} < S_{\Delta t}^{n-k}(F(X_k, Y_{k,m_1,j}) - \overline{F}^\tau(X_k)), S_{\Delta t}^{n-k}(F(X_k, Y_{k,m_2,j}) - \overline{F}^\tau(X_k)) > .$$

For the first part, we can directly use the exponential convergence to equilibrium result of (3.22), on each factor, to get a bound with

$$(\frac{1}{N}\sum_{m=n_T}^{m_0} e^{-cm\tau})^2 \leq (\frac{Ce^{-cn_T\tau}}{N\tau+1})^2.$$

For the second part, with no loss of generality we can treat the case $m_1 \leq m_2$, and we introduce the conditional expectation $\mathbb{E}_{k,m_1,j}$ with respect to the $\sigma$-field generated by $\mathcal{G}_k$ and $(\zeta_{km_0+m}^{(j)})_{0\leq m\leq m_1-1}$, when $m_1 \leq m_2$. The current time appearing in the exponential convergence estimate is $(m_2 - m_1)\tau$, and we get a bound with

$$\frac{2}{MN^2}\sum_{n_T\leq m_1\leq m_2\leq m_0} e^{-c(m_2-m_1)\tau} \leq \frac{C}{M(N\tau+1)}.$$

We therefore get

$$\Sigma_1 \leq C\Delta t\big((\frac{e^{-cn_T\tau}}{N\tau+1})^2 + \frac{1}{M(N\tau+1)}\big). \tag{3.32}$$

*(ii)* We now consider $\Sigma_2$, which corresponds to the cross-terms in the expansion of the square of the norm of the quantity $\sum_k S_{\Delta t}^{n-k}(\tilde{F}_k - \overline{F}^\tau(X_k))$. By the definition of $\tilde{F}_k$, the general term with indices $k_1 < k_2$ in $|\Sigma_2|$ is bounded by

$$|\mathbb{E} < S_{\Delta t}^{n-k_1}(\tilde{F}_{k_1} - \overline{F}^\tau(X_{k_1})), S_{\Delta t}^{n-k_2}(\tilde{F}_{k_2} - \overline{F}^\tau(X_{k_2})) > |$$
$$\leq \frac{\Delta t^2}{M^2N^2}|\sum_{m_i,j_i}\mathbb{E} < S_{\Delta t}^{n-k_1}(F(X_{k_1}, Y_{k_1,m_1,j_1}) - \overline{F}^\tau(X_{k_1})), S_{\Delta t}^{n-k_2}(F(X_{k_2}, Y_{k_2,m_2,j_2}) - \overline{F}^\tau(X_{k_2})) > |$$
$$\leq C\frac{\Delta t^2}{MN}\sum_{m=n_T}^{m_0}\mathbb{E}|\mathbb{E}_{k_2}[S_{\Delta t}^{n-k_2}(F(X_{k_2}, Y_{k_2,m,j}) - \overline{F}^\tau(X_{k_2})]|,$$

using conditional expectation $\mathbb{E}_{k_2}$ and the boundedness of $F$.

Using the exponential convergence result of (3.22) and Lemma 3.21, we get the bound

$$\mathbb{E}|\mathbb{E}_{k_2}[S_{\Delta t}^{n-k_2}(F(X_{k_2}, Y_{k_2,m,j}) - \overline{F}^\tau(X_{k_2})]| \leq Ce^{-m\tau},$$

so that the previous quantity is bounded by

$$C\frac{\Delta t^2}{MN}\sum_{m=n_T}^{m_0}\sum_{j=1}^{M}e^{-cm\tau} \leq C\Delta t^2\frac{e^{-cn_T\tau}}{N\tau+1}.$$

Summing on $k_1 < k_2$, we can now conclude that

$$\Sigma_2 \leq C\frac{e^{-cn_T\tau}}{N\tau+1}; \tag{3.33}$$

then by (3.32) and (3.33)

$$\mathbb{E}|\Delta t\sum_{k=0}^{n-1} S_{\Delta t}^{n-k}\tilde{F}_k - S_{\Delta t}^{n-k}\overline{F}^{\tau}(X_k)|^2 \leq C\Big(\frac{e^{-cn_T\tau}}{N\tau+1} + \Delta t(\frac{e^{-cn_T\tau}}{N\tau+1})^2 + \frac{\Delta t}{M(N\tau+1)}\Big),$$

and the result of Theorem 3.11 now follows from (3.30).

### 3.7. Proof of the weak convergence Theorem 3.12

In order to get a better bound for the weak error than for the strong error, we use an auxiliary function which is solution of a Kolmogorov equation.

The proof below only requires the weak dissipativity Assumption 3.8.

We divide the proof in two parts: the first one contains the elements of the proof, while the second one is devoted to two technical lemmas.

3.7.1. **Proof of the Theorem.** According to the decomposition (3.13), we want to control for any $0 \leq n \leq n_0$

$$\begin{aligned}|\mathbb{E}\Phi(X^\epsilon(n\Delta t)) - \mathbb{E}\Phi(X_n)| \leq\ & |\mathbb{E}\Phi(X^\epsilon(n\Delta t)) - \mathbb{E}\Phi(\overline{X}(n\Delta t))| \\ & + |\Phi(\overline{X}(n\Delta t)) - \Phi(\overline{X}_n)| \\ & + |\Phi(\overline{X}_n) - \mathbb{E}\Phi(X_n)|.\end{aligned} \tag{3.34}$$

Thanks to the averaging Theorem of [4], which is proved under the weak dissipativity assumption (WD), the first term above can be controlled by $C_r\epsilon^{1-r}$, where $C_r$ is a constant - depending on $r, \Phi, x, y, T$, for any $0 < r < 1$.

For the second term, since we look at the error made by using a deterministic scheme to approximate a deterministic equation, there is no difference between the strong and the weak orders - since the test function $\Phi$ is Lipschitz continuous; so we again use Proposition 3.26.

For the third term, we see that we have to control some error between two different numerical schemes, in a weak sense. The usual strategy is to decompose this error by means of an auxiliary function satisfying some kind of Kolmogorov equation.

More precisely, we use the deterministic scheme defining $\overline{X}_k$ in order to define for any $0 \leq k \leq n$

$$u_n(k,x) = \Phi(\overline{X}_{n-k}(x)), \tag{3.35}$$

where we explicitly mention dependence of the numerical solution $\overline{X}_k$ on the initial condition $x$.

**Remark 3.27.** *We can easily prove that we have*

$$u_n(n,x) = \Phi(x)$$
$$u_n(k,x) = u(k+1, S_{\Delta t}x + \Delta t S_{\Delta t}\overline{F}(x)) \text{ for any } k < n.$$

*This is the way this function is defined in* [25].

We now analyse the error by identifying a telescoping sum:

$$\begin{aligned}|\Phi(\overline{X}_n) - \mathbb{E}\Phi(X_n)| &= |u_n(0,x) - \mathbb{E}u_n(n,X_n)| \\ &= |\sum_{k=0}^{n-1}(\mathbb{E}u_n(k,X_k) - \mathbb{E}u_n(k+1,X_{k+1}))| \\ &\leq \sum_{k=0}^{n-1}|\mathbb{E}u_n(k,X_k) - \mathbb{E}u_n(k+1,X_{k+1})|.\end{aligned} \tag{3.36}$$

According to Lemma 3.28 below, $u_n$ is of class $\mathcal{C}_b^2$, and we can control the first and second order derivatives:

**Lemma 3.28.** *For any $0 < T < +\infty$, there exists $C_T > 0$ such that for any $0 \leq n \leq n_0 = \lfloor \frac{T}{\Delta t} \rfloor$ and $0 \leq k \leq n$, we have, for any $x \in H$, $h \in H$, $h_1, h_2 \in H$:*

$$|D_x u_n(k,x).h| \leq C_T|h|,$$
$$|D_{xx}^2 u_n(k,x).(h_1,h_2)| \leq C_T|h_1||h_2|.$$

*Moreover for any* $0 \leq k \leq n-1$ *and any* $x \in H$, $h \in H$,

$$|D_x u_n(k,x).h| \leq C_T(\Delta t|h| + \frac{|h|_{(-A)^{-\eta}}}{((n-k)\Delta t)^{\eta}}),$$

*where* $\eta$ *is defined in Assumption 3.3.*

Since the auxiliary functions $u_n$ are linked to the deterministic discrete-time process $(\overline{X}_n)_{n\geq 0}$, the proof does not use stochastic tools.

Moreover the second and the third estimates of this Lemma reveal some smoothing effect in the equation, due to the semi-group $(S^n_{\Delta t})_{n\in\mathbb{N}}$. The necessity for such results is specific to the infinite dimensional framework; we can also remark that a control of the second derivative with $C|h_1||h_2|_{(-A)^{\eta}}$ is not sufficient.

To make the proof of Theorem 3.12 clearer, we postpone the proof of Lemma 3.28 in Section 3.7.2.

In the general term of (3.36), we proceed with a Taylor expansion and using the estimates of Lemma 3.28 we get

$$\begin{aligned}
&|\mathbb{E}u_n(k,X_k) - \mathbb{E}u_n(k+1,X_{k+1})| \\
&\qquad = |\mathbb{E}u_n(k+1, S_{\Delta t}X_k + \Delta t S_{\Delta t}\overline{F}(X_k)) - \mathbb{E}u_n(k+1, S_{\Delta t}X_k + \Delta t S_{\Delta t}\tilde{F}_k)| \\
&\qquad \leq \Delta t|\mathbb{E}D_x u_n(k+1, S_{\Delta t}X_k + \Delta t S_{\Delta t}\overline{F}(X_k)).(S_{\Delta t}\overline{F}(X_k) - S_{\Delta t}\tilde{F}_k)| \\
&\qquad\quad + C\Delta t^2 \mathbb{E}|\tilde{F}_k - \overline{F}(X_k)|^2,
\end{aligned} \tag{3.37}$$

Since $F$ is bounded, the last term is of order $\mathrm{O}(\Delta t^2)$, and when we sum over $0 \leq k \leq n-1$, we get a $\mathrm{O}(\Delta t)$ term, which is already dominated in the final estimate.

When $k = n-1$, since $u_n(n,.) = \Phi$,

$$|\mathbb{E}u_n(n-1,X_{n-1}) - \mathbb{E}u_n(n,X_n)| \leq C\Delta t. \tag{3.38}$$

In the rest of the proof, we focus on the general case $0 \leq k < n-1$.

For the first term, we do not exactly follow the proof of [25]; we rather define auxiliary functions for $0 \leq k \leq n$ in order to keep on looking at a weak error term:

$$\Psi_n(k,x,y) = D_x u_n(k, S_{\Delta t}x + \Delta t S_{\Delta t}\overline{F}(x)).(S_{\Delta t}F(x,y)). \tag{3.39}$$

Then if we define $\overline{\Psi}_n(k+1,x) = \int_H \Psi_n(k+1,x,y)\mu^x(dy)$ we have

$$\begin{aligned}
&|\mathbb{E}D_x u_n(k+1, S_{\Delta t}X_k + \Delta t S_{\Delta t}\overline{F}(X_k)).(S_{\Delta t}\overline{F}(X_k) - S_{\Delta t}\tilde{F}_k)| \\
&\qquad\qquad \leq \frac{1}{MN}\sum_{m=n_T}^{n_T+N-1}\sum_{j=1}^{M}|\mathbb{E}\Psi_n(k+1,X_k,Y_{k,m,j}) - \mathbb{E}\overline{\Psi}_n(k+1,X_k)|.
\end{aligned}$$

By using conditional expectation $\mathbb{E}_k$ with respect to $\mathcal{G}_k$ defined by (3.31),

$$\mathbb{E}\Psi_n(k+1,X_k,Y_{k,m,j}) = \mathbb{E}\mathbb{E}_k\Psi_n(k+1,X_k,Y_{k,m,j})$$

does not depend on $j$, so that in the sequel we fix $j \in \{1,\dots,M\}$.

The following Lemma gives the regularity results for the auxiliary functions:

**Lemma 3.29.** *For any* $0 < T < +\infty$ *and* $\Delta t_0 > 0$, *there exists a constant* $C$, *such that for any* $0 < \Delta t \leq \Delta t_0$, *any* $1 \leq n \leq n_0 = \lfloor \frac{T}{\Delta t} \rfloor$ *and any* $0 \leq k \leq n-1$, *the following derivatives exist and are controlled: for any* $x, y \in H$, $h, k \in H$,

$$|D_y\Psi_n(k,x,y).h| \leq C|h|_H,$$

$$|D^2_{yy}\Psi_n(k,x,y).h| \leq C(1 + \frac{C}{((n-k)\Delta t)^{\eta}})|h|_H|k|_H.$$

Like in Lemma 3.28, the proof relies smoothing effect of the semi-group $(S^n_{\Delta t})_{n\in\mathbb{N}}$. The Lemma is proved below in Section 3.7.2.

We can then apply Theorem 3.24, using the conditional expectation $\mathbb{E}_k$: for any $n_T \leq m \leq m_0$ and $k < n-1$

$$|\mathbb{E}\Psi_n(k+1,X_k,Y_{k,m,j}) - \mathbb{E}\overline{\Psi}_n(k+1,X_k)| \leq Ce^{-cm\tau} + C(1 + \frac{1}{((m-1)\tau)^{1/2-\kappa}})\tau^{1/2-\kappa}.$$

Therefore

$$|\mathbb{E}D_x u_n(k+1, S_{\Delta t}X_k + \Delta t S_{\Delta t}\overline{F}(X_k)).(S_{\Delta t}\overline{F}(X_k) - S_{\Delta t}\tilde{F}_k)|$$
$$\leq \frac{1}{MN}\sum_{m=n_T}^{n_T+N-1}\sum_{j=1}^{M}|\mathbb{E}\Psi_n(k+1, X_k, Y_{k,m,j}) - \mathbb{E}\overline{\Psi}_n(k+1, X_k)|$$
$$\leq C\frac{e^{-cn_T\tau}}{N\tau+1} + C\tau^{1/2-\kappa} + C\frac{\tau^{1/2-\kappa}}{((n_T-1)\tau)^{1/2-\kappa}}.$$

To conclude, it remains to use (3.36) and (3.34).

**Remark 3.30.** *When the strict dissipativity Assumption is satisfied, we can indeed obtain a bound without* $\frac{1}{((m-1)\tau)^{1/2-\kappa}}$*: we can control the distance between the invariant measures of the continuous and discrete time processes, thanks to the second part of Corollary 3.25.*

3.7.2. **Proof of the auxiliary Lemmas 3.28 and 3.29.**

Proof of Lemma 3.28 We use the following expression for $\overline{X}_n(x)$:

$$\overline{X}_n(x) = S_{\Delta t}^n x + \Delta t\sum_{k=0}^{n-1} S_{\Delta t}^{n-k}\overline{F}(\overline{X}_k(x)).$$

By definition, for any $0 \leq k \leq n$ we have $u_n(k,x) = \Phi(\overline{X}_{n-k}(x))$; we see that the derivatives in directions $h, h_1, h_2 \in H$ are given by

$$D_x u_n(k,x).h = D\Phi(\overline{X}_{n-k}(x)).(\frac{d}{dx}\overline{X}_{n-k}(x).h),$$

and

$$D_{xx}^2 u_n(k,x).(h_1,h_2) = D\Phi(\overline{X}_{n-k}(x)).(\frac{d^2}{dx^2}\overline{X}_{n-k}(x).(h_1,h_2))$$
$$+ D^2\Phi(\overline{X}_{n-k}(x)).(\frac{d}{dx}\overline{X}_{n-k}(x).h_1, \frac{d}{dx}\overline{X}_{n-k}(x).h_2).$$

$\Phi$ is of class $\mathcal{C}^2$ on $H$, with bounded derivatives; therefore we just need to control $\frac{d}{dx}\overline{X}_{n-k}(x).h$ and $\frac{d^2}{dx^2}\overline{X}_{n-k}(x).(h_1,h_2)$. We use the following estimates of the derivatives of $\overline{F}$, given in Proposition 3.18: for any $x \in H$, $h \in H$, $h_1, h_2 \in H$,

$$|D\overline{F}(x).h| \leq C|h|$$
$$|(-A)^{-\eta}D^2\overline{F}(x).(h_1,h_2)| \leq C|h_1||h_2|.$$

*(i)* For any $0 \leq n \leq n_0$, we can write

$$\frac{d}{dx}\overline{X}_n(x).h = S_{\Delta t}^n h + \Delta t\sum_{k=0}^{n-1} S_{\Delta t}^{n-k}D\overline{F}(\overline{X}_k(x))(\frac{d}{dx}\overline{X}_k(x).h);$$

therefore

$$|\frac{d}{dx}\overline{X}_n(x).h| \leq |h| + C\Delta t\sum_{k=0}^{n-1}|\frac{d}{dx}\overline{X}_k(x).h|,$$

and a discrete Gronwall Lemma then yields

$$|\frac{d}{dx}\overline{X}_n(x).h| \leq |h|e^{Cn\Delta t} \leq |h|e^{CT}.$$

*(ii)* For any $0 \leq n \leq n_0$, we can write

$$\frac{d^2}{dx^2}\overline{X}_n(x).(h_1,h_2) = \Delta t \sum_{k=0}^{n-1} S_{\Delta t}^{n-k}(-A)^{\eta}(-A)^{-\eta} D^2\overline{F}(\overline{X}_k(x)).(\frac{d}{dx}\overline{X}_k(x).h_1, \frac{d}{dx}\overline{X}_k(x).h_2)$$
$$+ \Delta t \sum_{k=0}^{n-1} S_{\Delta t}^{n-k} D\overline{F}(\overline{X}_k(x)).(\frac{d^2}{dx^2}\overline{X}_k(x).(h_1,h_2)).$$

Since $|S_{\Delta t}^{n-k}(-A)^{\eta}| \leq \frac{C}{((n-k)\Delta t)^{\eta}}$ when $k < n$, and thanks to the previous estimates on $\frac{d}{dx}\overline{X}_k(x).h$ and $D^2\overline{F}$, we get

$$|S_{\Delta t}^{n-k}(-A)^{\eta}(-A)^{-\eta} D^2\overline{F}(\overline{X}_k(x)).(\frac{d}{dx}\overline{X}_k(x).h_1, \frac{d}{dx}\overline{X}_k(x).h_2)| \leq \frac{C}{((n-k)\Delta t)^{\eta}}|h|_1|h|_2.$$

Therefore

$$|\frac{d^2}{dx^2}\overline{X}_n(x).(h_1,h_2)| \leq C|h_1||h_2|$$
$$+ C\Delta t \sum_{k=0}^{n-1} |\frac{d^2}{dx^2}\overline{X}_k(x).(h_1,h_2)|,$$

and a discrete Gronwall Lemma then yields

$$|\frac{d^2}{dx^2}\overline{X}_n(x).(h_1,h_2)| \leq C|h_1||h_2|.$$

*(iii)* To prove the last estimate of the Lemma, we write

$$|\frac{d}{dx}\overline{X}_n(x).h| = |S_{\Delta t}^n h + \Delta t \sum_{k=0}^{n-1} S_{\Delta t}^{n-k} D\overline{F}(\overline{X}_k(x))(\frac{d}{dx}\overline{X}_k(x).h)|$$
$$\leq \frac{C}{(n\Delta t)^{\eta}}|h|_{(-A)^{-\eta}} + C\Delta t \sum_{j=0}^{n-1} |\frac{d}{dx}\overline{X}_k(x).h|.$$

We obtain

$$(n\Delta t)^{\eta}|\frac{d}{dx}\overline{X}_n(x).h| \leq C|h|_{(-A)^{-\eta}} + C\Delta t(n\Delta t)^{\eta}|h|_H$$
$$+ C(n\Delta t)^{\eta}\Delta t \sum_{j=1}^{n-1} \frac{1}{(j\Delta t)^{\eta}}(j\Delta t)^{\eta}|\frac{d}{dx}\overline{X}_j(x).h|;$$

To conclude, we use Gronwall Lemma, to get

$$(n\Delta t)^{\eta}|\frac{d}{dx}\overline{X}_n(x).h| \leq C_T(|h|_{(-A)^{-\eta}} + \Delta t(n\Delta t)^{\eta}).$$

□

Proof of Lemma 3.29

*(i)* The first derivative with respect to $y$ is easy to control: we have for any $h \in H$

$$D_y\Psi_n(k,x,y).h = D_x u_n(k, S_{\Delta t}x + \Delta t S_{\Delta t}\overline{F}(x)).(S_{\Delta t}D_yF(x,y).h),$$

and we get $|D_y\Psi_n(k,x,y).h| \leq C|h|_H$.

*(ii)* When we look at the second-order derivative, we see that

$$D^2_{yy}\Psi_n(k,x,y).(h,k) = D_x u_n(k, S_{\Delta t}x + \Delta t S_{\Delta t}\overline{F}(x)).(S_{\Delta t}D^2_{yy}F(x,y).(h,k)).$$

Thanks to the last estimate of Lemma 3.28, we can control the expression by

$$C\left(\Delta t|S_{\Delta t}D^2_{yy}F(x,y).(h,k)|_H + \frac{|S_{\Delta t}D^2_{yy}F(x,y).(h,k)|_{(-A)^{-\eta}}}{((n-k)\Delta t)^{\eta}}\right),$$

We then notice that

$$\begin{aligned}\Delta t|S_{\Delta t}D^2_{yy}F(x,y).(h,k)|_H &\leq C\Delta t^{1-\eta}|(-A)^{-\eta}D^2_{yy}F(x,y).(h,k)|\\ &\leq C|h|_H|k|_H,\end{aligned}$$

since $\eta < 1$ and $\Delta t$ is bounded, and thanks to Assumption 3.4; the other part is controlled thanks to Assumption 3.4. □

# Chapitre 4

# Schéma hybride semi-lagrangien: présentation et analyse

Résumé

On s'intéresse à l'approximation de solutions d'EDP déterministes de type parabolique, à l'aide de méthodes combinant un principe semi-lagrangien et la discrétisation d'une expérance (apparaissant dans une formule de représentation probabiliste de la solution) par une méthode de Monte-Carlo. Dans un cas simple, on montre une estimation de l'erreur, signifiant que l'erreur sur la variance est dominée par le pas de temps de la méthode, sous une condition de type anti-CFL.

## 4.1. Introduction

We consider a class of numerical methods which combine the principles of semi-lagrangian and Monte-Carlo methods. We present their construction and their analysis in a simple case: we consider a diffusion equation on the domain $D = \mathbb{R}^d$

$$\begin{aligned} \frac{\partial u(t,x)}{\partial t} &= \frac{1}{2}\Delta u(t,x) + c(x).\nabla u(t,x), \text{ for any } 0 < t \leq T \text{ and } x \in \mathbb{R}^d \\ u(0,x) &= u_0(x), \text{ for any } x \in \mathbb{R}^d, \end{aligned} \tag{4.1}$$

with a given initial condition $u_0$, assumed to be smooth. The function $c$ is assumed to be smooth. We notice that with $D = \mathbb{R}^d$ there is no boundary condition, so that the presentation is simplified. Later in this Chapter, we study the case with $c = 0$, but with periodic boundary conditions on the domain $D = (0,1)$. In Chapter 5, we give elements for more general situations.

The solution $u$ of this PDE admits a probabilistic representation: for any $t \geq 0$ and $x \in \mathbb{R}^d$ we have

$$u(t,x) = \mathbb{E}u_0(X_t^x),$$

where $(X_t^x)_{t\geq 0}$ is solution of the SDE

$$dX_t^x = c(X_t^x)dt + dB_t, \quad X_0^x = x, \tag{4.2}$$

where $(B_t)_{t\geq 0}$ is a standard $d$-dimensional Brownian Motion.

The infinitesimal generator of this diffusion process is $\mathcal{L} = c.\nabla + \frac{1}{2}\Delta$, and we see that $\frac{\partial u}{\partial t} = \mathcal{L}u$.

In general, the law of the random variable $X_t^x$ is not explicitly known, and we are not able to compute the expectation. The classical approximation procedures for such a problem are Monte-Carlo methods.

We recall that in order to approximate the expectation of a random variable $Y$, we simulate a sample $(Y_1, \ldots, Y_N)$ of size $N$ - i.e. $N$ independent realizations of $Y$ - and we define the empirical average

$$\overline{Y} = \frac{1}{N}\sum_{m=1}^{N} Y_m.$$

The Law of Large Numbers ensures that if $Y$ is integrable, almost surely when $N \to \infty$ we have the convergence $\overline{Y} \to \mathbb{E}Y$; moreover we have a control of the error when $Y$ is assumed to have a finite second moment: we get $\left(\mathbb{E}|\overline{Y} - \mathbb{E}Y|^2\right)^{1/2} \leq \frac{\sqrt{\mathrm{Var}(Y)}}{\sqrt{N}}$.

Therefore if we assume that we are able to compute $N$ independent realizations $(X_t^{x,m})_{1\leq m\leq N}$ of with the law of $X_t^x$, we can approach $u(t,x)$ with

$$\frac{1}{N}\sum_{m=1}^{N} u_0(X_t^{x,m}). \tag{4.3}$$

In general, the variance of the random variables $u_0(X_t^{x,m})$ is of size $t$.

If a global knowledge of the solution is required, the above operation must be repeated for different values of $x$, for instance on a spatial grid $(x_j = j\delta x)_{j\in\mathbb{N},j\in\mathbb{N}}$ with mesh size $\delta x > 0$.

A difficulty arises since the law of random variables $X_t^x$ is not known, and since there is no direct method to compute the random variables $X_t^{x,m}$ used in (4.3). An approximation of the law of the random variables can be obtained thanks to a numerical scheme, for instance of Euler type: if $\tau > 0$ denotes the time step of

the method, we define recursively

$$X_0(x) = x,$$
$$X_{n+1}(x) = X_n(x) + \tau c(X_n(x)) + (B_{(n+1)\tau} - B_{n\tau}), \text{ for } n \geq 0. \tag{4.4}$$

The error induced by this scheme is in a weak sense of order 1 with respect to $\tau$: if $x \in \mathbb{R}^d$ and if $\varphi : \mathbb{R}^d \to \mathbb{R}$ is a smooth function, there exists a constant $C > 0$ such that for any $n \geq 0$ such that $n\tau \leq t$ we have

$$|\mathbb{E}\varphi(X^x_{n\tau}) - \mathbb{E}\varphi(X_n(x))| \leq C\tau.$$

We can see that these discretization methods can be used, but that the associated variances in the Monte-Carlo method are not small, even if $\delta t$ and $\delta x$ are small. Moreover, for each new computed value the procedure must be started again. The method we now propose uses more carefully the time and space dependence of the solutions, and aims at reducing the variances. Moreover, we directly obtain a global approximation defined on the whole spatial domain, in an intrinsic way.

The construction relies on a generalization of the probabilistic representation formula which links the values of the solution at different times $t_1 < t_2$, thanks to the Markov property of the associated stochastic processes: in our situation we have for any $x \in \mathbb{R}^d$

$$u(t_2, x) = \mathbb{E}u(t_1, X^x_{t_2 - t_1}), \tag{4.5}$$

Given a time step $\delta t > 0$ and a mesh size $\delta x > 0$, (4.5) gives a way for defining approximations of $u((n+1)\delta t, j\delta x)$ from the knowledge of the solution $u(n\delta t, .)$ at time $n\delta t$. Moreover to get an approximation of the law of $X^x_{\delta t}$ we can use the Euler method (4.4) with $\tau = \delta t$.

It thus remains to define an interpolation procedure, which allows to recover a function defined on the closure of the domain $\overline{D}$ from given values at the nodes $(j\delta x)_j$: we need an operator $\mathcal{J}$ which maps a vector $w = (w_j)_j$ to a function $\mathcal{J}(w) : \overline{D} \to \mathbb{R}$. In the analysis below, we consider linear interpolation.

We can now detail the recursion of the method: we start from the initial vector $u^0 = (u_0(k\delta x))_k$. Then given $u^n = (u^n_k)_k$, we first approach

$$u((n+1)\delta t, j\delta x) = \mathbb{E}u(n\delta t, X^{j\delta x}_{\delta t})$$

with

$$\mathbb{E}u(n\delta t, j\delta x + \delta t c(j\delta x) + B_{\delta t}) = \mathbb{E}u(n\delta t, j\delta x + \delta t c(j\delta x) + \sqrt{\delta t}\mathcal{N}),$$

thanks to an iteration of the Euler method, where $\mathcal{N}$ denotes a standard Gaussian random variable, with mean 0 and variance 1; then the interpolation procedure gives the next approximation with

$$v^{n+1}_j := \mathbb{E}\big(\mathcal{J}(u^n)\big)(j\delta x + \delta t c(j\delta x) + \sqrt{\delta t}\mathcal{N});$$

finally the Monte-Carlo approximation is used to define

$$u^{n+1}_j := \frac{1}{N}\sum_{m=1}^{N} \big(\mathcal{J}(u^n)\big)(j\delta x + \delta t c(j\delta x) + \sqrt{\delta t}\mathcal{N}^{n,m,j}),$$

where for fixed $n$ and $j$ the random variables $(\mathcal{N}^{n,m,j})_{1 \leq m \leq M}$ are independent standard Gaussian random variables. Moreover a fundamental assumption is that independence with respect to the parameters $n$ and $j$ also holds. On the one hand, we then break a symmetry of the problem in the space variable with a noise which is white in space; on the other hand, the noise has an effect on the convergence which is essential in the proof of Theorem 4.1.

The principle of using (random) characteristic curves in (4.5) over a time interval of size $\delta t$ and to use an interpolation procedure to get functions defined on the whole domain fits in the semi-lagrangian framework. The addition of the Monte-Carlo approximation then justifies the use of the hybrid terminology.

We can notice that we have defined explicit methods, and that they do not need CFL conditions: we rather require an anti-CFL condition - see the estimate (4.8) - due to the semi-lagrangian construction of the scheme.

Now that we have presented the method, we focus on a simple case: we consider the heat equation - $c = 0$ - in dimension $d = 1$. We moreover assume that periodic boundary conditions are satisfied on $D = (0, 1)$:

$$\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2}$$
$$u(0, .) = u_0,$$

with periodic boundary conditions on the domain $D = (0, 1)$, and a smooth, periodic initial condition $u_0 : \mathbb{R} \to \mathbb{R}$.

We also introduce in Definition 4.2 a discrete norm $\|.\|_{l^2}$ and a discrete semi-norm $|.|_{h^1}$, which are discrete counterparts of the norm in $L^2(0, 1)$ and the semi-norm in $H^1(0, 1)$. We consider the full-discretization error, which can be decomposed into an error due to randomness - we control its variance in the $l^2$ norm, depending on the $h^1$ semi-norm of the initial condition - and an error only due to deterministic effects - corresponding to the accumulation in time of the interpolation error. The error estimate is given in the following Theorem:

**Theorem 4.1.** *Assume that the initial condition $u_0$ is of class $\mathcal{C}^2$.*

*For any $p \in \mathbb{N}$ and any final time $T > 0$, there exists a constant $C_p > 0$, such that for any $\delta t > 0$, $\delta x > 0$ and $N \in \mathbb{N}^*$ we have*

$$(u(n\delta t, k\delta x))_k - u^n = (u(n\delta t, k\delta x))_k - v^n) + (v^n - u^n), \tag{4.6}$$

*with*

$$\sup_{j\in\mathbb{N};0\leq j\delta x<1} |u(n\delta t, x_j) - v_j^n| \leq C\frac{\delta x^2}{\delta t} \sup_{x\in[0,1]} |u_0^{''}(x)| \tag{4.7}$$

*and*

$$\begin{aligned} \mathbb{E}\|u^n - v^n\|_{l^2}^2 &= \delta x \sum_{j;0\leq j\delta x<1} \mathbb{E}|u_j^n - v_j^n|^2 \\ &\leq C_p|u^0|_{h^1}^2(1 + \frac{\delta x^2}{\delta t}) \left(1 + \frac{\delta x}{\delta t} + \frac{\delta x^2}{\delta t^2}(1 + |\log(\delta t)|)\right)^p \left(\frac{\delta t}{N} + \frac{1}{N^{p+1}}\right). \end{aligned} \tag{4.8}$$

The control of the first part of the error is rather classical, while the estimate on the Monte-Carlo error given by (4.8) is more original and requires more attention in its analysis and in its proof.

First, we observe that the estimate is only interesting if a condition of anti-CFL type is satisfied: for some constant $c > 0$ we require

$$\frac{\delta x}{\delta t}\max(1, \sqrt{\log(\delta t)}) < c.$$

We can then identify in (4.8) a leading term of size $\frac{\delta t}{N}$, which corresponds to the statistical error in a Monte-Carlo method for random variables of variance $\delta t$, and a remaining term, which goes to 0 with arbitrary order of convergence with respect to the number of realizations $N$. This second term is obtained via a bootstrap argument. Indeed it is easy to get the classical estimate with $p = 0$. The core of the proof is contained in the recursion which allows to increase the order from $p$ to $p + 1$; it heavily relies on the spatial structure of the noise and on the choice of the $l^2$-norm.

We can give a nice interpretation of the case $p = 1$: we have

$$\frac{\delta t}{N} + \frac{1}{N^2} \leq \frac{\delta t^2}{2} + \frac{3}{2N^2},$$

which can be interpreted as a bound on a classical Monte-Carlo estimate with the variance $\delta t$: we have for any sample $(Y_1, \ldots, Y_N)$ of a random variable $Y$

$$\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^N Y_i\right) = \frac{\mathrm{Var}(Y)}{N} \leq \frac{\mathrm{Var}(Y)^2}{2} + \frac{1}{2N^2}.$$

Indeed, the estimate (4.8) is better, and shows that the interpretation with a variance $\delta t$ is not exact but quite precise.

The control of the Monte-Carlo error in Theorem 4.1 relies on various arguments. First, the first factor corresponds to the accumulation of the variances appearing at each time step - where two sources of error can be identified: the random variables involves a stochastic diffusion process evaluated at time $\delta t$, and an error

is introduced by the interpolation procedure. To obtain another factor, we observe that the independence of the random variables appearing for different nodes implies that only diagonal entries of some matrices appear. However, this independence property also complicates the proof: the solutions are badly controlled with respect to the $h^1$ semi-norm. We then propose a decomposition of the error where the number of realizations $N$ appears in the variance with the different orders 1 and 2: the first part is controlled by $\delta t$ and $\delta x$, while the second one is only bounded. We finally use recursively this decomposition in order to improve the estimate, with a bootstrap argument.

We can mention that Monte-Carlo simulations are not always necessary, and deterministic methods using the probabilistic representation formula for the solutions of the PDE have been developed: another way of approximating the expectations $\mathbb{E}\Psi(X)$ is via quantization of the law, where the idea is to approximate the random variable $X$ by another one $\tilde{X}$ with a finite support, such that the error between $\mathbb{E}\Psi(X)$ and $\mathbb{E}\Psi(\tilde{X})$ is controlled. For example, for a smooth test function $\phi$, if $(B_t)$ is a one-dimensional Brownian Motion, we have by Itô formula

$$\begin{aligned}\mathbb{E}\phi(B_t) &= \frac{1}{2}(\phi(\sqrt{t}) + \phi(-\sqrt{t})) + \mathrm{O}(t^2) \\ &= \mathbb{E}\phi(\hat{B}_t) + \mathrm{O}(t^2),\end{aligned}$$

if $\hat{B}_t$ is random variable with the law $\frac{1}{2}(\delta_{\sqrt{t}} + \delta_{-\sqrt{t}})$. This idea leads to the definition of the layer methods - see [54] and references therein - which are deterministic methods based on a probabilistic interpretation. Various well-known finite difference methods are then recovered. In [28], this idea leads to the definition and the analysis of semi-lagrangian methods. Nevertheless, the Monte-Carlo method seems simpler and more general; moreover the result we obtain in the present article seems to be specific in the context of Monte-Carlo methods.

