

Quantification d'incertitude pour l'Approximation Stochastique

Stéphane CREPEY¹, Gersende FORT², Emmanuel GOBET³, Uladzislau STAZHYNSKI³

¹Laboratoire de Mathématiques et Modélisation d'Evry
91037 Evry, France

²Institut de Mathématiques de Toulouse
118 route de Narbonne, 31062 Toulouse Cedex, France

³Centre de Mathématiques Appliquées de l'Ecole Polytechnique
route de Saclay, 91128 Palaiseau, France

stephane.crepey@univ-evry.fr, gersende.fort@math.univ-toulouse.fr
emmanuel.gobet@polytechnique.edu, uladzislau.stazhynski@polytechnique.edu

Résumé – L'Approximation Stochastique est une procédure itérative pour le calcul d'un zéro θ^* d'une fonction non explicite mais définie comme une espérance. C'est par exemple un outil numérique pour le calcul du maximum de vraisemblance dans des modèles à données latentes "réguliers". Si la définition du modèle statistique est entachée d'une incertitude τ , dont on ne connaît qu'un a priori $d\pi(\tau)$, alors les zéros dépendent de τ et la question naturelle est d'explorer leur distribution lorsque $\tau \sim d\pi$. Dans ce papier, nous proposons un algorithme itératif basé sur un schéma d'Approximation Stochastique qui, à la limite, calcule $\theta^*(\tau)$ pour tout τ et produit une caractérisation de sa distribution; et nous énonçons des conditions suffisantes pour la convergence de cet algorithme.

Abstract – Stochastic Approximation is an iterative procedure for the computation of a root θ^* of a non explicit function defined as an expectation. It is for example a numerical tool for the computation of the Maximum Likelihood in "regular" latent variable models. When the definition of the statistical model is uncertain, depending on a quantity τ for which only a prior $\pi(d\tau)$ is known, then the roots also depend on τ ; a natural question is to explore their distribution when $\tau \sim d\pi$. In this paper, we propose a Stochastic Approximation-based algorithm which, in its limiting behavior, provides a computation of $\theta^*(\tau)$ for any τ and a characterization of its distribution; we also state sufficient conditions for the convergence of this algorithm.

1 Incertitude en Approximation Stochastique

L'Approximation Stochastique (AS) est une procédure itérative qui permet de répondre au problème numérique suivant : calcul d'un zéro d'une fonction $h : \mathbb{R}^q \rightarrow \mathbb{R}^q$ non explicite mais de la forme

$$h(\theta) := \int_{\mathbb{X}} H(x, \theta) d\mu(x) \quad (1)$$

où $d\mu$ est une mesure de probabilité sur \mathbb{X} (disons un ensemble mesurable de \mathbb{R}^d pour faire simple).

Afin de motiver ce cadre général, considérons le problème computationnel suivant issue de l'Apprentissage Statistique : estimation par maximum de vraisemblance dans un modèle à données cachées. Dans cet exemple, on cherche à résoudre

$$\operatorname{argmax}_{\theta \in \Theta} \log \int_{\mathbb{X}} f(x, Y_{1:N}; \theta) d\nu(x) \quad (2)$$

où $\Theta \subseteq \mathbb{R}^q$, $f(x, Y_{1:N}; \theta)$ est la densité de la loi jointe des données cachées x et des observations $Y_{1:N}$ dans le modèle statistique indexé par θ , et $d\nu$ est une mesure de domination sur l'espace des données cachées. Hormis dans les exemples jouets, l'intégrale n'est pas explicite. Si le modèle est suff-

isamment régulier, le calcul de ce maximum revient à chercher le zéro du gradient qui lui-même revient à résoudre

$$\theta \text{ tel que } \int_{\mathbb{X}} \partial_{\theta} \log f(x, Y_{1:N}; \theta) \frac{f(x, Y_{1:N}; \theta) d\nu(x)}{\int_{\mathbb{X}} f(u, Y_{1:N}; \theta) d\nu(u)} = 0.$$

Là encore, l'intégrale est non explicite pour les modèles les plus pertinents. On cherche donc bien le zéro d'une fonction non explicite et de la forme (1) en ayant posé

$$H(x, \theta) := \partial_{\theta} \log f(x, Y_{1:N}; \theta) \quad (3)$$

$$d\mu(x) \propto f(x, Y_{1:N}; \theta) d\nu(x); \quad (4)$$

ici, l'apprentissage se faisant pour un jeu d'observations $Y_{1:N}$ fixé, il n'est pas nécessaire de reporter la dépendance en ces observations dans la notation de H et μ . A noter que la probabilité $d\mu$ donnée par (4) n'est autre que la loi a posteriori des données cachées, sachant les observations dans le modèle statistique indexé par θ .

