Analyse de la force de bruitage dans les modèles génératifs basés sur le score.

Stanislas Strasman*, Antonio Ocello†, Claire Boyer*,‡, Sylvain Le Corff* & Vincent Lemaire*

* *LPSM, Sorbonne Université, UMR CNRS 8001, France,*
stanislas.strasman@sorbonne-universite.fr
prenom.nom@sorbonne-universite.fr
† *CMAP, École Polytechnique, Institut Polytechnique de Paris, France,*
prenom.nom@polytechnique.edu
‡ *Institut Universitaire de France (IUF)*

Résumé. Les modèles génératifs basés sur le score (SGMs) visent à estimer une distribution de données cible en apprenant des fonctions de score (correspondant au gradient du logarithme de densités de probabilité) uniquement à partir d'échantillons bruités de la cible. La littérature récente s'est largement concentrée sur l'évaluation de l'erreur entre les distributions cible et estimée, en mesurant la qualité des données générées à travers la divergence de Kullback-Leibler (KL) ou encore des distances de Wasserstein. Néanmoins, les résultats existants ont été obtenus pour une vitesse de bruitage homogène dans le temps. Sous des hypothèses faibles sur la distribution des données, nous établissons une borne supérieure pour la divergence KL entre les distributions cible et estimée qui dépend explicitement de la force de bruitage utilisée au cours du temps. De plus, en supposant que le score est Lipschitz continu, et avec des capacités d'approximation du score et de discrétisation parfaites, nous montrons une borne d'erreur améliorée en distance de Wasserstein, tirant parti des mécanismes de contraction sous-jacents des équations différentielles stochastiques en jeu. Enfin, nous proposons un algorithme pour ajuster automatiquement la fonction de bruitage au cours de la diffusion en utilisant la borne supérieure théorique établie.

Mots-clés. Méthodes générative diffusives, modèles génératifs basés sur le score, force de bruitage.

Abstract. Score-based generative models (SGMs) aim at estimating a target data distribution by learning score functions using only noise-perturbed samples from the target. Recent literature has focused extensively on assessing the error between the target and estimated distributions, gauging the generative quality through the Kullback-Leibler (KL) divergence and Wasserstein distances. All existing results have been obtained so far for time-homogeneous speed of the noise schedule. Under mild assumptions on the data distribution, we establish an upper bound for the KL divergence between the target and the estimated distributions, explicitly depending on any time-dependent noise schedule. Assuming that the score is Lipschitz continuous, we provide an improved error bound in Wasserstein distance, taking advantage of favourable underlying contraction mechanisms. We also propose an algorithm to automatically tune the noise schedule using the proposed upper bound. We illustrate empirically the performance of the noise schedule optimization in comparison to standard choices in the literature.

Keywords. Generative diffusion models, score-based generative models, noise schedule.

1 Introduction

Les modèles génératifs visent à générer de nouveaux échantillons synthétiques à partir d'échantillons dits d'entraînement, supposés être issus d'une distribution π_{data} inconnue. Ces modèles incluent des approches reposant sur des diffusions (Sohl-Dickstein et al., 2015; Song and Ermon, 2019; Ho et al., 2020; Song et al., 2021), qui ont récemment permis des résultats prometteurs dans de nombreuses applications, comme par exemple dans le cadre de la génération d'image à partir de description textuelle (Ramesh et al., 2022), ou encore de la génération de langage naturel (Gong et al., 2023). Nous référons le lecteur à Yang et al. (2023) qui offre un aperçu complet des dernières avancées sur ce sujet. Il convient de noter que dans ces applications réelles, la complexité des données empêche en général de représenter la distribution π_{data} à travers un modèle paramétrique classique et, proscrit dès lors toute estimation par des méthodes de maximum de vraisemblance traditionnelles. Les modèles génératifs basés sur le score (SGMs) représentent alors une alternative implémentable.