The Chapter is organized as follows. In Section 4.2, we present the method and introduce the various notations. We give there the definition of the discrete norms and of the matrices which define the numerical method. In Section 4.3, we provide the proof of Theorem 4.1, where we divide the control of the Monte-Carlo error of (4.8) into three steps: first, we explain how the error can be decomposed; second, we study the error induced by one step of the scheme; third, we explain how the independence of the random variables for different nodes implies the improved bound. A proof of (4.7) is given in Section 4.3.4.

Extensions of the numerical method with numerical experiments are given in Chapter 5.

## 4.2. Presentation of the numerical method

This section is devoted to the description of the numerical method and its application to the heat equation in dimension 1, in the domain $(0,1)$ with periodic boundary conditions. This simple example is chosen for several reasons: first, the standard Brownian motion is the associated stochastic process. Second, due to the periodic setting the problem is invariant by translation in space and calculations below do not contain additional boundary terms - for example in integration by parts.

However, this simple problem does not have a great practical interest. We have seen in the Introduction that the hybrid method can be used for more complex and interesting situations, but the analysis of the simplest case is already challenging.

Below, we first briefly present the PDE, and then we explain how space and time are discretized. We also define the discrete norms and give approximation results with respect to the continuous counterparts. Next, we precise the definition of the numerical method, and introduce its matrix-description. Various properties of these matrices are given.

4.2.1. **The continuous setting.** We consider the linear heat equation on $(0,1)$, with a smooth initial condition and with periodic boundary conditions; more precisely, we want to approximate the unique periodic solution of the following partial differential equation:

$$\frac{\partial u}{\partial t} = \frac{1}{2}\frac{\partial^2 u}{\partial x^2},$$

such that $u(0,.) = u_0 : \mathbb{R} \to \mathbb{R}$ is a smooth periodic function of period 1, and such that for any $t \geq 0$ the function $u(t,.)$ is periodic of period 1.

Our concern is the approximation of the solution of this equation at a fixed time $T > 0$, using a hybrid numerical scheme, which mixes a semi-lagrangian method, and a Monte-Carlo approximation associated with a probabilistic representation of the solution.

4.2.2. **The discrete setting.**

4.2.2.1. *Discretization and interpolation procedure.* We need to discretize time and space, and to define an interpolation procedure allowing to recover functions defined on the line from given values at the nodes.

First, we consider a final time $T > 0$, and an integer $M_T$, such that we divide the interval $[0, T]$ into $M_T$ intervals of size $\delta t := \frac{T}{M_T}$. We are thus interested in approximating the solution $u$ at times $t_n = n\delta t$ for $n \leq M_T$. The constants in the error bounds may depend on the finite time $T$ but not on the discretization time step $\delta t$.

We discretize the space interval $[0, 1]$ - which represents a canonical continuous spatial period - with the introduction of nodes $x_j = j\delta x$ for $j \in \{0, \dots, M_S\}$, with the condition $x_{M_S} = M_S\delta x = 1$. By periodicity, the definition of $x_j$ may be extended for indices in $\mathbb{Z}$. In sums written below, we may forget to mention the set of indices: it must be then considered that we sum over a period: for example we often choose a canonical discrete period $S = \{0, \dots, M_S - 1\}$, which represents the nodes in $[0, 1)$, since by periodicity the information contained in values at index $j = M_S$ already appears for index $j = 0$.

We use linear interpolation to reconstruct functions on the whole interval from values at the nodes. We can define an appropriate basis made of periodic and piecewise linear functions for $k \in S = \{0, \dots, M_S - 1\}$. They are defined using the following shape function defined on $\mathbb{R}$ by

$$\hat{\phi}(x) = \begin{cases} 0 \text{ if } |x| > 1, \\ 1 - |x| \text{ if } |x| \leq 1. \end{cases} \tag{4.9}$$

For any $k \in S$, the function $\phi_k$ is piecewise linear, periodic, and satisfies

$$\phi_k(x_j) = \begin{cases} 1 \text{ if } k = j, \\ 0 \text{ if } k \neq j. \end{cases} \tag{4.10}$$

As a consequence, for any $k \in S$ we have on the period $I_k := [x_k - 1/2; x_k + 1/2]$ centered in $x_k$

$$\phi_k(x) = \hat{\phi}(\frac{x - x_k}{\delta x}), \tag{4.11}$$

We remark that for any $x \in [0, 1]$

$$\phi_k(x) \geq 0 \quad \text{and} \quad \sum_{k \in S} \phi_k(x) = 1. \tag{4.12}$$

Moreover we have the following useful property: for any $k \in S$ and any $x \in \mathbb{R}$

$$\phi_{k+1}(x) = \phi_k(x - \delta x) \tag{4.13}$$

We finally define two important operators associated with the interpolation. The first one is projection $\mathcal{P}$: given some periodic function defined on the whole space, it associates the values obtained at the nodes. The second one is the interpolation function $\mathcal{I}$, which gives a piecewise linear and periodic function from the values specified at the nodes. More precisely, if we denote by $\mathcal{B}_{p,1}(\mathbb{R})$ the set of all 1-periodic functions with values in $\mathbb{R}$, and by $\mathcal{C}_{p,1}(\mathbb{R})$ the set of the continuous 1-periodic functions, we can define the following operators

$$\mathcal{P} : \begin{cases} \mathcal{B}_{p,1}(\mathbb{R}) \to \mathbb{R}^{M_S} \\ f \mapsto (f(x_j))_{0 \leq j \leq M_S - 1}, \end{cases} \tag{4.14}$$

and

$$\mathcal{I} : \begin{cases} \mathbb{R}^{M_S} \to \mathcal{C}_{p,1}(\mathbb{R}) \\ u = (u_k)_{0 \leq k \leq M_S - 1} \mapsto \sum_{k=0}^{M_S - 1} u_k \phi_k. \end{cases} \tag{4.15}$$

This definition can be extended in an obvious way to define periodic sequences - of period $M_S$ - $(f(x_j))$ or $(u_k)$.

Clearly, $\mathcal{P}\circ\mathcal{I}$ is the identity on $\mathbb{R}^{M_S}$; nevertheless the distance between the identity and the composition of the operators $\mathcal{I}\circ\mathcal{P}$ depends on the functional spaces and on the norms. Below, we give the estimates that are useful in our setting. We just notice that (4.12) means that interpolation is exact for constant functions, which is simply written $\mathcal{I}\circ\mathcal{P}(\mathbb{1}) = \mathbb{1}$.

4.2.2.2. *Discrete norms.* We define appropriate norms on $\mathbb{R}^{M_S}$, which are related to norms for functions defined on continuous intervals. We moreover state some important properties of the operators $\mathcal{P}$ and $\mathcal{I}$ related to the norms, which give estimates of the error made by interpolation.

Below $J$ denotes any period; for definiteness, one can take $J = S := \{0, \dots, M_S - 1\}$.

**Definition 4.2.** *The discrete $L^2$ norm, which we denote by $\| \, . \, \|_{l^2}$, is defined for any $u = (u_j)_{j\in S}$ by*

$$\|u\|_{l^2}^2 = \delta x \sum_{j\in S} u_j^2.$$

*The discrete $H^1$ semi-norm, which we denote by $| \, . \, |_{h^1}$, is defined for any $u = (u_j)_{j\in S}$ by*

$$|u|_{h^1}^2 = \delta x \sum_{j\in S} \frac{(u_{j+1} - u_j)^2}{\delta x^2}.$$

In the definition of $|u|_{h^1}^2$, one must use the extension by periodicity of the sequence $(u_j)$: we thus have $u_{M_S} = u_0$. As in the continuous case, with $| \, . \, |_{h^1}$ we only define a semi-norm: if $|u|_{h^1} = 0$, then there exists $c \in \mathbb{R}$ such that for any $0 \le j \le M_S - 1$ we have $u_j = \frac{1}{M_S}\sum_{k=0}^{M_S-1} u_k$.

In order to define an appropriate norm, we can consider $\|u\|_{h^1} = (\|u\|_{l^2}^2 + |u|_{h^1}^2)^{1/2}$.

Both discrete norm and semi-norm are approximations of the corresponding quantities on the whole interval $[0, 1]$:

$$\|f\|_{L^2(0,1)}^2 = \int_0^1 |f(x)|^2 dx, \text{ for } f \in L^2(0,1),$$
$$|f|_{H^1(0,1)}^2 = \int_0^1 |f'(x)|^2 dx, \text{ for } f \in H^1(0,1).$$

We have the following properties:

**Proposition 4.3.** *There exists a constant $c > 0$ such that for any mesh size $\delta x = 1/M_S$, and any sequence $u = (u_j)$ we have:*

$$|u|_{h^1} = |\mathcal{I}u|_{H^1(0,1)}$$
$$\|u\|_{l^2}^2 = \|\mathcal{I}u\|_{L^2(0,1)}^2 + c\delta x^2 |u|_{h^1}^2.$$

*Moreover, for any function $f \in H^1(0,1)$ we have*

$$\|f - (\mathcal{I}\circ\mathcal{P})f\|_{L^2(0,1)}^2 \le c\delta x^2 \left(|f|_{H^1(0,1)}^2 + |(\mathcal{I}\circ\mathcal{P})f|_{H^1(0,1)}^2\right).$$

**Remark 4.4.** *Thanks to a usual Sobolev embedding, $H^1(0,1) \subset C^0(0,1)$ the set of continuous functions, and therefore $\mathcal{P}f$ is well-defined.*

Proof It is important to remark that the first equality of the Proposition above is specific to the use of linear interpolation.

The second one is proved expanding in the $L^2$ scalar product $\mathcal{I}u = \sum_k u_k\phi_k$, and rewriting the sums in order to make the $h^1$ semi-norm appear: we have

$$\begin{aligned}\|\mathcal{I}u\|_{L^2(0,1)}^2 &= \|\sum_{k\in S} u_k\phi_k\|_{L^2(0,1)}^2 \\ &= \sum_{k,\ell\in S} u_k u_\ell \langle \phi_k, \phi_\ell\rangle_{L^2(0,1)} \\ &= \frac{2\delta x}{3}\sum_{k\in S} u_k^2 + \frac{\delta x}{6}\sum_{k\in S}(u_k u_{k+1} + u_k u_{k-1}).\end{aligned}$$

Indeed, the interpolation basis satisfies the following properties:

$$\langle \phi_k, \phi_\ell \rangle_{L^2(0,1)} = 0 \text{ if } \ell \notin \{k-1, k, k+1\}$$
$$\langle \phi_k, \phi_k \rangle_{L^2(0,1)} = \frac{2\delta x}{3}$$
$$\langle \phi_k, \phi_{k-1} \rangle_{L^2(0,1)} = \langle \phi_k, \phi_{k+1} \rangle_{L^2(0,1)} = \frac{\delta x}{6}.$$

Now, the equality contains $\|u\|_{h^1}^2$ which appears with natural integration by parts - using periodicity:

$$\begin{aligned}
\|u\|_{l^2}^2 - \|\mathcal{I}u\|_{L^2(0,1)}^2 &= \frac{\delta x}{6} \sum_{k \in S} (u_k(u_k - u_{k-1}) + u_k(u_k - u_{k+1})) \\
&= \frac{\delta x}{6} \sum_{k \in S} (u_{k+1}(u_{k+1} - u_k) + u_k(u_k - u_{k+1})) \\
&= \frac{\delta x}{6} \sum_{k \in S} (u_{k+1} - u_k)^2 \\
&= \frac{1}{6} \delta x^2 \|u\|_{h^1}^2.
\end{aligned}$$

We now focus on the proof of the last estimate.

We begin with the case of a smooth periodic function $f \in \mathcal{C}^1(\mathbb{R})$. By a density argument, we can easily extend the result for functions in $H^1(0,1)$.

For any $j \in S = \{0, 1, \dots M_S - 1\}$, for any $x \in [x_j, x_{j+1}]$, we have

$$\begin{aligned}
|f(x) - (\mathcal{I} \circ \mathcal{P} f)(x)|^2 &\leq 2(\int_{x_j}^{x} f'(t) dt)^2 + 2(\int_{x_j}^{x} \frac{f(x_{j+1}) - f(x_j)}{\delta x} dt)^2 \\
&\leq 2(x - x_j) \int_{x_j}^{x_{j+1}} |f'(t)|^2 dt + 2\delta x^2 \frac{|f(x_{j+1}) - f(x_j)|^2}{\delta x^2},
\end{aligned}$$

using the Cauchy-Schwarz inequality. Now we integrate over $[x_j, x_{j+1}]$, and then it remains to take the sum over $j \in S$ of the following quantities:

$$\int_{x_j}^{x_{j+1}} |f(x) - (\mathcal{I} \circ \mathcal{P} f)(x)|^2 \leq \delta x^2 \int_{x_j}^{x_{j+1}} |f'(x)|^2 dx + 2\delta x^2 \frac{|f(x_{j+1}) - f(x_j)|^2}{\delta x^2} \delta x. \tag{4.16}$$

The first term of the right-hand side is controlled with $|f|_{H^1(0,1)}^2$, while the second term involves $|\mathcal{P}f|_{h^1}^2 = |(\mathcal{I} \circ \mathcal{P})f|_{H^1(0,1)}^2$. □

One can also notice that if the function $f$ is of class $\mathcal{C}^1$ on $[0,1]$, we get

$$|f|_{H^1(0,1)}^2 + |(\mathcal{I} \circ \mathcal{P})f|_{H^1(0,1)}^2 \leq \sup_{x \in [0,1]} |f'(t)|^2.$$

4.2.3. **Definition of the numerical method.** The method is based on the following representation formula for the solution of the PDE:

$$u(t,x) = \mathbb{E} u_0(x + B_t).$$

The idea of the hybrid semi-lagrangian method is to combine the interpolation procedure with the evolution at each-time step based on this formula. The Monte-Carlo error appears since in full generality we cannot compute exactly the expectation in the formula above for every test function $u_0$, or when other stochastic processes are involved. As a consequence, we compute an approximation with statistical averages - which converge to the expected value when the number of simulations goes to infinity.

The number of independent realizations we need for this approximation is denoted by $N$. According to the general Monte-Carlo methodology, we expect an error of size $\frac{1}{\sqrt{N}}$; the main result here is that this error also goes to 0 with respect to the time-step.

We start with an initial condition $u^0 = (u_k^0 = u_0(x_k))$, which contains the values of the initial condition at the nodes. To obtain simple expressions with products of matrices, we consider that vectors like $u^0$ are column vectors.

Having calculated the approximation at discrete time $n$, we can define the values at time $n+1$: for any $j \in S$ the canonical set $S$

$$u_j^{n+1} = \frac{1}{N} \sum_{m=1}^{N} \left( \sum_{k \in S} u_k^n \phi_k(x_j + \sqrt{\delta t} \mathcal{N}^{n,m,j}) \right), \tag{4.17}$$

where the random variables $\mathcal{N}^{n,m,j}$, indexed by $0 \leq n \leq M_T - 1$, $1 \leq m \leq N$ and $j \in S$ are independent standard normal variables.

More precisely, to avoid an error term due to the approximation of Brownian Motion at discrete times, we require that

$$\sqrt{\delta t} \mathcal{N}^{n,m,j} = B^{(m,j)}_{(n+1)\delta t} - B^{(m,j)}_{n\delta t} \tag{4.18}$$

for some independent Brownian Motions $(B^{(m,j)})$ for $1 \leq m \leq N$ and $0 \leq j \leq M_S - 1$.

This definition can be rewritten with matrix notations: for column vectors of size $M_S$ such that $(u^n)_j = u_j^n$, we see that

$$u^{n+1} = P^{(n)} u^n, \tag{4.19}$$

where the entries of square matrix satisfy for any $1 \leq j, k \leq M_S$

$$P^{(n)}_{j,k} = \frac{1}{N} \sum_{m=1}^{N} \phi_k(x_j + \sqrt{\delta t} \mathcal{N}^{n,m,j}). \tag{4.20}$$

Moreover we can decompose these matrices into $N$ independent parts: for $1 \leq m \leq N$

$$P^{(n)} = \frac{1}{N} \sum_{m=1}^{N} P^{(n,m)}, \tag{4.21}$$

with the entries $(P^{(n,m)})_{j,k} = \phi_k(x_j + \sqrt{\delta t} \mathcal{N}^{n,m,j})$.

We observe that the matrices $P^{(n,m)}$ are independent; in each one, the rows are independent, however in a row indexed by $j$ two different entries are never independent, since they depend on the same random variable $\mathcal{N}^{n,m,j}$; moreover, the sum of coefficients in a row is 1.

All matrices $P^{(n,m)}$ have the same law; we define a matrix $Q = \mathbb{E}P^{(n,m)} = \mathbb{E}P^{(n)}$, by taking the expectations of each entry: for any $j, k \in S$

$$Q_{j,k} = \mathbb{E}[\phi_k(x_j + \sqrt{\delta t} \mathcal{N}^{n,m,j})]. \tag{4.22}$$

The right-hand side above does not depend on $n, m$ since we take expectation. It only depends on $j$ through the position $x_j$, not through the random variable $\mathcal{N}^{n,m,j}$.

We can then define the important auxiliary sequence $v^n$, which satisfies the following relations:

$$\begin{aligned} v_j^{n+1} &= \frac{1}{N} \sum_{m=1}^{N} \left( \sum_{k \in S} v_k^n \mathbb{E}[\phi_k(x_j + \sqrt{\delta t} \mathcal{N}^{n,m,j})] \right) \\ &= \sum_{k \in S} v_k^n \mathbb{E}[\phi_k(x_j + \sqrt{\delta t} \mathcal{N}^{n,m,j})], \end{aligned} \tag{4.23}$$

with the initial condition $v^0 = u^0$.

Indeed, for any $0 \leq n \leq M_T$ the vector $v^n$ is the expected value - defined component-wise - of the random vector $u^n$.

Thanks to these notations, the scheme can be written with the simple following expression: for any $n \geq 0$

$$u^n = \prod_{i=0}^{n-1} P^{(i)} u^0 = P^{(n-1)} \dots P^{(0)} u^0. \tag{4.24}$$

The expected value satisfies the same kind of property:

$$v^n = Q^n u^0. \tag{4.25}$$

4.2.4. **Basic properties of the matrices.** We only present a few basic properties of the matrices $P^{(n,m)}$, $P^{(n)}$ and $Q$. First, we show that they are stochastic matrices. Second, we control their behavior with respect to the discrete norms and semi-norms. In order to prove the convergence result, we need other more technical properties which are developed during the proof.

First, all the matrices are stochastic matrices. This is an easy consequence of their definition with the interpolation functions.

**Proposition 4.5.** *For any $0 \leq n \leq M_T - 1$ and for any $1 \leq m \leq N$, almost surely $P^{(n,m)}$) is a stochastic matrix: for any indices $j, k \in S$ we have $P^{(n,m)}_{j,k} \geq 0$, and for any $j \in S$ we have $\sum_{k\in S} P^{(n,m)}_{j,k} = 1$.*

*For any $0 \leq n \leq M_T - 1$, $P^{(n)}$ is also a random stochastic matrix.*

*The matrix $Q$ is stochastic and symmetric - and therefore is bistochastic.*

Proof The stochasticity of the random matrices $P^{(n,m)}$ is a simple consequence of their definition (4.20) and of the relations in (4.12). Since $P^{(n)}$ is a convex sum of the $P^{(n,m)}$, the property for those matrices also holds.

Finally, by taking expectation $Q$ is obviously stochastic; symmetry is a consequence of (4.22), and of the property (4.13) of the functions $\phi_k$:

$$\begin{aligned} Q_{j,k} &= \mathbb{E}[\phi_k(x_j + \sqrt{\delta t}\mathcal{N}^{n,m,j})] \\ &= \mathbb{E}[\phi_0(x_j - x_k + \sqrt{\delta t}\mathcal{N}^{n,m,j})] \\ &= \mathbb{E}[\phi_0(x_k - x_j - \sqrt{\delta t}\mathcal{N}^{n,m,j})] \\ &= \mathbb{E}[\phi_0(x_k - x_j + \sqrt{\delta t}\mathcal{N}^{n,m,k})] \\ &= Q_{k,j}, \end{aligned}$$

since $\phi_0$ is an even function, and since the law of $\mathcal{N}^{n,m,j}$ is symmetric and does not depend on $j$. However this symmetry property is not satisfied by the $P$-matrices, because the trajectories of these random variables are different when $j$ changes. □

Thanks to the chain of equalities in the proof above, we see that $Q_{j,k}$ only depends on $k - j$, but we observe that no similar property holds for the matrices $P^{(n,m)}$.

We now focus on the behavior of the matrices with respect to the $l^2$-norm $\|.\|_{l^2}$. The following proposition is a discrete counterpart of the decreasing of the $L^2$-norm of solutions of the heat equation.

**Proposition 4.6.** *For any $0 \leq n \leq M_T - 1$ and for any $1 \leq m \leq N$, and for any $u \in \mathbb{R}^{M_S}$ we have*

$$\mathbb{E}\|P^{(n,m)}u\|^2_{l^2} \leq \|u\|^2_{l^2} \text{ and } \mathbb{E}\|P^{(n)}u\|^2_{l^2} \leq \|u\|^2_{l^2}.$$

Proof According to the definitions above (4.19) and (4.21), we have for any index $j$ $(P^{(n,m)}u)_j = \sum_{k\in S} P^{(n,m)}_{j,k} u_k$. Thanks to the previous Proposition 4.5, we can use the Jensen inequality to get

$$\begin{aligned} \mathbb{E}\|P^{(n,m)}u\|^2_{l^2} &= \delta x \sum_{j\in S} \mathbb{E}|(P^{(n,m)}u)_j|^2 \\ &\leq \delta x \sum_{j\in S}\sum_{k\in S} \mathbb{E}P^{(n,m)}_{j,k}|u_k|^2 \\ &\leq \delta x \sum_{k\in S}\left(\sum_{j\in S} Q_{j,k}\right)|u_k|^2; \end{aligned}$$

now we use the properties of the matrix $Q$ - it is a bistochastic matrix according to Proposition 4.5 - to conclude the proof, since $\sum_{j\in S} Q_{j,k} = 1$.

The extension to matrices $P^{(n)}$ is straightforward. □

The matrix $Q$ satisfies the same decreasing property in the $l^2$-norm; moreover we can easily obtain a bound relative to the $h^1$-semi norm:

**Proposition 4.7.** *For any $u \in \mathbb{R}^{M_S}$, we have $\|Qu\|_{l^2} \leq \|u\|_{l^2}$ and $|Qu|_{h^1} \leq |u|_{h^1}$.*

Proof The proof of the first inequality is similar to the previous situation for the random matrices. To get the second one, it suffices to define a sequence $\tilde{u}$ such that for any $0 \leq j \leq M_S - 1$ we have $\tilde{u}_j = \frac{u_{j+1} - u_j}{\delta x}$ - with the convention $u_{M_S} = u_0$. Then thanks to the properties of $Q$ we have $\widetilde{Qu} = Q\tilde{u}$: for any $j \in S$

$$\begin{aligned}
\delta x \widetilde{Qu}_j &= (Qu)_{j+1} - (Qu)_j \\
&= \sum_{k \in S} Q_{j+1,k} u_k - \sum_{k \in S} Q_{j,k} u_k \\
&= \sum_{k \in S} Q_{j,k-1} u_k - \sum_{k \in S} Q_{j,k} u_k \\
&= \sum_{k \in S} Q_{j,k} (u_{k+1} - u_k) \\
&= \delta x (Q\tilde{u})_j,
\end{aligned}$$

using a translation of indices with periodic conditions, and the equality $Q_{j+1,k} = Q_{j,k-1}$ as explained above.

As a consequence, we have $|Qu|_{h^1} = \|\widetilde{Qu}\|_{l^2} = \|Q\tilde{u}\|_{l^2} \leq \|\tilde{u}\|_{l^2} = |u|_{h^1}$. □

It is worth noting that the previous argument can not be used to control $\mathbb{E}|P^{(n,m)}u|_{h^1}$: for a matrix $P = P^{(n,m)}$, the corresponding quantity $\widetilde{Pu}$ can not be easily expressed with $\tilde{u}$. Indeed, given a deterministic $u$, then $(P^{(n,m)}u)_j$ and $(P^{(n,m)}u)_{j+1}$ are independent random variables - since they are defined respectively with $\mathcal{N}^{(n,m,j)}$ and $\mathcal{N}^{(n,m,j+1)}$. Indeed, we can only prove Proposition 4.8: since it requires Proposition 4.9, and since (4.26) is not used in the sequel of the proof of Theorem 4.1, we give the proof in Section 4.4.

**Proposition 4.8.** *There exists a constant $C$, such that for any discretization parameters $N \geq 1$, $\delta t = \frac{T}{M_T}$ and $\delta x = \frac{1}{M_S}$, we have for any vector $u \in \mathbb{R}^{M_S}$*

$$\mathbb{E}|P^{(0)}u|_{h^1}^2 \leq (1 + C\frac{\delta t + \delta x^2}{N \delta x^2})|u|_{h^1}^2. \tag{4.26}$$

We can remark that in (4.26) the control depends on the number of Monte-Carlo simulations: as we can guess from the expression of $P^{(0)}u$, the treatment of random variables corresponding to identical or distinct realizations is different. On the one hand, Monte-Carlo independence implies the same behavior as for $Qu$; on the other hand, for a same Monte-Carlo realization independence of positions $j+1$ and $j$ leads to an estimate which does not take into account the smallness of $\delta x$.

Due to independence of matrices involved at different steps of the scheme, the previous inequalities can be used in chain.

We thus observe that the matrices $P^{(k)}$ and $Q$ are quite different, even if $Q = \mathbb{E}P^{(k)}$. On the one hand, the matrix $Q$ is symmetric, and therefore respects the structure of the heat equation - the Laplace operator is also symmetric with respect to the $L^2$-scalar product. On the other hand, the structure of the noise destroys this symmetry for matrices $P^{(k)}$, while it introduces many other properties due to independence - in some sense noise is white in space and implies first that solutions are not regular, but that on the average a better estimate can be obtained.

## 4.3. Proof of Theorem 4.1

We begin with a detailed proof of (4.8). A proof of the other part of the error (4.7) is given below in Section 4.3.4.

We recall that the error estimate of (4.8) is given in the $l^2$-norm.

Easy computations give the following expression for the part corresponding to the Monte-Carlo error: for any $0 \leq n \leq M_T$

$$\begin{aligned}
\delta x \sum_{j=0}^{M_S - 1} \mathrm{Var}(u_j^n) &= \delta x \sum_{j=0}^{M_S - 1} \mathbb{E}|u_j^n - v_j^n|^2 \\
&= \mathbb{E}\|u^n - v^n\|_{l^2}^2 \\
&= \delta x \mathbb{E}(u^n - v^n)^*(u^n - v^n),
\end{aligned}$$

where the superscript $*$ denotes transposition of matrices.

Since the vectors $u^n$ and $v^n$ satisfy respectively (4.24) and (4.25), with the same deterministic initial condition $u^0$, we have

$$\begin{aligned}\mathbb{E}\|u^n - v^n\|_{l^2}^2 &= \mathbb{E}\|(P^{(n-1)}\dots P^{(0)} - Q^n)u^0\|_{l^2}^2 \\ &= \delta x (u^0)^* \mathbb{E}\left((P^{(n-1)}\dots P^{(0)} - Q^n)^*(P^{(n-1)}\dots P^{(0)} - Q^n)\right) u^0 \\ &= \delta x (u^0)^* \mathbb{E}\left((P^{(0)})^*\dots(P^{(n-1)})^* P^{(n-1)}\dots P^{(0)} - (Q^n)^* Q^n\right) u^0,\end{aligned}$$

where the last inequality is a consequence of the relation $\mathbb{E}P^{(k)} = Q$ and of the independence of the matrices $P^{(k)}$.

Therefore we need to study the matrix $S_n = \mathbb{E}\left((P^{(0)})^*\dots(P^{(n-1)})^* P^{(n-1)}\dots P^{(0)} - (Q^n)^* Q^n\right)$ given by the expression above, such that

$$\mathbb{E}\|u^n - v^n\|_{l^2}^2 = \delta x (u^0)^* S_n u^0.$$

4.3.1. **Decompositions of the error.** We propose two decompositions of $S_n$ into sums of $n$ terms, involving products of matrices $P^{(k)}$, of $Q$ and of the difference between two matrices $P^{(k)}$ and $Q$, which corresponds to a one-step error:

$$P^{(n-1)}\dots P^{(0)} - Q^n = \sum_{k=0}^{n-1} P^{(n-1)}\dots P^{(k+1)}(P^{(k)} - Q)Q^k, \tag{4.27}$$

and

$$P^{(n-1)}\dots P^{(0)} - Q^n = \sum_{k=0}^{n-1} Q^{n-1-k}(P^{(k)} - Q)P^{(k-1)}\dots P^{(0)}. \tag{4.28}$$

These decompositions lead to the following expressions for $S_n$ - where we use the independence of the matrices $P^{(k)}$ for different values of $k$:

$$\begin{aligned}S_n &= \mathbb{E}\sum_{k=0}^{n-1}(Q^k)^*(P^{(k)} - Q)^*(P^{(k+1)})^*\dots(P^{(n-1)})^* P^{(n-1)}\dots P^{(k+1)}(P^{(k)} - Q)Q^k \\ &= \mathbb{E}\sum_{k=0}^{n-1}(P^{(0)})^*\dots(P^{(k-1)})^*(P^{(k)} - Q)^*(Q^{n-1-k})^* Q^{n-1-k}(P^{(k)} - Q)P^{(k-1)}\dots P^{(0)}\end{aligned}$$

Therefore we obtain the following expressions for the error:

$$\begin{aligned}\mathbb{E}\|u^n - v^n\|_{l^2}^2 &= \delta x (u^0)^* S_n u^0 \\ &= \sum_{k=0}^{n-1}\mathbb{E}\|P^{(n-1)}\dots P^{(k+1)}(P^{(k)} - Q)Q^k u^0\|_{l^2}^2 \\ &= \sum_{k=0}^{n-1}\mathbb{E}\|Q^{n-1-k}(P^{(k)} - Q)P^{(k-1)}\dots P^{(0)}u^0\|_{l^2}^2.\end{aligned} \tag{4.29}$$

Before we show how each decomposition can be used to obtain a convergence result, we focus on the variance induced by one step of the scheme. In fact, only the second one can give the improved estimate of Theorem 4.1. Nevertheless, we can also get a useful error bound thanks to the first one.

4.3.2. **One-step variance.** In the previous Section, we have introduced decomposition of the error, and we observed that we need a bound on the error made after each time-step. The following Proposition states that the variance after one step of the scheme is of size $\delta t$ if we consider the $l^2$ norm, and that a residual term of size $\delta x^2$ appears due to the interpolation procedure. If we consider $N$ independent realizations, Corollary 4.10 below states that the variance is divided by $1/N$ if we look at the full matrix of the scheme.

**Proposition 4.9.** *There exists a constant $C$, such that for any discretization parameters $\delta t = \frac{T}{M_T}$ and $\delta x = \frac{1}{M_S}$, and for any $1 \leq m \leq N$ and $0 \leq n \leq M_T - 1$, we have for any vector $u \in \mathbb{R}^{M_S}$*

$$\mathbb{E}\|(P^{(n,m)} - Q)u\|_{l^2}^2 \leq C(\delta t + \delta x^2)|u|_{h^1}^2. \tag{4.30}$$

**Corollary 4.10.** *For any $0 \leq n \leq M_T - 1$ and for any vector $u \in \mathbb{R}^{M_S}$, we have*

$$\mathbb{E}\|(P^{(n)} - Q)u\|_{l^2}^2 \leq C\frac{(\delta t + \delta x^2)}{N}|u|_{h^1}^2.$$

The proof of the corollary is straightforward, since $P^{(n)} = \frac{1}{N}\sum_{m=1}^{N} P^{(n,m)}$ with independent and identically distributed matrices $P^{(n,m)}$. However, the proof of Proposition 4.9 is very technical.

Heuristically, the error is of order 1/2 with respect to the time step since the left-hand side can be interpreted as the variance of a functional of a diffusion process, evaluated at time $\delta t$. Due to the use of discrete norms and to the interpolation procedure, an error of size $\delta x$ also appears. Finally, the error is controlled with respect to the $h^1$- semi norm of the initial condition $u$, which contains information on the regularity of this vector - and of a function $f$ such that $\mathcal{P}f = u$.

One difficulty of the proof is the dependence of the noise on the position $j$: for different indices $j_1$ and $j_2$, the random variables $(P^{(n,m)}u)_{j_1}$ and $(P^{(n,m)}u)_{j_2}$ are independent. To deal with this problem, we need to introduce various auxiliary functions - basically we need one for each $j$ - and we analyze the error on each interval $[x_j, x_{j+1}]$ separately. We also need to take care of some regularity properties of the functions - they are $H^1$ functions, piecewise linear, but they are not in general of class $\mathcal{C}^1$ - and to obtain bounds involving the $h^1$ and $H^1$ semi-norms.

Proof of Proposition 4.9 To simplify the notations, we assume that $n = 0$ and that $m = 1$; using (4.18), we could write explicitly similar expressions with the true indices, but the final result remains the same. Therefore we only work with one matrix $P$ with entries

$$P_{j,k} = \phi_k(x_j + B^j_{\delta t}),$$

where the $B^j$ are independent Brownian Motions.

We define the following auxiliary periodic functions: for any $x \in \mathbb{R}$

$$V(x) = \mathbb{E}\mathcal{I}u(x + B^j_{\delta t}), \tag{4.31}$$

and for any index $0 \leq j \leq M_S - 1$

$$U^{(j)}(x) = \mathcal{I}u(x + B^j_{\delta t}). \tag{4.32}$$

We can observe that since we take expectation in (4.31) the index $j$ plays no role there. Moreover we have the following relations for any $j \in S$:

$$V(x_j) = (Qu)_j$$
$$U^{(j)}(x_j) = (Pu)_j$$
$$U^{(j)}(x_{j+1}) \neq (Pu)_{j+1}.$$

The last relation is the reason why we need to introduce different auxiliary functions $U^{(j)}$ for each index $j$.

We finally introduce the following function depending on two variables: for any $0 \leq t \leq \delta t$ and $x \in \mathbb{R}$,

$$\mathcal{V}(t, x) = \mathbb{E}\mathcal{I}u(x + B_t), \tag{4.33}$$

for some standard Brownian Motion $B$.

Then $\mathcal{V}$ defined by (4.33) is solution of the backward Kolmogorov equation associated with the Brownian Motion, with the initial condition $\mathcal{V}(0, .) = \mathcal{I}u$, and for $t > 0$

$$\partial_t \mathcal{V} = \frac{1}{2}\partial^2_{xx}\mathcal{V}.$$

Moreover we have $\mathcal{V}(\delta t, .) = V$. Indeed, $\mathcal{V}$ is solution of the heat equation in a periodic setting, which we are trying to discretize with a semi-lagrangian method, but with a different initial condition obtained by interpolation.