Etant donnée une suite de réels strictement positifs $\{\gamma_n, n \geq 0\}$ dits *pas d'apprentissage*, et un point initial $\theta_0 \in \mathbb{R}^q$, l'AS définit une suite de points $\{\theta_n, n \geq 1\}$ de \mathbb{R}^q par

$$\theta_{n+1} = \theta_n - \gamma_{n+1} H(X_{n+1}, \theta_n) \quad (5)$$

où $\{X_n, n \geq 1\}$ sont des variables aléatoires de même loi $d\mu$ et indépendantes [1]. La procédure remplace donc l'espérance

$h(\theta_n)$ par une approximation Monte Carlo à un seul point; dans le cas de l'exemple (2)-(4), elle remplace le gradient de la log-vraisemblance des observations par le gradient de la log-vraisemblance complète dans laquelle la donnée manquante x est fixée à une réalisation de sa loi a posteriori.

Des conditions suffisantes sur les pas d'apprentissage, la fonction H et la loi $d\mu$ entraînant la convergence de chaque trajectoire $\{\theta_n, n \geq 0\}$ de l'algorithme vers un zéro θ^* de h (voir (1)) dans le cas basique où $X_n \sim d\mu$ ou le cas plus compliqué où X_n est tiré sous une loi biaisée, peuvent être trouvées dans [1, 6] par exemple.

La quantification d'incertitude en AS correspond au cas où une incertitude sur la distribution $d\mu$ et/ou la fonction H veut être prise en compte, et l'on souhaite quantifier l'influence de cette méconnaissance sur les points limites de schémas d'AS associés. Nous allons dans la suite considérer le cas où (i) cette incertitude s'exprime par une dépendance en une quantité notée $\tau \in \mathcal{T}$, de sorte que l'on pose

$$h(\theta, \tau) := \int_{\mathcal{X}} H(x, \theta, \tau) \mu(\tau, dx); \quad (6)$$

et (ii) sur cette quantité τ , on ne connaît qu'une loi a priori $\pi(d\tau)$. L'objectif est de trouver un algorithme, et d'en prouver le bien-fondé, pour répondre conjointement aux questions

Q1 trouver une fonction $\theta^* : \mathcal{T} \rightarrow \mathbb{R}^q$, mesurable et telle que pour π -presque-tout τ , $h(\theta^*(\tau), \tau) = 0_{\mathbb{R}^q}$,

Q2 caractériser la loi de $\theta^*(\tau)$ lorsque $\tau \sim d\pi$.

2 Procédures algorithmiques

2.1 Procédure naïve I

Pour traiter la question **Q1**, une première idée est de calculer pour tout τ , une solution $\theta^*(\tau)$ comme la limite de l'algorithme (5) appliqué avec $H(\cdot, \theta_n) \leftarrow H(\cdot, \theta_n, \tau)$ et $X_{n+1} \sim \mu(\tau, dx)$. A l'exception du cas où Θ est fini de cardinal raisonnable, cette idée est trop coûteuse pour être intéressante.

Pour traiter la question **Q2**, on peut (i) tirer M points τ_1, \dots, τ_M indépendants de loi $\pi(d\tau)$, (ii) pour chacun, approcher $\theta^*(\tau_i)$ par la procédure (5), (iii) puis conclure en construisant la loi empirique associée. L'inconvénient de cette approche est qu'elle fait appel à un double Monte Carlo : un Monte Carlo externe pour l'échantillonnage des points τ_i , et un Monte Carlo interne pour le calcul d'une approximation de $\theta^*(\tau_i)$.

