

Plans d'Expériences Optimaux Pour Réseaux De Neurones

Sébastien ISSANCHOU - NETRAL

Jean-Pierre GAUCHI - INRA

Introduction (1/2)

- **Contexte et Objectifs :**

- Partenariat de recherche : Netral, CEA, IFP, Peugeot, Rhodi
- Développement d'un logiciel de construction de **plans d'expériences pour réseaux de neurones**
- Objectifs similaires aux plans pour surfaces de réponse :
 - Optimiser la **précision des prévisions** du modèle
 - **Caractère économique** marqué du plan d'expériences

- **Etude présentée :**

- **Confirmer l'intérêt des plans proposés sur un « cas idéal »**
 - Quantifier la précision des prévisions en fonction du plan utilisé
 - Études par la voie de **simulations de Monte-Carlo** (pas de théorie)

Introduction (2/2)

- Démarche mise en place :

- Remplacement du réseau de neurones dans le cadre de **modélisation statistique paramétrique non-linéaire** :

- Modèle statistique : $(\mathcal{Y}, \mathcal{P}(\mathbf{w}^*))$ avec $\mathcal{P}(\mathbf{w}^*) \equiv \mathcal{N}(E(y), \Sigma)$
- Un réseau de neurones, **approximateur universel**, est utilisé pour modéliser l'espérance des observations y en fonction des facteurs \mathbf{x} , comme cela peut être fait avec un modèle de connaissance, un polynôme, ... :

$$E(y) = RN(\mathbf{x}, \boldsymbol{\theta}^*)$$

- Utilisation de la théorie des **plans d'expériences optimaux**

Plans optimaux en linéaire (1/3)

- Répondre aux limites des plans classiques :
 - Domaine expérimental tronqué
 - Réparation de plans d'expériences
 - **Modèle particulier** (linéaire)
- Retrouver une certaine **optimalité** du plan d'expériences
- Principales propriétés statistique recherchées :
 - **Précision des estimateurs** (grandeurs scalaires de $V(\hat{\theta})$)
 - ⇒ « Minimiser » la distance (au sens de $\| \cdot \|_2$) entre θ^* et $\hat{\theta}$
 - (\Leftrightarrow) **Région de confiance des paramètres** (ellipsoïde)
 - ⇒ « Minimiser » l'ensemble des valeurs possibles pour θ^*

Plans optimaux en linéaire (2/3)

- Critères d'optimalité et interprétation géométrique :

- Modèle statistique linéaire : θ_2

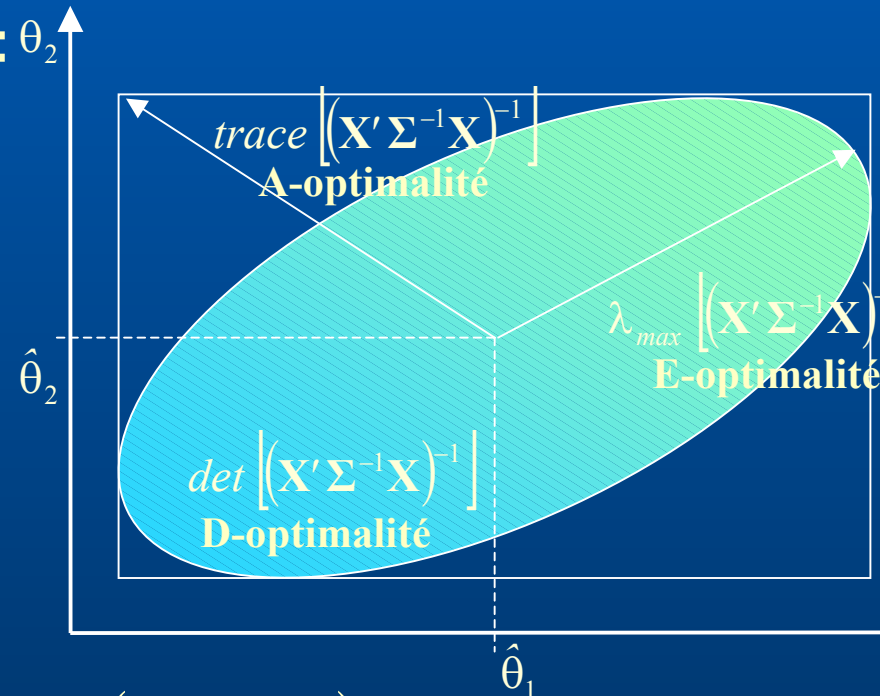
$$y = \mathbf{X}\boldsymbol{\theta}^* + \varepsilon, \quad \varepsilon \rightarrow \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