We have the following expression for the mean-square error, integrated over an interval $[x_j, x_{j+1}]$: for any index $j \in S$

$$\int_{x_j}^{x_{j+1}} \mathbb{E}|U^{(j)}(x) - V(x)|^2 dx = \int_0^{\delta t} \int_{x_j}^{x_{j+1}} \mathbb{E}|\partial_x \mathcal{V}(\delta t - s, x + B_s^j)|^2 dx ds. \tag{4.34}$$

The proof of this identity is as follows. First, thanks to smoothing properties of the heat semi-group, for any $t > 0$ the function $\mathcal{V}(t, .)$ is smooth. Using Itô formula, with the Brownian Motion $B^{(j)}$ corresponding to the function $U^{(j)}$,

$$d\mathcal{V}(\delta t - s, x + B_s^j) = \partial_x \mathcal{V}(\delta t - s, x + B_s^j) dB_s^j,$$

for $0 \leq s \leq \delta t - \epsilon$ and for any $\epsilon \in (0, \delta t)$, and the isometry property implies

$$\mathbb{E}|\mathcal{V}(\delta t, x) - \mathcal{V}(\epsilon, x + B_{\delta t - \epsilon}^j)|^2 = \int_0^{\delta t - \epsilon} |\partial_x \mathcal{V}(\delta t - s, x + B_s^j)|^2 ds.$$

We integrate over $x \in [x_j, x_{j+1}]$, and we can then pass to the limit $\epsilon \to 0$, since $\mathcal{V}(0, .) = \mathcal{I}u$ is a piecewise linear function. Moreover, we use the identity $\mathcal{V}(\delta t, .) = V$. We observe that in the right-hand side of the last equality we take expectation, so that we can replace $B^j$ with the Brownian Motion $B$, which does not depend on $j$.

Summing over indices $j \in S$, we then get

$$\begin{aligned}
\sum_{j \in S} \int_{x_j}^{x_{j+1}} \mathbb{E}|U^{(j)}(x) - V(x)|^2 dx &= \int_0^{\delta t} \int_0^1 \mathbb{E}|\partial_x \mathcal{V}(\delta t - s, x + B_s)|^2 dx ds \\
&= \int_0^{\delta t} \int_0^1 |\partial_x \mathcal{V}(\delta t - s, x)|^2 dx ds \\
&= \int_0^{\delta t} |\mathcal{V}(\delta t - s, .)|_{H^1}^2 ds \\
&\leq \int_0^{\delta t} |\mathcal{V}(0, .)|_{H^1}^2 ds = \delta t |\mathcal{I}u|_{H^1}^2 = \delta t |u|_{h^1}^2.
\end{aligned}$$

We used periodicity of the functions - with an affine change of variables $y = x + B_s$ in the space integral - and the decreasing of the $H^1$ semi-norm for solutions of the heat equation with periodic boundary conditions on $(0, 1)$.

Now we claim that for some constant $C$ - which does not depend on the parameters or on $u$ - we have

$$\left| \mathbb{E}\|Pu - Qu\|_{l^2}^2 - \sum_{j \in S} \int_{x_j}^{x_{j+1}} \mathbb{E}|U^{(j)}(x) - V(x)|^2 dx \right| \leq C \delta x^2 |u|_{h^1}^2. \tag{4.35}$$

The proof of (4.35) is done in two steps: first we show that

$$\sum_{j \in S} \int_{x_j}^{x_{j+1}} \mathbb{E}|U^{(j)}(x) - V(x) - \mathcal{I} \circ \mathcal{P}(U^{(j)} - V)(x)|^2 dx \leq C \delta x^2 |u|_{h^1}^2, \tag{4.36}$$

and second we show that

$$|\mathbb{E}\|Pu - Qu\|_{l^2}^2 - \sum_{j \in S} \int_{x_j}^{x_{j+1}} \mathbb{E}|\mathcal{I} \circ \mathcal{P}(U^{(j)} - V)(x)|^2 dx| \leq C \delta x^2 |u|_{h^1}^2. \tag{4.37}$$

For each realization of the random variable $B_{\delta t}^j$, the function $U^{(j)}$ belongs to $H^1(0, 1)$. We can use the inequality (4.16) from the proof of Proposition 4.3 on each interval $[x_j, x_{j+1}]$ and we have

$$\begin{aligned}
\int_{x_j}^{x_{j+1}} |U^{(j)}(x) - V(x) - \mathcal{I} \circ \mathcal{P}(U^{(j)} - V)(x)|^2 dx &\leq C \delta x^2 \int_{x_j}^{x_{j+1}} |\partial_x (U^{(j)} - V)(x)|^2 dx \\
&\quad + C \delta x \delta x^2 \frac{|[U^{(j)}(x_{j+1}) - V(x_{j+1})] - [U^{(j)}(x_j) - V(x_j)]|^2}{\delta x^2}.
\end{aligned}$$

We treat separately each term in the right-hand side above. Taking the sum over indices $j \in S$ and expectation, we see that

$$\begin{aligned}\sum_{j\in S}\int_{x_j}^{x_{j+1}} \mathbb{E}|(U^{(j)}-V)'(x)|^2dx &\le 2\sum_{j\in S}\int_{x_j}^{x_{j+1}} \mathbb{E}|\partial_x(\mathcal{I}u)(x+B^j_{\delta t})|^2dx\\ &+\sum_{j\in S}\int_{x_j}^{x_{j+1}} |\partial_x V(x)|^2dx\\ &\le 2(|\mathcal{I}u|^2_{H^1}+|\mathcal{V}(\delta t,.)|^2_{H^1})\\ &\le 4|\mathcal{I}u|^2_{H^1}=4|u|^2_{h^1},\end{aligned}$$

since $V=\mathcal{V}(\delta t,.)$, and using the fact that taking expectation implies that we can take a Brownian motion $B$ which does not depend on $j$, and then by an obvious change of variables we obtain the result.

We treat the other part in the same way:

$$\begin{aligned}\frac{|[U^{(j)}(x_{j+1})-V(x_{j+1})]-[U^{(j)}(x_j)-V(x_j)]|^2}{\delta x^2} &\le 2\frac{|U^{(j)}(x_{j+1})-U^{(j)}(x_j)|^2}{\delta x^2}\\ &+2\frac{|V(x_{j+1})-V(x_j)|^2}{\delta x^2}.\end{aligned}$$

With the second part, using Proposition 4.7 we see that

$$\begin{aligned}\delta x\sum_{j\in S}\frac{|V(x_{j+1})-V(x_j)|^2}{\delta x^2} &= \delta x\sum_{j\in S}\frac{|(Qu)_{j+1}-(Qu)_j|^2}{\delta x^2}\\ &=|Qu|^2_{h^1}\\ &\le |u|^2_{h^1}.\end{aligned}$$

To treat the first part, we make the fundamental observation that for a fixed $j\in S$, the same noise process $B^j$ is used to compute all values $U^{(j)}(x)$ when $x$ varies. As a consequence, we can use a pathwise, almost sure version of the argument leading to the proof of Proposition 4.7 which concerns the behavior of $Q$ with respect to the $h^1$ semi norm. We recall that the obstruction for a similar result on matrices $P$ is spatial roughness of the noise process, due to independence of the Gaussian random variables - see Proposition 4.8.

$$\begin{aligned}U^{(j)}(x_{j+1})-U^{(j)}(x_j) &= \sum_{k\in S}u_k[\phi_k(x_{j+1}+B^j_{\delta t})-\phi_k(x_j+B^j_{\delta t})]\\ &=\sum_{k\in S}u_k[\phi_{k-1}(x_j+B^j_{\delta t})-\phi_k(x_j+B^j_{\delta t})]\\ &=\sum_{k\in S}[u_{k+1}-u_k]\phi_k(x_j+B^j_{\delta t}),\end{aligned}$$

using (4.13) and an integration by parts.

Now summing over indices $j\in S$ and using the Jensen inequality - thanks to Proposition 4.5 - we obtain

$$\begin{aligned}\delta x\sum_{j\in S}\mathbb{E}\frac{|U^{(j)}(x_{j+1})-U^{(j)}(x_j)|^2}{\delta x^2} &\le \delta x\sum_{k\in S,j\in S}\mathbb{E}\phi_k(x_j+B^j_{\delta t})\frac{|u_{k+1}-u_k|^2}{\delta x^2}\\ &\le \delta x\sum_{k\in S}\frac{|u_{k+1}-u_k|^2}{\delta x^2}=|u|^2_{h^1}.\end{aligned}$$

Having proved (4.36), we now focus on (4.37). We have, since $\mathcal{P}(U^{(j)}-V)_j=[(P-Q)u]_j$

$$\begin{aligned}&|\sum_{j\in S}\int_{x_j}^{x_{j+1}}|\mathcal{I}\circ\mathcal{P}(U^{(j)}-V)(x)|^2dx-\delta x\sum_j|[(P-Q)u]_j|^2|\\ &\qquad\le C\delta x^2\delta x\sum_{j\in S}\frac{|(U^{(j)}-V)(x_{j+1})-(U^{(j)}-V)(x_j)|^2}{\delta x^2}.\end{aligned}$$

It remains to take expectation and to conclude like for (4.36).

□

4.3.3. **Proof of Theorem 4.1.** As we have explained in the introduction, we can consider that $\delta x$ is controlled by $\delta t$ thanks to a anti-CFL condition. Roughly, from Proposition 4.9 we thus see that the variance obtained after one step of the scheme is of size $\delta t$, and that the error depends on the solution through the $h^1$ semi-norm. Moreover, from Propositions 4.7 and 4.8 we can remark that the behaviors of the matrices $Q$ and $P^{(n)}$ with respect to this semi-norm are quite different.

Using the first decomposition of the error in (4.29), we can use in chain the bounds given above in Propositions 4.6, 4.7 and 4.9 and Corollary 4.10:

$$\begin{aligned}
\mathbb{E}\|u^n - v^n\|_{l^2}^2 &= \sum_{k=0}^{n-1} \mathbb{E}\|P^{(n-1)} \dots P^{(k+1)}(P^{(k)} - Q)Q^k u^0\|_{l^2}^2 \\
&\leq \sum_{k=0}^{n-1} \mathbb{E}\|(P^{(k)} - Q)Q^k u^0\|_{l^2}^2 \\
&\leq \sum_{k=0}^{n-1} C\frac{(\delta t + \delta x^2)}{N}|Q^k u^0|_{h^1}^2 \\
&\leq \sum_{k=0}^{n-1} C\frac{(\delta t + \delta x^2)}{N}|u^0|_{h^1}^2 \\
&\leq C\frac{1 + \delta x^2/\delta t}{N}|u^0|_{h^1}^2.
\end{aligned}$$

If the continuous problem is initialized with the function $u_0$, which is periodic and of class $\mathcal{C}^1$, then $u^0 = \mathcal{P}u_0$ satisfies $|u^0|_{h^1} \leq \sup_{x\in[0,1]} |u_0'(x)|$. Moreover we can assume that an anti-CFL condition is satisfied, so that the term $\delta x^2/\delta t$ is bounded. As a consequence, we find a classical Monte-Carlo estimate, where the error does not decrease when $\delta t$ goes to 0 and is only controlled with the number of realizations:

$$\mathbb{E}\|u^n - v^n\|_{l^2}^2 \leq C\frac{1 + \delta x^2/\delta t}{N}|u^0|_{h^1}^2. \tag{4.38}$$

In fact, (4.38) shows that the variances obtained at each time step can be summed to obtain some control of the variance at the final time. To get an improved bound, we thus need other arguments.

The main observation is that using independence of rows in the $P$-matrices, we only need to focus on diagonal terms

$$\sup_{j\in S}\left((Q^\ell)^* Q^\ell\right)_{jj} = \sup_{j\in S}\left(Q^{2\ell}\right)_{jj},$$

for indices $0 \leq \ell = n-k-1 \leq n-1$. We recall that indeed $Q$ is a symmetric matrix, so that $(Q^\ell)^* Q^\ell = Q^{2\ell}$.

More precisely, the error can be written

$$\begin{aligned}
\mathbb{E}\|u^n - v^n\|_{l^2}^2 &= \delta x (u^0)^* S_n u^0 = \delta x \sum_{i,j\in S} u_i^0 (S_n)_{i,j} u_j^0 \\
&= \delta x \sum_{k=0}^{n-1} \sum_{i,j\in S} u_i^0 \mathbb{E}\big((A_k)^*(P^{(k)} - Q)Q^{2(n-1-k)}(P^{(k)} - Q)A_k\big)_{i,j} u_j^0.
\end{aligned}$$

where for simplicity we use the notation $A_k := P^{(k-1)} \dots P^{(0)}$. We can compute for any $i, j \in S$, using the independence properties at different steps

$$\begin{aligned}
&\mathbb{E}((A_k)^*(P^{(k)} - Q)Q^{2(n-1-k)}(P^{(k)} - Q)A_k)_{i,j} \\
&\quad = \sum_{k_1,k_2,k_3,k_4\in S} \mathbb{E}[(A_k)_{k_1,i}(P^{(k)} - Q)_{k_2,k_1}(Q^{2(n-1-k)})_{k_2,k_3}(P^{(k)} - Q)_{k_3,k_4}(A_k)_{k_4,j}] \\
&\quad = \sum_{k_1,k_2,k_3,k_4\in S} \mathbb{E}[(A_k)_{k_1,i}(A_k)_{k_4,j}]\mathbb{E}[(P^{(k)} - Q)_{k_2,k_1}(P^{(k)} - Q)_{k_3,k_4}](Q^{2(n-1-k)})_{k_2,k_3}.
\end{aligned}$$

The observation is now that if $k_2 \neq k_3$, then the independence of the random variables for different nodes implies that

$$\mathbb{E}[(P^{(k)} - Q)_{k_2,k_1}(P^{(k)} - Q)_{k_3,k_4}] = 0,$$

since it is the covariance of two independent random variables - see (4.20). Moreover, when $k_2 = k_3$ we see that $((Q^{(n-1-k)})^* Q^{(n-1-k)})_{k_2,k_3}$ only depends on $n-k-1$, due to invariance properties of the equation. Therefore we can rewrite the former expansion in the following way:

$$\begin{aligned}
\mathbb{E}\|u^n - v^n\|_{l^2}^2 &= \delta x \sum_{k=0}^{n-1} \sum_{i,j\in S} u_i^0 \mathbb{E}\big((A_k)^*(P^{(k)} - Q)Q^{2(n-1-k)}(P^{(k)} - Q)A_k\big)_{i,j} u_j^0 \\
&= \delta x \sum_{k=0}^{n-1} \sum_{i,j\in S} u_i^0 u_j^0 (Q^{2(n-1-k)})_{1,1} \sum_{k_2\in S} \mathbb{E}[\Big((P^{(k)} - Q)A_k\Big)_{k_2,i} \Big((P^{(k)} - Q)A_k\Big)_{k_2,j}] \qquad (4.39) \\
&= \sum_{k=0}^{n-1} \Big(Q^{2(n-1-k)}\Big)_{1,1} \mathbb{E}\|(P^{(k)} - Q)P^{(k-1)} \dots P^{(0)} u^0\|_{l^2}^2.
\end{aligned}$$

We thus have to control $(Q^{2\ell})_{1,1} = (Q^{2\ell})_{j,j}$ for any $j \in S$. The following Lemma 4.11 gives a control of this expression. The first estimate means that the coefficients $Q^{2\ell}_{j_1,j_2}$ are approximations of the solution of the PDE at time $2\ell\delta t$, at position $j_2$, starting from the initial condition $\phi_{j_1}$, with an error due to interpolation. The second estimate is fundamental in the proof of the Theorem, since it allows to introduce an additional factor $\delta x$; however, we need to treat carefully the denominator.

**Lemma 4.11.** *There exists a constant $C$ such that for any discretization parameters $\delta t = \frac{T}{M_T}$ and $\delta x = \frac{1}{M_S}$, we have for any $1 \leq \ell \leq M_T - 1$ and for any $0 \leq j_1, j_2 \leq M_S - 1$*

$$|Q^{2\ell}_{j_1,j_2} - \mathbb{E}\phi_{j_1}(x_{j_2} + B_{2\ell\delta t})| \leq C\frac{\delta x^2}{\delta t}(1 + |\log(\delta t)|). \qquad (4.40)$$

*Moreover, for any $j \in S$, we have for any $1 \leq \ell \leq M_T$*

$$\mathbb{E}\phi_j(x_j + B_{2\ell\delta t}) \leq C\frac{\delta x}{\sqrt{2\ell\delta t}}. \qquad (4.41)$$

**Remark 4.12.** *The estimate (4.41) only holds for $\ell \geq 1$, since for $\ell = 0$ we have $\phi_j(x_j) = 1$. This case is exceptional and we can consider that the common behavior is given by the Lemma.*

*In the first estimate (4.40), the singularity when $\delta t \to 0$ with a fixed $\delta x$ is linked to the necessity of proving a uniform bound to control the interpolation error, and to the use of regularization properties of the heat equation: the test functions used here as initial conditions $\phi_j$ are bounded but non smooth, while the solution for positive times becomes smooth.*

*For the second estimate (4.41), two important remarks can be made. First, the constant $C$ depends on the final time $T$, and we cannot directly let $\ell$ tend to $+\infty$: we have*

$$\lim_{\ell\to+\infty} \mathbb{E}\phi_j(x_j + B_{2\ell\delta t}) = \int_{x_j-1/2}^{x_j+1/2} \phi_j(x)dx = \delta x \neq 0.$$

*Second, from (4.41) we get for any $\ell > 0$ and for any fixed $\delta t$*

$$\lim_{\delta x\to 0} \mathbb{E}\phi_j(x_j + B_{2\ell\delta t}) = 0,$$

*while we know that for a fixed $\delta x > 0$ and a fixed $\ell$, we have*

$$\lim_{\delta t\to 0} \mathbb{E}\phi_j(x_j + B_{2\ell\delta t}) = \phi_j(x_j) = 1.$$

*These two behaviors are different and from (4.41) we see the kind of relations that the parameters $\delta x$ and $\delta t$ must satisfy for obtaining one convergence or the other.* □

Proof of Lemma 4.11 For any $0 \leq \ell \leq 2M_T$, we define

$$M_\ell = \sup_{i,j\in S} |(Q^\ell)_{i,j} - \mathbb{E}\phi_j(x_i + B_{\ell\delta t})|,$$

where $(B_t)_{t\geq 0}$ is a standard Brownian Motion.

We have $M_0 = 0$, and by definition of $Q$ we also have $M_1 = 0$.

We define some auxiliary functions $W_j$, for any index $j \in S$: for any $x \in \mathbb{R}$ and any $t \geq 0$

$$W_j(t,x) = \mathbb{E}\phi_j(x + B_t).$$

$W_j$ is solution of the heat equation, with periodic boundary conditions and initial condition $\phi_j$. For any $t > 0$, $W_j(t,.)$ is therefore a smooth function - thanks to regularization properties of the heat semi-group - and since $\phi_j$ is bounded by 1 we easily see that we have the following estimates, for some constant $C$:

$$\begin{aligned} \|\partial_x W_j(t,.)\|_\infty &\leq \frac{C}{\sqrt{t}} \\ \|\partial^2_{xx} W_j(t,.)\|_\infty &\leq \frac{C}{t}. \end{aligned} \tag{4.42}$$

We now prove the following estimate on the sequence $(M_\ell)$: for any $1 \leq \ell \leq M_T - 1$

$$M_{\ell+1} \leq M_\ell + C\frac{\delta x^2}{\ell \delta t}. \tag{4.43}$$

The error comes from the interpolation procedure which is made at each time step.

For any $i, j \in S$, Markov property implies that

$$\begin{aligned} (Q^{\ell+1})_{i,j} - \mathbb{E}\phi_j(x_i + B_{(\ell+1)\delta t}) &= \sum_{k \in S} Q_{i,k}(Q^\ell)_{k,j} - \mathbb{E}W_j(\ell\delta t, x_i + B_{\delta t}) \\ &= \sum_{k \in S} Q_{i,k}(Q^\ell)_{k,j} - \mathbb{E}\mathcal{I} \circ \mathcal{P}(W_j(\ell\delta t,.))(x_i + B_{\delta t}) \\ &+ \mathbb{E}[\mathcal{I} \circ \mathcal{P}(W_j(\ell\delta t,.)) - W_j(\ell\delta t,.)](x_i + B_{\delta t}). \end{aligned}$$

For the first term, we remark that it can be bounded by $M_\ell$; indeed we see that

$$\begin{aligned} \mathbb{E}\mathcal{I} \circ \mathcal{P}(W_j(\ell\delta t,.))(x_i + B_{\delta t}) &= \sum_{k \in S} W_j(\ell\delta t, x_k)\mathbb{E}\phi_k(x_i + B_{\delta t}) \\ &= \sum_{k \in S} Q_{i,k}\mathbb{E}\phi_j(x_k + B_{\ell\delta t}). \end{aligned}$$

To conclude, it remains to use the stochasticity of the matrix $Q$: entries are positive, and their sum over each line is equal to 1.

The second term is bounded using the following argument:

$$\|\mathcal{I} \circ \mathcal{P}(W_j(\ell\delta t,.)) - W_j(\ell\delta t,.)]\|_\infty \leq C\delta x^2 \|\partial^2_{xx} W_j(\ell\delta t,.)\|_\infty \leq C\frac{\delta x^2}{\ell\delta t},$$

according to well-known interpolation estimates and to (4.42).

From (4.43), using $M_1 = 0$ we obtain for any $1 \leq \ell \leq M_T$

$$\begin{aligned} M_\ell &\leq C\frac{\delta x^2}{\delta t}\sum_{k=1}^{M_T - 1}\frac{1}{k} \\ &\leq C\frac{\delta x^2}{\delta t}(|\log(T)| + |\log(\delta t)|). \end{aligned}$$

which gives the result, with a constant depending on $T$.

Now we prove the second estimate of the Lemma. Thanks to (4.13) we see that the left-hand side does not depend on $j \in S$; moreover we can expand the calculation of the expectation using the periodicity of the function $\phi_j$ and relation (4.9), the description of its support as $x_j + \bigcup_{k \in \mathbb{Z}}[k - \delta x, k + \delta x]$: we get for

$1 \leq \ell \leq M_T$

$$\begin{aligned}
\mathbb{E}\phi_j(x_j + B_{2\ell\delta t}) &= \sum_{k\in\mathbb{Z}} \frac{1}{\sqrt{2\pi\ell\delta t}} \int_{k-\delta x}^{k+\delta x} \hat{\phi}(\frac{z-k}{\delta x}) e^{-|z|^2/(2\ell\delta t)} dz \\
&\leq \frac{1}{\sqrt{2\pi\ell\delta t}} \sum_{k\in\mathbb{Z}} \int_{k-\delta x}^{k+\delta x} e^{-|z|^2/(2\ell\delta t)} dz \\
&\leq \frac{1}{\sqrt{2\pi\ell\delta t}} \sum_{k\in\mathbb{Z}} \int_{k-\delta x}^{k+\delta x} e^{-|z|^2/(2T)} dz \\
&\leq \frac{1}{\sqrt{2\pi\ell\delta t}} \sum_{k\in\mathbb{Z}} C\delta x e^{-k^2/(2T)} \\
&\leq C\frac{\delta x}{\sqrt{2\ell\delta t}}.
\end{aligned}$$

□

The estimate of Lemma 4.11 can be used in (4.39), and we obtain:

$$\begin{aligned}
\sum_{k=0}^{n-1} \left(Q^{2(n-1-k)}\right)_{1,1} &\leq C \sum_{k=0}^{n-2} \left( \frac{\delta x^2}{\delta t}(1 + |\log(\delta t)|) + \frac{\delta x}{\sqrt{(n-1-k)\delta t}} \right) + 1 \\
&\leq C \left( \frac{\delta x^2}{\delta t^2}(1 + |\log(\delta t)|) + \frac{\delta x}{\delta t} + 1 \right) =: C\mathcal{A}.
\end{aligned} \tag{4.44}$$

To conclude one more argument is necessary: we need to apply Proposition 4.9 in order to sum the variances. However this involves the quantity $\mathbb{E}|P^{(k-1)} \dots P^{(0)} u^0|_{h^1}^2$, which is badly controlled according to Proposition 4.8: for example, when $\delta t = \delta x$ the accumulation only implies that

$$\begin{aligned}
\mathbb{E}|P^{(k-1)} \dots P^{(0)} u|_{h^1}^2 &\leq (1 + C\frac{\delta t}{N\delta x^2})^k |u|_{h^1}^2 \\
&\leq e^{\frac{CT}{N\delta x^2}} |u|_{h^1}^2,
\end{aligned}$$

for any $k \leq \frac{T}{\delta t}$. We recall that this bad behavior of the matrices $P^{(n)}$ with respect to the $h^1$-semi norm is a consequence of the independence of the random variables for different nodes, whereas this independence property is essential to get the improved estimate, since it allows to use the second estimate of Lemma 4.11.

The solution we propose relies on the following idea: if above we can replace $P^{(k-1)} \dots P^{(0)}$ with $Q^k$, we can easily conclude. Another error term appears, which is controlled by $1/N$ instead of $1/\sqrt{N}$. More precisely, independence properties yield for $k \geq 1$

$$\begin{aligned}
\mathbb{E}\|(P^{(k)} - Q)P^{(k-1)} \dots P^{(0)} u^0\|_{l^2}^2 &= \mathbb{E}\|(P^{(k)} - Q)Q^k u^0\|_{l^2}^2 \\
&\quad + \mathbb{E}\|(P^{(k)} - Q)\left(P^{(k-1)} \dots P^{(0)} - Q^k\right) u^0\|_{l^2}^2.
\end{aligned} \tag{4.45}$$

The roles of the different terms are as follows. On the one hand, the first term gives the part of size $\frac{\delta t}{N}$, thanks to Lemma 4.11: according to Corollary 4.10 and to Proposition 4.7, we have for any $k \geq 1$ with $k\delta t \leq T$

$$\begin{aligned}
\mathbb{E}\|(P^{(k)} - Q)Q^k u^0\|_{l^2}^2 &\leq C\frac{\delta t + \delta x^2}{N} |Q^k u^0|_{h^1}^2 \\
&\leq C\frac{\delta t + \delta x^2}{N} |u^0|_{h^1}^2.
\end{aligned} \tag{4.46}$$

On the other hand, the second term can be used to improve recursively the error estimate, since we have

$$\mathbb{E}\|(P^{(k)} - Q)\left(P^{(k-1)} \dots P^{(0)} - Q^k\right) u^0\|_{l^2}^2 \leq \frac{C}{N} \mathbb{E}\|\left(P^{(k-1)} \dots P^{(0)} - Q^k\right) u^0\|_{l^2}^2. \tag{4.47}$$

The independence of realizations at step $k$ gives the factor $\frac{1}{N}$; we remark that we cannot use the estimation of the one-step variance given by Corollary 4.10: otherwise we would need to control $\mathbb{E}\|\left(P^{(k-1)} \dots P^{(0)} - Q^k\right) u^0\|_{h^1}^2$.

Using also (4.44) and (4.45) into (4.39), we see that we can prove that

$$\sup_{n\in\mathbb{N},n\delta t\leq T}\mathbb{E}\|u^n-v^n\|_{l^2}^2\leq C\delta t\mathcal{A}\frac{1+\frac{\delta x^2}{\delta t}}{N}|u^0|_{h^1}^2 + C\frac{\mathcal{A}}{N}\sup_{n\in\mathbb{N},n\delta t\leq T}\mathbb{E}\|u^n-v^n\|_{l^2}^2. \tag{4.48}$$

The proof of the Theorem now reduces to the study of the following recursive inequalities, for $p\geq 0$

$$E^{(p+1)}\leq C\delta t\mathcal{A}\frac{1+\frac{\delta x^2}{\delta t}}{N}|u^0|_{h^1}^2+\frac{C\mathcal{A}}{N}E^{(p)},$$

with an initialization $E^{(0)}=C\frac{\mathcal{B}}{N}$, according to (4.38), with the notation $\mathcal{B}:=(1+\frac{\delta x^2}{\delta t})|u^0|_{h^1}^2$. We can remark that the control of the matrices $P^{(k)}$ and $Q$ with respect to the $l^2$-norm leads to another possibility for the initialization: $E^{(0)}=2\|u^0\|_{l^2}^2$; we observe that the recursion then yields the same kind of estimate.

We finally easily prove that for any $p\geq 0$ there exists a constant $C_p\geq 1$ such that

$$\sup_{n\in\mathbb{N},n\delta t\leq T}\mathbb{E}\|u^n-v^n\|_{l^2}^2\leq C_p\left(\frac{\mathcal{A}^p\mathcal{B}}{N^{p+1}}+\mathcal{A}\mathcal{B}\frac{\delta t}{N}\right), \tag{4.49}$$

and the proof of Theorem 4.1 is finished.

**Remark 4.13.** *If we consider the equation $\frac{\partial u}{\partial t}=\frac{\nu}{2}\frac{\partial^2 u}{\partial x^2}$ with a viscosity parameter $\nu>0$, the quantities $\mathcal{A}$ and $\mathcal{B}$ appearing in the proof are transformed into*

$$\mathcal{A}_\nu=(1+\frac{\delta x}{\sqrt{\nu}\delta t}+\frac{\delta x^2}{\nu\delta t^2}(1+|\log(\delta t)|))$$
$$\mathcal{B}_\nu=(\nu+\frac{\delta x^2}{\delta t})|u^0|_{h^1}^2.$$

*where the constant $C$ does not depend on $\nu$.*

*The first change in the proof concerns the analysis of the one-step variance: in* (4.30)*, the right-hand side is replaced by $C(\nu\delta t+\delta x^2)$. We observe that the error due to interpolation remains the same.*

*The second change concerns Lemma 4.11, where we use some regularization properties thanks to gaussian noise: when $\nu$ goes to* 0 *the estimates degenerates.*

*As a consequence, we may observe that the first estimate gives a bound valid for a fixed value of $\nu$, while the second one becomes more interesting when $\nu$ is small compared with the discretization parameters.*

4.3.4. **Accumulation of the interpolation error.** The conclusion of Theorem 4.1 concerns the variance in the numerical scheme: under appropriate conditions on the parameters $\delta t$ and $\delta x$, the Monte-Carlo error decreases when $\delta t$ and $\delta x$ tend to 0. However, the comparison of the numerical and the exact solutions involves other sources of error, as explained in the Introduction: the interpolation and the SDE discretization induce error terms which need to be controlled. For instance, we propose a uniform estimate: for any $n\in\mathbb{N}$ such that $n\delta t\leq T$, and for any $j\in\mathbb{N}$ with $0\leq x_j=j\delta x<1$, we have

$$|u(n\delta t,x_j)-v_j^n|\leq C\frac{\delta x^2}{\delta t}\sup_{x\in[0,1]}|u_0^{''}(x)|, \tag{4.50}$$

where $u$ is the exact solution and where $v^n$ is defined by (4.23). In other words, this estimate corresponds to the error in the theoretical situation when we assume that expectations are computed exactly.

Since $\|u(n\delta t,x_.)-v^n\|_{l^2}\leq\sup_j|u(n\delta t,x_j)-v_j^n|$, we easily obtain an estimate in the $l^2$-norm. Therefore, the conditions imposed on $\delta x$ and $\delta t$ by (4.8) are not restrictive, and can be seen as consequences of the semi-lagrangian framework.

The proof of (4.50) in our context is as follows: using the exact representation formula and its discrete counterpart (4.23), we have

$$\begin{aligned} u((n+1)\delta t, x_j) - v_j^{n+1} &= \mathbb{E}u(n\delta t, x_j + B_{\delta t}) - \mathbb{E}\sum_{k\in S} v_k^n \phi_k(x_j + B_{\delta t}) \\ &= \sum_{k\in S}(u(n\delta t, x_k) - v_k^n)\mathbb{E}\phi_k(x_j + B_{\delta t}) \\ &+ \mathbb{E}\left(u(n\delta t, x_j + B_{\delta t}) - \sum_{k\in S} u(n\delta t, x_k)\phi_k(x_j + B_{\delta t})\right), \end{aligned}$$

where $B_{\delta t}$ is a Brownian Motion at time $\delta t$.

It is easy to see that

$$|\sum_{k\in S}(u(n\delta t, x_k) - v_k^n)\mathbb{E}\phi_k(x_j + B_{\delta t})| \leq \sup_{k\in S}|u(n\delta t, x_k) - v_k^n|,$$

and we see that the other term depends on the interpolation error:

$$\begin{aligned} |\mathbb{E}[u(n\delta t, x_j + B_{\delta t}) - \sum_{k\in S} u(n\delta t, x_k)\phi_k(x_j + B_{\delta t})]| &\leq \sup_{x\in[0,1]} |u(n\delta t, x) - \mathcal{I}\circ\mathcal{P}u(n\delta t, .)(x)| \\ &\leq C\delta x^2 \sup_{x\in[0,1]} |\frac{\partial^2 u}{\partial x^2}(n\delta t, x)| \\ &\leq C\delta x^2 \sup_{x\in[0,1]} |\frac{\partial^2 u}{\partial x^2}(0, x)|. \end{aligned}$$

To conclude, we remark that for $n = 0$ we have $u(0, x_j) = v_j^0$.

## 4.4. Appendix: Proof of Proposition 4.8

Proof

For any index $j \in S$, we have by definition of the scheme

$$(P^{(0)}u)_j = \frac{1}{N}\sum_{m=1}^{N}\sum_{k\in S}\phi_k(x_j + \sqrt{\delta t}\mathcal{N}^{m,j})u_k.$$

As a consequence, if we decompose the sum below with respect to $m_1 = m_2$ or $m_1 \neq m_2$, we get

$$\begin{aligned} ((P^{(0)}u)_{j+1} - (P^{(0)}u)_j)^2 &= \frac{1}{N^2}\sum_{1\leq m_1\neq m_2\leq N} A(m_1)A(m_2) \\ &+ \frac{1}{N^2}\sum_{1\leq m\leq N} A(m)^2, \end{aligned}$$

where $A(m) = \sum_{k\in S} u_k\left(\phi_k(x_{j+1} + \sqrt{\delta t}\mathcal{N}^{m,j+1}) - \phi_k(x_j + \sqrt{\delta t}\mathcal{N}^{m,j})\right)$.

Since the random variables $A(m)$ are independent and have the same distribution, the first part of the expression above gives

$$\begin{aligned} \mathbb{E}\left(\frac{1}{N^2}\sum_{1\leq m_1\neq m_2\leq N} A(m_1)A(m_2)\right) &= \frac{1}{N^2}\sum_{1\leq m_1\neq m_2\leq N} \mathbb{E}A(m_1)\mathbb{E}A(m_2) \\ &= (1 - \frac{1}{N})(\mathbb{E}A(1))^2. \end{aligned}$$

We can see that

$$\begin{aligned} \mathbb{E}(A(1)) &= \sum_{k\in S} u_k\left(\mathbb{E}\phi_k(x_{j+1} + \sqrt{\delta t}\mathcal{N}^{m,j+1}) - \mathbb{E}\phi_k(x_j + \sqrt{\delta t}\mathcal{N}^{m,j})\right) \\ &= (Qu)_{j+1} - (Qu)_j. \end{aligned}$$

We can therefore use Proposition 4.7 to obtain an estimate for the part with independent Monte-Carlo realizations.

In the other part of the estimate, we have

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{N^2}\sum_{1\le m\le N} A(m)^2\right) &= \frac{1}{N}\mathbb{E}A(1)^2 \\
&= \frac{1}{N}\mathbb{E}((P^{0,1}u)_{j+1} - (P^{0,1}u)_j)^2 \\
&= \frac{1}{N}\mathrm{Var}\left((P^{0,1}u)_{j+1} - (P^{0,1}u)_j\right) + \frac{1}{N}\left(\mathbb{E}(P^{0,1}u)_{j+1} - \mathbb{E}(P^{0,1}u)_j\right)^2 \\
&= \frac{1}{N}\mathrm{Var}\left((P^{0,1}u)_{j+1}\right) + \frac{1}{N}\mathrm{Var}\left((P^{0,1}u)_j\right) + \frac{1}{N}\left((Qu)_{j+1} - (Qu)_j\right),
\end{aligned}$$

by independence of random variables $\mathcal{N}^{0,1,j}$ and $\mathcal{N}^{0,1,j+1}$.