2.2 Procédure naïve II - décomposition en chaos

Puisque la question **Q1** consiste à identifier une fonction, une seconde approche est de restreindre la recherche à des espaces de fonctions à base dénombrable et apprendre un élément de cet espace en apprenant ses coefficients sur la base. Ici, nous considérons l'espace des fonctions $f = (f_1, \dots, f_q) : \mathcal{T} \rightarrow \mathbb{R}^q$ de carré intégrable sous π . Dans cette optique, soit une famille de fonctions $\{B_i, i \geq 0\}$, $B_i : \mathcal{T} \rightarrow \mathbb{R}$, famille orthonormée pour

le produit scalaire

$$\langle b_1, b_2 \rangle_\pi := \int_{\mathcal{T}} b_1(\tau) b_2(\tau) \pi(d\tau);$$

voir [3] pour des exemples de telles familles. Alors nous avons

$$f(\cdot) = \sum_{i \geq 0} u_i B_i(\cdot), \quad (7)$$

où $u_i \in \mathbb{R}^q$ et est donné par

$$u_i := \int_{\mathcal{T}} f(\tau) B_i(\tau) d\pi(\tau). \quad (8)$$

Nous insistons sur le fait que dans (7), les coefficients u_i sont à valeur \mathbb{R}^q et les éléments de la base B_i sont à valeur dans \mathbb{R} .

Dans ce contexte dit de *décomposition en chaos*, on peut répondre à **Q1** en (i) tronquant la recherche à l'espace des fonctions engendré par (B_0, \dots, B_m) pour un entier m choisi par l'utilisateur, (ii) calculer $\widehat{u}_0, \dots, \widehat{u}_m$ par une approximation Monte Carlo du produit scalaire (8) à l'aide de tirages τ_1, \dots, τ_M sous π , qui elle-même nécessite (iii) d'avoir calculé une approximation de $\theta^*(\tau_k)$ pour tout $k \in \{1, \dots, M\}$ par exemple par l'algorithme (5). Cette approche a au moins trois défauts: d'abord elle présente une erreur d'estimation de θ^* due à la projection sur un espace fonctionnel de dimension finie (choix de m); ensuite, elle présente une erreur Monte Carlo dans l'approximation du produit scalaire (8); enfin, elle nécessite elle aussi un double Monte Carlo (voir section 2.1). Cet algorithme travaillant à m et M fixés, les deux sources de biais d'approximation n'ont aucune chance de disparaître dans une certaine asymptotique du déroulement de l'algorithme.

La décomposition en chaos a l'avantage de donner une réponse simple à la question **Q2**. Lorsque la fonction B_0 est la fonction constante 1, on déduit de (7) et de la π -orthogonalité de la famille $\{B_i, i \geq 0\}$ que l'espérance et la matrice de covariance-covariance de $\theta^*(\tau)$, $\tau \sim d\pi$ sont données par u_0 et $\sum_{i \geq 1} u_i u_i^*$; dans le cas d'une famille polynomiale $\{B_i, i \geq 0\}$ (voir [8, Appendix C]), les moments d'ordre plus élevés se déduisent des coefficients $\{u_i, i \geq 0\}$; quant à des statistiques plus générales (quantiles, intervalles de confiance, etc), elles peuvent toujours être approchées à l'aide de $\sum_{i \geq 0} u_i B_i(\tau_k)$, où $\{\tau_k, 1 \leq k \leq M\}$ sont des réalisations indépendantes de la loi $d\pi$.

Ainsi, l'approche par décomposition en chaos répond aux questions **Q1** et **Q2** dès lors que l'on dispose d'une méthode numérique efficace pour le calcul d'une approximation de la famille dénombrable (non nécessairement finie) $\{u_i, i \geq 0\}$ dans (7).

2.3 L'algorithme UQSA

L'algorithme UQSA, acronyme de Uncertainty Quantification for Stochastic Approximation, que nous proposons, produit une telle approximation des coefficients $\{u_i, i \geq 0\}$. L'intuition est la suivante : résoudre en $\{u_i^*, i \geq 0\}$ l'équation

$$h \left(\sum_{i \geq 0} u_i^* B_i(\tau), \tau \right) = 0_{\mathbb{R}^q} \quad (9)$$

pour π -presque tout τ , est équivalent à résoudre

$$\forall i \geq 0, \int_{\mathcal{T}} h \left(\sum_{j \geq 0} u_j^* B_j(\tau), \tau \right) B_i(\tau) \pi(d\tau) = 0_{\mathbb{R}^q},$$

soit encore en utilisant (6) et Fubini, pour tout $i \geq 0$,

$$\int_{\mathcal{T} \times \mathcal{X}} H \left(x, \sum_{j \geq 0} u_j^* B_j(\tau), \tau \right) B_i(\tau) \pi(d\tau) \mu(\tau, dx) = 0_{\mathbb{R}^q}. \quad (10)$$