Modèles génératifs utilisant le score (SGM). Les modèles génératifs basés sur le score (SGMs) sont des modèles probabilistes s'articulant en deux phases. La première phase consiste à bruiter les données (également appelée phase *forward*), c'est-à-dire que l'on perturbe progressivement la distribution empirique en ajoutant du bruit aux données d'entraînement jusqu'à ce que leur distribution atteigne approximativement une distribution facilement échantillonnable π_∞. D'autre part, la seconde phase vise à inverser cette dynamique en débruitant séquentiellement des réalisations de π_∞. On appelle cette phase, la phase d'échantillonnage (ou phase *backward*). Inverser la dynamique nécessite en principe la connaissance de la fonction de score, c'est-à-dire, le gradient du logarithme de la densité du processus *forward* à chaque étape temporelle de la diffusion. Cependant, connaître le score exactement revient à connaître la distribution au temps $t = 0$, c'est-à-dire, à connaître la distribution cible π_{data} selon laquelle nous souhaitons simuler de nouveaux exemples. Pour contourner ce problème, la fonction de score est donc apprise sur la base de l'évolution des échantillons de données bruités, en utilisant une architecture de réseau de neurones profond. Une fois le score appris, nous pouvons l'utiliser dans la dynamique inverse appliquée aux échantillons tirés selon π_∞, nous obtenons ainsi un modèle génératif, approchant des tirages selon π_{data}.

Une attention significative a été accordée à la compréhension des sources d'erreurs qui affectent la qualité de la génération de données associée aux SGM (Chen et al., 2023a,b; De Bortoli, 2022; Lee et al., 2023). Pour ce faire, des bornes supérieures pour des (pseudo-)distances entre les distributions d'échantillons d'entraînement et générés ont été établies.

Contributions. Dans ce travail, nous procédons à une analyse mathématique approfondie de la force de bruitage dans les modèles génératifs basés sur le score.

— Nous établissons une borne supérieure pour la divergence de Kullback-Leibler (KL) entre la distribution des données et la loi du SGM. Cette borne est valide sous des hypothèses minimales et dépend explicitement de la force de bruitage utilisée pour entraîner le SGM.
— À travers des expériences numériques, nous illustrons la borne supérieure obtenue en pratique en ce qui concerne les divergences KL empiriques effectives. Ces simulations mettent en évidence la pertinence de la borne supérieure, reflétant en pratique l'effet de la force de bruitage au cours du temps sur la qualité de la distribution générée.
— En faisant une hypothèse supplémentaire sur la régularité de la fonction de score, nous établissons une borne plus précise de l'erreur due au temps de mélange en termes de distance de Wasserstein, en tirant parti de la contraction du terme de dérive non seulement en phase forward, mais aussi en en phase backward de la diffusion stochastique.
— Enfin, nous proposons d'exploiter la borne théorique obtenue pour guider et améliorer la mise en œuvre des SGM en pratique. Nous suggérons en effet une procédure pour optimiser conjointement le réseau de neurones approchant le score et la force de bruitage en utilisant comme fonction objectif la borne supérieure établie.

2 Analyse théorique de l'impact de la force de bruitage dans les SGMs

2.1 Notation et définitions

Processus *forward*. Posons $\beta : [0, T] \mapsto \mathbb{R}_{>0}$ la fonction de bruitage, supposée continue et croissante. Bien que développés initialement en utilisant un nombre fini d'étapes de bruitage, les analyses les plus récentes considèrent des perturbations continues à l'aide d'équations différentielles stochastiques (EDS) (Song et al., 2021). Considérons donc un processus *forward* donné par

$$\mathrm{d}\overrightarrow{X}_t = -\frac{\beta(t)}{2\sigma^2}\overrightarrow{X}_t \mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}B_t, \quad \overrightarrow{X}_0 \sim \pi_{\text{data}}\,. \tag{1}$$

Notons p_t la densité de $\overrightarrow{X}_t$ au temps $t \in (0, T]$. Comme la dérive est linéaire par rapport à $(X_t)_{t\geq 0}$, une simulation exacte de ce processus est possible. De plus, la distribution stationnaire du processus *forward* est la distribution gaussienne avec moyenne 0 et variance $\sigma^2 \mathrm{I}_d$ que l'on note π_∞.

Lorsque $\beta(t)$ est constant, égal à 2, (c'est-à-dire dans le cas homogène en temps), ce processus de diffusion est connu sous le nom de *Variance Preserving SDE* (VPSDE, De Bortoli et al., 2021; Conforti et al., 2023; Chen et al., 2023b), dont la discrétisation permet de retrouver les *Denoising Diffusion Probabilistic Models* (DDPM, Ho et al., 2020).