Using Proposition 4.9 and combining the expressions, we obtain

$$\begin{aligned}
\mathbb{E}\|P^{(0)}u\|_{h^1}^2 &= \delta x \sum_{j\in S} \frac{|(P^{(0)}u)_{j+1} - (P^{(0)}u)_j|^2}{\delta x^2} \\
&= (1 - \frac{1}{N})\|Qu\|_{h^1}^2 \\
&+ \frac{2}{N\delta x^2}\mathbb{E}\|(P^{(0,1)} - Q)u\|_{l^2}^2 \\
&+ \frac{1}{N}\|Qu\|_{h^1}^2 \\
&\le (1 + \frac{C}{N\delta x^2}(\delta t + \delta x^2))\|u\|_{h^1}^2.
\end{aligned}$$

□

# Chapitre 5

# Applications de la méthode hybride semi-lagrangienne


Résumé

On présente certaines extensions de la méthode définie et analysée dans un cas simple dans le Chapitre 4: équations en dimension supérieure, avec différents choix de conditions au bord. On considère également des équations non-linéaires, de type Burgers ou Navier-Stokes.

L'objectif est de montrer le potentiel de la méthode, et de mettre en avant certaines de ses caractéristiques. Une étude détaillée des algorithmes, ainsi que d'éventuelles améliorations, sont susceptibles de travaux futurs.

Les simulations numériques ont été réalisées grâce au logiciel Matlab.

# Chapitre 5. Applications of the hybrid semi-lagrangian method

This Chapter is devoted to extensions and illustrations of the hybrid semi-lagrangian method described in Chapter 4. The objective is not a convergence analysis, nor a comparison with other methods which are efficient and built for specific problems: our aim is to show that the principle of the method is general and simple, and to suggest interesting applications.

The definition of more efficient methods related to specific problems, as well as an analysis of the schemes, should be part of future work.

We first present the general principle of the method in Section 5.1. Then in Sections 5.2, 5.3, 5.4 and 5.5 we describe applications for computing approximate solutions of the Heat equation in dimension 1 and 2, and then of the Burgers and the Navier-Stokes equations in various contexts.

Finally we gather the numerical simulations we have obtained in Section 5.6.

## 5.1. General description of the method

We describe the method of discretization by the Monte-Carlo semi-lagrangian scheme for various kinds of equations. We focus on equations on a spatial domain $D \subset \mathbb{R}^d$ written

$$\frac{\partial u}{\partial t} = \mathcal{L}u,$$

where $\mathcal{L}$ is the generator of a diffusion process and satisfies appropriate regularity and ellipticity conditions. More complicated equations can be approximated: for instance, we can solve equations like

$$\frac{\partial u}{\partial t} = \mathcal{L}u + f(u),$$

and give approximations of stationary solutions of such equations. For instance, $f$ can be a non-linear function of the solution $u$, like in the case of Burgers and Navier-Stokes considered in Sections 5.4 and 5.5, with $f(u) = u.\nabla u$; we can also consider reaction-diffusion equations, for instance the Fisher-KPP equation with $f(u) = u - u^2$.

The simplest example of operator $\mathcal{L}$ is $\mathcal{L}_0 = \frac{1}{2}\Delta$, which is the generator of the standard $d$-dimensional Brownian motion.

We provide numerical simulations for various kinds of equations. We will choose $d = 1$ or 2. The domain $D$ will be $(a, b)^d$, for some real numbers $a < b$. In this situation, boundary conditions need to be given: the case of periodic, Dirichlet and Neumann conditions can be treated. Finally, we present simulations for more complicated equations: simulations for Burgers equations in dimension $d = 1$ and 2 are presented. We also implement a projection method for the approximation of incompressible Navier-Stokes equations, and apply it in particular for simulations of the driven cavity flow with different values of the Reynolds number.

Before giving precise examples, we recall the principle of the method: we first describe an algorithm where we do not take into account the Monte-Carlo approximations. We then describe two general problems linked to the numerical approximation in this kind of methods. First, in general the associated stochastic processes can not be exactly simulated, and we need to use numerical integrators to compute approximated solutions of the SDEs. Second, we focus on the boundary conditions, and we show how algorithms can be derived for each specific type of conditions.

The advantage of this method is that it is explicit and requires no CFL condition. Moreover the implementation of the algorithms is very simple.

5.1.1. **A theoretical scheme.** We present the principle of the semi-lagrangian scheme, without the Monte-Carlo approximation. The two ingredients are an interpolation method and a representation formula to link the value of the solution at different moments.

The value of the solution at time $t$ and at a given point $x$ depends on the evolution of a stochastic process starting at $x$ and evaluated at time $t$. The general idea is that the solution $u$ at different times $t_1 < t_2$ satisfies a relation $u(t_2, x) = \mathbb{E}[u(t_1, \mathcal{X}(t_2; t_1, x))]$, thanks to the so-called characteristics $\mathcal{X}$, defined according to the coefficients of the PDE - more details are given in the next sections.

The discretization of the equation is then as follows: given a mesh $(x_j)_{j\in J}$ and a time step of size $\delta t$, we build an approximation $u^n$ of the solution $u(n\delta t, .)$ at time $n\delta t$ thanks to:

$$u^n = \mathcal{I}(u_j^n)$$

where $\mathcal{I}$ is an interpolation operator, such that $u^n(x_j) = u_j^n$, and

$$u_j^{n+1} = \mathbb{E}[u^n(\mathcal{X}((n+1)\delta t; n\delta t, x_j))].$$

At each time step we thus compute values at the nodes using the representation formula, and obtain a function defined on the whole domain thanks to an interpolation.

This scheme is only theoretical, since in general the computation of $\mathbb{E}[u^n(\mathcal{X}((n+1)\delta t; n\delta t, x_j))]$ is not exact and requires an approximation procedure: the approximation of the expectation relies on a Monte-Carlo method, and in the next Subsection 5.1.2 we present methods for the approximation of the law of the random variables $\mathcal{X}((n+1)\delta t; n\delta t, x_j)$.

5.1.2. **Approximation of the solutions of SDEs.** As explained in the previous section, the numerical method requires the knowledge of expectations of functionals of the solution evaluated at some fixed deterministic times of a stochastic differential equation, with coefficients related to the generator $\mathcal{L}$. More precisely, the generator satisfies the following relation for some regular coefficients $b$ and $\sigma$

$$\mathcal{L} =< b(x), \partial > +\frac{1}{2}\mathrm{Tr}(\sigma(x)\sigma(x)^T\partial^2) = \sum_{i=1}^{d} b_i(x)\partial_i + \frac{1}{2}\sum_{i,j=1}^{d} a_{i,j}(x)\partial_i\partial_j,$$

where $a = \sigma\sigma^T$ is a square matrix of size $d$. We can then associate the following SDE to the equation

$$dX_t^x = b(X_t^x)dt + \sigma(X_t^x)dW_t, \qquad X_0^x = x.$$

The associated probabilistic representation formula states that the solution of the PDE satisfies

$$u(t, x) = \mathbb{E}u_0(X_t^x),$$

when the initial condition $u_0$ is sufficiently regular.

For this representation formula and the other ones given below, we refer to [13].

In the semi-lagrangian method, we use the formula with $t = \delta t$: more precisely we need to approximate $\mathbb{E}v(X_{\delta t}^x)$, where the function $v$ is piecewise linear and obtained by interpolation, and $X_{\delta t}^x$ is the solution of the SDE starting from $x$. Expectations are approximated thanks to the Monte-Carlo idea, using simulated independent realizations of the random variable and calculating the empirical average.

However in general the solution of the SDE at a given time $t$ can not be exactly solved or simulated, and we need to use an approximation based on a numerical integrator, like the Euler scheme. We recall that given a step size $\delta t$, the values of $X^x$ at discrete times $n\delta t$, for $n \in \mathbb{N}$, are approximated by the random variables $X_n$ defined by the following recursion:

$$X_0 = x,$$
$$X_{n+1} = X_n + \delta t b(X_n) + \sigma(X_n)(B_{(n+1)\delta t} - B_{n\delta t}).$$

Under suitable assumptions on the coefficients $b$ and $\sigma$, the numerical scheme is convergent, and order of convergence is known to be $1/2$ in a strong sense - approximation of the trajectories in the $L^2$ sense - and 1 in a weak sense - approximation of the laws at any fixed time, for smooth enough test functions. Here, the weak convergence can be used, since we consider the approximation of the functionals of the stochastic process. The theory on numerical schemes for SDEs is developed for instance in the books [43] and [54].

A classical Monte-Carlo method for equations with diffusive part would use the scheme until time $T = N_{\text{time}}\delta t$, taking into account the corresponding representation formula. In order to obtain approximated values of the solution at the final time for different points in space, the scheme must be used starting successively from various initial conditions $x_j = j\delta x$, for integers $j$. This strategy gives rise to methods for

which the Monte-Carlo error does not depend on the discretization parameters $\delta t$ and $\delta x$, and is of order $1/2$ with respect to the number of independent realizations used to compute the empirical average.

The hybrid semi-lagrangian method is based on an interpolation at each time step of the method. As a consequence, the integrator is only used for one time step $\delta t$. The number of random variables acting in the definition of the scheme is the same for both kinds of method, but the way they are used changes: if we discretize time in $N_{\text{time}}$ intervals, introduce $N_{\text{space}}$ grid points for space, and use $N_{\text{MC}}$ independent realizations, in a naive, classical method for each point we need $N_{\text{time}} \times N_{\text{MC}}$ independent Gaussian random variables, which gives a total number $(N_{\text{time}} \times N_{\text{MC}}) \times N_{\text{space}}$ of simulated independent random variables. For the semi-lagrangian method, each time step requires $N_{\text{space}} \times N_{\text{MC}}$ independent random variables, and at the end we obtain the same number $(N_{\text{space}} \times N_{\text{MC}}) \times N_{\text{time}}$.

The simulation of the Euler method over a number of time steps greater than 1 may be needed to approximate more accurately the behavior of the solution near the boundary of the domain, depending on the type of boundary conditions. We thus introduce a separation of the domain into interior and boundary points, for which a more precise treatment is required as explained in next Section.

Below, we introduce the various types of boundary conditions, and we explain in more details how and why we treat differently the interior and the boundary points.

5.1.3. **Presentation of the boundary conditions.**

First we recall that for the equation $\partial_t u = \mathcal{L}u$ on the whole domain $D = \mathbb{R}^d$, with the initial condition $u_0$ assumed to be smooth and bounded, with bounded derivatives, the unique solution has the expression $u(t,x) = \mathbb{E}u_0(X_t^x)$, where the diffusion process $X$ is solution of the stochastic differential equation defined on $\mathbb{R}^d$:

$$dX_t^x = b(X_t^x)dt + \sigma(X_t^x)dB_t$$
$$X_0^x = x.$$

The coefficients $b$ and $\sigma$ are such that the generator of this diffusion is $\mathcal{L}$.

When we study the same equation in a bounded domain $D \subset \mathbb{R}^d$, for well-posedness some conditions are imposed on the boundary. This changes both the representation formula and the definition of the numerical approximation of the new associated processes.

(i) Periodic boundary conditions

We assume that the domain is a product set in $\mathbb{R}^d$ of one-dimensional intervals, and without loss of generality we consider the case $D = (0,1)^d$. We can then impose the following boundary conditions for the solution: for any $1 \leq i \leq d$, for any $t \geq 0$, for any $(x_1, \ldots, x_d) \in (0,1)^d$, we have

$$u(t, x_1, \ldots, x_{i-1}, 0, x_{i+1}, \ldots, x_d) = u(t, x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_d).$$

We can also think of the problem as looking for a solution defined on $\mathbb{R}^d$, which is 1-periodic. This is possible when we can extend the initial condition $u_0$ and the coefficients $b$ and $\sigma$ to periodic functions defined on $\mathbb{R}^d$, and preserving the necessary smoothness properties - for example for constant functions $b$ and $\sigma$ it is obviously the case. Then the representation formula $u(t,x) = \mathbb{E}u_0(X_t^x)$ is satisfied - the stochastic process is well-defined on $\mathbb{R}^d$ and its values can be reduced modulo 1.

(ii) Dirichlet boundary conditions

The value of the solution on the boundary $\partial D$ - assumed to be of class $\mathcal{C}^2$ - of the bounded domain can also be imposed, given the so-called Dirichlet boundary conditions. Thus the equation is

$$\partial_t u = \mathcal{L}u$$
$$u(t,.) = u_0$$
$$u(t,x) = 0, \text{ for } t > 0 \text{ and } x \in \partial D.$$

The representation formula then involves the family of the first-exit times of the process starting from the different points of the domain: if the process $X$ satisfies the stochastic differential equation written above, we define, for any $x \in \overline{D}$, $\tau^x = \inf\{t > 0; X(t,x) \in D^c\}$.

Then the solution satisfies

$$u(t,x) = \mathbb{E}\left[u_0(X_t^x)\mathbb{1}_{t \leq \tau^x}\right];$$

the stochastic process is killed when it reaches the boundary.

If we replace the previous equation by $\partial_t u = \mathcal{L}u + f$ for some function $f$ depending only in the space variable, we still get a representation formula:

$$u(t,x) = \mathbb{E}[u_0(X_t^x)\mathbb{1}_{t<\tau^x} + \int_0^{t\wedge\tau^x} f(X_s^x)ds].$$

The numerical approximation becomes more complicated, since we also need an accurate approximation of the stopping times. This problem is well-known, and solutions have been proposed in [32] and [49].

The main point of the method focuses on the following situation: it is possible that at the times $n\delta t$ and $(n+1)\delta t$ the simulated values $X_n$ and $X_{n+1}$ both belong to the domain $D$, however an exit of the domain has been possible in continuous time between $n\delta t$ and $(n+1)\delta t$. An additional test is introduced in the scheme to take into account this situation, so that convergence is improved. This test is described more precisely below for the one-dimensional case; it is based on the knowledge on the law of exit of the diffusion process between $n\delta t$ and $(n+1)\delta t$, conditionally to its values at these times - see formula (5.3).

Nevertheless, in practice this technique requires that the time step $\delta t$ is small; such a restriction is not compatible with the anti-CFL conditions which are naturally associated with semi-lagrangian methods. We propose to refine the time intervals $[n\delta t, (n+1)\delta t]$ in order to improve the approximation of the processes together with the associated stopping times. The interpolation procedure is not concerned with this refinement, and only appears at times $n\delta t$.

Finally, we introduce a decomposition of the domain into an "interior" zone and a "boundary" zone, with different treatments. In the boundary zone, we refine in time and use a subdivision of $[n\delta t, (n+1)\delta t]$ of mesh size $\tau \leq \delta t$; moreover we can also use a value $N_b$ for the number of Monte-Carlo realizations. In the interior part, less care is necessary and we can take $\tau = \delta t$; we also introduce a value $N_i < N_b$ for the size of the sample. Moreover, the test is only performed for the boundary points.

The number of boundary points, the refined time step $\tau$, and the sample sizes $N_b$ and $N_i$ are specified for each simulation.

(iii) Neumann boundary conditions We can finally impose the value of some directional derivative of the solution for the points on the boundary $\partial D$ - assumed to be of class $\mathcal{C}^3$: we then get the following evolution equation:

$$\begin{aligned} \partial_t u &= \mathcal{L}u \\ u(t,.) &= u_0 \\ \gamma(x).\nabla u(t,x) &= 0, \text{ for } t \geq 0 \text{ and } x \in \partial D, \end{aligned}$$

where for simplicity $\gamma$ is a vector field related to the operator $\mathcal{L}$. More precisely, if $\mathcal{L}\phi = b.D\phi + \frac{1}{2}\text{Trace}(aD^2\phi)$, then for any $x \in \partial D$ $\gamma(x) = \frac{1}{2}a(x)n(x) \in \mathbb{R}^d$, if we denote by $n(x)$ the outward normal vector to $\partial D$ at the point $x$.

The construction of the stochastic process used in the probabilistic representation formula is more difficult: while for Dirichlet boundary conditions we needed to kill the diffusion process when it reaches the boundary, we now require a reflexion of the process at the boundary, in the direction given by $\gamma$. The stochastic differential equation is then replaced by a stochastic equation with two unknown processes $X$ and $\eta$, such that $\eta$ is an increasing process, with the following properties:

$$\begin{aligned} X_0 = x, dX_t = b(X_t)dt + \sigma(X_t)dB_t + 1_{\partial D}(X_t)\gamma(X_t)d\eta_t, \\ \int_0^t \mathbb{1}_D(X_s)d\eta_s = 0, \forall t \geq 0. \end{aligned}$$

The last equality means that the process $\eta$ only increases when the process touches the boundary.

The associated representation formula is then $u(t,x) = \mathbb{E}[u_0(X_t); X_0 = x]$.

We thus need to define a discrete time process $(X_n)_n$ which approximates $X$: we can use the Euler method until we reach the boundary, where we decide to reflect the process. More precisely, if $X_n \in D$ and $X_{n+1} \notin D$ - obtained by one step of the Euler scheme - we compute a new value of $X_{n+1}$ taking the symmetric element in $D$ with respect to the boundary, in direction given by $\gamma$. To obtain a more accurate approximation, like in the previous case of Dirichlet boundary conditions we decompose the domain, and we refine in time near

the boundary. Moreover, we mention that in [32] efficient schemes for reflected diffusions are proposed, but we have not used them.

## 5.2. The Heat Equation in dimension 1

We consider the heat equation in dimension 1 with various types of boundary conditions.

5.2.1. **With periodic boundary conditions.** We study the heat equation in the interval $[0,1]$ with periodic boundary conditions:

$$\begin{aligned} \partial_t u &= \nu \partial^2_{xx} u, \\ u(t,1) &= u(t,0) \text{ for any } t \geq 0, \\ u(0,x) &= u_0(x) \text{ for any } x \in [0,1]. \end{aligned} \tag{5.1}$$

When the initial condition is a trigonometric polynomial function, the exact solution at any time $t \geq 0$ is known: if for some nonnegative integer $N$ and some real numbers $a_0, a_1, \dots a_N, b_1, \dots, b_N$

$$u_0(x) = a_0 + \sum_{n=1}^{N} a_n \cos(2\pi n x) + \sum_{n=1}^{N} b_n \sin(2\pi n x),$$

then the solution is given at time $t$ by

$$u(t,x) = a_0 + \sum_{n=1}^{N} a_n \exp(-4\pi^2 n^2 \nu t) \cos(2\pi n x) + \sum_{n=1}^{N} b_n \exp(-4\pi^2 n^2 \nu t) \sin(2\pi n x).$$

The theory of Fourier series allows to extend such a formula for more general initial conditions; we do not develop this well-known idea here.

The stochastic process used to define the representation formula for the solution is the Brownian Motion, with a change in time due to the parameter $\nu$:

$$u(t,x) = \mathbb{E} u_0(x + B_{2\nu t}).$$

To interpret this formula with respect to periodicity, we have two points of view: either we consider that $u_0$ is defined on $\mathbb{R}$, and we obtain for any $t > 0$ a function $u(t,.)$ also defined on $\mathbb{R}$ and periodic, or either we need to reduce the value of the process $x + B_{2\nu t}$ modulo 1. The simulation of the process is easily done with the Euler scheme - approximation is exact in the weak sense: we have the equality in law $B_{2\nu\delta t} = \sqrt{2\nu\delta t}\mathcal{N}(0,1)$.

To compute an approximate solution at time $T$, the algorithm is the following. We divide the space interval $[0,1]$ into $N_S = 1/\delta x$ intervals of size $\delta x$, with the nodes $x_j = j\delta x$ for $0 \leq j \leq N_S$. From the initial condition we build the initial discrete space approximation: $u^0_j = u_0(x_j)$, for $1 \leq j \leq N_S$ - with $u^0_{N_S} = u^0_0$ by periodicity.

Now if for some integer $n \leq N_T - 1$, with $N_T \delta t = T$, we know the vector $u^n = (u^n_j)_{0 \leq j \leq N_S}$ with the periodicity condition $u^n_{N_S} = u^n_0$, we obtain the vector $u^{n+1} = (u^{n+1}_j)_{0 \leq j \leq N_S}$ as follows: for $1 \leq j \leq N_S$

$$u^{n+1}_j = \frac{1}{N} \sum_{m=1}^{N} (\mathcal{I} u^n)(x_j + \sqrt{2\nu\delta t}\mathcal{N}^{n,m,j}),$$

with independent standard Gaussian random variables $\mathcal{N}^{n,m,j}$.

Finally we define $u^{n+1}_0 = u^{n+1}_{N_S}$ so that the periodic boundary condition are satisfied.

The first simulation is given in Figure 1: we represent the value of the numerical solution at different times $t = 0.5, 1, 1.5, 2$. The initial condition is $u_0(x) = \sin(2\pi x)$, the viscosity is $\nu = 0.01$. The discretization parameters satisfy $\delta t = 0.01$, $\delta x = 0.005$ and $N = 100$. We see that the method gives a good approximation of the exact solution.

In Figure 2, we show simulations for different values of $\delta t = \delta x$ and $N$; the initial condition is $u_0(x) = \sin(2\pi x)$, the final time is $T = 0.1$, and the viscosity is $\nu = 0.1$. We observe that even with small values of the parameters we can get accurate approximations. Some oscillations are present due to the randomness in the scheme, and below we try to analyze their dependence with respect to the parameters.

In Figure 3, we fix a value $\delta t = 0.04$ for the time step and a value $N = 50$ for the Monte-Carlo parameter, and we give simulations for different values of the mesh size $\delta x \in \{0.2, 0.1, 0.05, 0.025\}$. The oscillations depend heavily on the mesh size. From Figure 2, we see that they also depend on $N$. These observations traduce a lack of regularity in space of the numerical solution. It seems that the approximation is only accurate in the $L^2$ sense, not in the $H^1$ sense. A connection can be made with the result of Proposition 4.8, where the stability of one iteration of the scheme with respect to the discrete $h^1$ norm is studied: we have proved that

$$\mathbb{E}\|u^{n+1}\|_{h^1}^2 \leq C(1 + \frac{\delta t}{N\delta x^2})\mathbb{E}\|u^n\|_{h^1}^2.$$

We focus on the dependence of the maximum of $\|u^n\|_{h^1}$ during a realization of the scheme, with respect to the discretization parameters: in Figure 4 we draw curves in logarithmic scales. The initial condition is $u_0(x) = \cos(2\pi x)$ and the final time is $T = 0.1$.

On the left, we choose 3 different values of $N$ and consider the evolution when $\delta t = \delta x$. On the right, $\delta t = 10^{-2}$ is fixed, and for different values of $N$ we study the dependence with respect to $\delta x$. The observation of the convergence rates confirms the theoretical prediction. The irregularity of trajectories is a property inherent to the hybrid semi-lagrangian method.

The method can be extended to an equation with an additional term $g$ depending only in space:

$$\begin{aligned} \partial_t u &= \nu \partial_{xx}^2 u + g(x), \\ u(t,1) &= u(t,0) \text{ for any } t \geq 0, \\ u(0,x) &= u_0(x) \text{ for any } x \in [0,1]. \end{aligned} \tag{5.2}$$

We assume that $g$ is a periodic function defined on $\mathbb{R}$, and then the representation formula is:

$$u(t,x) = \mathbb{E}\left(u_0(x + B_{2\nu t}) + \int_0^t g(x + B_{2\nu s})ds\right).$$

In the algorithm, the formula becomes:

$$u_j^{n+1} = \frac{1}{N}\sum_{m=1}^{N}\left((\mathcal{I}u^n)(x_j + \sqrt{2\nu\delta t}\mathcal{N}^{n,m,j}) + \delta t(\mathcal{I}\circ\mathcal{P}g)(x_j + \sqrt{2\nu\delta t}\mathcal{N}^{n,m,j})\right).$$

This numerical technique is useful to approximate by relaxation in time the stationary solution of the PDE when it exists, which satisfies $\nu\partial_{xx}^2 u_\infty + g(x) = 0$ and the periodic boundary conditions. We can give easy examples with trigonometric polynomial function: if

$$g(x) = \sum_{n=1}^{N}\alpha_n\cos(2\pi nx) + \sum_{n=1}^{N}\beta_n\sin(2\pi nx),$$

then we have

$$u_\infty(x) = \nu^{-1}\sum_{n=1}^{N}\frac{\alpha_n}{4\pi^2n^2}\cos(2\pi nx) + \nu^{-1}\sum_{n=1}^{N}\frac{\beta_n}{4\pi^2n^2}\sin(2\pi nx).$$

We show an example in the case $g(x) = \nu 4\pi^2\cos(2\pi x)$ and $u_0(x) = 0$, with the parameters

$$\begin{aligned} \nu &= 0.05, \\ \delta t &= 0.01, \\ \delta x &= 0.01, \\ N &= 50. \end{aligned}$$

The simulation is given in Figure 5, where we represent the solution at different times, and we observe the convergence to the stationary solution, which is $u_\infty(x) = \cos(2\pi x)$ in this simple example.

Finally, we study the convergence of the scheme, with a numerical simulation which confirms the order of convergence with respect to the parameters $\delta t = \delta x$ of the Monte-Carlo error. The final time is $T = 0.1$, the viscosity is $\nu = 0.1$ and the initial condition is $u_0(x) = \cos(2\pi x)$. We compare the numerical solution

$u^n$ with the exact solution; we only observe the Monte-Carlo error, which is dominant with respect to the deterministic part of the error according to Theorem 4.1. The mean-square error in the $l^2$ norm is estimated with a sample of size 20.

The error in Figure 6 is represented in logarithmic scales. The parameters $\delta t$ and $\delta x$ are equal and satisfy $\delta t = \delta x = \frac{1}{n}$ for the following values $n = 50, 100, 200, 400, 800, 1600, 3200$. Each line is obtained when we draw the logarithm of the Error as a function of $\log_{10}(n)$, for a fixed value of $N \in \{10, 20, 40, 80\}$. The dot-line represents a straight-line with slope $-1/2$.

This experiment confirms that the Monte-Carlo error is of order $1/2$ with respect to the parameters when $\delta t = \delta x$, as (4.8) claims. Indeed, the shift between the lines when $N$ varies also corresponds to the size $1/\sqrt{N}$ of the Monte-Carlo error.

5.2.2. **With Dirichlet boundary conditions.** We now consider the heat equation on the interval $[a, b]$, with homogeneous Dirichlet boundary conditions:

$$\partial_t u = \nu \partial^2_{xx} u$$
$$u(t, b) = u(t, a) = 0 \text{ for any } t \geq 0$$
$$u(0, x) = u_0(x) \text{ for any } x \in [a, b].$$

In this situation, the theory of Fourier series also gives explicit formulas for the solution at any time if the initial condition is a trigonometric polynomial function, which is a sum of the basis functions satisfying the boundary conditions: if for a nonnegative integer $N$ and real numbers $a_1, \ldots, a_N$ we have for any $x \in [a, b]$

$$u_0(x) = \sum_{n=1}^{N} a_n \sin(n\pi \frac{(x-a)}{b-a}),$$

then the solution u is for any $t \geq 0$ and $x \in [a, b]$

$$u(t, x) = \sum_{n=1}^{N} a_n \exp(-\frac{\pi^2 n^2}{(b-a)^2} \nu t) \sin(n\pi \frac{(x-a)}{b-a}).$$

As explained above, the stochastic process giving the probabilistic representation formula is a killed diffusion: it is the solution of a stochastic differential equation until it reaches the boundary of the domain. Here the diffusion process depends on the Brownian Motion as follows: if $X^x_t = x + B_{2\nu t}$, then

$$u(t, x) = \mathbb{E}[u_0(X^x_t) \mathbb{1}_{t<\tau^x}],$$

where the stopping time $\tau^x$ is defined by

$$\tau^x = \inf \{s > 0; x + B_{2\nu s} \notin (a, b)\}.$$

We have explained in Section 5.1.3 that the estimation of this stopping time during the numerical scheme is a difficult problem, and there we have proposed a method which combines an acceptance-rejection test and a refinement near the boundary.

The principle of the test introduced in [32] and [49] is based on the observation that the law of $(B_t)_{0 \leq t \leq \delta t}$ conditionally to $B_{n\delta t} = x$ and $B_{(n+1)\delta t} = y$ is the law of the process

$$(\tilde{B}^{x,y,\delta t}_t := ty + (\delta t - t)x + W_t - \frac{t}{\delta t} W_{\delta t})_{0 \leq t \leq \delta t}$$

where $W$ is a Brownian Motion. When $x = y = 0$ and $\delta t = 1$, this process is called the Brownian Bridge, and we denote it by $\overline{B}$: in this situation, if we define the exit time of $\tilde{B}^{x,y,\delta t}$ from the domain $(-\infty, b)$ with

$$T^{x,y,\delta t,b} = \inf \left\{ t \in (0, 1); \tilde{B}^{x,y,\delta t}_t \geq b \right\},$$

then for any $b$ such that $b \geq x$ and $b \geq y$ we have - see for instance [3]:

$$\mathbb{P}(T^{x,y,\delta t,b} \leq \delta t) = \mathbb{P}(\sup_{0 \leq t \leq \delta t} \tilde{B}^{x,y,\delta t} \leq b) = \exp\left(-2 \frac{(b-x)(b-y)}{\delta t}\right). \tag{5.3}$$

The test is then based on the comparison on the value of a uniform random variable $U$ on $(0, 1)$ and of this probability denoted by $p$: if $U \leq p$ we decide that $T^{x,y,\delta t,b} \leq \delta t$ so that the numerical process as leaved the boundary; when $U > p$ we rather decide that the process has not touched the barrier $b$.

We now describe in detail the algorithm.

First, the space interval is discretized with a mesh size $\delta x = (b-a)/N_S$, for some nonnegative integer $N_S$: $x_j = a + j\delta x$ for $0 \le j \le N_S$. The time is discretized into $N_T$ intervals of size $\delta t$, with $N_T \delta t = T$.

Space is then divided into 2 parts, thanks to a nonnegative integer parameter $BD$: the scheme is different for the interior indices

$$Ii = \{j; BD \le j \le N_S - BD\}$$

and the boundary indices

$$Ib = \{j; 0 \le j \le BD - 1 \text{ or } N_S - BD + 1 \le j \le N_S\}.$$

Both sets refer to positions $x_j$ inside or outside an interval centered on $\frac{a+b}{2}$ of size $2BD\delta x$.

Corresponding with this decomposition, we also introduce new nonnegative integers: the Monte-Carlo parameters $N_i$ and $N_b$, as well as the subdivision parameter $SUB \ge 1$. The value $SUB = 1$ is allowed, and is used in the interior part.

The algorithm is used to compute a vector $u^n = (u_j^n)_{0 \le j \le N_S}$ for each $0 \le n \le N_T$. The initial vector $u^0$ satisfies for any index $0 \le j \le N_S$ the relation $u_j^0 = u_0(x_j)$, where $u_0$ is the initial condition of the PDE. We notice that thanks to the boundary conditions $u_0^n = u_{N_S}^n = 0$ for any time $n$.

Having calculated the value $u^n$ at time $n \le N_T - 1$, we need to explain how to compute $u^{n+1}$. The definition of $u_j^{n+1}$ changes whether $j \in Ii$ or $j \in Ib$.

**CASE 1** If the index $j \in Ii$ is in the interior, we compute for $1 \le m \le N_i$

$$\tilde{X}_j^{n,m} = x_j + \sqrt{\nu \delta t} \mathcal{N}^{n,m,j},$$

where the $\mathcal{N}^{n,m,j}$ are independent standard Gaussian random variables. It remains to test if $\tilde{X}$ has reached the boundary: we simply define

$$X_j^{n,m} = \begin{cases} \tilde{X}_j^{n,m} \text{ if } \tilde{X}_j^{n,m} \in [a,b], \\ b \quad \text{if } \tilde{X}_j^{n,m} > b, \\ a \quad \text{if } \tilde{X}_j^{n,m} < a. \end{cases}$$

Then we compute by interpolation and average the value at time $n+1$ and position $j$:

$$u_j^{n+1} = \frac{1}{N_i} \sum_{m=1}^{N_i} \mathcal{I}(u^n)(X_j^{n,m}).$$

**CASE 2** If the index is 0 or $N_S$, we just set $u_0^{n+1} = u_{N_S}^{n+1} = 0$.

**CASE 3** If $j \in Ib$, with $j \ne 0$ and $j \ne N_S$, the corresponding point is in the part near the boundary, and more work must be done: the interval $[n\delta t, (n+1)\delta t]$ is divided into $SUB$ subintervals of size $\tau = \frac{\delta t}{SUB}$.

For each $j \in Ib$, $j \notin \{0, N_S\}$, and each $1 \le m \le N_b$, we define $\tilde{X}_j^{n,m}(0) = x_j$. Then for any $0 \le k \le SUB - 1$, we define recursively $\tilde{X}_j^{n,m}(k+1)$ by applications of sub-steps of the Euler scheme of size $\tau$, taking into account the exits of the domain.

If $\tilde{X}_j^{n,m}(k) \notin (a,b)$, we have already reached the boundary so that we set $\tilde{X}_j^{n,m}(k+1) = \tilde{X}_j^{n,m}(k)$.

Otherwise, we define an auxiliary quantity

$$\tilde{Y}_j^{n,m}(k+1) = \tilde{X}_j^{n,m}(k) + \sqrt{\tau} \mathcal{N}_{n,m,j,k},$$

with some independent Gaussian random variables $\mathcal{N}_{n,m,j,k}$.

Now we make a first test: if $\tilde{Y}_j^{n,m}(k+1) \ge b$, then we set $\tilde{X}_j^{n,m}(k+1) = b$. In the case $\tilde{Y}_j^{n,m}(k+1) \le a$, then we set $\tilde{X}_j^{n,m}(k+1) = a$.

It remains a possibility: $\tilde{Y}_j^{n,m}(k+1) \in (a,b)$. According to the discussion above, we add a possibility of exit, taking into account the knowledge of the distribution of the exit time in an interval of size $t$ of the Brownian Motion conditioned to its initial and initial value. According to the position of $\tilde{Y}_j^{n,m}(k+1)$ in the interval, we test an exit through $a$ or $b$.

We simulate random variables $U^{n,m,j,k}$ for the required choices of indices, which are independent from the other random variables used so far, and have uniform distribution on $(0,1)$. We use the law of the exit time of the Brownian Bridge, with the barriers $a$ and $b$:

If $\tilde{Y}_j^{n,m}(k) \in [\frac{a+b}{2}, b)$, the barrier is $b$, and thanks to (5.3) we see that we need to compare $U^{n,m,j,k}$ with

$$H^{n,m,j,k} = \exp\Big(-\frac{\Big(b-\tilde{Y}_j^{n,m}(k+1)\Big)\Big(b-\tilde{Y}_j^{n,m}(k)\Big)}{\nu\delta t}\Big) :$$

if $U^{n,m,j,k} \leq H_j^{n,m,j,k}$ then we consider that the process has reached the boundary and we set $\tilde{X}_j^{n,m}(k+1) = b$; otherwise we set $\tilde{X}_j^{n,m}(k+1) = \tilde{Y}_j^{n,m}(k+1)$.