Plutôt que d'associer à ce système de conditions, un algorithme d'AS mettant à jour à chaque itération une quantité homogène à une suite dénombrable (algorithme irréaliste), UQSA met à jour à l'itération k , les m_k premiers coefficients u_0, \dots, u_{m_k-1} où $\{m_k, k \geq 0\}$ est une suite d'entiers telle que $\lim_k m_k = +\infty$. De plus, l'approximation Monte Carlo du champs moyen (10) est faite à l'aide de M_k tirages. Ainsi, dans l'asymptotique du nombre d'itérations (i.e. quand $k \rightarrow +\infty$), UQSA évite l'écueil de la troncation de l'espace fonctionnel (m fixe) commentée en section 2.2, mais peut rester dans l'esprit de l'AS où la fonction objectif est approchée par Monte Carlo à l'aide d'un seul tirage ($M_k = 1$ possible).

Algorithm 1 The UQSA algorithm

Input: Suites $\{\gamma_k, k \geq 1\}$, $\{m_k, k \geq 0\}$, $\{M_k, k \geq 1\}$, $K \geq 1$, $\{u_i^0, i = 0, \dots, m_0\}$.

for $k = 1, \dots, K$ **do**

for $s = 1, \dots, M_k$ **do**

Indépendamment du passé, simuler (τ_k^s, X_k^s) sous la loi $\pi(d\tau)\mu(\tau, dx)$

calculer $\hat{f}_k^s = \sum_{j=0}^{m_k-1} u_j^{k-1} B_j(\tau_k^s)$

end for

Pour tout $i > m_{k-1}$, poser $u_i^k = 0$

for $i = 0, \dots, m_k$ **do**

$u_i^k = u_i^{k-1} - \gamma_k M_k^{-1} \sum_{s=1}^{M_k} H \left(X_k^s, \hat{f}_k^s, \tau_k^s \right) B_i(\tau_k^s)$

end for

end for

return le vecteur $\{u_i^K, i = 0, \dots, m_K\}$.

end

3 Bien-fondé de UQSA

L'algorithme UQSA est une version simplifiée de la méthode plus générale présentée en [5, section 2.3]; cette simplification se fait au prix de conditions suffisantes fortes pour la convergence. On trouvera en [5, section 3.1] un jeu de conditions plus faibles que celles énoncées ici.

Pour une fonction $\theta : \mathcal{T} \rightarrow \mathbb{R}^q$ de carré intégrable sous π (l'espace L^2_π), définissons sa norme

$$\|\theta\|_\pi := \left(\int_{\mathcal{T}} |\theta(\tau)|^2 \pi(d\tau) \right)^{1/2}.$$

Notons \mathcal{S} l'ensemble des solutions de la question **Q1**. Nous étudions la convergence de l'algorithme dans L^p et presque-sûrement sous les conditions suivantes. **A1** l'ensemble des solutions \mathcal{U}^* de (9) est un espace compact et non-vide de l'espace des suites à valeur \mathbb{R}^q de norme de carré sommable. **A2** les paramètres d'implémentation γ_k, m_k, M_k vérifient

$$\begin{aligned} \sum_k \gamma_k &= +\infty, & \sum_k \gamma_k^{1+\kappa} &< \infty, \\ \sum_k \gamma_k^2 Q_{m_k} M_k^{-1} &< \infty & \sum_k \gamma_k^{1-\kappa} q_{m_k} &< \infty \end{aligned}$$

où $\kappa \in [0, 1]$ et

$$q_m := \sup_{u^* \in \mathcal{U}^*} \sum_{i > m} |u_i^*|^2 \quad Q_m := \sup_{\tau \in \mathcal{T}} \sum_{i \leq m} |B_i(\tau)|^2.$$

A3 Il existe une constante C telle que pour tout $\theta \in \mathbb{R}^q$

$$\sup_{\tau \in \mathcal{T}} \int_{\mathcal{X}} |H(\theta, x, \tau)|^2 \mu(\tau, dx) \leq C (1 + |\theta|^2).$$

A4 L'application $\theta \mapsto h(\theta(\cdot), \cdot)$ de l'espace des fonctions de L^2_π dans lui-même, est continue. **A5** Pour π -presque tout τ , pour tout $\theta, \theta^* \in \mathbb{R}^q$ tels que $h(\theta, \tau) \neq 0$ et $h(\theta^*, \tau) = 0$, on a $(\theta - \theta^*)' h(\theta, \tau) > 0$. **A6** Pour tout B , il existe une constante C_B telle que pour tout $\theta \in L^2_\pi$ et $\theta^* \in \mathcal{S}$ et vérifiant $\|\theta - \theta^*\|_\pi \leq B$, on a

$$\int_{\mathcal{T}} (\theta(\tau) - \theta^*(\tau))' h(\theta^*(\tau), \tau) \pi(d\tau) \geq C_B \min_{\bar{\theta} \in \mathcal{S}} \|\theta - \bar{\theta}\|_\pi^2.$$