Processus *backward*. Le processus *backward* correspondant est initialisé à la distribution stationnaire π_∞ et peut être écrit

$$\mathrm{d}\overleftarrow{X}_t = \eta(t, \overleftarrow{X}t)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim \pi_\infty,$$

où

$$\begin{cases} \bar{\beta}(t) & := \beta(T-t) \\ \eta(t, \overleftarrow{X}_t) & := \bar{\beta}(t)\overleftarrow{X}_t/(2\sigma^2) + \bar{\beta}(t)\nabla \log p_{T-t}\left(\overleftarrow{X}_t\right). \end{cases}$$

Notons $\mathbb{Q}_T \in \mathcal{P}(C([0,T], \mathbb{R}^d))$ la mesure de chemin associée à la diffusion *backward*. Nous introduisons $\tilde{p}_t(x)$ la distribution marginale (en temps) du processus *forward* renormalisée par la densité de sa distribution stationnaire :

$$\forall x \in \mathbb{R}^d, \quad \tilde{p}_t(x) = \frac{p_t(x)}{\varphi_{\sigma^2}(x)}, \tag{2}$$

où φ_{σ^2} désigne la fonction de densité de π_∞, une distribution gaussienne avec moyenne 0 et variance $\sigma^2 \mathrm{I}_d$. Ainsi, le processus *backward* peut être réécrit

$$\mathrm{d}\overleftarrow{X}_t = \bar{\eta}\left(t, \overleftarrow{X}t\right)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t, \quad \overleftarrow{X}_0 \sim \pi_\infty, \tag{3}$$

où $\bar{\eta}(t, \overleftarrow{X}_t) := -\frac{\bar{\beta}(t)}{2\sigma^2}\overleftarrow{X}_t + \bar{\beta}(t)\nabla \log \tilde{p}_{T-t}(\overleftarrow{X}_t)$. Considérer $\tilde{p}_t$ dans notre analyse permet de ré-interpréter le processus *backward* comme une perturbation d'un processus OU. Cette astuce est cruciale pour mettre en valeur le rôle central de l'information de Fisher dans la performance du SGM. Cette renormalisation a déjà été utilisée par Conforti et al. (2023).

Estimation du score. Simuler le processus *backward* requiert de connaître le score. Cependant, la fonction de score (modifiée) $\nabla \log \tilde{p}_t(x) = \nabla \log p_t(x) + x/\sigma^2$ ne peut pas être évaluée directement, car elle dépend de la distribution des données inconnues. Pour contourner ce problème, la fonction de score $\nabla \log p_t$ doit être estimée. Dans Hyvärinen and Dayan (2005), les auteurs proposent d'estimer la fonction de score associée à une distribution en minimisant la distance au carré L^2 entre la vraie fonction de score et l'approximation proposée. Dans le contexte des modèles de diffusion, cela se fait généralement avec l'utilisation d'une architecture de réseau de neurones profond $s_\theta : [0,T] \times \mathbb{R}^d \mapsto \mathbb{R}^d$ paramétrée par $\theta \in \Theta$, et visant à minimiser :

$$\mathcal{L}_{\text{explicit}}(\theta) = \mathbb{E}\left[\left\| s_\theta(\tau, \overrightarrow{X}_\tau) - \nabla \log p_\tau(\overrightarrow{X}_\tau)\right\|^2\right],$$

avec $\tau \sim \mathcal{U}(0,T)$ indépendant du processus *forward* $(\overrightarrow{X}_t)_{t\geq 0}$. Cependant, ce problème d'estimation souffre du fait que la cible de régression n'est pas explicitement connue. Un problème d'optimisation tractable partageant les mêmes optima peut cependant être défini à travers la marginalisation de π_{data} par p_τ (voir Vincent, 2011; Song et al., 2021) :

$$\mathcal{L}_{\text{score}}(\theta) = \mathbb{E}\left[\| s_\theta(\tau, \overrightarrow{X}_\tau) - \nabla \log p_\tau(\overrightarrow{X}_\tau | X_0)\|^2\right],$$

où τ est uniformément distribué sur $[0,T]$, indépendant de $X_0 \sim \pi_{\text{data}}$ et $\overrightarrow{X}\tau \sim p_\tau(\cdot|X_0)$. Cette fonction de coût ne requiert que la connaissance du noyau de transition du processus *forward*. Dans le cadre classique des modèles de diffusion donné par (1), il s'agit d'un noyau gaussien avec moyenne et variance explicites.