- EMC (non biaisé, efficace) :

$$\hat{\boldsymbol{\theta}} \rightarrow \mathcal{N}\left(\boldsymbol{\theta}^*, \underline{(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}}\right)$$

- Région de confiance :

$$\underline{(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})} \leq ps_{N-p}^2 F_{\alpha}(p, N-p)$$



Plans optimaux en linéaire (3/3)

- Pour un objectif prévisionnel ...
 - Les **critères dans l'espace des observations** semblent plus naturels que les critères dans l'espaces des paramètres
 - On s'intéresse à la **variance des prévisions** :
 - G-optimalité : critère d'optimalité au sens de minmax
 - Q-optimalité (ou V) : critère d'optimalité au sens de la moyenne
- ... Malheureusement :
 - Difficulté d'optimisation
 - Pas ou peu d'algorithmes de construction de tels plans
 - La **D-optimalité** apporte une solution détournée (**TEG**)

Plans D-optimaux en non-linéaire (1/7)

- Il ne s'agit pas d'une véritable optimalité :

- On utilise une **approximation asymptotiquement convergente**

- **Linéarisation du modèle non-linéaire** par développement Taylor limité au premier terme au voisinage d'une valeur approchée θ^0 de θ^* :

$$\eta(\mathbf{x}_i, \boldsymbol{\theta}) = \eta(\mathbf{x}_i, \boldsymbol{\theta}^0) + \sum_{j=1}^p \frac{\partial \eta(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \theta_j} \Big|_{\theta_j = \theta_j^0} (\theta_j - \theta_j^0)$$

- Avec θ^0 fixé, on a un modèle linéaire : $\{\mathbf{Z}_{\theta^0}\}_{ij} = \frac{\partial \eta(\mathbf{x}_i, \boldsymbol{\theta})}{\partial \theta_j} \Big|_{\theta_j = \theta_j^0} \equiv \{\mathbf{X}_{\theta^0}\}_{ij}$
- **Optimalité locale** : D(θ^0)-optimalité

Plans D-optimaux en non-linéaire (2/7)

- Pour une **distance finie** (outre l'approximation $\theta^0 \equiv \theta^*$), linéarisation du modèle induit ...
- ... Du point de vue de la région de confiance :
 - On utilise un **ellipsoïde**
 - En réalité, les régions exactes sont de **forme irrégulière** peuvent être **disjointes**, voire **ouvertes** (volume infini)
- ... Du point de vue de la distribution de l'EMC :
 - On utilise une **distribution gaussienne** $\hat{\theta} \rightarrow \mathcal{N}(\theta^*, (\mathbf{Z}'\Sigma^{-1}\mathbf{Z})^{-1})$
 - En réalité, $E(\hat{\theta}) \neq \theta^*$ et $(\mathbf{Z}'\Sigma^{-1}\mathbf{Z})^{-1}$ n'est qu'une **approximation d'ordre un** de la matrice de variance-covariance de l'EMC

Plans D-optimaux en non-linéaire (3/7)

- Malgré ces approximations, **résultats très probants** dans le cas des **modèles non-linéaires de connaissance**
- **Démarche séquentielle** possible pour accroître robustesse du plan vis-à-vis de l'approximation $\theta^0 \equiv \theta^*$:
 - Utilisation de $\hat{\theta}$ comme valeur approchée de θ^*
 - Alternance des phases de planification d'expérience D($\hat{\theta}$)-optimales et d'estimation des paramètres pour affiner l'approximation $\hat{\theta} \equiv \theta^*$
 - Initialisation de la stratégie séquentielle par un plan de type « space filling » pour obtenir une première estimation $\hat{\theta}^{(1)}$ de θ^*

Plans D-optimaux en non-linéaire (4/7)

- **Exemple pour modèle de connaissance :**

- Tiré de [Carr, 1960 – Ind. & Eng. Chem.]

- Isomérisation du N-pentane (A) en iso-pentane (B) : $A \xrightarrow{H_2 / Cat.}$

- Réponse observée : vitesse de la réaction (y)

- 3 facteurs : pressions partielles en H₂ (x₁), A (x₂) et B (x₃)

- Modèle de type Langmuir-Hinshelwood-Hougen-Watson :

$$y_i = \frac{\theta_1 * \theta_3 * (x_2 - x_3 / 1.632)}{1 + \theta_2 * x_1 + \theta_3 * x_2 + \theta_4 * x_3} + \varepsilon_i, \quad \varepsilon_i \rightarrow \mathcal{N}(0, \sigma^2)$$