A1 autorise en particulier la non unicité des solutions au problème posé. La première condition de **A2** est classique en AS : elle assure que l'algorithme en temps discret hérite du comportement en temps long d'une ODE associée; les deux suivantes s'interprètent comme des conditions sur la sommabilité des erreurs dues à l'AS; tandis que la dernière guide la vitesse à laquelle la troncation (en m) doit être relâchée pour que l'algorithme converge vers un point $u^* \in \mathcal{U}^*$. La condition **A3** est une condition assez forte pour assurer notamment la stabilité d'un algorithme d'AS qui n'est pas reprojété à chaque itération (voir [1, Eq (1.10.4) Partie 2] pour une condition analogue dans le cas usuel de l'AS (5)) - elle peut être affaiblie si on rajoute une étape de projection dans l'algorithme UQSA (voir [5]). **A5** est en écho à l'inégalité clé à la base de l'existence d'une fonction de Lyapunov pour les algos d'AS (voir [1, Eq. (1.6.1) Partie 2]) et **A6** est une version plus forte, intégrée sous la loi de l'incertitude $d\pi$ et avec une minoration minimale (voir [1, Eq. (1.10.5) Partie 2]); la condition **A6** n'est pas nécessaire si on introduit une projection dans l'algorithme UQSA (voir [5]).

Posons $\theta^{(k)}(\cdot) := \sum_{i \geq 0} u_i^k B_i(\cdot)$, la fonction de L^2_π associée à la suite u^k construite par l'algorithme UQSA à l'itération k (on rappelle que $u_i^k = 0$ pour tout $i > m_k$). Nous prouvons dans [5, Section 3.3.] le théorème suivant. Il établit que (i) pour presque toutes les trajectoires de l'algorithme UQSA, il existe une solution $\theta^* \in \mathcal{S}$ telle que la trajectoire converge vers cette solution, (ii) la convergence a lieu aussi dans L^p ; (iii) UQSA est stable i.e. bornitude des trajectoires dans L^p .

Théorème 1 *Supposons A1 à A5. Alors pour toute solution $\theta^* \in \mathcal{S}$, $\lim_k \|\theta^{(k)} - \theta^*\|_\pi$ existe et est finie avec probabilité un; et nous avons*

$$\sup_k \mathbb{E} \left[\|\theta^{(k)} - \theta^*\|_\pi^2 \right] < \infty.$$

Si de plus A6 est vérifiée, alors il existe une v.a. θ^ à valeur dans \mathcal{S} telle que avec probabilité un*

$$\lim_k \|\theta^{(k)} - \theta^*\|_\pi = 0,$$

et pour tout $p \in]0, 2[$

$$\lim_k \mathbb{E} \left[\|\theta^{(k)} - \theta^*\|_\pi^p \right] = 0.$$

Une illustration numérique de cette convergence et une discussion du rôle des paramètres d'implémentation γ_k, m_k, M_k peuvent être trouvées dans [5, Section 5]. L'ingrédient clé de la preuve de convergence est la décroissance d'une fonction de Lyapunov, qui ici est $\|\theta^{(k)} - \theta^*\|_\pi^2$ pour une solution $\theta^* \in \mathcal{S}$.

4 Conclusion

Nous avons proposé un algorithme itératif, de type Approximation Stochastique, qui résoud conjointement les questions **Q1** et **Q2**; l'ingrédient clé est l'utilisation de la décomposition en chaos.

La procédure UQSA est un algorithme d'AS qui à chaque itération met à jour un vecteur de dimension finie, pour répondre à un problème d'AS en dimension infinie (voir la question **Q1**). On trouve dans la littérature des algorithmes qui traitent de cette question, et nous concluons ce papier par une comparaison de UQSA à ces méthodes. Tout d'abord un exemple très répandu en Apprentissage Statistique, relevant de la recherche d'une fonction, est celui de la régression fonctionnelle, sur un espace de Hilbert de type RKHS (Reproducing Kernel Hilbert spaces). Dans ce contexte, il s'agit de minimiser sur cet espace de Hilbert un critère somme d'une attache aux N données et d'un terme de pénalité contraignant la fonction à estimer. Dès lors que ces termes ont des bonnes propriétés, le Théorème de Représentation (voir e.g. [10, Théorème 4.2]) montre que toute solution est une combinaison linéaire finie du noyau évalué en les points de régression; par suite, déterminer cette solution revient à déterminer N coefficients. Le problème, bien que formulé en dimension infinie, se ramène pour sa partie computationnelle à la dimension finie.