Discrétisation. Une fois la fonction de score apprise, il reste que la dynamique *backward* ne présente plus une dérive linéaire. Cela rend sa simulation exacte difficile. Pour résoudre ce problème, une solution consiste à discrétiser la dynamique continue du processus *backward*. Ainsi, Song et al. (2021) propose un schéma de discrétisation d'Euler-Maruyama (EM) dans lequel les coefficients de dérive et de diffusion sont discrétisés récursivement. En particulier, introduisons $\tilde{s}_\theta(t,x) := s_\theta(t,x) + x/\sigma^2$ et considérons la discrétisation temporelle $0 =: t_0 \leq t_1 \leq \cdots \leq t_N := T$, le schéma EM correspond à

$$\mathrm{d}\overleftarrow{X}_t^{EM} = \left(-\frac{\bar{\beta}(t_k)}{2\sigma^2}\overleftarrow{X}_{t_k}^{EM} + \bar{\beta}(t)\tilde{s}_\theta\left(T-t_k, \overleftarrow{X}_{t_k}^{EM}\right)\right)\mathrm{d}t + \sqrt{\bar{\beta}(t_k)}\mathrm{d}B_t\,.$$

Autrement, l'intégrateur exponentiel d'Euler (EI), déjà utilisé dans Conforti et al. (2023), nécessite seulement de discrétiser la partie associée à la fonction de score modifiée. Soit $\left(\overleftarrow{X}_t^\theta\right) t \in [0,T]$ tel que, pour $t \in [t_k, t_{k+1}]$,

$$\mathrm{d}\overleftarrow{X}_t^\theta = \bar{\beta}(t)\left(-\frac{1}{2\sigma^2}\overleftarrow{X}_t^\theta + \tilde{s}_\theta\left(T-t_k, \overleftarrow{X}_{t_k}^\theta\right)\right)\mathrm{d}t + \sqrt{\bar{\beta}(t)}\mathrm{d}B_t\,.$$

Ce schéma peut être considéré comme un raffinement du schéma classique Euler-Maruyama car il intègre explicitement le terme de dérive linéaire. Nous considérons donc un tel schéma dans nos développements théoriques ultérieurs. Enfin, notons $\bar{\mathbb{Q}}_N^{\beta,\theta} \in \mathcal{P}(C([0,T],\mathbb{R}^d))$ la mesure de chemin associée à cette version discrétisée de la diffusion *backward* et par $\widehat{\pi}_N^{(\beta,\theta)}$ la densité de probabilité marginale de $\overleftarrow{X}_T^\theta$ avec une discrétisation à N pas de temps.

2.2 Une borne supérieure explicite en la force de bruitage

La distribution des données π_{data} est supposée absolument continue par rapport à la mesure gaussienne π_∞. L'information de Fisher relative $\mathcal{I}(\pi_{\text{data}}|\pi_\infty)$ est donnée par

$$\mathcal{I}(\pi_{\text{data}}|\pi_\infty) := \int \left\|\nabla \log\left(\frac{\mathrm{d}\pi_{\text{data}}}{\mathrm{d}\pi_\infty}\right)\right\|^2 \mathrm{d}\pi_{\text{data}}\,.$$

Nous considérons les hypothèses suivantes :

H1 La fonction de bruitage est continue, non décroissante et telle que $\int_0^\infty \beta(t)\mathrm{d}t = \infty$.

H2 La distribution des données a une information de Fisher finie par rapport à la distribution gaussien, c'est-à-dire, $\mathcal{I}(\pi_{\text{data}}|\pi_\infty) < \infty$.

H3 Le paramètre $\theta \in \Theta$ et la fonction β satisfont

$$\mathbb{E}\left[\exp\left\{\frac{1}{2}\int_0^T \bar{\beta}(t)\left\|\left(\tilde{s}\left(T-t, \overleftarrow{X}_t\right) - \tilde{s}_\theta\left(T-t_k, \overleftarrow{X}_{t_k}\right)\right)\right\|^2 \mathrm{d}t\right\}\right] < \infty,$$

où $\tilde{s}(t,x) := \nabla \log \tilde{p}_t(x)$ correspond à la fonction de score modifiée (2).