The treatment of the case $\tilde{Y}_j^{n,m}(k) \in (a, \frac{a+b}{2})$ is similar, with the barrier $a$, and a comparison between $U^{n,m,j,k}$ and

$$H^{n,m,j,k} = \exp\Big(-\frac{\Big(a-\tilde{Y}_j^{n,m}(k+1)\Big)\Big(a-\tilde{Y}_j^{n,m}(k)\Big)}{\nu\delta t}\Big) :$$

if $U^{n,m,j,k} \leq H_j^{n,m,j,k}$ then we set $\tilde{X}_j^{n,m}(k+1) = a$, and otherwise we set $\tilde{X}_j^{n,m}(k+1) = \tilde{Y}_j^{n,m}(k+1)$.

We finally set $X_j^{n,m} = \tilde{X}_j^{n,m}(SUB)$. And it remains to compute the $u_j^{n+1}$ by interpolation and average:

$$u_j^{n+1} = \frac{1}{N_b}\sum_{m=1}^{N_b} \mathcal{I}(u^n)(X_j^{n,m}).$$

The first example in Figure 7 is the approximation of the solution at time $T = 1$ when the initial condition is $u_0(x) = \sin(\pi\frac{(x+1)}{2})$, on the domain $(-1,1)$. The parameters satisfy

$$\delta t = 0.01,$$
$$\delta x = 0.01,$$
$$\nu = 0.05.$$

The decomposition of the domain depends on the parameter $BD = 20$; in the interior part, the Monte-Carlo parameter is $N_i = 50$, while in the exterior part we have $N_b = 50$ realizations, and a subdivision in time with $SUB = 20$ subintervals.

This example shows that the method gives accurate approximations.

In Figure 8, we propose simulations with different values of the subdivision parameter $SUB = 1, 5, 10, 20$, where the other parameters remain unchanged:

$$\delta t = 0.02,$$
$$\delta x = 0.02,$$
$$\nu = 0.1,$$
$$BD = 10,$$
$$N_i = N_b = 100.$$

This Figure shows the importance of the refinement near the boundary. The case $SUB = 1$ shows that the additional exit test is not sufficient when the time step is not small enough, and that the error is not localized only near the boundary. The values $SUB = 5$ and $SUB = 10$ already give satisfactory results, for a reasonable computational cost.

To give other examples, we now suppose that the equation contains an additional term $g(x)$:

$$\begin{aligned}\partial_t u(t,x) &= \nu \partial^2_{xx} u(t,x) + g(x)\\ u(t,-1) = u(t,1) &= 0 \text{ for any } t \geq 0\\ u(0,x) &= u_0(x) \text{ for any } x \in [-1,1].\end{aligned}$$

The representation formula contains an additional integral term where the stopping time appears:

$$u(t,x) = \mathbb{E}[u_0(X_t^x)\mathbb{1}_{t<\tau^x} + \int_0^{t\wedge\tau^x} g(X_s^x)ds].$$

Below, we consider that the initial condition is null, and we approximate the stationary solution $u_\infty$ in some particular cases where we have an explicit formula:

$$\begin{aligned}\nu \partial^2_{xx} u_\infty + g &= 0\\ u_\infty(-1) = u_\infty(1) &= 0.\end{aligned}$$

In the algorithm, the computation of the $X_j^{n,m}$ is unchanged; we just remark that we can also easily obtain values for the stopping times on each interval. In the interior part, we define $\tau_j^{n,m} = 0$ when the exit test is positive and $\tau_j^{n,m} = \delta t$. In the exterior part, this quantity is approximated more precisely thanks to the subdivision: $\tau_j^{n,m} = k\frac{\delta t}{SUB}$, where $k = \sup\left\{l; \tilde{X}_j^{n,m}(k) \in (-1,1)\right\}$.

We then define recursively for $j \in Ii$

$$u_j^{n+1} = \frac{1}{N_i}\sum_{m=1}^{N_i}\Big(\mathcal{I}(u^n)(X_j^{n,m}) + \tau_j^{n,m}\mathcal{I}(g)(X_j^{n,m})\Big),$$

and for $j \in Ib$

$$u_j^{n+1} = \frac{1}{N_b}\sum_{m=1}^{N_b}\Big(\mathcal{I}(u^n)(X_j^{n,m}) + \tau_j^{n,m}\mathcal{I}(g)(X_j^{n,m})\Big).$$

In the numerical simulation for this problem, we take the initial condition $u_0 = 0$. In Figure 9, the function is $g^1(x) = -\nu x$, which gives the stationary solution $u^1_\infty(x) = \nu\frac{x^3-x}{6}$.

The parameters are

$$\begin{aligned}\nu &= 0.1,\\ \delta t = \delta x &= 0.02,\\ N_i = N_b &= 100,\\ SUB &= 10.\end{aligned}$$

We finally give in Figure 10 the result of investigations on the convergence of the method. We draw in logarithmic scales the Error in terms of $n = 1/\delta t = 1/\delta x$, with $n = 50, 100, 200, 400, 800$, with different values of the Monte-Carlo parameter $N = 10, 20, 40, 80$. We have chosen on the interval $(-1,1)$ the initial function $u_0(x) = \sin(\pi\frac{x+1}{2})$, with the viscosity $\nu = 0.1$. There is no additional term $g$, and the solutions are simulated until time $T = 0.1$. Like in the case of periodic boundary conditions, the statistical error is dominant with respect to the other error terms; we compare with the exact solution, and to estimate the variance we use a sample of size 100.

It is important to notice that we have considered $SUB = 10$ and that $N_b = 10N_i = 10N$.

The observation of Figure 10 shows that the Monte-Carlo error depends on the parameter $\delta t = \delta x$; the comparison with the "theoretical" line with slope $-1/2$ indicates a conjecture that the error is also of order $1/2$, like for the periodic case. The shift between the curves for different values of $N$ corresponds in the error to a factor $1/\sqrt{N}$.

5.2.3. **With Neumann boundary conditions.** The third type of boundary conditions exposed in the introductory part are the Neumann boundary conditions: the value of the derivative at the left and right points of the interval is assumed to be 0 - the case of non-homogeneous Neumann boundary conditions is more complicated: we now consider the equation

$$\partial_t u = \nu \partial^2_{xx} u$$
$$u'(t,b) = u'(t,a) = 0 \text{ for any } t \geq 0$$
$$u(0,x) = u_0(x) \text{ for any } x \in [a,b].$$

We assume that the initial condition satisfies the Neumann boundary conditions.

The case of the equation $\partial_t u = \nu \partial^2_{xx} u + g$ could also be studied, but for simplicity we reduce our presentation to the case $g = 0$.

In some special cases, we can give an explicit solution: if for a nonnegative integer $N$ and real numbers $a_0, a_1, \ldots, a_N$ we have for any $x \in [a,b]$

$$u_0(x) = a_0 + \sum_{n=1}^{N} a_n \cos(n\pi \frac{(x-a)}{b-a}),$$

then the solution u is for any $t \geq 0$ and $x \in [a,b]$

$$u(t,x) = a_0 + \sum_{n=1}^{N} a_n \exp(-\frac{\pi^2 n^2}{(b-a)^2}\nu t) \cos(n\pi \frac{(x-a)}{b-a}).$$

The representation formula requires to solve a system of stochastic differential equations involving a reflexion of the process at the boundary.

The construction of the method follows the same strategy as for Dirichlet boundary conditions: we decompose the domain into two parts, and we use a refinement of the time intervals near the boundary to improve the approximation.

The space interval is discretized with a mesh size $\delta x = (b-a)/N_S$, for some nonnegative integer $N_S$: $x_j = a + j\delta x$ for $0 \leq j \leq N_S$. The time is discretized into $N_T$ intervals of size $\delta t$, with $N_T \delta t = T$.

Space is then divided into 2 parts, thanks to a nonnegative integer parameter $BD$: the scheme is different for the interior indices

$$Ii = \{j; BD \leq j \leq N_S - BD\}$$

and the boundary indices

$$Ib = \{j; 0 \leq j \leq BD - 1 \text{ or } N_S - BD + 1 \leq j \leq N_S\}.$$

Corresponding with this decomposition, we also introduce new nonnegative integers: the Monte-Carlo parameters $N_i$ and $N_b$, as well as the subdivision parameter $SUB \geq 1$. The value $SUB = 1$ is allowed.

The algorithm is used to compute a vector $u^n = (u^n_j)_{0 \leq j \leq N_S}$ for each $0 \leq n \leq N_T$. The initial vector $u^0$ satisfies for any index $0 \leq j \leq N_S$ the relation $u^0_j = u_0(x_j)$, where $u_0$ is the initial condition of the PDE. According to the presentation in Section 5.1.3 the probabilistic representation formula involves the reflexion of the diffusion process when it reaches the boundary: this reflexion must be done at the discrete time level.

**CASE 1** If the index $j \in Ii$ is in the interior, we compute for $1 \leq m \leq N_i$

$$\tilde{X}^{n,m}_j = x_j + \sqrt{\nu \delta t} \mathcal{N}^{n,m,j},$$

where the $\mathcal{N}^{n,m,j}$ are independent standard Gaussian random variables. It remains to test if $\tilde{X}$ has reached the boundary, in which case we need to reflect the process: we simply define

$$X^{n,m}_j = \begin{cases} \tilde{X}^{n,m}_j & \text{if } \tilde{X}^{n,m}_j \in [a,b], \\ 2b - \tilde{X}^{n,m}_j & \text{if } \tilde{X}^{n,m}_j > b, \\ 2a - \tilde{X}^{n,m}_j & \text{if } \tilde{X}^{n,m}_j < a. \end{cases}$$

Then we can compute by interpolation and average the value at time $n+1$ and position $j$:

$$u^{n+1}_j = \frac{1}{N_i} \sum_{m=1}^{N_i} \mathcal{I}(u^n)(X^{n,m}_j).$$

**CASE 2** If $j \in Ib$ the corresponding point is in the part near the boundary, and more work must be done: we subdivide $[n\delta t, (n+1)\delta t]$ into $SUB$ subintervals of size $\tau = \frac{\delta t}{SUB}$. This allows to compute a value of the discrete time process using $SUB$ steps, with a smaller time step; a simple exit test-reflexion is done at each step.

For each $j \in Ib$ and each $1 \leq m \leq N_b$, we define $\tilde{X}_j^{n,m}(0) = x_j$. Then for any $0 \leq k \leq SUB - 1$, we define $\tilde{X}_j^{n,m}(k+1)$.

We first set an auxiliary random variable

$$\tilde{Y}_j^{n,m}(k+1) = \tilde{X}_j^{n,m}(k) + \sqrt{\tau}\mathcal{N}_{n,m,j,k},$$

with independent Gaussian random variables $\mathcal{N}_{n,m,j,k}$. We then define

$$\tilde{X}_j^{n,m}(k+1) = \begin{cases} \tilde{Y}_j^{n,m}(k+1) & \text{if } \tilde{Y}_j^{n,m}(k+1) \in [a,b], \\ 2b - \tilde{Y}_j^{n,m}(k+1) & \text{if } \tilde{Y}_j^{n,m}(k+1) > b, \\ 2a - \tilde{Y}_j^{n,m}(k+1) & \text{if } \tilde{Y}_j^{n,m}(k+1) < a. \end{cases}$$

We finally $X_j^{n,m} = \tilde{X}_j^{n,m}(SUB)$. It then remains to compute the $u_j^{n+1}$ by interpolation and average:

$$u_j^{n+1} = \frac{1}{N_b} \sum_{m=1}^{N_b} \mathcal{I}(u^n)(X_j^{n,m}).$$

We consider an example in the domain $(-1,1)$, where we approximate the solution at time $T = 1$ when the initial condition is $u_0(x) = \cos(\pi \frac{x+1}{2})$. The parameters are

$$\nu = 0.1,$$
$$\delta t = \delta x = 0.01,$$
$$N_i = N_b = 100,$$
$$SUB = 20.$$

The simulation is given in Figure 11, and we observe that the method seems performant.

5.2.4. **Simulations of boundary layers.** We end the treatment of equations in dimension 1 with an example where the behavior of the solutions near the boundary must be treated carefully.

Given $\epsilon > 0$, we define $u^\epsilon$ the solution of the equation for $x \in (-1,1)$

$$\epsilon \partial_{xx}^2 u^\epsilon - (g(u^\epsilon) - 1) = 0, u^\epsilon(1) = u^\epsilon(-1) = 0.$$

It corresponds to the evolution problem

$$\partial_t u = \epsilon \partial_{xx}^2 u^\epsilon - (g(u^\epsilon) - 1), u(t,1) = u(t,-1) = 0, \forall t \geq 0, u(0,.) = u_0.$$

The simplest choice for the function $g$ is $g(x) = x$. In this situation, we observe that when $\epsilon$ goes to 0 the solution $u^\epsilon$ needs to satisfy two constraints: first, the homogeneous Dirichlet boundary conditions impose that $u^\epsilon$ is 0 at the points $-1$ and 1, while the PDE itself implies that $u^\epsilon \to 1$. These two conditions are not compatible, and we observe the formation of boundary layers when $\epsilon$ goes to 0, which allow that inside the domain the second constraint is satisfied without breaking the boundary conditions.

The exact solution of the stationary problem is

$$u^\epsilon(x) = 1 - \frac{\cosh(\frac{x}{\epsilon})}{\cosh(\frac{1}{\epsilon})}.$$

It is then easy to check that if $|x| < 1$ we have $u^\epsilon(x) \to 1$ when $\epsilon \to 0$.

This is a simple case, and we rather study a generalized situation, with for instance some polynomial functions $g_\alpha$ such that

$$g_\alpha(x) = x - \alpha x^3.$$

When $\alpha$ is small enough, we define $x_\alpha$ such that $g_\alpha(x_\alpha) = 1$, and the solution $u^\epsilon$ for $g = g_\alpha$ satisfies

$$u^\epsilon(x) \to x_\alpha$$

when $\epsilon \to 0$ and $|x| < 1$.

In the case $\alpha \neq 0$, we do not have an explicit formula for the solution $u^\epsilon$: the numerical hybrid scheme becomes useful.

Numerically, we approximate the solution of the evolution equation for some large final time $T$ - depending on the values of $\epsilon$. We recall that the method is explicit. The use of classical finite difference schemes would require more refined grids when $\epsilon$ goes to 0.

We use the algorithm proposed above for the heat equation with Dirichlet boundary conditions. The change is that the function $g$ now depends on the solution $u$. To deal with this situation, we propose an explicit scheme, where we replace $g_\alpha(u(t,x))$ with $g_\alpha(u(n\delta t,x))$ on the interval $[n\delta t,(n+1)\delta t]$.

We notice that in the interior of the domain we do not need to simulate many gaussian random variables for each point in the interior part of the domain: there $\epsilon$ only plays the role of a viscosity, and the Monte-Carlo error should then decrease when $\epsilon$ goes to 0.

In Figure 12, we present the simulations of the stationary solutions for different values of $\epsilon$ and $\alpha = -0.1, 0, 0.1$. The other parameters are $\delta t = \delta x = 0.01$, $BD = 20$, $SUB = 10$. For $\epsilon = 0.001$, we take $N_i = N_b = 40$, while for $\epsilon = 0.0001$ we choose $N_i = 1, N_b = 10$.

In the case $\alpha = 0$, the exact solution is represented in dot lines.

We observe the expected formation of the boundary layers. We notice that since the viscosity is small, the number of Monte-Carlo simulations does not need to be high; nevertheless a careful treatment near the boundary, with a refinement in time and additional exit tests, remains essential.

### 5.3. The Heat Equation in dimension 2

The aim of this section is to give first, easy examples of applications in dimension greater than one, with three types of boundary conditions on a square domain: Dirichlet, Neumann and "mixed" conditions - i.e. with Dirichlet conditions on left and right and Neumann conditions on top and bottom. In Section 5.3.4, we present a simulation where boundary layers appear.

We do not give the expressions of the algorithms: they can be easily guessed from the general presentation in Section 5.1.3 and from the corresponding algorithms in dimension 1 given above in Section 5.2.

The numerical simulations are given in Section 5.6, and they all confirm the applicability of the method.

5.3.1. **With Dirichlet boundary conditions.** We consider the heat equation on the domain $(-1,1)^2$, with homogeneous Dirichlet boundary conditions, and with a given initial condition $u_0$.

$$\frac{\partial u}{\partial t} = \nu \Delta u + \mu g(x),$$
$$u(t,x,y) = 0 \text{ if } |x| = 1 \text{ or } |y| = 1, \forall t \geq 0,$$
$$u(0,x,y) = u_0(x,y) \forall (x,y) \in [-1,1]^2.$$

In Figure 13, we present a case with $\mu = 0$, and $u_0(x,y) = \sin(\pi \frac{x+1}{2}) \sin(\pi y)$. We compute the numerical solution at time $T = 1$, and make a comparison with the exact solution such that $u(T,x,y) = \exp(-(1 + 1/4)\pi^2 \nu T) u_0(x,y)$. The parameters are

$$\nu = 0.01,$$
$$\delta t = \delta x = 0.05,$$
$$BD = 10, SUB = 20,$$
$$N_i = N_b = 100.$$

In Figure 14, we assume that the initial condition satisfies $u_0(x,y) = 0$, and we consider $\mu = \nu$ and $g(x,y) = -\cosh(x) \left(\frac{y^3 - y}{6}\right) - (\cosh(x) - \cosh(1))\, y$. The stationary solution - such that $\nu \Delta u + \mu g = 0$ - then satisfies $u(x,y) = (\cosh(x) - \cosh(1)) \left(\frac{y^3 - y}{6}\right)$. The parameters are

$$\nu = \mu = 0.1,$$
$$\delta t = \delta x = 0.05,$$
$$BD = 10, SUB = 20,$$
$$N_i = N_b = 500.$$

In Figure 15, we assume that the initial condition satisfies $u_0(x,y) = 0$, and we consider $\mu = \nu$ and $g(x,y) = \big((x^2-1)\pi^2 - 2\big)\sin(\pi y)$. The stationary solution - such that $\Delta u + g = 0$ - then satisfies $u(x,y) = (x^2-1)\sin(\pi y)$. The parameters are

$$\nu = \mu = 0.1,$$
$$\delta t = \delta x = 0.05,$$
$$BD = 10, SUB = 20,$$
$$N_i = N_b = 500.$$

We have chosen large values for the number of Monte-Carlo realizations in order to improve the regularity of the numerical solutions. The three examples of simulations show that the method can be adapted to higher dimensional cases.

5.3.2. **With Neumann boundary conditions.** With homogeneous Neumann boundary conditions, we study the following equation in the domain $(-1,1)^2$:

$$\frac{\partial u}{\partial t} = \nu \Delta u,$$
$$\partial_x u(t,x,y) = 0 \text{ if } |x| = 1,$$
$$\partial_y u(t,x,y) = 0 \text{ if } |y| = 1,$$
$$u(0,x,y) = u_0(x,y) \forall (x,y) \in [-1,1]^2.$$

In Figure 16, we present a case with the initial condition $u_0(x,y) = \cos(\pi \frac{x+1}{2})\cos(\pi y)$. We approximate the exact solution at time $T = 1$, which has the expression $u(T,x,y) = \exp(-(1+1/4)\pi^2 \nu T) u_0(x,y)$. The parameters are

$$\nu = 0.01,$$
$$\delta t = \delta x = 0.05,$$
$$BD = 10, SUB = 20,$$
$$N_i = N_b = 100.$$

We mention that this case of the heat equation in dimension 2 with homogeneous Neumann boundary conditions is important: we meet it again in the definition of projections methods for Stokes or Navier-Stokes equations in Section 5.5.

5.3.3. **With mixed boundary conditions.** We can also mix the boundary conditions: when $|x| = 1$, we assume that Dirichlet boundary conditions are given, while when $|y| = 1$ we consider Neumann boundary conditions:

$$\frac{\partial u}{\partial t} = \nu \Delta u + \mu g(x),$$
$$u(t,x,y) = 0 \text{ if } |x| = 1, \partial_y u(t,x,y) = 0 \text{ if } |y| = 1, u(0,x,y) = u_0(x,y) \forall (x,y) \in [-1,1]^2.$$

When $\mu = 0$ and $u_0(x,y) = \sin(\pi \frac{x+1}{2})\cos(\pi y)$, the exact solution satisfies $u(t,x,y) = \exp(-(1 + 1/4)\pi^2 \nu t) u_0(x,y)$. In Figure 17, we draw the numerical approximation at time $T = 1$ and the corresponding exact solution. The parameters are

$$\nu = 0.01,$$
$$\delta t = \delta x = 0.05,$$
$$BD = 10, SUB = 20,$$
$$N_i = N_b = 100.$$

We also consider the situation $\mu = \nu$ and $g(x) = \big(\pi^2(x^2-1) - 2\big)\cos(\pi y)$, with the initial condition $u_0(x,y) = 0$. The associated stationary solution is $u(x,y) = (x^2-1)\cos(\pi y)$. Figure 18 is obtained with the

parameters

$$\nu = \mu = 0.1,$$
$$\delta t = \delta x = 0.05,$$
$$BD = 10, SUB = 20,$$
$$N_i = N_b = 500.$$

We observe that the method is also applicable for this example of boundary conditions, with a very simple algorithm.

5.3.4. **Simulation of boundary layers.** Like in the one-dimensional case, the hybrid semi-lagrangian method can be applied to a situation where boundary layers appear - see Section 5.2.4. Here for a given $\epsilon > 0$ we study the equation in the domain $D = (-1, 1)^2$

$$\epsilon \Delta u^\epsilon - (u^\epsilon - 1) = 0, u^\epsilon{}_{|\partial D} = 0,$$

and we look at the behavior of the solutions when $\epsilon$ goes to 0.

The simulations are done with the parameters

$$\delta t = \delta x = 0.02,$$
$$BD = 10, SUB = 20,$$
$$N_i = 20, N_b = 100,$$
$$T = 5.$$

In Figure 19, we first consider $\epsilon = 0.001$ on the left, and $\epsilon = 0.0001$ on the right. We observe the presence of boundary layers, which are well-approximated by the method. We recall that the method is explicit and simple.

## 5.4. The Burgers Equation

We now consider the case of the viscous Burgers equation $\frac{\partial u}{\partial t} + (u.\nabla)u = \nu \Delta u + f$. We assume that on the boundary homogeneous Dirichlet conditions are satisfied, and the initial condition $u(0, .) = u_0$ is given.

More precisely, the velocity field $u = (u_1, \ldots, u_d)$ is a $d$-dimensional function of time variable $t$ and space variable $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$. The notation $\Delta u$ stands for the vector $(\Delta u_1, \ldots, \Delta u_d)$, and we also have $\frac{\partial u}{\partial t} = (\frac{\partial u_1}{\partial t}, \ldots, \frac{\partial u_d}{\partial t})$. The differential operator $u.\nabla$ stands for $u.\nabla = \sum_{i=1}^d u_i \frac{\partial}{\partial x_i}$. As a consequence, $(u.\nabla)u = (\sum_{i=1}^d u_i \frac{\partial u_1}{\partial x_i}, \ldots, \sum_{i=1}^d u_i \frac{\partial u_d}{\partial x_i})$.

The coefficient $f$ represents the effect of a force applied to the fluid described by the Eulerian velocity field $u$. The differential operator $u.\nabla$ corresponds to transport with velocity $u$, and gives a nonlinear term in the equation. The fluid is assumed to have a viscosity $\nu > 0$. Then the PDE - of parabolic type - admits a unique solution which is smooth. When the viscosity tends to 0, the solution tends to the unique entropic solution of the inviscid Burgers equation $\frac{\partial u}{\partial t} + (u.\nabla)u = f$, which is now an hyperbolic system of conservation laws and may have non smooth solutions.

One important property of the Burgers equation is the nonlinear term. In the next section, we will describe the relations of this PDE with the Navier-Stokes equations for incompressible fluids; the Burgers equation corresponds to a simpler model - without the incompressibility constraint - but with the same nonlinearity.

First, we describe the construction of the hybrid semi-lagrangian algorithm associated with the Burgers equation. We then explain how exact solutions can be computed thanks to the Hopf-Cole transformation. Finally, we present various examples of numerical solutions.

5.4.1. **The algorithm.** We must face a new difficulty due to the nonlinear term, which does not allow to construct directly a diffusion process. The strategy is to consider that during a time step the transport occurs with a constant velocity computed with the previous step, so that we obtain a linear parabolic equation on each subinterval, which admits a probabilistic representation formula. We assume that the force satisfies $f = 0$ - otherwise we need to introduce other terms, which are similar to the cases treated before.

More precisely, we construct approximations $u^n$ of the solution at discrete times $n\delta t$, introducing functions $v^n$ such that for any $n \geq 0$ with the following semi-implicit scheme:

$$\frac{\partial v^{n+1}}{\partial t} + (u^n.\nabla)v^{n+1} = \nu\Delta v^{n+1}, \tag{5.4}$$

for any time $n\delta t \leq t \leq (n+1)\delta t$ and $x \in D$. The initial condition is $v^{n+1}(n\delta t, .) = u^n = v^n(n\delta t, .)$. The discrete-time approximation then satisfies $u^0 = u_0$ and $u^n = v^n(n\delta t, .)$.

On each subinterval $[n\delta t, (n+1)\delta t]$, we have

$$v^{n+1}(t,x) = \mathbb{E}[v^{n+1}(n\delta t, X_t^x)\mathbb{1}_{t<\tau^x}],$$

where the diffusion process $X$ satisfies

$$dX_t^x = -u^n(X_t^x)dt + \sqrt{2\nu}dB_t, X_{n\delta t}^x = x.$$

The stopping times $\tau^x$ represent the first exit time of the process in the time interval $[n\delta t, (n+1)\delta t]$. Since $v^{n+1}(n\delta t, .) = u^n$, the scheme only requires the knowledge of the approximations $u^n$. For other types of boundary conditions, other representation formulas are also available.

It is then easy to deduce a hybrid semi-lagrangian algorithm. The only change is that the diffusion used depends on the solution at the previous step in an explicit way.

5.4.2. **Exact solutions.** To check that the method is accurate, it is useful to know a way to compute the exact solution. We assume that $f = 0$. Thanks to the Hopf-Cole transformation, if the dimension is $d = 1$, if the initial condition $u_0$ can be expressed as $u_0 = -2\nu\partial_x \log(\Psi_0)$ for some function $\Psi_0$, then $u = -2\nu\partial_x \log(\Psi)$, where $\Psi$ is solution of the heat equation $\frac{\partial\Psi}{\partial t} = \nu\partial_{xx}^2\Psi$, with the initial condition $\Psi(0, .) = \Psi_0$. If the domain $D$ is bounded, we must impose boundary conditions for $u$ and $\Psi$. If $u$ is assumed to satisfy periodic boundary conditions, the same holds for $\Psi$. If $u$ satisfies homogeneous Dirichlet boundary conditions, we see that $\Psi$ satisfies homogeneous Neumann boundary conditions.

It is easy to give examples, thanks to the expansion in Fourier series of the general solution of the heat equation with Neumann boundary conditions: if for a nonnegative integer $N$ and real numbers $a_0, a_1, \ldots, a_N$ we have for any $x \in [a, b]$

$$\Psi_0(x) = a_0 + \sum_{n=1}^{N} a_n \cos(n\pi\frac{(x-a)}{b-a}),$$

then the solution $\Psi$ satisfies for any $t \geq 0$ and $x \in [a, b]$

$$\Psi(t,x) = a_0 + \sum_{n=1}^{N} a_n \exp(-\frac{\pi^2 n^2}{(b-a)^2}\nu t)\cos(n\pi\frac{(x-a)}{b-a}).$$

If we choose the coefficients such that $\Psi_0$ is everywhere non zero, we can define $u_0(x) = -2\nu\frac{(\Psi_0)'(x)}{\Psi_0(x)}$, and the solution of the problem with the initial condition $u_0$ is

$$u(t,x) = -2\nu\frac{\partial_x\Psi(t,x)}{\Psi(t,x)} = -2\nu\frac{\sum_{n=1}^{N}\frac{n\pi a_n}{b-a}\exp(-\frac{\pi^2n^2}{(b-a)^2}\nu t)\cos(n\pi\frac{(x-a)}{b-a})}{a_0 + \sum_{n=1}^{N} a_n \exp(-\frac{\pi^2n^2}{(b-a)^2}\nu t)\cos(n\pi\frac{(x-a)}{b-a})}.$$

In Figure 20, we show the numerical approximation at time $T = 1$, when the initial condition is $u_0(x) = \nu\frac{\pi\sin(\pi(x+1)/2)}{2+\cos(\pi(X+1)/2)}$. The parameters are

$$\nu = \mu = 0.1,$$
$$\delta t = \delta x = 0.05,$$
$$BD = 10, SUB = 20,$$
$$N_i = N_b = 500.$$

We observe that the method gives a good approximation of the exact solution.

5.4.3. **Examples with formation of shocks.** We consider a few examples with formation of shocks in the solution when viscosity decreases. The corresponding numerical solutions are given in Section 5.6.

We first consider the case when the initial condition is $u_0^+(x) = \sin(\pi x)$. In Figure 21, we draw the numerical solution at times $t \in \{0, 0.5, 1, 1.5, 2\}$, with the viscosity $\nu = 0.1$, while in Figure 22 the viscosity is $\nu = 0.01$.

The other discretization parameters are:

$$\delta t = \delta x = 0.02,$$
$$BD = 10, SUB = 20,$$
$$N_i = N_b = 100.$$

We also consider the initial condition $u_0^-(x) = -\sin(\pi x)$. Due to the nonlinearity of the equation, changing the sign of the initial condition totally the changes the behavior of the solution. In Figure 23, we draw the numerical solution at times $t \in \{0, 0.5, 1, 1.5, 2\}$, with the viscosity $\nu = 0.1$, while in Figure 24 the viscosity is $\nu = 0.01$.

We observe that the qualitative behavior of the solutions respects the entropy conditions - see for instance [26].

However, the interpolation procedure introduces a numerical viscosity, depending on the method: the viscosity of the numerical experiment is greater than $\nu$. To overcome this problem, higher order interpolation methods should be used.

We finally make the following remark: if formally we choose $\nu = 0$ the numerical scheme we obtain is not consistent with the inviscid Burgers equation. To improve the method, we could try to build a scheme in the case $\nu > 0$ from an existing efficient scheme when $\nu = 0$.

## 5.5. The Navier-Stokes Equations

In this section, we focus on an example of computation of numerical solutions for the incompressible Navier-Stokes equations: more precisely, we present simulations of flows in a two-dimensional driven cavity.

In this Section, we first recall a few general and basic elements about the incompressible Navier-Stokes system. Then we explain how hybrid semi-lagrangian schemes can be combined with the projection methods for the discretization of the equations.

The numerical simulations we have obtained are given in Section 5.6.

5.5.1. **General facts about the incompressible Navier-Stokes equations.** This part is devoted to a short presentation of the PDE: we briefly explain the origins and the main fundamental properties of the equations. We refer to [63] for a more detailed description.

The equations govern the behavior in a time interval $[0, T]$ of a fluid in a spatial domain $D \in \mathbb{R}^d$, where for physical situations the dimension $d$ is equal to 2 or 3. The domain $D$ which contains the fluid is assumed to be a fixed bounded open set, with a boundary $\partial D$ satisfying appropriate smoothness conditions; in practice, we treat the case of a box $D = ]\alpha, \beta[^d$. Considering the Eulerian description of the fluid, the unknown of the problem is the velocity field $u : (t, x) \in [0, T] \times \mathbb{R}^d \mapsto u(t, x) \in \mathbb{R}^d$, such that $u(t, x)$ is the velocity of a particle of the fluid being in position $x$ at time $t$. The configuration at time $t = 0$ is given: the initial condition on the velocity is $u(0, .) = u_0$. Following various physical assumptions and principles, $u$ is solution of the following PDE problem:

$$\begin{aligned} \frac{\partial u}{\partial t} + (u.\nabla)u - \nu\Delta u + \nabla p &= f \\ \mathrm{div}(u) &= 0, \\ u(0, .) &= u_0, \\ +\text{boundary conditions.} & \end{aligned} \tag{5.5}$$

The differential operators appearing above are defined in the previous Section 5.4 on the Burgers equation. The divergence operator satisfies $\mathrm{div}(\Psi) = \sum_{i=1}^d \frac{\partial \Psi_i}{\partial x_i}$ for a $d$-dimensional field $\Psi$ with components $\Psi_1, \ldots, \Psi_d$.

The first equation represents the conservation of momentum, while the second one corresponds to mass conservation. More precisely, we assume that the density of the fluid does not depend on $t$ and $x$, and is normalized to 1. Then the equation $\mathrm{div}u = 0$ expresses the incompressibility of the fluid.

The term $f$ represents the effect of an external force applied to the fluid.

Due to the Eulerian representation, the acceleration term in the momentum equation is $\frac{\partial u}{\partial t} + (u.\nabla)u$ and contains a nonlinear term due to convection effects in the fluid. The presence of this nonlinearity represents an important aspect for the mathematical study of fluids. If we suppress the convection term $(u.\nabla)u$ in (5.5), we obtain the Stokes problem for incompressible fluids; its main advantage is linearity.

The fluid is assumed to be viscous: due to internal friction forces, we obtain the term $\nu\Delta u$ in (5.5), with $\nu > 0$. A more appropriate parameter is the Reynolds number $Re = \frac{1}{\nu}$, which gives a comparison of the effects of diffusion and of advection on the fluid. When advection dominates diffusion, $Re$ is high, while small values of $Re$ correspond to strong diffusion with respect to advection.

We need conditions on the velocity on the boundary $\partial D$ of the domain. For example, in the case of Dirichlet boundary conditions the value of $u$ on $\partial D$ is imposed at each time. In the sequel, we also consider examples with periodic boundary conditions, in order to check the accuracy of the method.

Finally, the system (5.5) contains an additional unknown $p$ called the pressure field. The opposite of the gradient of $p$ acts as a force on the fluid. This extra necessary term is interpreted a consequence of the incompressibility constraint $\mathrm{div}(u) = 0$.

The difficulties in the mathematical analysis of (5.5) rely on two elements: the nonlinearity and the presence of the pressure due to the incompressibility constraint. In dimension $d = 2$, existence and uniqueness results have been obtained, while in dimension $d = 3$ the theory has not been successfully constructed so far: for instance uniqueness, or global existence are still open problems. Another challenging theoretical and numerical issue in both dimensions is turbulence, which appears for high Reynolds numbers. We do not consider such difficult questions in the sequel.

5.5.2. **Projection methods and hybrid semi-lagrangian framework.** In this section, we present the projection methods used for numerical computations of solutions of Navier-Stokes and Stokes equations with the incompressibility constraint, and we explain how the hybrid schemes can be useful for their practical realizations.

First, we remark a link between Navier-Stokes and Burgers equations: the momentum equations has the same form. Indeed, if we forget the incompressibility constraint, the pressure disappears and we recover a Burgers equation, for which we have developed an hybrid scheme. When we then take into account the incompressibility of the fluid, we get an extra pressure term; the problems of evolution and of preserving the incompressibility may thus be treated separately. In the case of Stokes equations, the evolution is even simpler, since it is governed by a heat equation.

Even if the previous idea cannot be applied easily and directly, this methodology has been adapted to the case of incompressible Navier-Stokes equations, in order to derive the so-called Projection Methods. For example, if we discretize time, for each step we deal with two problems: in a first sub-step, the fluid evolves according to the momentum equation without any incompressibility constraint; in the second sub-step, we correct the velocity thanks to the computation of a pressure, so that we obtain a divergence-free velocity field.

Mathematically, in this way the velocity calculated by the first sub-step is projected onto the space of divergence free fields thanks to the second sub-step, according to the following orthogonal decomposition of Hilbert spaces - called the Helmholtz decomposition, see for instance [63].