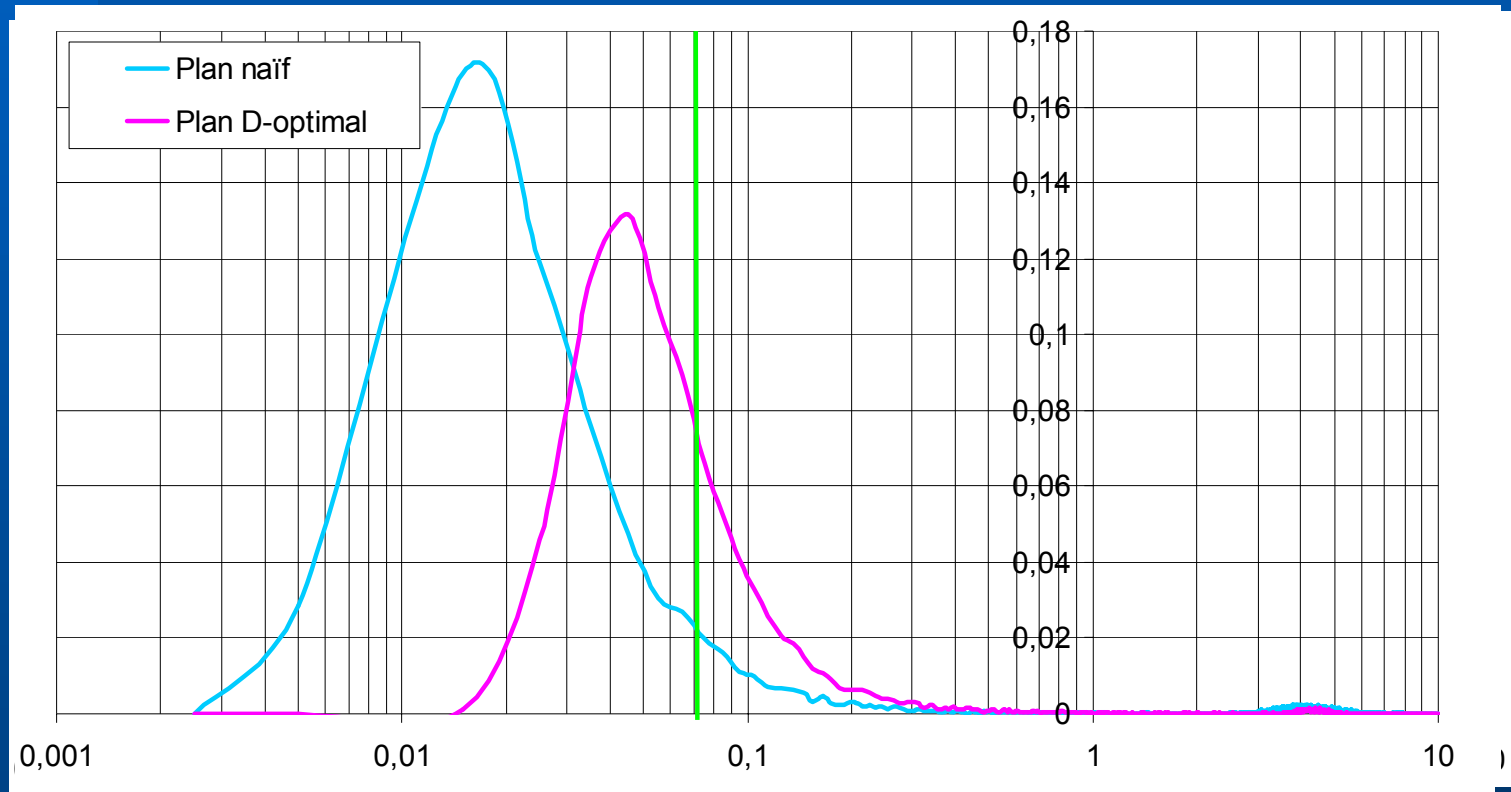
- 24 expériences publiées par Carr

Plans D-optimaux en non-linéaire (5/7)

- **Comparaison plan naïf vs. plan D-optimal séquentiel :**
 - **Plan naïf : 24 expériences publiées par Carr**
 - **Plan D-optimal séquentiel (sur le même domaine d'action) :**
 - 12 expériences aléatoires parmi les 24 de Carr
 - estimation des paramètres : $\hat{\theta}^{(1)}$
 - + 12 expériences supplémentaires $D(\hat{\theta}^{(1)})$ -optimales
 - **Evaluation des propriétés statistiques de l'EMC par simulations de Monte-Carlo (10 000 simulations) :**
 - Fonction de densité de probabilité de l'EMC
 - Calcul des 2 premiers moments, modes, ...