L'hypothèse H1 est nécessaire pour garantir que le processus *forward* converge vers la distribution stationnaire lorsque le temps de diffusion tend vers l'infini. L'hypothèse H2 est inhérente à la distribution des données, car elle implique seulement l'intégrabilité L^2 de la fonction de score. Un tel type d'hypothèse a déjà été considéré dans la littérature, voir Conforti et al. (2023). Enfin, l'hypothèse H3 garantit une bonne approximation du score par le réseau de neurones $\tilde{s}_\theta$, pondérée par le niveau de bruitage.

Theorem 2.1. *Supposons que H1, H2 et H3 soient vérifiées. Alors,*

$$\mathrm{KL}\left(\pi_{\mathrm{data}} \middle|\middle| \widehat{\pi}_N^{(\beta,\theta)}\right) \leq \mathcal{E}_1(\beta) + \mathcal{E}_2(\theta,\beta) + \mathcal{E}_3(\beta)\,,$$

où

$$\mathcal{E}_1(\beta) = \mathrm{KL}\left(\pi_{\mathrm{data}} || \pi_\infty\right) \exp\left\{-\frac{1}{\sigma^2}\int_0^T \beta(s)\mathrm{d}s\right\}\,,$$

$$\mathcal{E}_2(\theta,\beta) = \sum_{k=1}^N \mathbb{E}\left[\left\|\nabla \log \tilde{p}_{T-t_k}\left(\overrightarrow{X}_{T-t_k}\right) - \tilde{s}_\theta\left(T-t_k, \overrightarrow{X}_{T-t_k}\right)\right\|^2\right] \int_{t_k}^{t_{k+1}} \beta(t)\mathrm{d}t\,,$$

$$\mathcal{E}_3(\beta) = 2h\beta(T)\max\left\{\frac{h\beta(T)}{4\sigma^2}; 1\right\}\mathcal{I}(\pi_{\mathrm{data}}|\pi_\infty)\,,$$

avec $h := \sup_{k\in 1,\ldots,N}(t_k - t_{k-1})$ *et* $t_0 := 0$.

La borne obtenue permet de retrouver des garanties existantes (De Bortoli et al., 2021; Conforti et al., 2023), mais va au-delà en faisant apparaître une dépendance explicite en la force de bruitage β, soit à travers sa version intégrée sur le temps de diffusion, soit à travers sa valeur finale au temps T.

Les différents termes de la borne correspondent à trois types d'erreurs affectant les performances des SGMs. Le terme $\mathcal{E}_1$ représente le *temps de mélange* du processus forward d'Ornstein-Uhlenbeck, résultant de la limitation pratique de considérer le processus forward jusqu'à un temps fini T. En effet, $\mathcal{E}_1$ tend vers 0 lorsque T tend vers l'infini. Notons que le terme multiplicatif dans $\mathcal{E}_1$ correspond à la divergence KL entre π_{data} et π_∞, finie par l'Hypothèse H2. Le second terme $\mathcal{E}_2$ correspond à l'*erreur d'approximation*, qui découle de l'utilisation d'un réseau de neurones pour estimer la fonction de score. Enfin, $\mathcal{E}_3$ est l'*erreur de discrétisation* du schéma de discrétisation EI. Ce dernier terme disparaît à mesure que la grille de discrétisation est de plus en plus fine (c'est-à-dire, $h \to 0$).

3 Sur la finesse de la borne supérieure

3.1 Une version raffinée

Dans cette section, nous nous concentrons sur le cadre d'une "approximation parfaite du score" et d'une discrétisation infiniment précise, c'est-à-dire, $\mathcal{E}_2(\theta,\beta) = \mathcal{E}_3(\theta,\beta) = 0$. Cela permet d'évaluer la précision du terme $\mathcal{E}_1(\beta)$ dans la borne supérieure du Théorème 2.1.