**Property 5.1** (Helmholtz decomposition)**.** *If $D$ is a Lipschitz bounded open set in $\mathbb{R}^d$, we have the following orthogonal decomposition*

$$(L^2(D))^d = H \oplus H^\perp,$$

*with $H = \left\{u \in (L^2(D))^d; div(u) = 0, u.n_{|\partial D} = 0\right\}$ and $H^\perp = \left\{u \in (L^2(D))^d; u = \nabla p, p \in H^1(D)\right\}$. Moreover, $H$ is the closure of $\left\{u \in (\mathcal{C}_c^\infty(D))^d; div(u) = 0\right\}$.*

In the definition of $H$, the divergence is understood in the distributional sense, and the condition at the boundary needs a precise definition of a trace operator for the normal component of $L^2$ functions on the boundary; $\mathcal{C}_c^\infty(D)$ is the set of infinitely differentiable functions on $D$ with compact support.

For special projection methods, an hybrid scheme can be used for each sub-step: the first sub-step computes the solution of a Burgers equation, with Dirichlet boundary conditions, while the pressure is computed as solution of a Poisson equation, seen as a stationary situation of a parabolic equation. However

the question of boundary conditions for the second sub-step is difficult: while the true pressure field satisfies no constraint on the boundary, the calculation of the pressure in the projection method is based on a Poisson equation which requires boundary conditions.

For an overview of the projection methods, we refer to [33].

5.5.3. **Description of the algorithm.** Time and space are discretized, and the aim is to build approximations of the solution at the grid points $x_j$ for any discrete time $n\delta t$. The initial function $u_0$ is given. The approximation is not direct: first, the projection method is built at the continuous space-time level. Then, each equation is discretized with the hybrid semi-lagrangian framework.

We build functions $u^n$ defined on $\overline{D}$ to be approximations of $u(n\delta t, .)$. If we assume that $u^n$ is known, the construction of $u^{n+1}$ is as follows.

First, we solve on $[n\delta t, (n+1)\delta t] \times D$ the partial differential equation with unknown $v^{n+1}$

$$\frac{\partial v^{n+1}}{\partial t} + (u^n.\nabla)v^{n+1} - \nu\Delta v^{n+1} = f(n\delta t, .), \tag{5.6}$$

with the initial condition $v^{n+1}(n\delta t, .) = u^n$, and the appropriate Dirichlet boundary conditions associated with (5.5). Notice that this equation is linear and parabolic.

On this equation, the force is frozen at its value $f(n\delta t, .)$ at time $n\delta t$, as well as the advection velocity $u^n$, like for the discretization of the Burgers equation - see (5.4). The pressure field does not explicitly appears in this evolution sub-step.

To recover the incompressibility property $\mathrm{div}(u^{n+1}) = 0$, according to the Helmholtz decomposition we introduce a correcting pressure $p^{n+1}$ such that

$$u^{n+1} = v^{n+1} - \nabla p^{n+1}.$$

Taking the divergence of this equation, we see that we need to solve a Poisson equation:

$$\Delta p^{n+1} = \mathrm{div}(v^{n+1}). \tag{5.7}$$

We warn the reader that the question of the boundary condition in such a method is very difficult - see for instance [33]. Here we choose the simplest projection scheme, with the simplest choice of boundary condition, of Neumann type:

$$\nabla p^{n+1}.\mathrm{n}|_{\partial D} = 0,$$

where the vector n is the outward normal vector.

Indeed a Dirichlet boundary condition for $u^{n+1}$ and $v^{n+1}$ naturally corresponds with a condition on both the normal and the tangential components of the gradient of $p^{n+1}$ on the boundary $\partial D$, while well-posedness requires that only the normal derivative should be prescribed. We remark that the solution $p^{n+1}$ is only defined up to a constant, but the knowledge of the gradient is sufficient. However, with such a method we can not avoid boundary layers, since we a priori only get $u^{n+1}.\mathrm{n}|_{\partial D} = 0$ instead of $u^{n+1}|_{\partial D} = 0$: for the next sub-step the initial condition may not satisfy the boundary conditions!

In a periodic case, the boundary conditions on (5.7) are also periodic.

For physical interpretations, it is important to remark that only $\delta t p^{n+1}$ can be considered as a pressure.

In the context of this work, the solution of (5.7) is not computed exactly; however it is approximated thanks to an evolution equation, for which the Poisson equation above defines stationary solutions:

$$\frac{\partial \tilde{p}^{n+1}}{\partial s} = \mu\Delta\tilde{p}^{n+1} - \mu\mathrm{div}(v^{n+1}), \tag{5.8}$$

with the homogeneous Neumann boundary conditions. The initial condition is $\tilde{p}^{n+1}(s = 0, .) = p^n$. The parameter $\mu > 0$ plays the role of a viscosity and also acts on the divergence term.

When time $s$ goes to infinity, $\tilde{p}^{n+1}(s, .)$ converges to $p^{n+1}$. In the numerical method, we choose a final time $s_\infty$, and we define $p^{n+1} = \tilde{p}^{n+1}(s_\infty, .)$.

The choice of the parameters $s_\infty$ and $\mu$ has an effect on the convergence of the scheme: the equilibrium is obtained in theory for $s_\infty = +\infty$, but a small value is more convenient. Moreover we can hope that the choice of the initial condition at each time step can accelerate convergence - instead of starting from 0 time and again. Higher values of $\mu$ correspond to a better convergence to equilibrium, but we need to remember that a priori the Monte-Carlo error appearing for probabilistic methods grows like $\sqrt{\mu}$.

It has been explained in other sections how hybrid semi-lagrangian schemes can be used to compute approximations of solutions of (5.6) and (5.8): the first equation is exactly the same intermediate equation used for the Burgers equation, and the second equation is a simple parabolic equation.

In the case of Stokes equations, the construction of the scheme is essentially the same - it is even simpler since we avoid the quadratic linearity, while the projection step remains the same.

5.5.4. **Choice of the parameters and related questions.** Both the evolution and the projection steps are based on probabilistic representation formulae, wherein an expectation must be approximated thanks to a Monte-Carlo method. The corresponding error is therefore of size $\frac{1}{\sqrt{N}}$ where $N$ is the number of independent realizations of the random variables.

We have explained in a simple case that this Monte-Carlo error also depends on the time and space discretization steps. As a consequence, a relatively small number of realizations $N$ could be chosen if $\delta x$ and $\delta t$ are also small enough. However, the solution is more irregular when $\delta x$ decreases. The balance between these two phenomena is not easy to find. Moreover, we need to compute the divergence of the vector field to compute the pressure correction, so regularity is essential: we should avoid strong oscillations between neighbor points.

It should be noticed that in principle the space grids used for the computation of the velocity and of the pressure are different. For instance, the grid for the velocity can be refined, or a staggered grid can be used. For simplicity, here we consider that the grids are the same.

The numerical computation of the divergence also appears as a technical problem, especially for the points at the boundary. Indeed we can check that even if the velocity field is exactly divergence free, the divergence function in Matlab produces some numerical unexpected error at the boundary. We could enforce a boundary condition on the divergence, whether of Dirichlet or Neumann type, but it then would not be compatible with the equation. At this stage, some boundary layers seem to appear, but it is not clear whether in this specific case they are only due to the realization of the scheme, or if they are intrinsic to the method.

We should also consider another problem linked to the computation of the divergence of the velocity field, used for the computation of the pressure in the projection step: if we are not careful we break the structure of the incompressible Navier-Stokes or Stokes equation. The divergence operator - applied to the pressure - and the gradient operator - applied on the velocity - as they appear both in the PDE problem and in the Helmholtz decomposition are linked by the following property: the divergence is the adjoint of the gradient. Such a property must be recovered when one wishes to define an efficient numerical method. Moreover, a so-called inf-sup condition is required for well-posedness of the continuous and discrete problems. In our situation, we should find a smart definition of a gradient and of a divergence operator related to the hybrid formulation of the numerical scheme, in order to analyze whether the method is efficient.

The Monte-Carlo error also depends on the viscosity parameters $\nu$ and $\mu$: when these quantities are small, a smaller $N$ can be used, and the method becomes faster. Unlike the true viscosity parameter $\nu$, $\mu$ is an artificial, non-physical quantity, that we can choose for the scheme. However, the greater it is, the faster the convergence to a pressure equilibrium should be. Once again, we face a balance between high and small $\mu$, and we need to decide which source of error is the worst: Monte-Carlo, or non-stationarity.

We can also have an action on $s_\infty$, or equivalently on the number of time steps necessary in the projection part. Indeed, we can hope that it is not too high, since at each step we are closer to equilibrium: if the change in the divergence term is small, the corresponding equilibria are also close, and maybe a few steps are enough form a former stationary pressure to obtain a new stationary one for the modified divergence. Indeed, the divergence seems to go to zero during the realization of the scheme.

5.5.5. **The driven cavity flow.** We focus on a famous test problem for Navier-Stokes equations: the driven cavity model. We present simulations for three different values of the Reynolds number. We recover the same kind of behavior as obtained by deterministic methods - see for instance [59].

We consider an incompressible fluid in a 2-dimensional square domain called the cavity - which is $D = (-1, 1)^2$ in the numerical simulation below - and we assume non-trivial Dirichlet boundary conditions: if we denote $u_1$ and $u_2$ the two components of the velocity field $u$, we assume that $u_2$ is null on the whole boundary $\partial D$, and that $u_1$ is equal to a non-zero constant speed on the upper part of the domain - corresponding to the set $x_2 = 1$ - and to be null on the rest of the boundary. This constant speed can be set to be equal to 1 without loss of generality. This kind of boundary conditions leads to singularities on the velocity at the

corners of the domain; instead we consider a regularized cavity, where on the upper part of the boundary we set $u_1(t, x_1, 1) = (1 - x_1)^2(1 + x_1)^2$.

Even if the Dirichlet conditions are not homogeneous, the principle of the numerical approximation remains the same: we do not change the pressure boundary condition, and the new velocity boundary values are only used for the evolution part of the scheme.

When time increases, vortices appear, in a way depending on the value of the Reynolds number. The first one appears on the top right of the cavity and moves towards the center when the Reynolds number increases. The second vortex appears on the bottom right corner, and a third one appears on the bottom left corner; these two secondary vortices also move from the corners to the interior of the cavity when the Reynolds number increases.

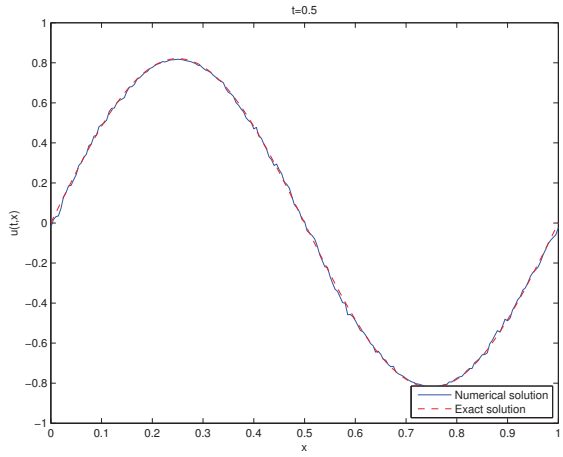
Two kinds of initial conditions can be considered, and both of them are not divergence-free. We can always choose $u_2(0, x) = 0$ for any $x \in \overline{D}$, and the first component $u_1$ must satisfy the boundary conditions. A first example is $u_1(0, x) = 0$, except on the upper boundary where $u_1(0, x1, 1) = (1 - x_1)^2(1 + x_1)^2$: then the initial condition is not smooth. Another example is $u_1(0, x) = \frac{(x_2+1)}{2}(1 - x_1)^2(1 + x_1)^2$, which provides more regularity at time $t = 0$. A correction must be added to obtain a divergence-free field thanks to the projection step of the scheme.

In Section 5.6, we give the streamlines of the flow obtained for large values of $t$. The time step and the mesh size satisfy $\delta t = \delta x = 0.01$, and $N = 100$ Monte-Carlo simulations are used. We have taken $\mu = 0.1$ and $s_\infty = 1$ - with an additional test on the relative error on the pressure between two steps to avoid unnecessary computations when it is quasi-stationary.

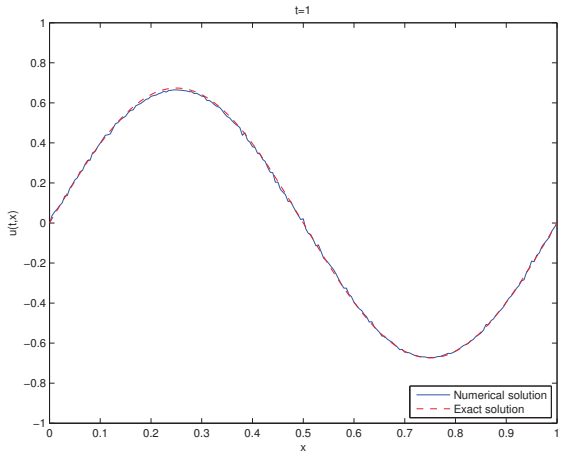
Three different values of the Reynolds number are used: $Re = 1000$ in Figure 25, $Re = 5000$ in Figure 26, $Re = 10000$ in Figure 27.

Thanks to a comparison with the simulations of [59], we observe the good behavior on the apparition of the vortices when the Reynolds number increases.

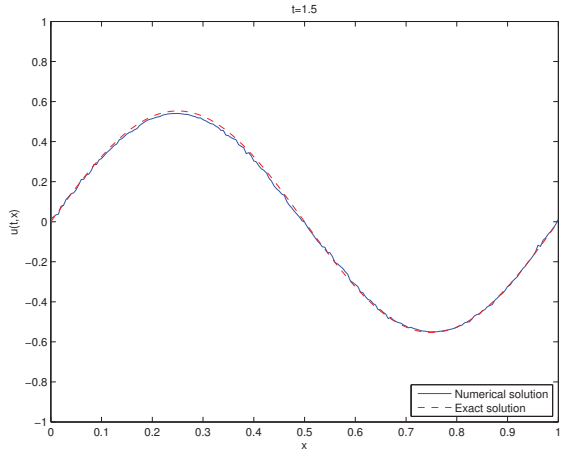
### 5.6. Numerical simulations



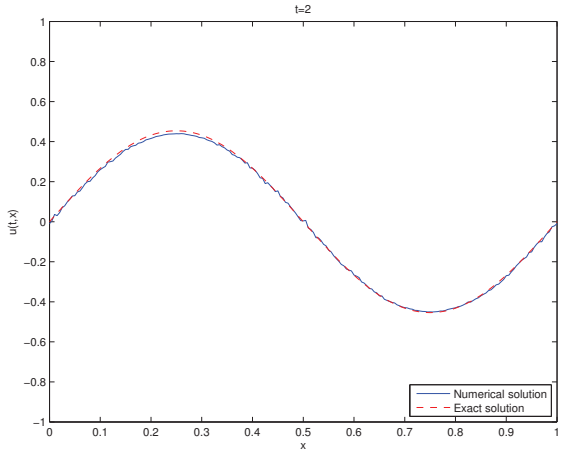

(a) $t = 0.5$




(b) $t = 1$

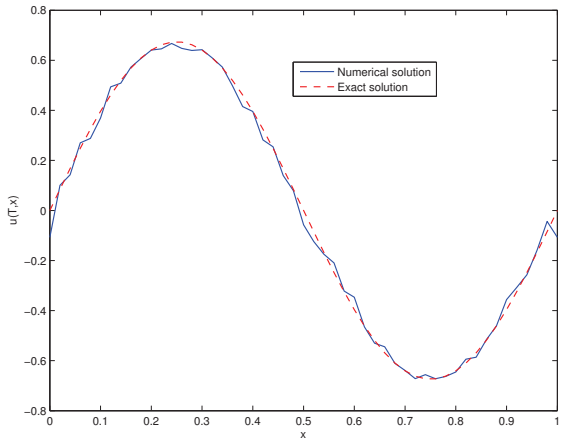



(c) $t = 1.5$

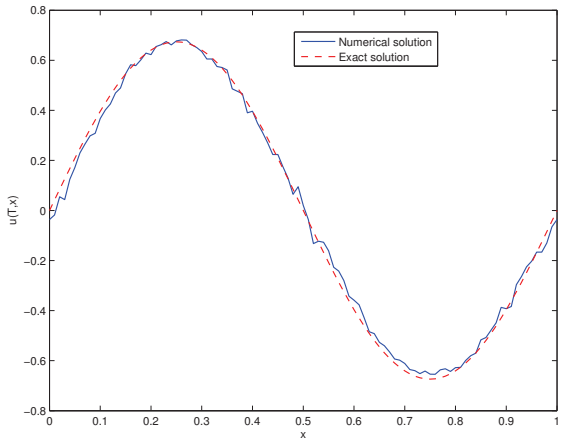



(d) $t = 2$

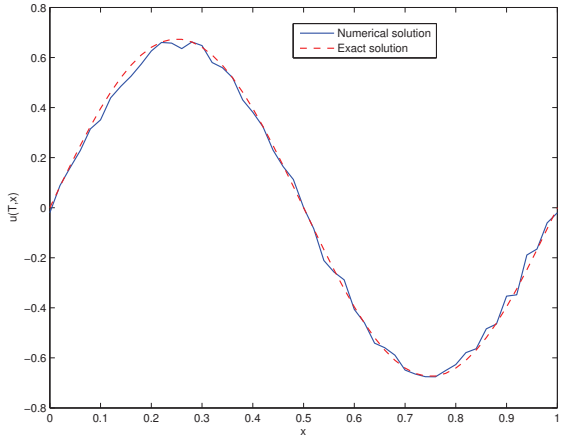
FIGURE 1. Simulations for (5.1) at different times, with $\delta t = 0.01$, $\delta x = 0.005$, $N = 100$, $\nu = 0.01$.



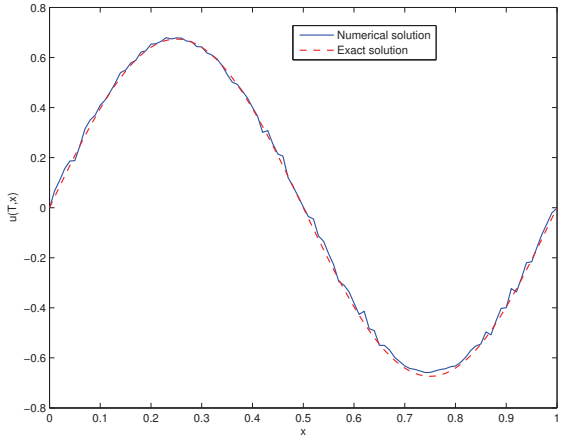

(a) $\delta t = \delta x = 0.02$, $N = 50$




(b) $\delta t = \delta x = 0.01$, $N = 50$

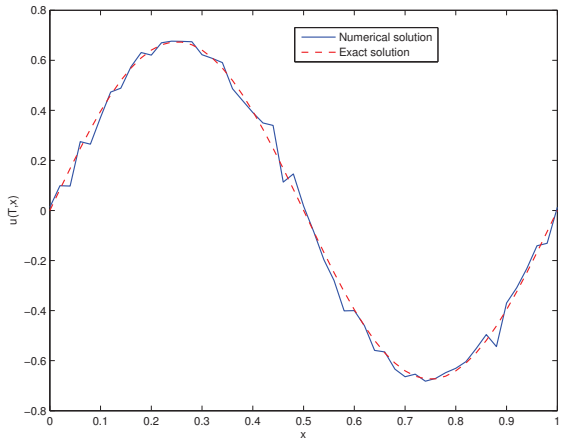



(c) $\delta t = \delta x = 0.02$, $N = 100$

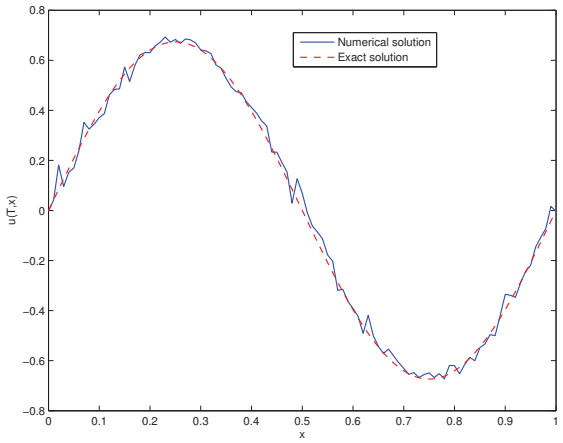



(d) $\delta t = \delta x = 0.01$, $N = 100$

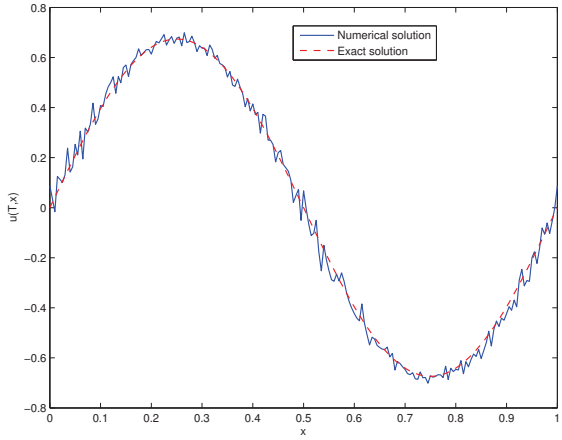
FIGURE 2. Simulations for (5.1) with different values of the discretization parameters



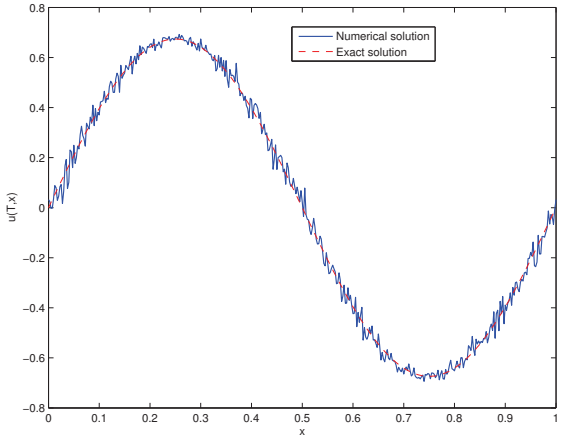

(a) $\delta x = 0.2$




(b) $\delta x = 0.1$

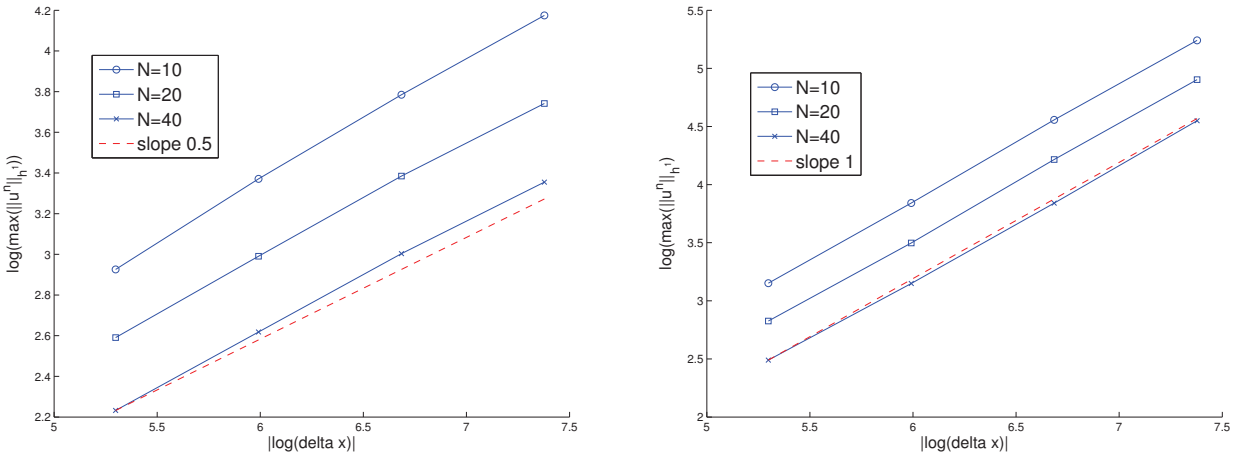



(c) $\delta x = 0.05$




(d) $\delta x = 0.025$

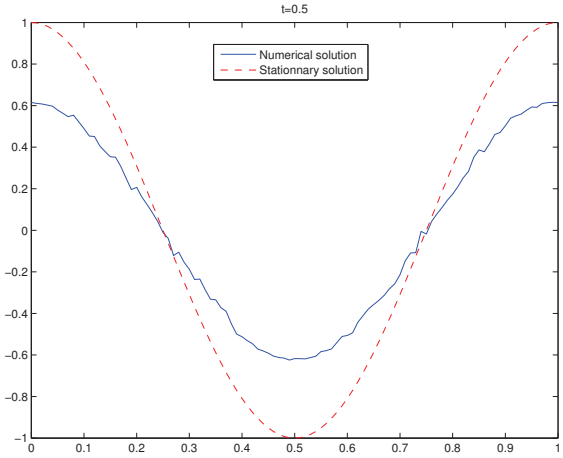
FIGURE 3. Simulations for (5.1) with different values of the mesh size, with $\delta t = 0.02$ and $N = 50$.



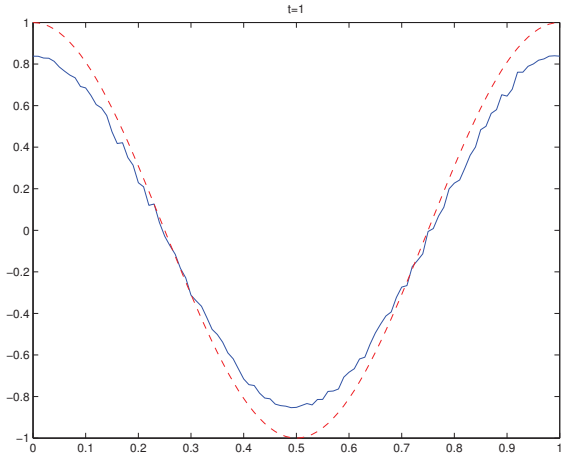

(a) When $\delta t = \delta x$

(b) When $\delta t$ is fixed

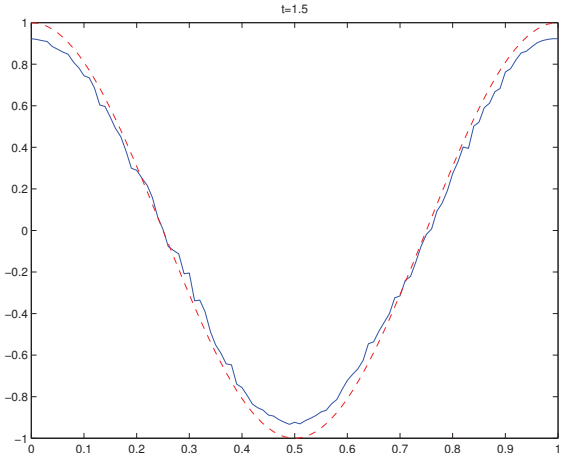
FIGURE 4. Behavior of the $h^1$ semi-norm with respect to the parameters, in logarithmic scales.



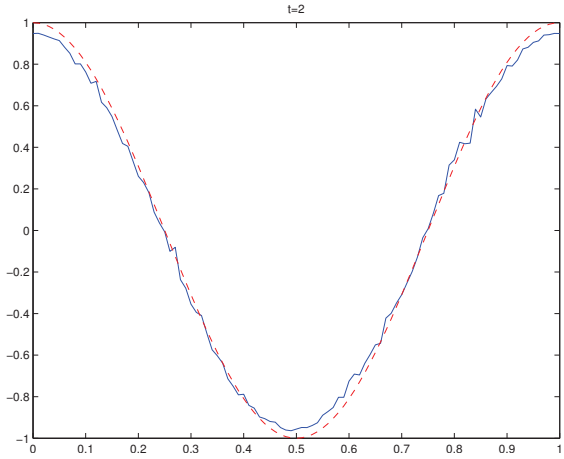

(a) $t = 0.5$



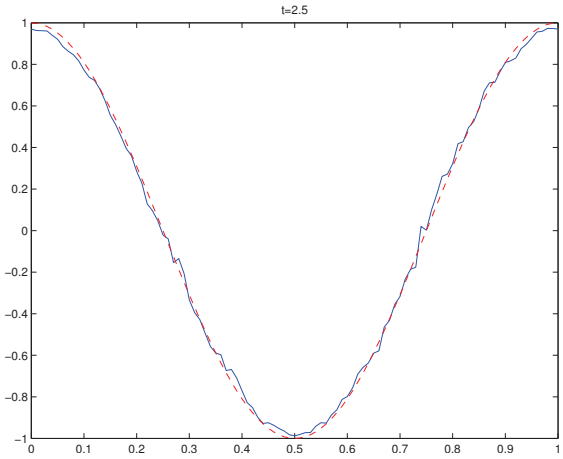

(b) $t = 1$



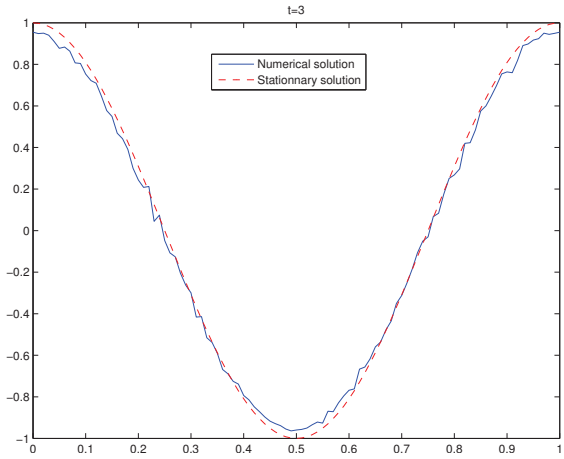

(c) $t = 1.5$




(d) $t = 2$




(e) $t = 2.5$




(f) $t = 3$

FIGURE 5. Solution at different times for (5.2) when $\nu = 0.05$, with $\delta t = \delta x = 0.01, N = 50$.

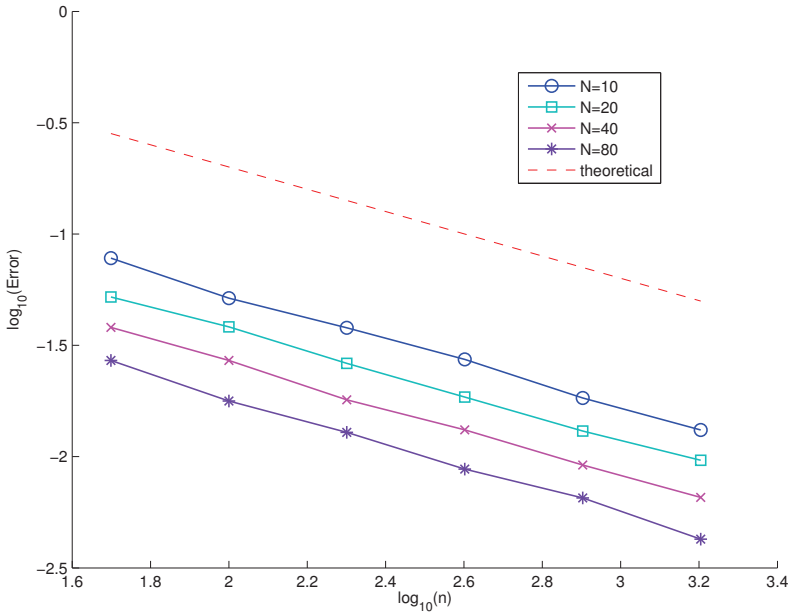



FIGURE 6. Error for periodic boundary conditions when $\delta t = \delta x = 1/n$, in logarithmic scales.

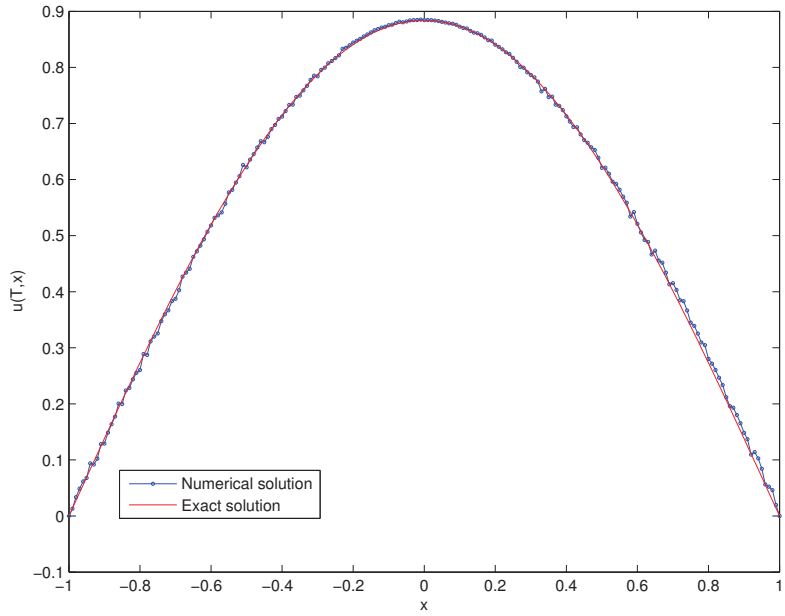
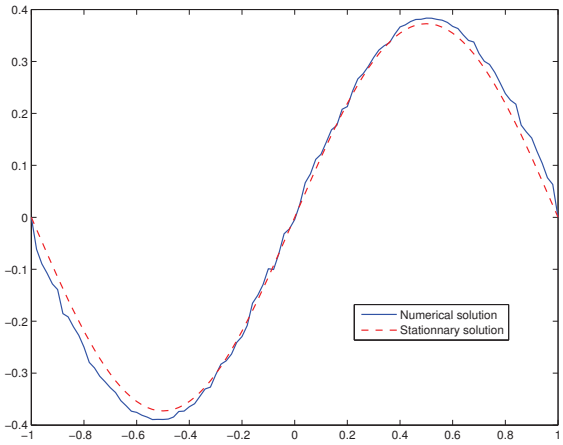
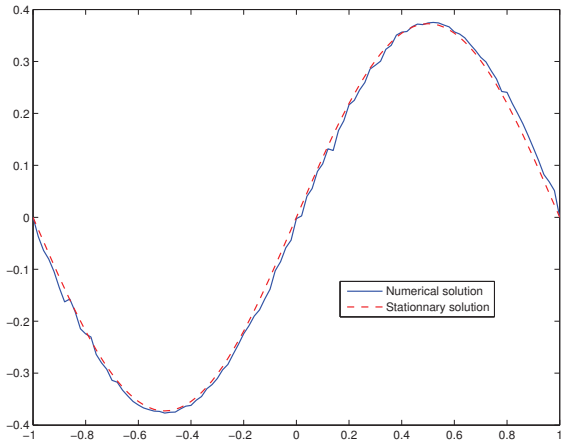


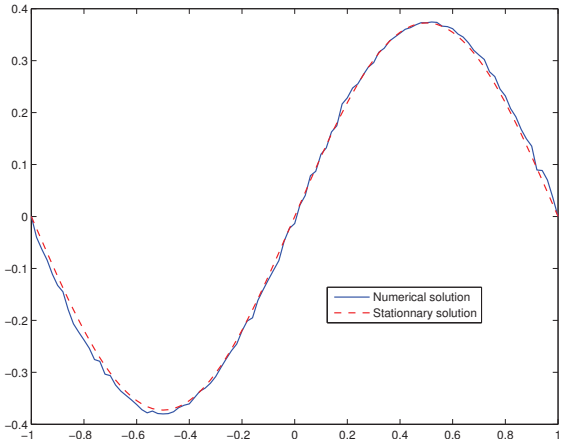

FIGURE 7. Solution at time $T = 1$ for Dirichlet boundary conditions, with $\delta t = 0.01, \delta x = 0.01, \nu = 0.05, N_i = N_b = 50$.