Plans D-optimaux en non-linéaire (6/7)

- Fonction de densité de probabilité pour $\hat{\theta}_H$:



Plans D-optimaux en non-linéaire (7/7)

- Résumés statistiques de la distribution de l'EMC :

Paramètre	Valeur théorique	Moyenne EMC		Biais EMC		Ecart-type EMC		Mode « principal »	
		naïf	D	naïf	D	naïf	D	naïf	D
θ_1	35.92	40.5	36.0	4.53	0.26	16.3	2.34	35.5	35.5
θ_2	0.071	1.62	0.50	1.55	0.43	2.13	1.24	0.015	0.045
θ_3	0.038	0.91	0.27	0.87	0.23	1.23	0.67	0.007	0.025
θ_4	0.167	3.72	1.16	3.55	0.99	4.72	2.81	0.035	0.105
σ^2	0.20	0.20	0.20	0.000	0.000	0.08	0.07	0.20	0.20

- Temps de calcul pour les 10 000 régressions :
 - **1422 secondes** pour le plan naïf
 - **419 secondes** pour le plan D-optimal (non-linéarité + faible)

Plans D-optimaux pour réseau de neurones (1/8)

- Étude de l'apport des plans D-optimaux sur un cas idéal
 - Utilisation d'un **modèle non biaisé**, i.e., d'une loi de distribution des observations gaussienne dont la moyenne est exactement un réseau à 2 neurones cachés et fonction de 2 facteurs expérimentaux \in au domaine d'action Ξ :

$$y_i = \theta_0 + \theta_1 \tanh(\theta_2 + \theta_3 x_i^{(1)} + \theta_4 x_i^{(2)}) + \theta_5 \tanh(\theta_6 + \theta_7 x_i^{(1)} + \theta_8 x_i^{(2)}) + \varepsilon_i \quad \varepsilon_i \rightarrow \mathcal{N}(0, \sigma^2)$$

- **Comparaison de plans aléatoires et de plans D-optimaux** :
 - **Précision du prédicteur** (EQM(x)) obtenu après estimation des paramètres de la loi (i.e., les poids du réseaux et la variance du bruit) par **simulations de Monte Carlo**

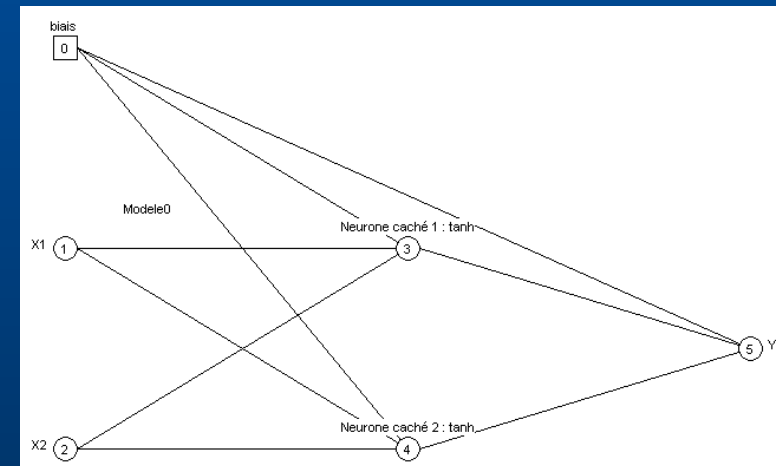
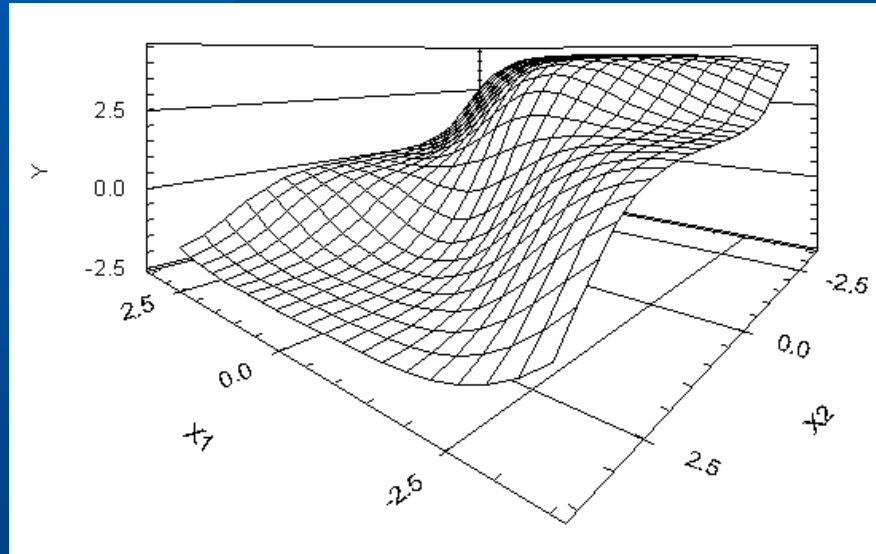
Plans D-optimaux pour réseau de neurones (2/8)