Lorsque la distribution des données est restreinte à une gaussienne $\mathcal{N}(\mu_0, \Sigma_0)$, on peut exploiter la contraction *backward* en supposant que $\lambda_{\max}(\Sigma_0) \leq \sigma^2$, où $\lambda_{\max}(\Sigma_0)$ désigne la plus grande valeur propre de Σ_0. Dans ce cas spécifique, nous pouvons obtenir une version raffinée pour $\mathcal{E}_1$, donnée par

$$\mathrm{KL}\left(\pi\mathrm{data}|\varphi_{\sigma^2}Q_T\right) \leq \mathrm{KL}\left(\pi_{\mathrm{data}}|\varphi_{\sigma^2}\right) \exp\left(-\frac{2}{\sigma^2}\int_0^T \beta(s)\mathrm{d}s\right).$$

En outre, dans la littérature, d'autres bornes supérieures ont été obtenues pour d'autres métriques telles que les distances de Wasserstein (Lee et al., 2023; De Bortoli, 2022). Nous proposons un contrôle en distance de Wasserstein sous l'hypothèse suivante :

H4 For any t, there exists $C_t \geq 0$ such that $\forall x, y \in \mathbb{R}^d$,

$$\left(\nabla \log \tilde{p}_t(x) - \nabla \log \tilde{p}_t(y)\right)^\top (x-y) \leq -C_t \left\|x-y\right\|^2 .$$

Proposition 3.1. *Supposons que $x \mapsto \nabla \log \tilde{p}_t(x)$ soit Lipschitz continu de coefficient L_t pour $t \in (0, T]$. Alors,*

$$\mathcal{W}_2(\pi_{\mathrm{data}}, \varphi_{\sigma^2}Q_T)^2 \leq \mathcal{W}_2(p_T, \varphi_{\sigma^2})^2 \times \exp\left(-\int_0^T \frac{\bar{\beta}(t)}{\sigma^2}\left(1 - 2L_t\sigma^2\right)\mathrm{d}t\right). \tag{4}$$

De plus, sous l'Hypothèse H4, nous avons

$$\mathcal{W}_2(\pi_{\mathrm{data}}, \varphi_{\sigma^2}Q_T)^2 \leq \mathcal{W}_2(p_T, \varphi_{\sigma^2})^2 \times \exp\left(-\int_0^T \frac{\bar{\beta}(t)}{\sigma^2}\left(1 + 2C_t\sigma^2\right)\mathrm{d}t\right). \tag{5}$$

Notons que l'Hypothèse H4 ou la propriété Lipschitz de la fonction de score sont toutes deux satisfaites lorsque la distribution cible est supposée être gaussienne avec une structure de covariance appropriée.

Lemma 3.2. *Supposons que π_{data} soit une distribution gaussienne $\mathcal{N}(\mu_0, \Sigma_0)$, telle que $\lambda_{\max}(\Sigma_0) \leq \sigma^2$. Alors, la borne d'erreur (5) est valable avec une contraction donnée par la constante suivante*

$$C_t := \frac{m_t^2\left(\sigma^2 - \lambda_{\max}(\Sigma_0)\right)}{m_t^2\lambda_{\max}(\Sigma_0) + \sigma^2\left(1 - m_t^2\right)}.$$

Ce résultat, restreint au cas gaussien, met l'accent sur l'importance de calibrer le paramètre σ^2 en fonction de la structure de covariance de la distribution des données, afin d'accélérer la vitesse de convergence de l'algorithme.

3.2 Illustration numérique

Pour illustrer la borne supérieure, nous considérons le cas où la vraie distribution est gaussienne en dimension $d = 50$ avec une moyenne $\mathbf{1}_d$ et différents choix de structure de covariance :

1. (Isotrope) $\Sigma^{(\mathrm{iso})} = 0.5\mathrm{I}_d$.
2. (Hétéroscédastique) $\Sigma^{(\mathrm{heterosc})} \in \mathbb{R}^{d\times d}$ est une matrice diagonale telle que $\Sigma^{(\mathrm{heterosc})}_{jj} = 10$ pour $1 \le j \le 5$, et $\Sigma^{(\mathrm{heterosc})}_{jj} = 0.1$ sinon.
3. (Corrélation) $\Sigma^{(\mathrm{corr})} \in \mathbb{R}^{d\times d}$ est une matrice pleine dont les entrées diagonales sont égales à un et les termes hors diagonale sont $\Sigma^{(\mathrm{corr})}_{jj'} = 1/\sqrt{|j-j'|}$ pour $1 \le j \ne j' \le d$.