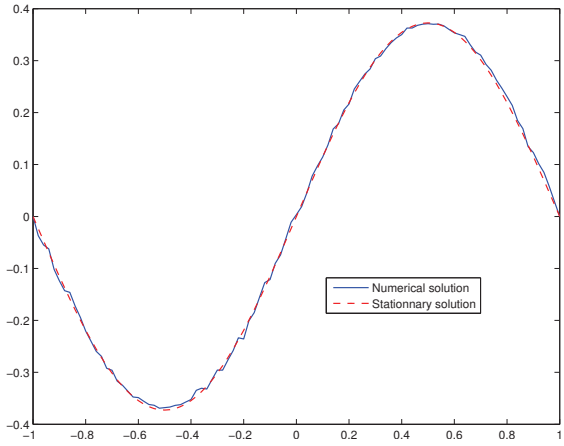

(a) $SUB = 1$




(b) $SUB = 5$

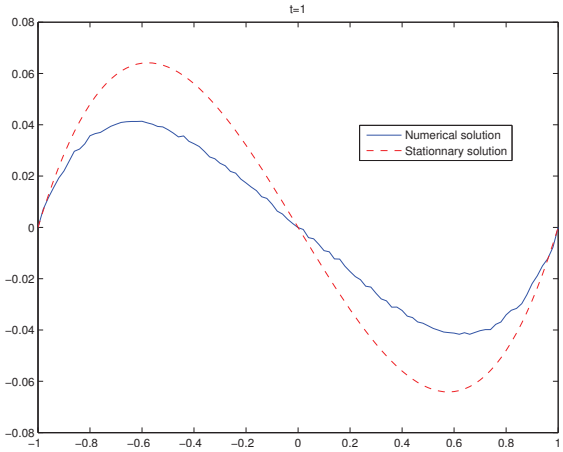



(c) $SUB = 10$

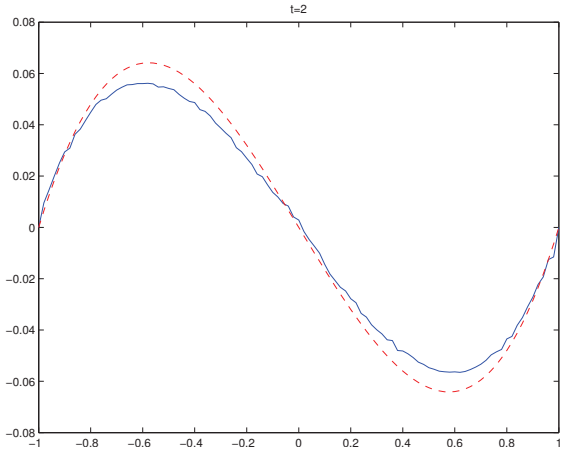



(d) $SUB = 20$

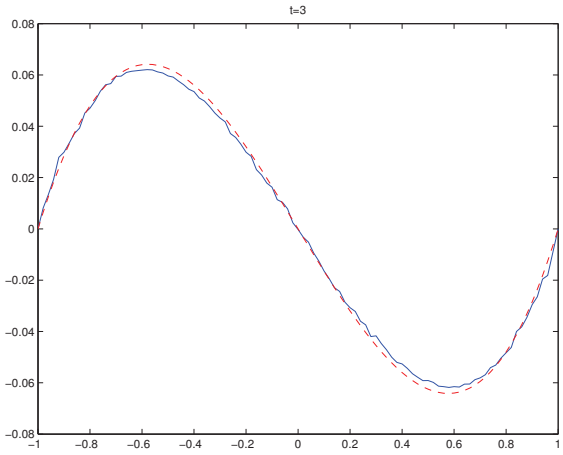
FIGURE 8. Solution with Dirichlet boundary conditions with different values of the subdivision parameter $SUB$.



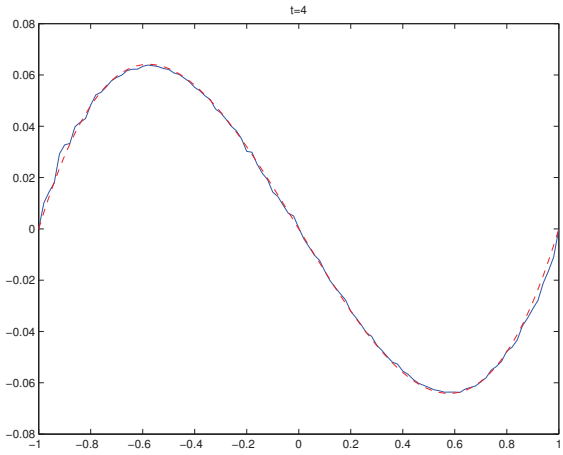

(a) $t = 1$



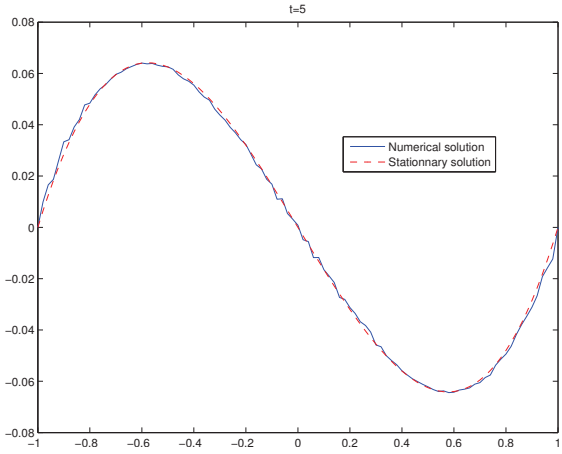

(b) $t = 2$




(c) $t = 3$




(d) $t = 4$




(e) $t = 5$

FIGURE 9. Solution at different times with Dirichlet boundary conditions when $\nu = 0.1$, with $\delta t = \delta x = 0.02, N = 100$.

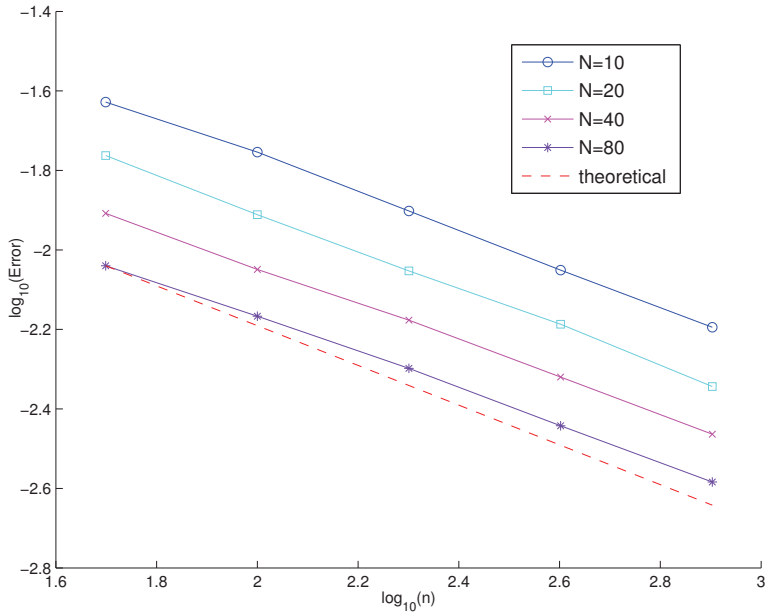



FIGURE 10. Error for Dirichlet boundary conditions when $\delta t = \delta x = 1/n$, in logarithmic scales.

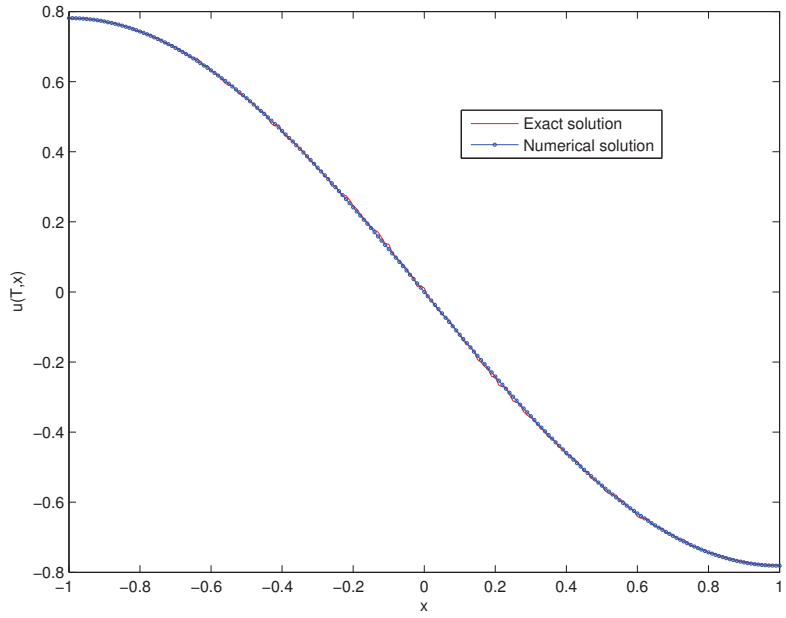
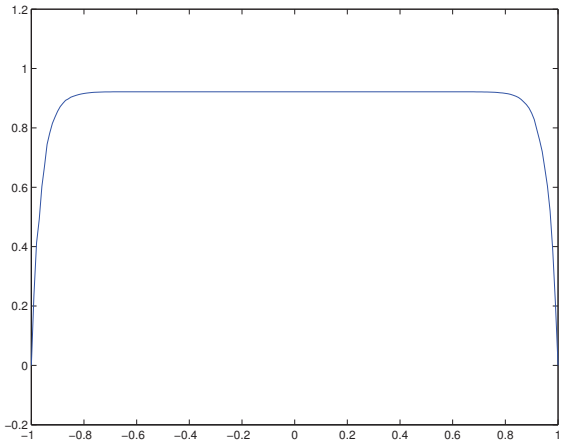
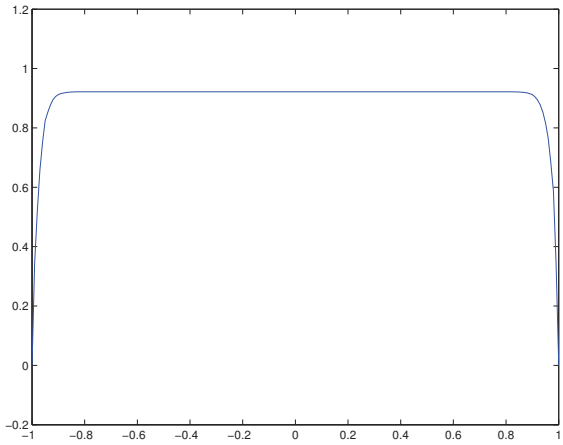


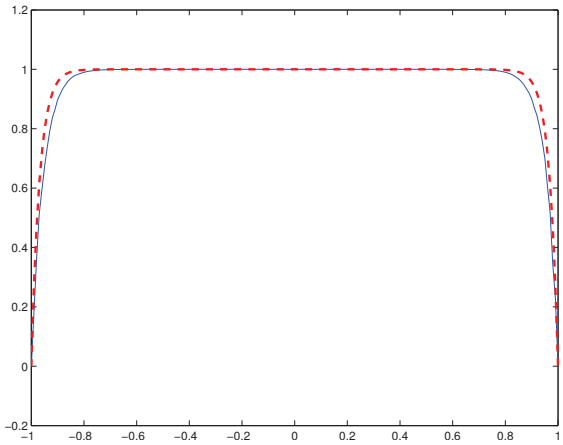

FIGURE 11. Solution at time $T = 1$ for Neumann boundary conditions, with $\nu = 0.1, \delta t = \delta x = 0.01, N_i = N_b = 100$.



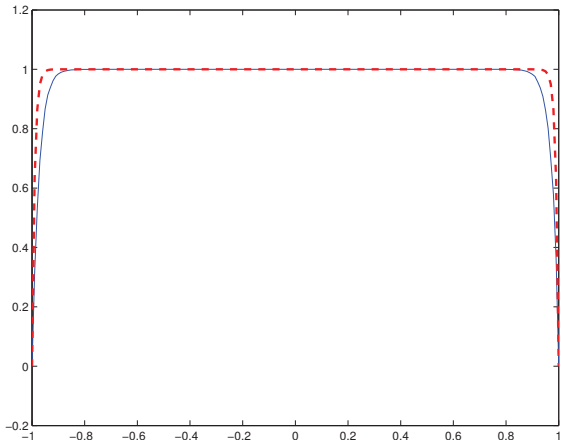

(a) $\epsilon = 0.001, \alpha = -0.1$



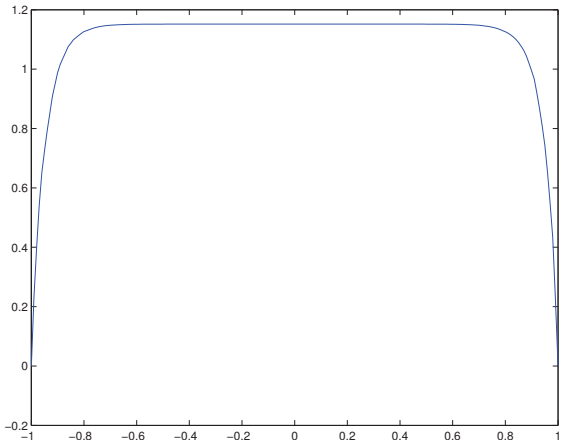

(b) $\epsilon = 0.0001, \alpha = -0.1$



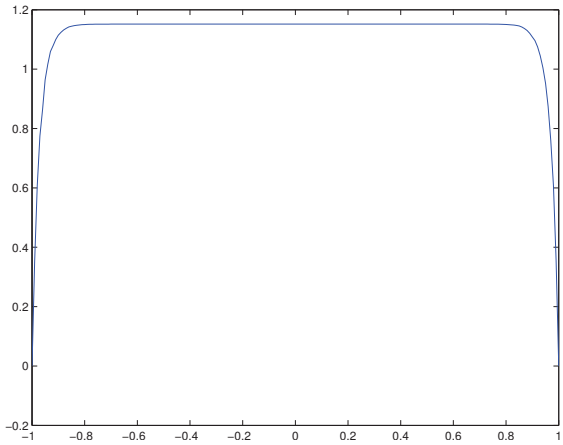

(c) $\epsilon = 0.001, \alpha = 0$




(d) $\epsilon = 0.0001, \alpha = 0$

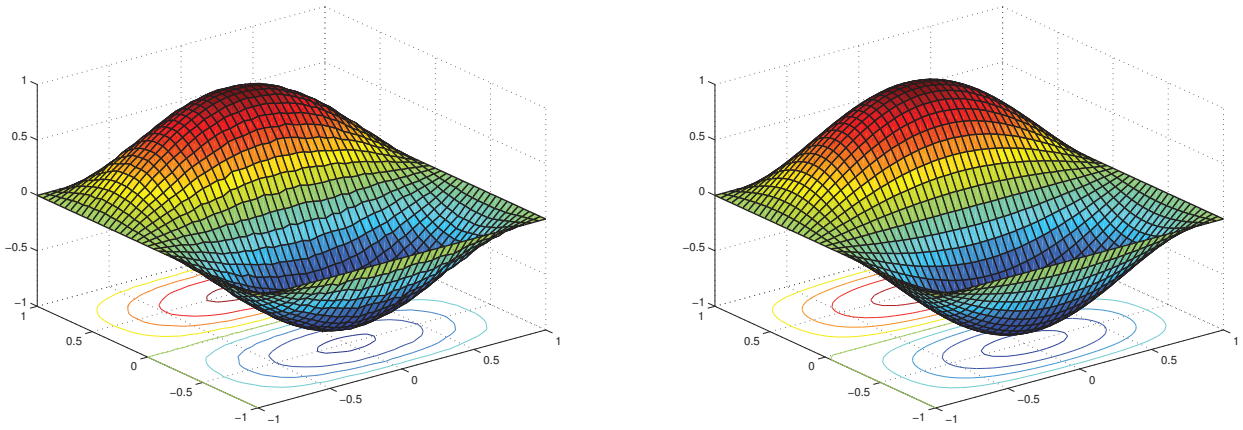



(e) $\epsilon = 0.001, \alpha = 0.1$




(f) $\epsilon = 0.0001, \alpha = 0.1$

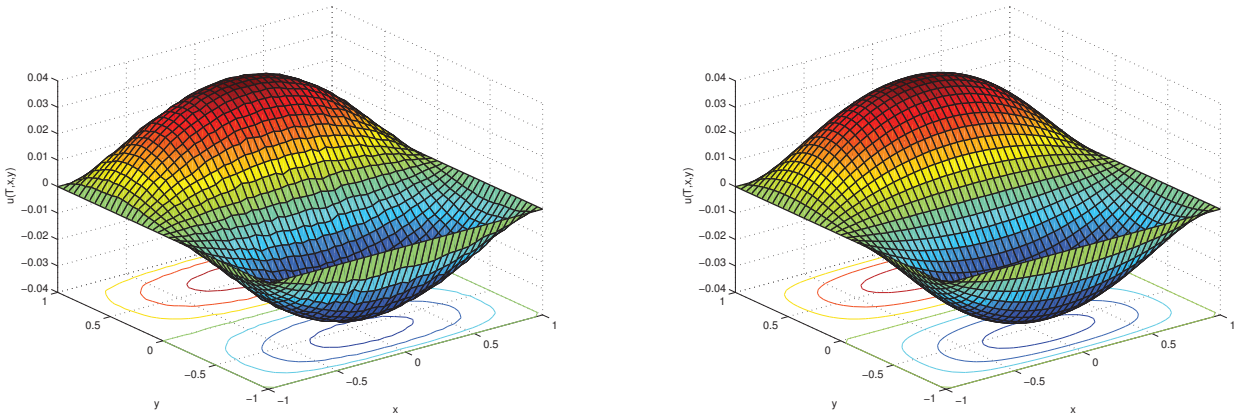
FIGURE 12. Observation of a boundary layer in dimension 1.




(a) numerical approximation

(b) exact solution

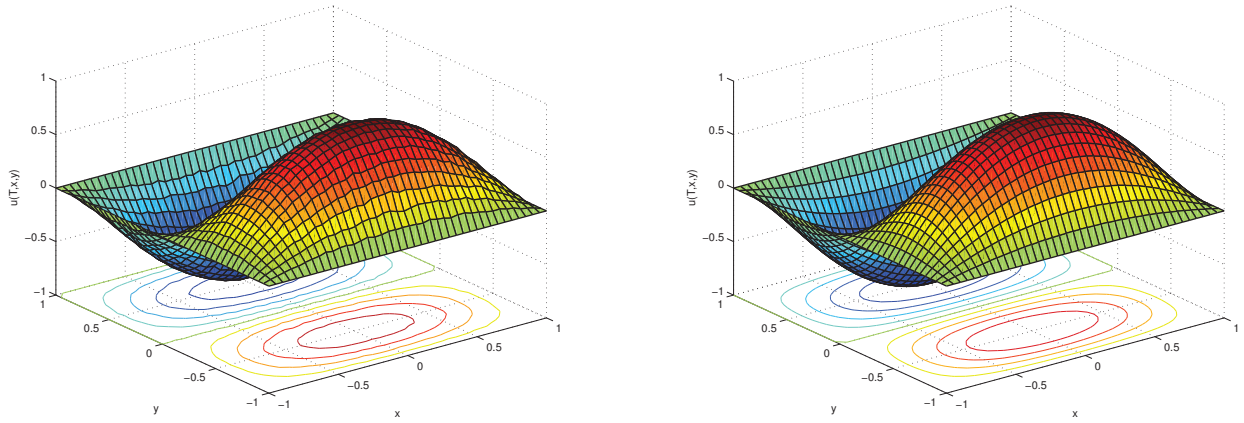
FIGURE 13. Solution at time $T = 0.1$ with $u_0(x, y) = \sin(\pi \frac{x+1}{2}) \sin(\pi y)$ (Dirichlet boundary conditions), with $\nu = 0.01, \delta t = \delta x = 0.05, N_i = N_b = 100$.




(a) numerical approximation

(b) exact solution

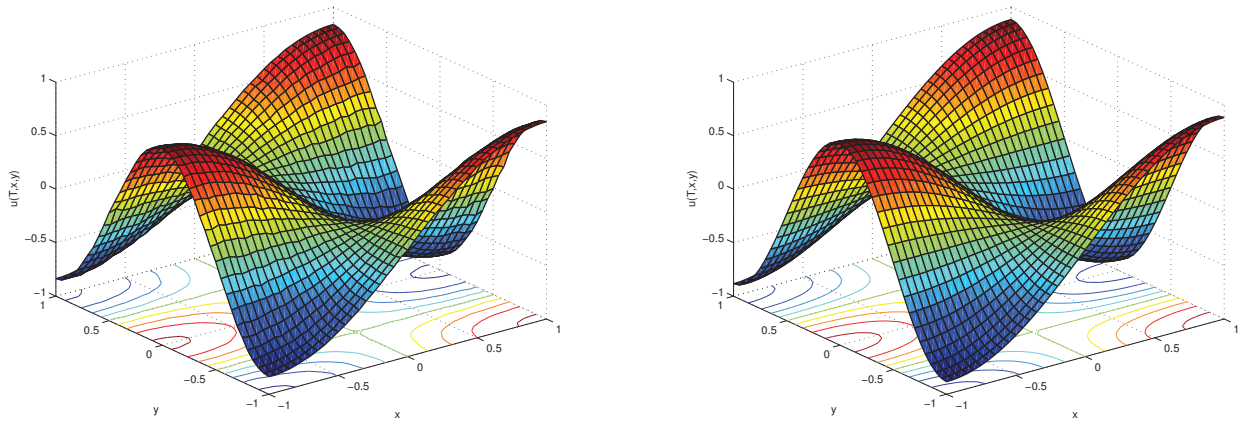
FIGURE 14. Stationary solution: first example (Dirichlet boundary conditions), with $\nu = \mu = 0.1, \delta t = \delta x = 0.05, N_i = N_b = 500$.




(a) numerical approximation

(b) exact solution

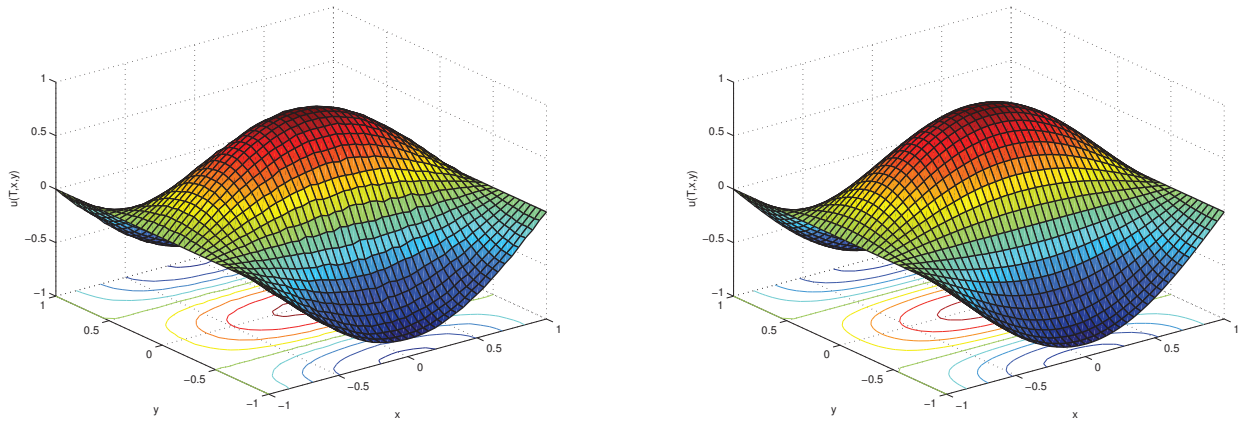
FIGURE 15. Stationary solution: second example (Dirichlet boundary conditions), with $\nu = \mu = 0.1, \delta t = \delta x = 0.05, N_i = N_b = 500$.




(a) numerical approximation

(b) exact solution

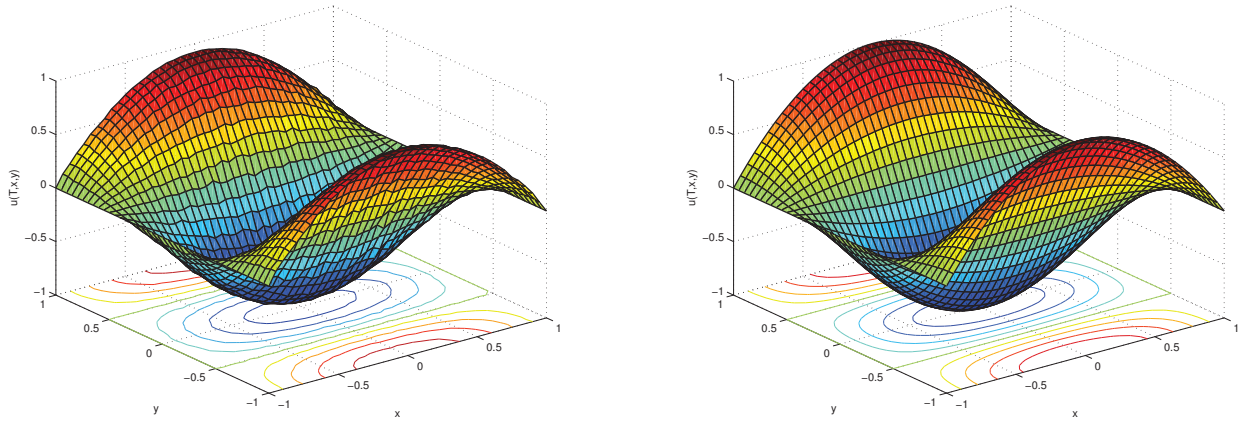
FIGURE 16. Solution at time $T = 1$ with $u_0(x, y) = \cos(\pi \frac{x+1}{2}) \cos(\pi y)$ (Neumann boundary conditions), with $\nu = 0.01, \delta t = \delta x = 0.05, N_i = N_b = 100$.




(a) numerical approximation

(b) exact solution

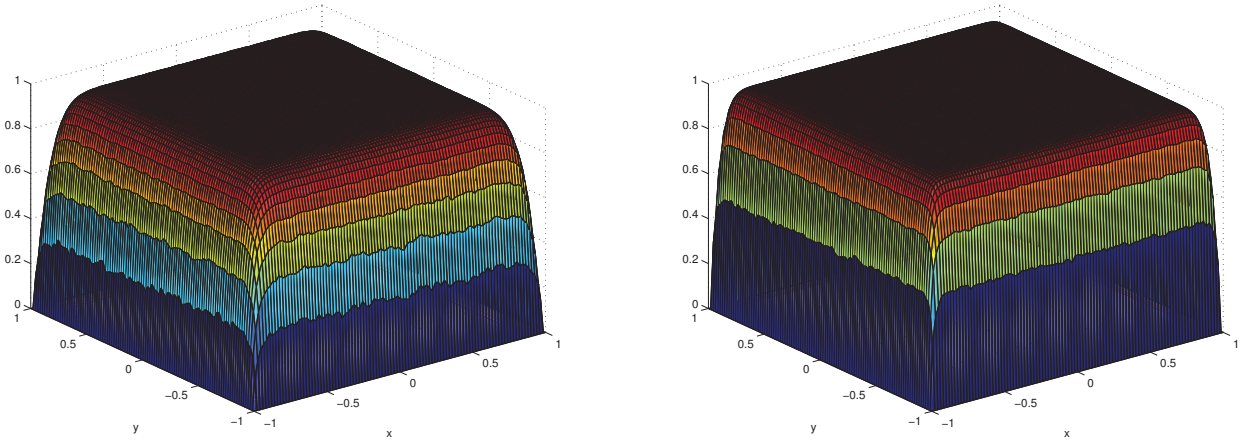
FIGURE 17. Solution at time $T = 1$ with $u_0(x, y) = \sin(\pi \frac{x+1}{2}) \cos(\pi y)$ (Mixed boundary conditions), with $\nu = 0.01, \delta t = \delta x = 0.05, N_i = N_b = 100$.




(a) numerical approximation

(b) exact solution

FIGURE 18. Stationary solution (Mixed boundary conditions), with $\nu = \mu = 0.1, \delta t = \delta x = 0.05, N_i = N_b = 500$.




(a) $\epsilon = 0.001$ (b) $\epsilon = 0.0001$

FIGURE 19. Observation of a boundary layer in dimension 2.

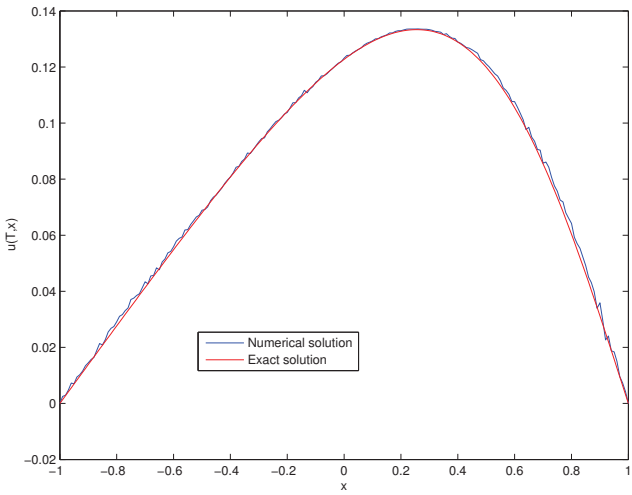
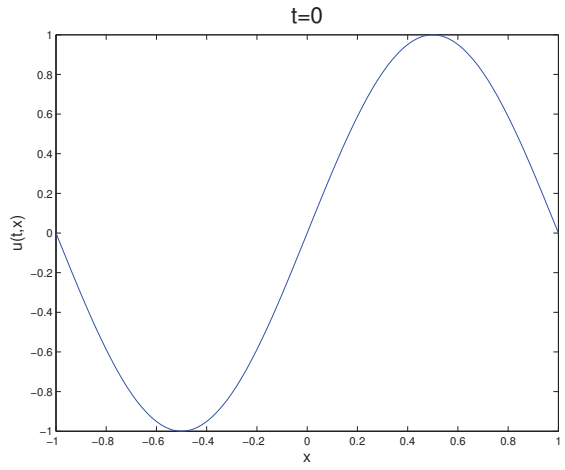
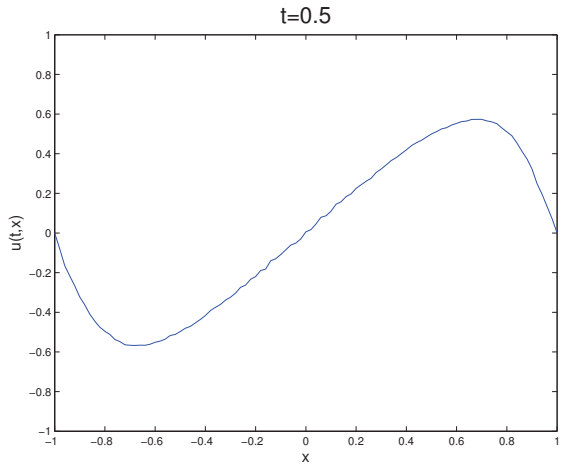


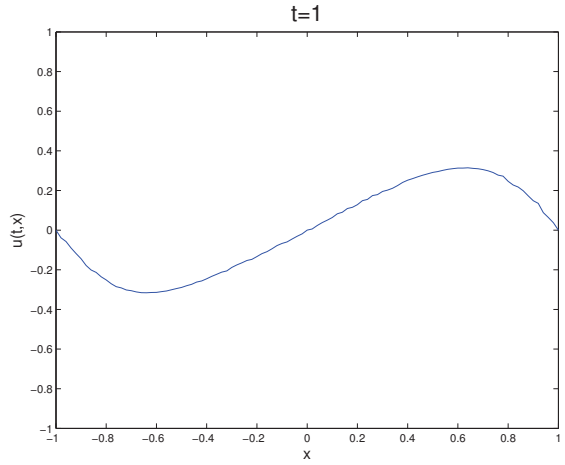

FIGURE 20. Hopf-Cole solution for the Burgers equation at time $T = 1$, with $\nu = \mu = 0.1, \delta t = \delta x = 0.05, N_i = N_b = 500$.



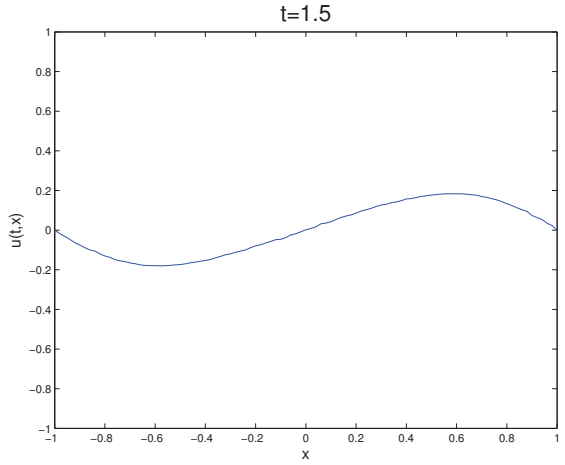

(a) $t = 0$



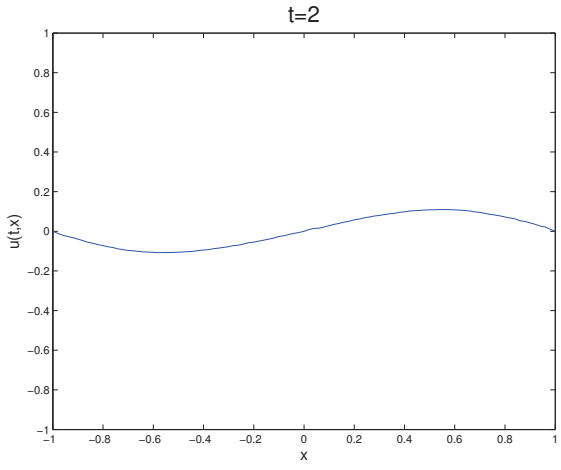

(b) $t = 0.5$




(c) $t = 1$

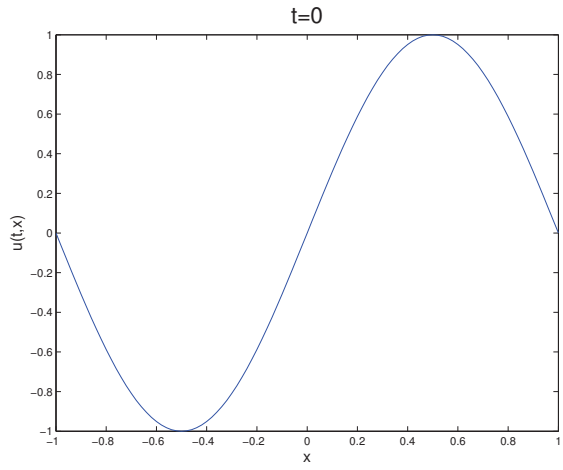



(d) $t = 1.5$

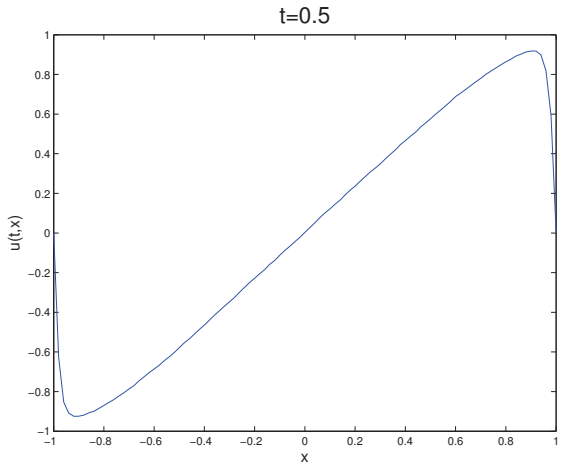



(e) $t = 2$

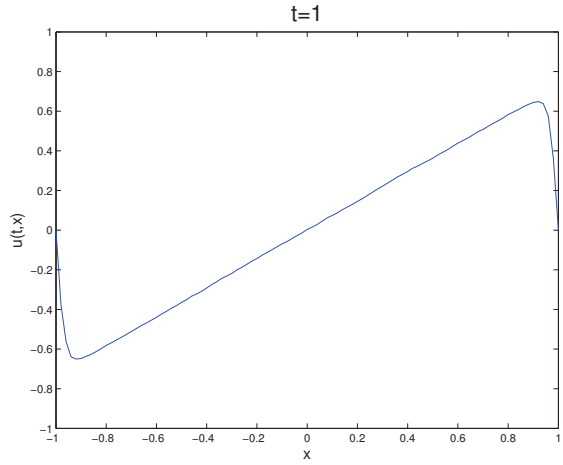
FIGURE 21. Solution at different times for $u_0^+$ when $\nu = 0.1$, with $\delta t = \delta x = 0.02, N_i = N_b = 100$.