- **Modèle « théorique » du processus :**

- Tiré de [Rivals & coll., 2000 – Neural Networks] :

$$y_i = 1 + \tanh(x_i^{(1)} - x_i^{(2)}) - 2 \tanh(x_i^{(1)} + x_i^{(2)}) + \varepsilon_i \quad \varepsilon_i \rightarrow \mathcal{N}(0, 0.1)$$

$$-3 \leq x^{(1)} \leq 3 \quad \text{et} \quad -3 \leq x^{(2)} \leq 3$$

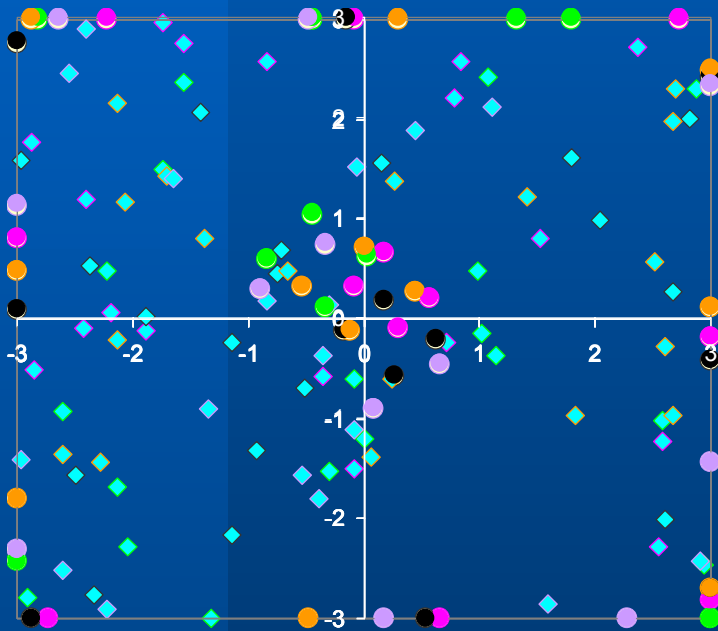


Plans D-optimaux pour réseau de neurones (3/8)

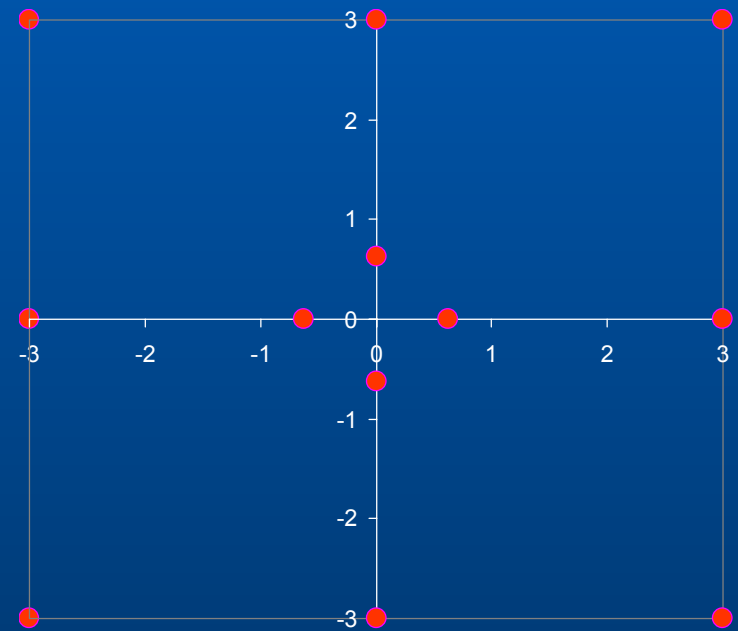
- **5 plans aléatoires vs. 5 plans D-optimaux :**
 - Pour les plans aléatoires : (18+12) expériences aléatoires
 - Pour les 5 plans D-optimaux : 18 expériences aléatoires
+ 12 expériences D($\hat{\theta}^{(1)}$)-optimales
- ⇒ comparaison de l'information apportée par les deux types d'expériences différentes suivant la stratégie aléatoire ou D-optimale séquentielle
- Pour chacun des plans : 10 000 simulations
- Comparaison de l'EQM obtenue en chacun des points (validation) d'une grille régulière (21x21)

Plans D-optimaux pour réseau de neurones (4/8)

- Visualisation des plans D-optimaux :