Les distributions de données résultantes sont respectivement désignées par $\pi^{(\mathrm{iso})}_{\mathrm{data}}$, $\pi^{(\mathrm{heterosc})}_{\mathrm{data}}$ et $\pi^{(\mathrm{corr})}_{\mathrm{data}}$. Le Théorème 2.1 fournit une borne supérieure générique de Kullback-Leibler :

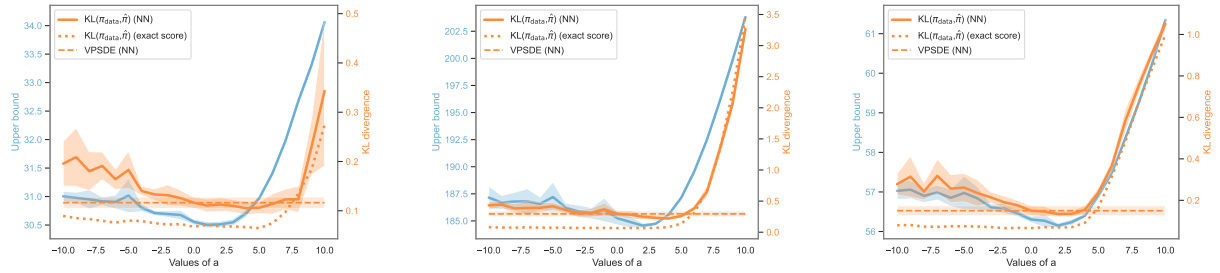
$$\mathcal{L}_{\mathrm{sched}}(\theta,\beta) = \mathcal{E}_1(\beta) + \mathcal{E}_2(\theta,\beta) + \mathcal{E}_3(\beta). \tag{6}$$

Nous proposons d'évaluer (6) pour les différentes distributions de données ci-dessus, et pour une fonction de bruitage de la forme

$$\beta_a(t) \propto (\mathrm{e}^{at} - 1)/(\mathrm{e}^{aT} - 1), \tag{7}$$

avec $a \in \mathbb{R}$ variant de -10 à 10. Pour ce faire, pour chaque valeur de a, et chaque distribution de données, nous entraînons un SGM avec 200 étapes de discrétisation de l'intervalle de temps $[0,1]$ avec $n = 10000$ échantillons gaussiens. Le score est appris en utilisant un réseau de neurones dense avec 3 couches cachées de largeur 256 sur 100 époques.

La Figure 1 met en évidence dans tous les scénarios que la force de bruitage utilisée dans les SGM impacte la valeur de $\mathrm{KL}(\pi_{\mathrm{data}} \| \widehat{\pi}_N^{(\beta,\theta)})$, et donc la qualité de la distribution apprise.



(a) Cas isotrope (b) Cas hétéroscédastique (c) Cas corrélé

FIGURE 1 – Comparaison de la divergence KL empirique (valeur moyenne ± écart-type sur 10 réalisations) entre π_{data} et $\widehat{\pi}_N^{(\beta,\theta)}$ (en orange) et la borne supérieure (6) (en bleu) par rapport au paramètre a utilisé dans la définition de la force de bruitage β_a, pour $d = 50$. Nous représentons également la divergence KL obtenue avec le modèle *VPSDE* (ligne *dashed*) et celle obtenue avec notre modèle (ligne *dotted*) lorsque le score n'est pas approché mais directement évalué.

4 Optimisation de la force de bruitage

Algorithme. Nous proposons d'exploiter la borne supérieure théorique (6) pour ajuster le choix de la force de bruitage. À cette fin, nous proposons une méthode itérative pour optimiser

conjointement les poids θ du réseau de neurones et la force de bruitage β, voir Algorithme 1. Les fonctions admissibles β_a pour la fonction de bruitage sont données par (7). Pour des comparaisons justes, nous entraînons à la fois le réseau *VPSDE* et le réseau adaptatif avec 10000 échantillons sur 200 époques en utilisant le même *learning rate*.