L'AS en dimension infinie est formulée sous la forme

$$\text{trouver } \theta^* \text{ tq } \int_{\mathcal{X}} H(x, \theta^*) \mu(dx) = 0_{\mathbb{R}^q}, \quad (11)$$

où θ^* vit dans un espace fonctionnel. Les articles [11, 13, 2] étudient l'AS dans des espaces de Hilbert pour des fonctions H de forme spécifique (à variable séparée typiquement). Ces travaux, bien que très intéressants d'un point de vue théorique, étudient des algorithmes itératifs type AS qui mettent à jour une quantité de dimension infinie à chaque itération, sans commenter comment cela se fait en pratique.

L'idée d'augmenter la taille du vecteur mis à jour à chaque itération de l'algorithme AS a déjà été proposée et étudiée, notamment là aussi pour des formes particulières de H [7, 9, 12]. L'algorithme TRMP de [4] reprend cette idée, chaque itération nécessitant une projection sur un sous-espace de dimension finie de l'espace de Hilbert d'intérêt; l'étude théorique est faite sous la condition que cette projection est exacte, ce qui est relativement irréaliste (voir le calcul de u_i par exemple dans (8), qui n'est pas exact en pratique, mais peut être approché par Monte Carlo comme dans UQSA).

Enfin, toutes ces approches "dimension croissante" répondent au problème (11) par une procédure type (5) et nécessitent donc à chaque itération (disons itération n) de calculer H évaluée en le tirage X_{n+1} et en un point de dimension infinie $\theta^{(k)}$ (une fonction). Ce n'est pas le cas de UQSA qui repose sur H évaluée en le tirage X_{n+1} , en l'incertitude τ_{n+1} et en $\theta^{(k)}(\tau_{k+1})$ - l'évaluation de la fonction en un point (soit un calcul dans \mathbb{R}^q); le coût computationnel de USQA reste donc réaliste.

References

- [1] A. Benveniste, M. Metivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics*. Springer-Verlag, 1990.
- [2] N. Berman and A. Shwartz. Abstract stochastic approximations and applications. *Stochastic Processes and their Applications*, 31(1):133–149, 1989.
- [3] C. Canuto, M.Y. Hussaini, A. Quarteroni, and T.A. Zang. *Spectral Methods: Fundamentals in Single Domains*. Springer Verlag, 2006.
- [4] X. Chen and H. White. Asymptotic properties of some projection-based Robbins-Monro procedures in a Hilbert space. *Studies in Nonlinear Dynamics and Econometrics*, 6(1), 2002.
- [5] S. Crépey, Gersende Fort, E. Gobet, and U. Staszynski. Uncertainty Quantification for Stochastic Approximation Limits Using Chaos Expansion. preprint, 2019.
- [6] G. Fort, E. Moulines, A. Schreck, and M. Vihola. Convergence of Markovian Stochastic Approximation with discontinuous dynamics. *SIAM J. Control Optim.*, 54(2):866–893, 2016.
- [7] L. Goldstein. Minimizing noisy functionals in Hilbert space: An extension of the Kiefer-Wolfowitz procedure. *Journal of Theoretical Probability*, 1(2):189–204, 1988.
- [8] O. Le Maître and O. Knio. *Spectral Methods for Uncertainty Quantification. With Applications to Computational Fluid Dynamics, Scientific Computation*. Springer Science & Business Media, 2010.
- [9] R. Nixdorf. An invariance principle for a finite dimensional stochastic approximation method in a Hilbert space. *Journal of Multivariate Analysis*, 15(2):252–260, 1984.
- [10] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.
- [11] H. Walk. An invariance principle for the Robbins-Monro process in a Hilbert space. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 39(2):135–150, 1977.
- [12] G. Yin. On H-valued stochastic approximation: Finite dimensional projections. *Stochastic Analysis and Applications*, 10(3):363–377, 1992.
- [13] G. Yin and Y.M. Zhu. On H-valued Robbins-Monro processes. *Journal of Multivariate Analysis*, 34(1):116–140, 1990.