Plans $D(\hat{\theta}^{(1)})$ -optimaux

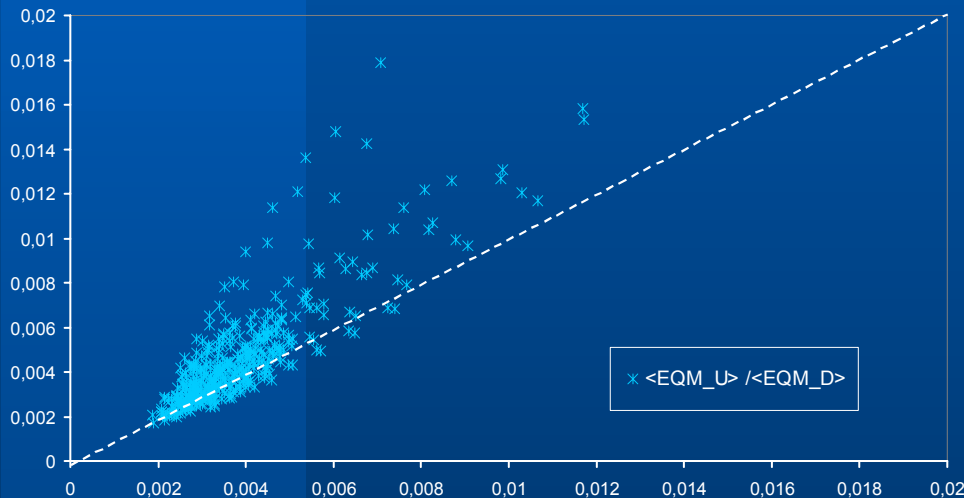


Plan $D(\theta^*)$ -optimal

Plans D-optimaux pour réseau de neurones (5/8)

$$\frac{1}{5} \sum_{i=1}^5 EQM_U^i(x^{(1)}, x^{(2)}) = f\left(\frac{1}{5} \sum_{i=1}^5 EQM_D^i(x^{(1)}, x^{(2)})\right)$$

(moyennes sur le type de plan)



$$C = 100 \frac{\langle EQM_U^1(\mathbf{x}) \rangle_{\Xi} - \langle EQM_D^1(\mathbf{x}) \rangle_{\Xi}}{\langle EQM_U^1(\mathbf{x}) \rangle_{\Xi}}$$

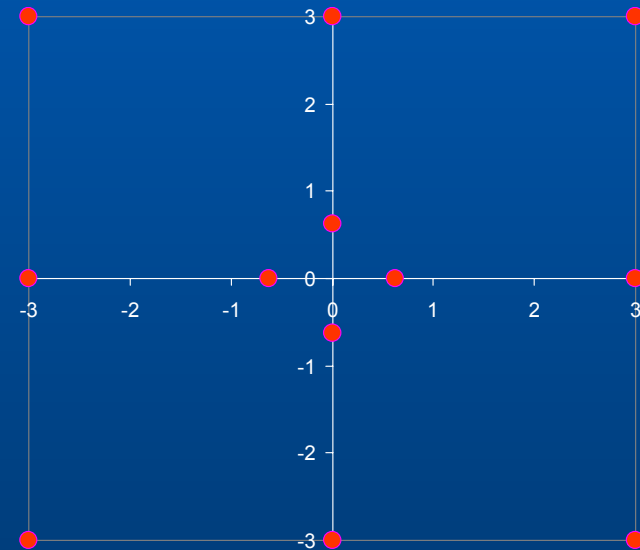
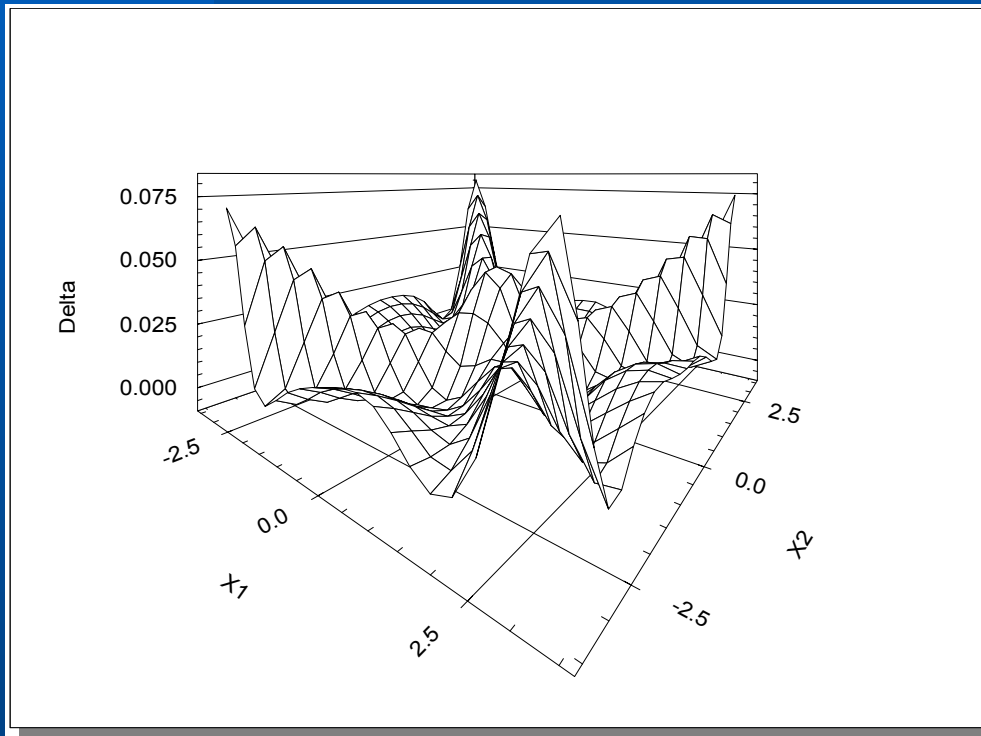
(moyennes sur le domaine d'action)