Algorithme 1 Optimisation itérative de la force de bruitage et de la fonction de score

Entrée : N échantillons d'entraînement, programme initial β_a avec $a = a^{(0)}$, paramètre initial $\theta^{(0)}$.
Définir $a^\star = a^{(0)}$
for $e = 0$ **to** nombre d'époques **do**
  Calculer $\theta^{(e+1)}$ en utilisant un *score matching* avec la fonction de bruitage $\beta_{a^\star}$ et l'estimation initiale $\theta^{(e)}$.
  if e mod10 $= 0$ **then**
    Mettre à jour

$$a^\star \in \operatorname{argmin}_a \mathcal{L}_{\text{sched}}(\theta^{(e+1)}, \beta_a).$$

  end if
end for

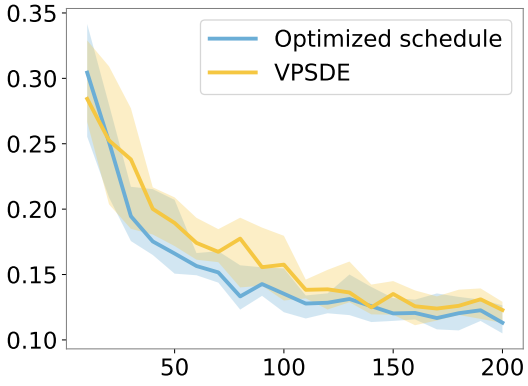


FIGURE 2 – Divergences KL empiriques (médiane et quartiles sur 10 exécutions) entre π_{data} et les distributions obtenues par l'Algorithme 1 (bleu) et le modèle VPSDE (jaune).

Résultats. Nous évaluons la performance de l'Algorithme 1 en considérant une distribution cible de données gaussienne $\pi_{\text{data}}^{(\text{corr})}$.

Sur la Figure 2, au fil des époques, nous affichons les divergences KL empiriques par rapport à la distribution générée via l'Algorithme 1 et par rapport à un algorithme *VPSDE* classique. Dès les premières époques, l'Algorithme 1 produit de meilleurs échantillons que le modèle *VPSDE* standard. Comme attendu, la valeur de a sélectionnée par l'Algorithme 1 tend à être décalée vers des valeurs positives avec une certaine stabilisation autour des valeurs optimales déjà observées sur la Figure 1.

Références

H. Chen, H. Lee, and J. Lu. Improved analysis of score-based generative modeling : User-friendly bounds under minimal smoothness assumptions. In International Conference on Machine Learning, pages 4735–4763. PMLR, 2023a.

S. Chen, S. Chewi, J. Li, Y. Li, A. Salim, and A. R. Zhang. Sampling is as easy as learning the score : theory for diffusion models with minimal data assumptions, 2023b.

G. Conforti, A. Durmus, and M. G. Silveri. Score diffusion models without early stopping : finite fisher information is all you need, 2023.

V. De Bortoli. Convergence of denoising diffusion models under the manifold hypothesis. arXiv preprint arXiv :2208.05314, 2022.

V. De Bortoli, J. Thornton, J. Heng, and A. Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. Advances in Neural Information Processing Systems, 34 :17695–17709, 2021.

S. Gong, M. Li, J. Feng, Z. Wu, and L. Kong. Diffuseq : Sequence to sequence text generation with diffusion models. In Proceedings of International Conference on Learning Representations, 2023.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems, 2020.

A. Hyvärinen and P. Dayan. Estimation of non-normalized statistical models by score matching. Journal of Machine Learning Research, 6(4), 2005.

H. Lee, J. Lu, and Y. Tan. Convergence of score-based generative modeling for general data distributions. In International Conference on Algorithmic Learning Theory, pages 946–985. PMLR, 2023.

A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv :2204.06125, 1(2) :3, 2022.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In F. Bach and D. Blei, editors, Proceedings of the 32nd International Conference on Machine Learning, volume 37 of Proceedings of Machine Learning Research, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In Advances in Neural Information Processing Systems, 2019.

Y. W. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In International Conference on Learning Representations, 2021.

P. Vincent. A connection between score matching and denoising autoencoders. Neural Computation, 23(7) :1661–1674, 2011. doi : 10.1162/NECO_a_00142.

L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui, and M.-H. Yang. Diffusion models : A comprehensive survey of methods and applications. ACM Computing Surveys, 56(4) :1–39, 2023.