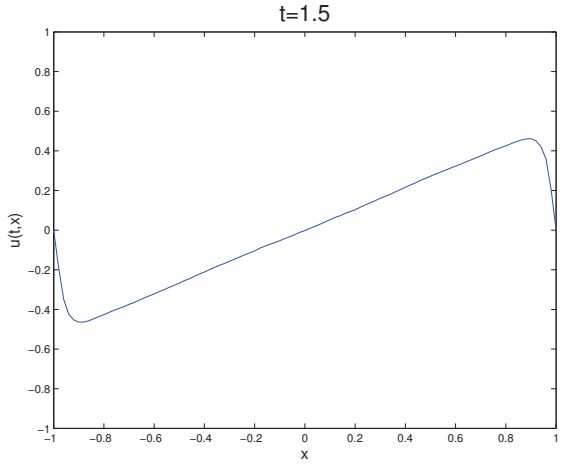

(a) $t = 0$



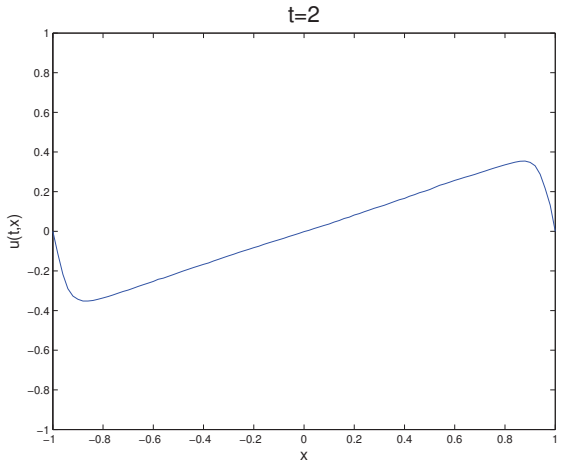

(b) $t = 0.5$




(c) $t = 1$

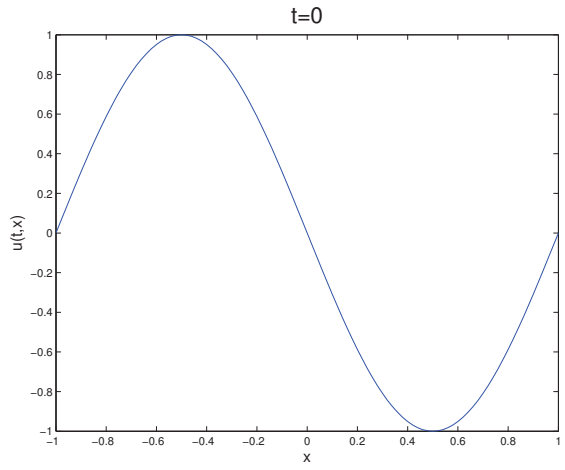



(d) $t = 1.5$

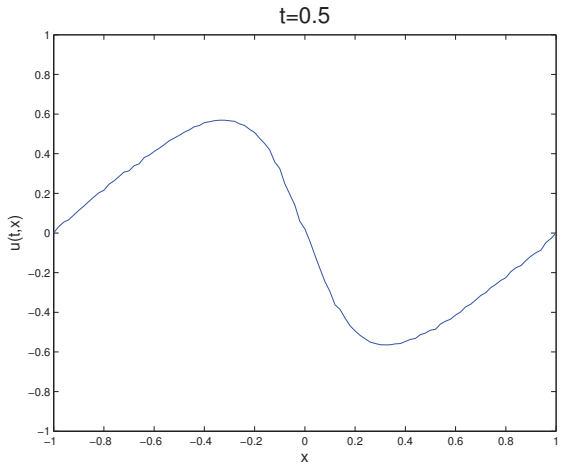



(e) $t = 2$

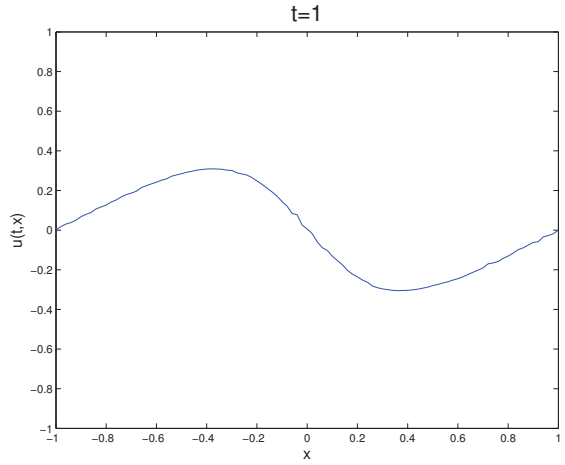
FIGURE 22. Solution at different times for $u_0^+$ when $\nu = 0.01$, with $\delta t = \delta x = 0.02, N_i = N_b = 100$.



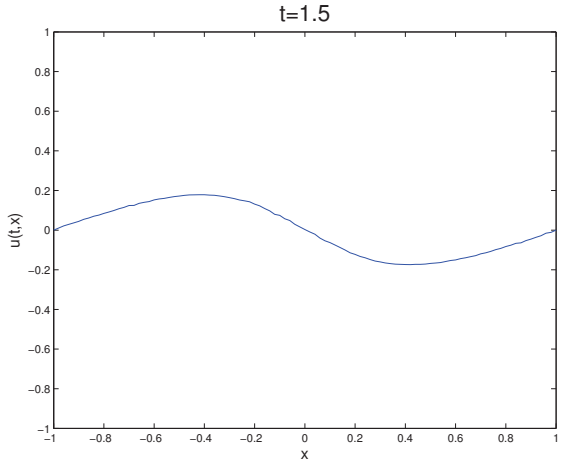

(a) $t = 0$



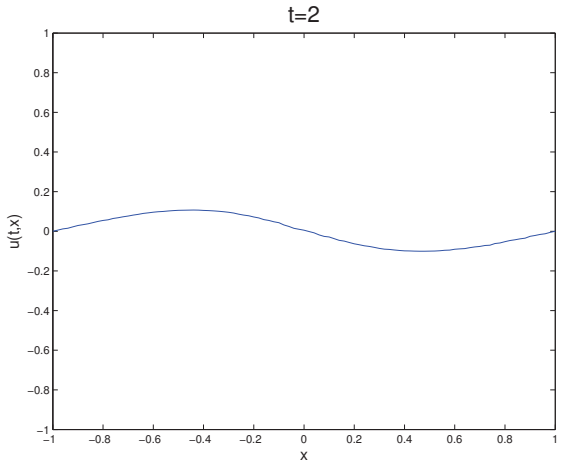

(b) $t = 0.5$




(c) $t = 1$

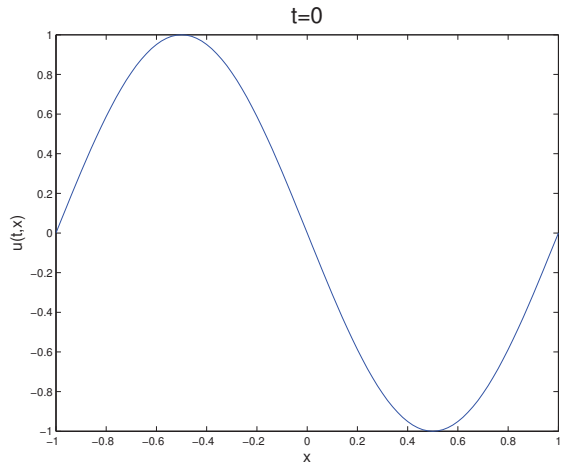



(d) $t = 1.5$

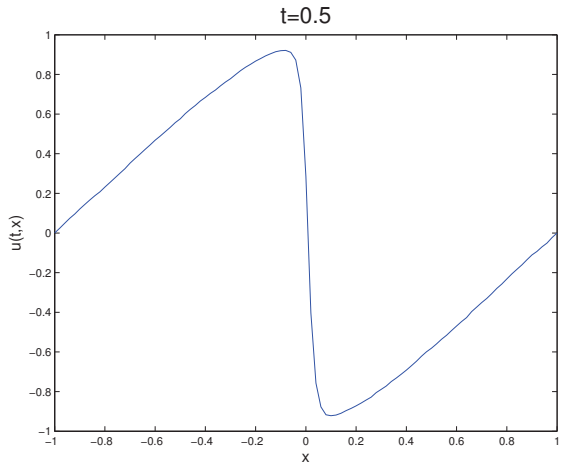



(e) $t = 2$

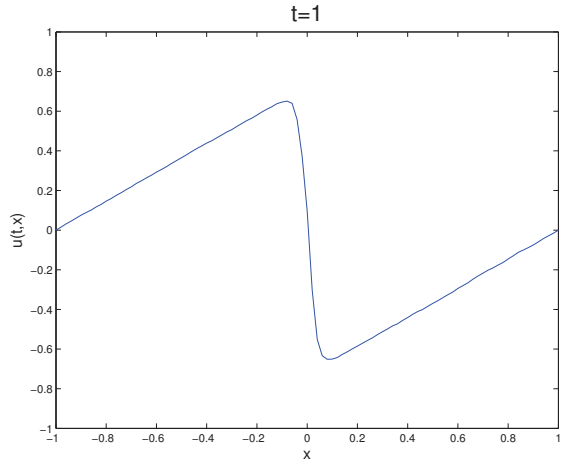
FIGURE 23. Solution at different times for $u_0^-$ when $\nu = 0.1$, with $\delta t = \delta x = 0.02, N_i = N_b = 100$.



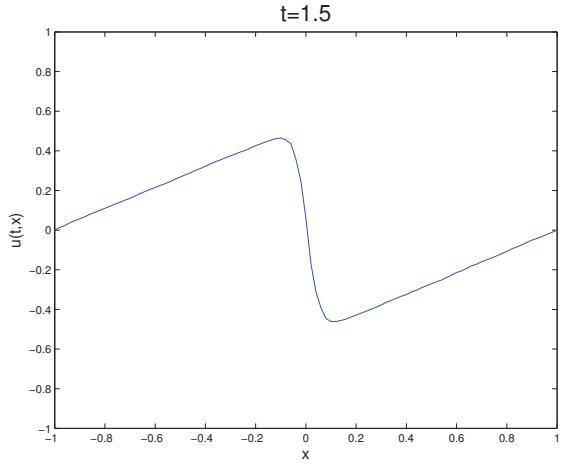

(a) $t = 0$



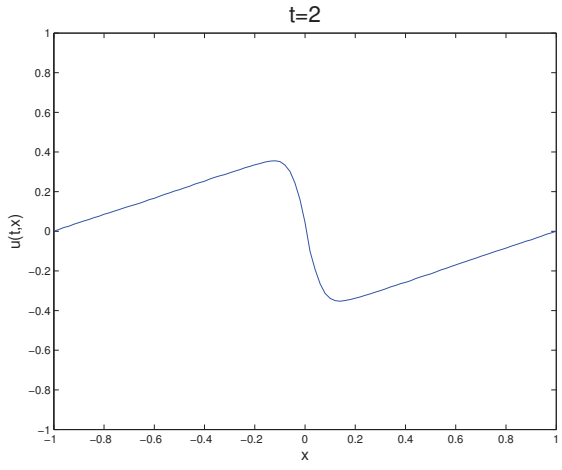

(b) $t = 0.5$




(c) $t = 1$




(d) $t = 1.5$




(e) $t = 2$

FIGURE 24. Solution at different times for $u_0^-$ when $\nu = 0.01$, with $\delta t = \delta x = 0.02, N_i = N_b = 100$.

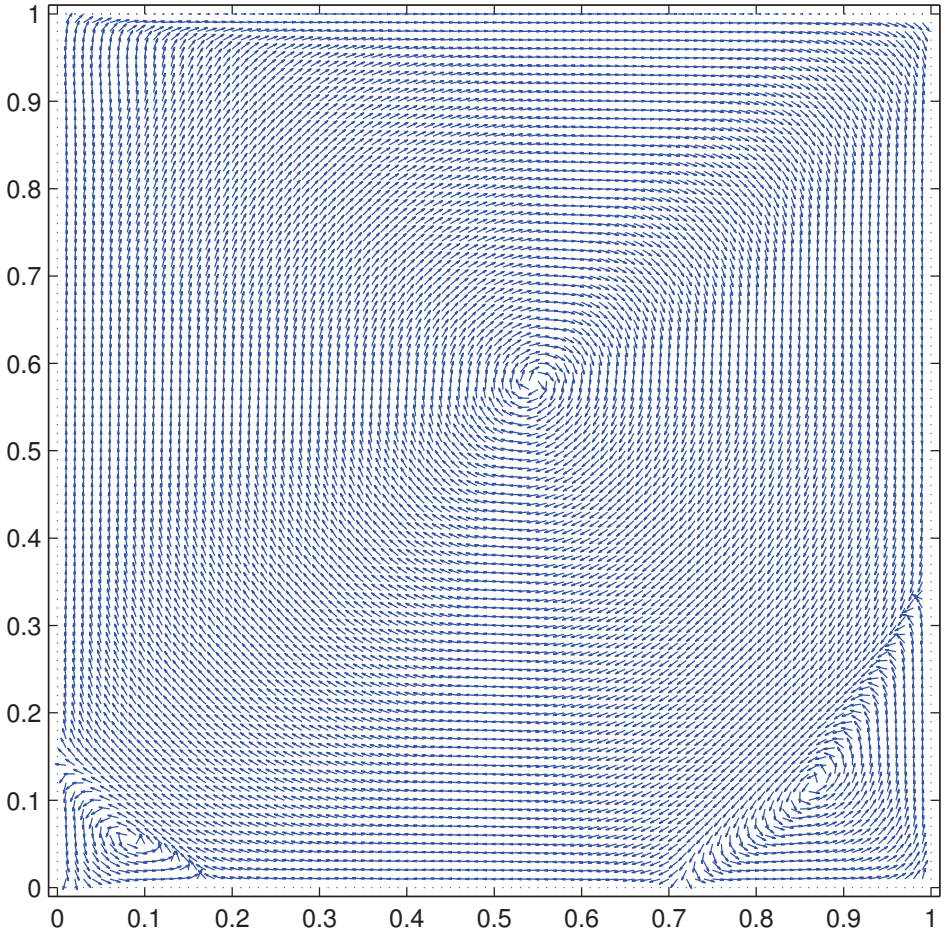



FIGURE 25. The driven cavity flow for $Re = 1000$.

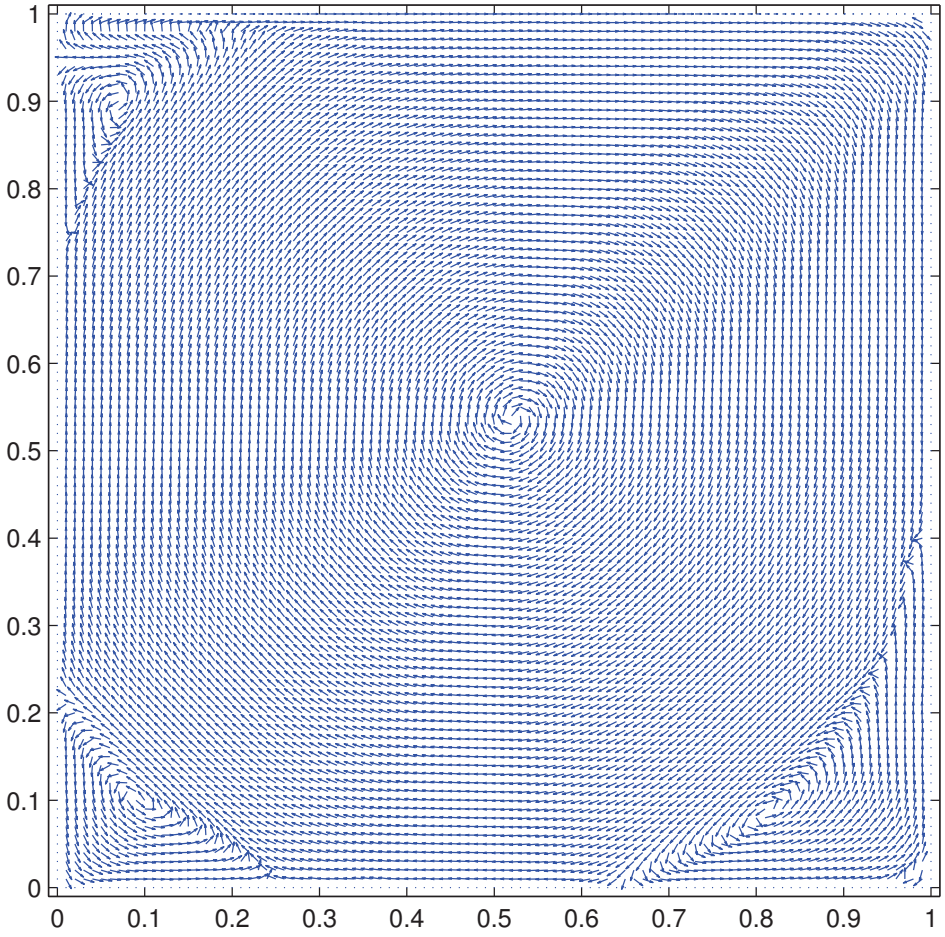



FIGURE 26. The driven cavity flow for $Re = 5000$.

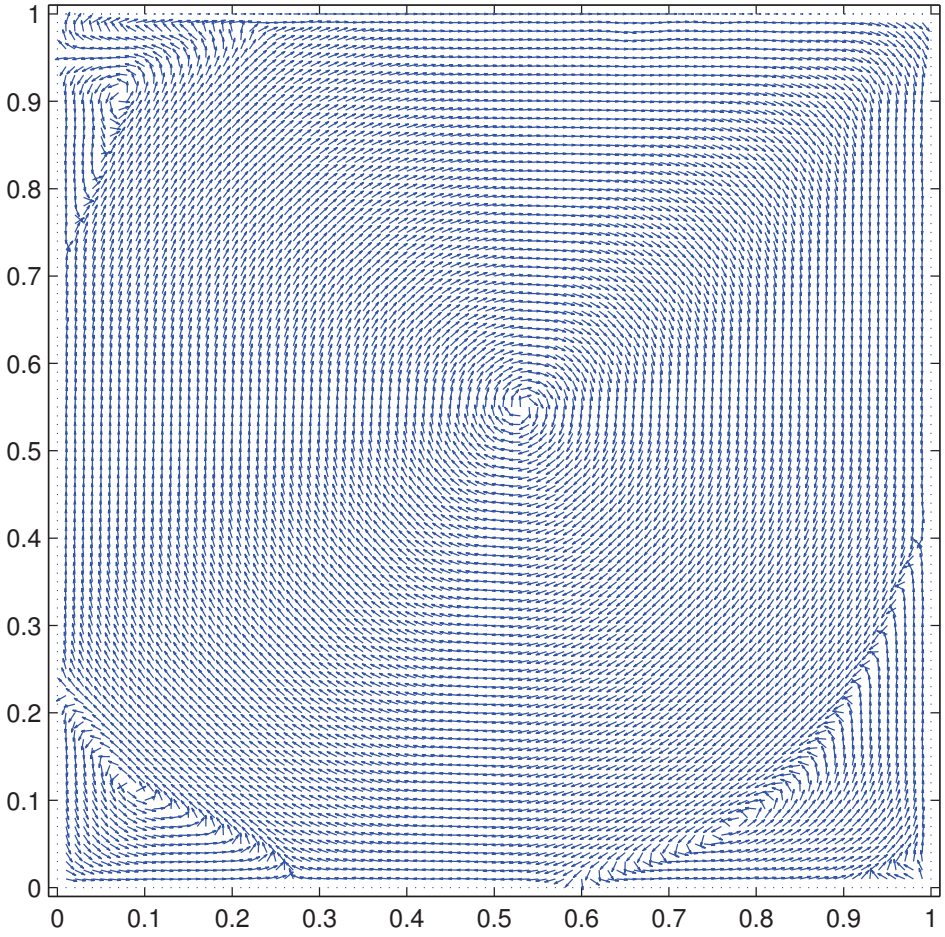



FIGURE 27. The driven cavity flow for $Re = 10000$.

# Bibliographie


[1] A. Abdulle, W. E and T. Li. Effectiveness of implicit methods for stiff stochastic differential equations. *Commun. Comput. Phys.*, 3(2): 295–307, 2008.

[2] L. Arnold. *Random Dynamical Systems.* Springer Monographs in Mathematics. Berlin: Springer. xi, 586 p., 1998.

[3] A.N. Borodin and P. Salminen. *Handbook of Brownian motion: Facts and formulae. 2nd ed.* Probability and Its Applications. Basel: Birkhäuser. xvi, 672 p., 2002.

[4] C.-E. Bréhier Strong and weak orders in averaging for SPDEs. Stochastic Processes Appl., 122(7), 2553–2593, 2012.

[5] C.-E. Brehier. Approximation of the invariant measure via a Euler scheme for stochastic PDEs driven by space-time white noise. *submitted.*

[6] C.-E. Brehier. Analysis of a HMM discretization scheme for SPDEs. *submitted.*

[7] P. Brémaud. *Markov chains. Gibbs fields, Monte Carlo simulation, and queues.* Texts in Applied Mathematics. New York, NY: Springer. xviii, 444 p. , 1999.

[8] H. Brézis. *Functional analysis. Theory and applications. (Analyse fonctionnelle. Théorie et applications.).* Collection Mathématiques Appliquées pour la Maîtrise. Paris: Masson. 248 p. , 1994.

[9] S. Cerrai. *Second order PDE's in finite and infinite dimension.* Lecture Notes in Mathematics. 1762. Berlin: Springer. ix, 330 p., 2001.

[10] S. Cerrai. A Khasminskii type averaging principle for stochastic reaction-diffusion equations. *Ann. Appl. Probab.*, 19(3):899–948, 2009.

[11] S. Cerrai. Normal deviations from the averaged motion for some reaction-diffusion equations with fast oscillating perturbation. *J. Math. Pures Appl.*, 91(6):614–647, 2009.

[12] S. Cerrai and M. Freidlin. Averaging principle for a class of stochastic reaction-diffusion equations. *Probab. Theory Relat. Fields*, 144(1-2):137–177, 2009.

[13] M. Cessenat, R. Dautray, G. Ledanois, P-L.Lions, E. Pardoux and R. Sentis. *Méthodes probabilistes pour les équations de la physique,.* Eyrolles, Paris, 1989.

[14] M. Crouzeix and V. Thomée. On the discretization in time of semilinear parabolic equations with nonsmooth initial data. *Math. Comput.*, 49:359–377, 1987.

[15] G. Da-Prato and J. Zabczyk. *Stochastic equations in infinite dimensions.* Encyclopedia of Mathematics and Its Applications. 44. Cambridge etc.: Cambridge University Press. xviii, 454 p. , 1992.

[16] G. Da Prato and J. Zabczyk. *Ergodicity for infinite dimensional systems.* London Mathematical Society Lecture Note Series. 229. Cambridge: Cambridge Univ. Press. xi, 339 p., 1996.

[17] A.M. Davie and J.G. Gaines. Convergence of numerical schemes for the solution of parabolic stochastic partial differential equations. *Math. Comput.*, 70(233):121–134, 2001.

[18] A. Debussche. Weak approximation of stochastic partial differential equations: the nonlinear case. *Math. Comput.*, 80(273):89–117, 2011.

[19] A. Debussche, Y. Hu, and G. Tessitore. Ergodic BSDEs under weak dissipative assumptions. *Stochastic Processes and their Applications*, 121(3):407–426, 2011.

[20] A. Debussche and J. Printems. Weak order for the discretization of the stochastic heat equation. *Math. Comput.*, 78(266):845–863, 2009.

[21] W. Doeblin. Expose de la théorie des chaînes simples constantes de Markoff à un nombre fini d'etats. *Rev. Math. Union Interbalkan.*, 2:77–105, 1938.

[22] W. E. Analysis of the heterogeneous multiscale method for ordinary differential equations. *Commun. Math. Sci.*, 1(3):423–436, 2003.

[23] W. E and B. Engquist. The heterogeneous multiscale methods. *Commun. Math. Sci.*, 1(1):87–132, 2003.

[24] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden. Heterogeneous multiscale methods: a review. *Commun. Comput. Phys.*, 2(3):367–450, 2007.

[25] W. E, D.Liu, and E. Vanden-Eijnden. Analysis of multiscale methods for stochastic differential equations. *Commun. Pure Appl. Math.*, 58(11):1544–1585, 2005.

[26] L.C. Evans. *Partial Differential Equations. 2nd ed.* Graduate Studies in Mathematics 19. Providence, RI: American Mathematical Society. xxi, 749 p., 2010.

[27] E. Faou. Analysis of splitting methods for reaction-diffusion problems using stochastic calculus. *Math. Comput.*, 78(267):1467–1483, 2009.

[28] R. Ferretti. A technique for high-order treatment of diffusion terms in semi-lagrangian schemes, 2000.

[29] J.-P. Fouque, J. Garnier, G. Papanicolaou, and K. Solna. *Wave propagation and time reversal in randomly layered media.* Stochastic Modelling and Applied Probability 56. New York, NY: Springer. xx, 612 p., 2007.

[30] M.I. Freidlin and A.D. Wentzell. *Random perturbations of dynamical systems. Transl. from the Russian by Joseph Szuecs. 2nd ed.* Grundlehren der Mathematischen Wissenschaften. 260. New York, NY: Springer. xi, 430 p., 1998.

[31] H.Fu and J.Liu. Strong convergence in stochastic averaging principle for two time-scales stochastic partial differential equations. J. Math. Anal. Appl., 384(1), 70–86, 2011.

[32] E. Gobet. Euler schemes and half-space approximation for the simulation of diffusion in a domain. *ESAIM, Probab. Stat.* 5, 261–297, 2001.

[33] J.L. Guermond, P. Minev and J. Shen. An overview of projection methods for incompressible flows. *Comput. Methods Appl. Mech. Eng.* , 195(44-47): 6011–6045, 2006.

[34] I. Gyongy. Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise. I. *Potential Anal.*, 9(1):1–25, 1998.
[35] I. Gyongy. Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise. II. *Potential Anal.*, 11(1):1–37, 1999.
[36] I. Gyongy and D. Nualart. Implicit scheme for stochastic parabolic partial differential equations driven by space-time white noise. *Potential Anal.*, 7(4):725–757, 1997.
[37] R.Z. Khasminskii. On an averaging principle for Itô stochastic differential equations. *Kibernetica*, (4):260–279, 1968.
[38] R.Z. Khasminskii and G.Yin. Limit behavior of two-time-scale diffusions revisited. *J. Differ. Equations*, 212(1):85–113, 2005.
[39] E. Hausenblas. Approximation for semilinear stochastic evolution equations. *Potential Anal.*, 18(2):141–186, 2003.
[40] J.C. Mattingly, A.M. Stuart, and D.J. Higham. Ergodicity for SDEs and approximations: locally Lipschitz vector fields and degenerate noise. *Stochastic Processes Appl.*, 101(2):185–232, 2002.
[41] B. Jourdain, C. Le Bris and T.Lelièvre. On a variance reduction technique for the micro-macro simulations of polymeric fluids. *J. Non-Newton. Fluid Mech.*, 122(1-3): 91-106, 2004.
[42] I.G. Kevrekidis, C.W. Gear, J.M. Hyman, P.G. Kevrekidis, O. Runborg, and C. Theodoropoulos. Equation-free, coarse-grained multiscale computation: enabling microscopic simulators to perform system-level tasks. *Communications in Mathematical Sciences*, 1(4):715–762, 2003.
[43] P.E. Kloeden and E. Platen. *Numerical solution of stochastic differential equations.* Applications of Mathematics. 23. Berlin: Springer-Verlag. xxxv, 632 p. , 1992.
[44] S. Kuksin and A. Shirikyan. A coupling approach to randomly forced nonlinear PDE's. I. *Commun. Math. Phys.*, 221(2):351–366, 2001.
[45] M.-N. Le Roux. Semidiscretization in time for parabolic problems. *Math. Comput.*, 33:919–931, 1979.
[46] F. Legoll, T. Lelievre and G.Samaey. A micro-macro parareal algorithm: application to singularly perturbed ordinary differential equations. preprint.
[47] E.T. Lindvall. *Lectures on the coupling method.* Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. New York, NY: Wiley. 272 p., 1992.
[48] D.Liu. Strong convergence of principle of averaging for multiscale stochastic dynamical systems. *Commun. Math. Sci.*, 8(4):999–1020, 2010.
[49] R. Mannella. Absorbing boundaries and optimal stopping in a stochastic differential equation. *Phys. Lett.,A* 254 (5), 257-262, 1999.
[50] J.C. Mattingly. Exponential convergence for the stochastically forced Navier-Stokes equations and other partially dissipative dynamics. *Commun. Math. Phys.*, 230(3):421–462, 2002.
[51] J.C. Mattingly, A.M. Stuart, and M.V. Tretyakov. Convergence of numerical time-averaging and stationary measures via Poisson equations. *SIAM J. Numer. Anal.*, 48(2):552–577, 2010.
[52] S. Meyn and R.L. Tweedie. *Markov chains and stochastic stability. Prologue by Peter W. Glynn. 2nd ed.* Cambridge Mathematical Library. Cambridge: Cambridge University Press. xviii, 594 p., 2009.
[53] G.N. Milstein. *Numerical integration of stochastic differential equations. Transl. from the Russian.* Mathematics and its Applications (Dordrecht). 313. Dordrecht: Kluwer Academic Publishers. vii, 169 p., 1994.
[54] G.N. Milstein, M.V. Tretyakov. *Stochastic numerics for mathematical physics.* Scientific Computation. Berlin: Springer. ixx, 594 p., 2004.
[55] D. Nualart. *The Malliavin calculus and related topics. 2nd ed.* Probability and Its Applications. Berlin: Springer. xiv, 382 p. , 2006.
[56] G.A. Pavliotis and A.M. Stuart. *Multiscale methods. Averaging and homogenization.* Texts in Applied Mathematics 53. New York, NY: Springer. xviii, 307 p., 2008.
[57] J. Printems. On the discretization in time of parabolic stochastic partial differential equations. *Monte Carlo Methods Appl.*, 7(3-4):359–368, 2001.
[58] M. Sanz-Solé. *Malliavin calculus with applications to stochastic partial differential equations.* Fundamental Sciences: Mathematics. Boca Raton, FL: CRC Press; Lausanne: EPFL Press. viii, 162 p., 2005.
[59] J.Shen. Hopf bifurcation of the unsteady regularized driven cavity flow. *J. Comput. Phys.* 95(1):228-245, 1991.
[60] D. Talay. Discrétisation d'une équation différentielle stochastique et calcul approché d'espérances de fonctionnelles de la solution. (Discretization of a stochastic differential equation and computation of expectations of functions of the solution). 1986.
[61] D. Talay. Second-order discretization schemes of stochastic differential systems for the computation of the invariant law. *Stochastics Stochastics Rep.*, 29(1):13–36, 1990.
[62] D. Talay and L. Tubaro. Expansion of the global error for numerical schemes solving stochastic differential equations. *Stochastic Anal. Appl.*, 8(4):483–509, 1990.
[63] R. Temam. *Navier-Stokes equations. Theory and numerical analysis.* Providence, RI: American Mathematical Society (AMS). xiv, 408 p. , 2001.
[64] E. Vanden-Eijnden. Numerical techniques for multi-scale dynamical systems with stochastic effects. *Commun. Math. Sci.*, 1(2):385–391, 2003.
[65] J.B. Walsh. Finite element methods for parabolic stochastic PDE's. *Potential Anal.*, 23(1):1–43, 2005.

## Résumé

### Analyse numérique d'EDP Stochastiques hautement oscillantes

Dans une première partie, on s'intéresse à un système d'EDP stochastiques variant selon deux échelles de temps, et plus particulièrement à l'approximation de la composante lente à l'aide d'un schéma numérique efficace. On commence par montrer un principe de moyennisation, à savoir la convergence de la composante lente du système vers la solution d'une équation dite moyennée. Ensuite on prouve qu'un schéma numérique de type Euler fournit une bonne approximation d'un coefficient inconnu apparaissant dans cette équation moyennée. Finalement, on construit et on analyse un schéma de discrétisation du système à partir des résultats précédents, selon la méthodologie dite HMM (Heterogeneous Multiscale Method).

On met en évidence l'ordre de convergence par rapport au paramètre d'échelle temporelle et aux différents paramètres du schéma numérique; on étudie les convergences au sens fort (approximation des trajectoires) et au sens faible (approximation des lois).

Dans une seconde partie, on étudie une méthode d'approximation de solutions d'EDP paraboliques, en combinant une approche semi-lagrangienne et une discrétisation de type Monte-Carlo. On montre d'abord dans un cas simplifié que la variance dépend des pas de discrétisation ; enfin on fournit des simulations numériques de solutions, afin de mettre en avant les applications possibles d'une telle méthode.



**Mots clés:**
Equations aux dérivées partielles stochastiques
Méthodes de Monte-Carlo
Méthodes numériques
Systèmes multi-échelles



## Abstract

### Numerical analysis of highly oscillatory Stochastic PDEs

In a first part, we are interested in the behavior of a system of Stochastic PDEs with two time-scales; more precisely, we focus on the approximation of the slow component thanks to an efficient numerical scheme. We first prove an averaging principle, which states that the slow component converges to the solution of the so-called averaged equation. We then show that a numerical scheme of Euler type provides a good approximation of an unknown coefficient appearing in the averaged equation. Finally, we build and we analyze a discretization scheme based on the previous results, according to the HMM methodology (Heterogeneous Multiscale Method).

We precise the orders of convergence with respect to the time-scale parameter and to the parameters of the numerical discretization; we study the convergence in a strong sense - approximation of the trajectories - and in a weak sense - approximation of the laws.

In a second part, we study a method for approximating solutions of parabolic PDEs, which combines a semi-lagrangian approach and a Monte-Carlo discretization. We first show in a simplified situation that the variance depends on the discretization steps. We then provide numerical simulations of solutions, in order to show some possible applications of such a method.



**Keywords:**
Stochastic partial differential equations
Monte-Carlo methods
Numerical methods
Multiscale systems



ENSC - 2012 n°370

**École normale supérieure de Cachan - Antenne de Bretagne**

Campus de Ker Lann - Avenue Robert Schuman - 35170 BRUZ

Tél : +33(0)2 99 05 93 00 - Fax : +33(0)2 99 05 93 29