Plans	$\langle EQM(\mathbf{x}) \rangle_{\Xi}$		C
	U	D	
U1 / D1	4.7 10^{-3}	3.9 10^{-3}	+ 17 %
U2 / D2	4.2 10^{-3}	3.6 10^{-3}	+ 14 %
U3 / D3	4.1 10^{-3}	3.3 10^{-3}	+ 18 %
U4 / D4	5.3 10^{-3}	4.5 10^{-3}	+ 14 %
U5 / D5	4.7 10^{-3}	3.8 10^{-3}	+ 20 %
Moyenne	4.6 10^{-3}	3.8 10^{-3}	+ 17 %

⇒ Les plans D-optimaux sont toujours meilleurs que les plans aléatoires respectifs : **précision accrue de 17 % en moyenne**

Plans D-optimaux pour réseau de neurones (6/8)

- **Tracé de** $\Delta(x^{(1)}, x^{(2)}) = \frac{1}{5} \sum_{i=1}^5 (EQM_U^i(x^{(1)}, x^{(2)}) - EQM_D^i(x^{(1)}, x^{(2)}))$



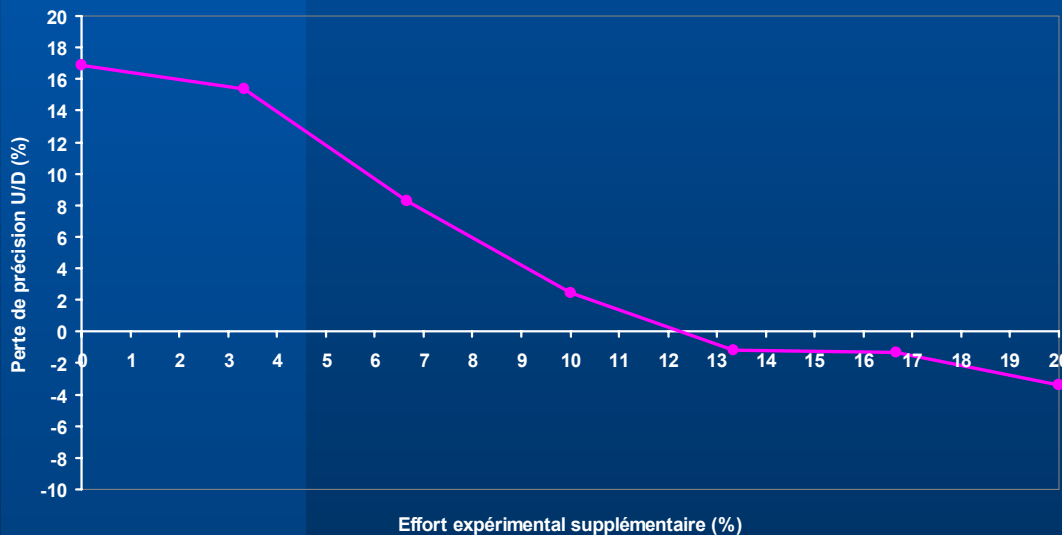
Zones les + significatives \equiv Points D-optimaux

Plans D-optimaux pour réseau de neurones (7/8)

- Estimation de la réduction de l'effort expérimental :

- Nombre d'expériences aléatoires (en %age) à rajouter au plan U1 pour avoir une précision comparable au plan D1

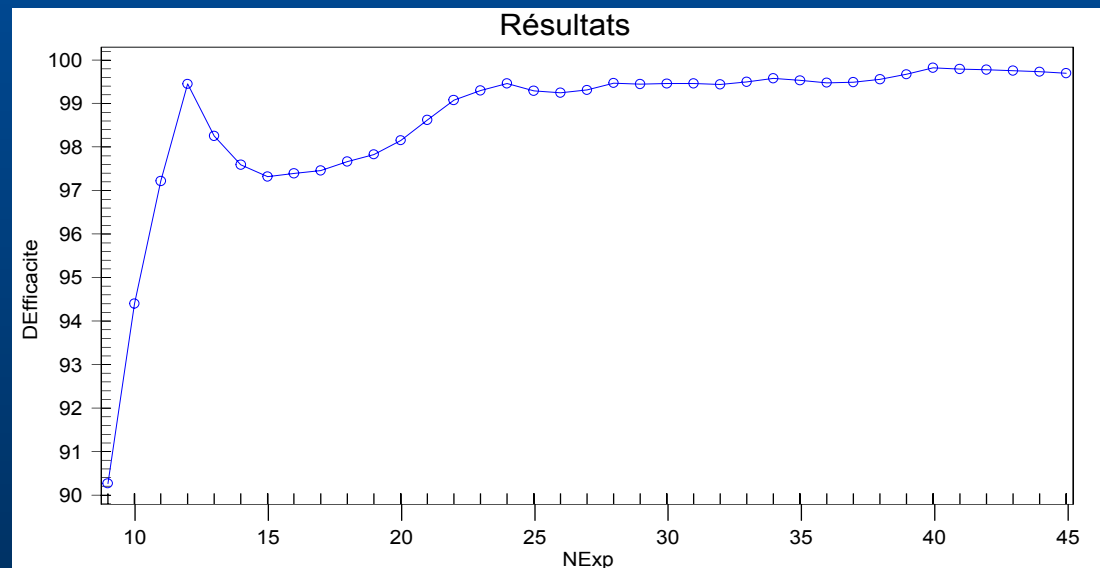
- critère de précision utilisé :
$$C = 100 \frac{\langle EQM_U^1(\mathbf{x}) \rangle_{\Xi} - \langle EQM_D^1(\mathbf{x}) \rangle_{\Xi}}{\langle EQM_U^1(\mathbf{x}) \rangle_{\Xi}}$$



effort expérimental réduit
de 12 % « en moyenne »

Plans D-optimaux pour réseau de neurones (8/8)

- **Choix du nombre d'expériences à planifier :**
 - 1^{ère} étape de « Space filling » : choix guidé par la **taille du réseau de neurones** (environ $2p$ expériences)
 - 2^{ème} étape (et suivantes) de D-optimalité : choix guidé par **D-efficacité** du plan d'expériences (de l'ordre de p)



Conclusions (1/3)

- Réseaux de neurones :
 - **Approximateurs universels**
 - Utilisables (comme les modèles de connaissance, les polynômes, ...) pour modéliser la réponse espérée d'un système
 - Peuvent être vus comme appartenant à la classe des modèles de **régression paramétrique non-linéaire**
 - Possibilité d'utiliser les théories développées dans le cadre de l' « **Optimum Design** » pour accroître leurs performances de **généralisation** (capacités prédictives)

Conclusions (2/3)

- ➤ **Précision des prévisions :**
 - Certaines zones du domaine expérimental insensibles au choix du plan d'expériences mais...
 - **Certaines zones très sensibles au plan d'expériences**
 - Les plans D-optimaux permettent de déterminer ces zones
 - Sur le cas étudié :
 - ➤ **Précision du modèle de prévision** : $\approx +15\%$
 - ➤ **Effort expérimental** (nombre d'expériences) : $\approx -15\%$
- ➤ **Temps de calcul** (- d'itérations pour convergence)
- ➤ **Optima locaux** (- d'initialisations pour optimum global)

Conclusions (3/3)

- **Perspectives pour réseaux de neurones :**

- **Prise en compte de la non-linéarité du modèle neuronal :**

- Critères de type « biais + variance »
- Utilisation d'approximations plus fines des moments de la loi de $\hat{\theta}$
- Utilisation d'une loi a priori pour les poids du réseau neuronal au lieu d'une valeur unique $\hat{\theta}^{(1)}$ (ELD-optimalité)

- **Perspectives pour modèles de connaissance :**

- Utilisation de **régions de confiances exactes** des paramètres
- Utilisation de la **densité de probabilité exacte de l'EMC**

Pour plus d'informations

Stand Netral/Camo

Chimio**m**étrie 2004

