Interpolation *vs* approximation

Étant donné un nuage de points $(M_k)_{0\leqslant k<N} = \big((x_k, y_k)\big)_{0\leqslant k<N}$ dans le plan, on peut :

- approcher ce nuage de points par une droite affine ;
- interpoler ce nuage de points par une courbe polynomiale — à condition que les abscisses des points soient deux à deux distinctes.

Des objectifs différents

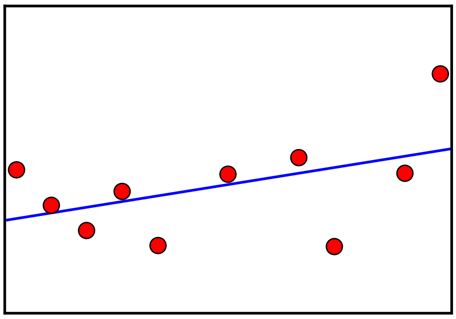
Ces deux opérations relèvent d'esprits très différents.

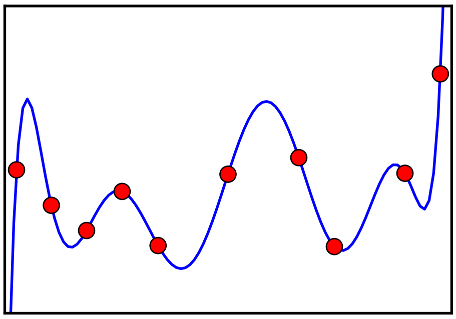
► La **régression linéaire** (figure de gauche) consiste à supposer que les points du nuage sont plus ou moins grossièrement alignés et à déterminer une droite qui permette de juger quantitativement de la qualité de cette hypothèse.

En général, la droite trouvée passe assez près de chacun des points, mais il est possible qu'aucun des points ne se situe exactement sur cette droite.

► L'**interpolation polynomiale**, ou **interpolation de Lagrange**, (figure de droite) consiste à déterminer un polynôme dont le graphe passe par chacun des points du nuage.

Plus le nombre de points est élevé, plus le degré du polynôme interpolateur est élevé et plus le graphe du polynôme varie brusquement entre les points du nuage.





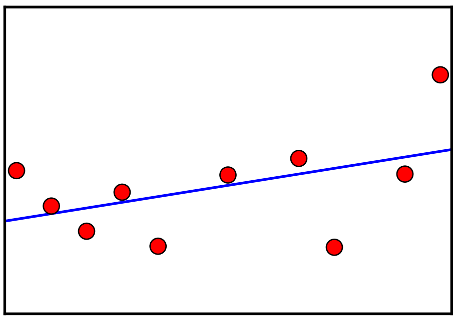
► En résumé,

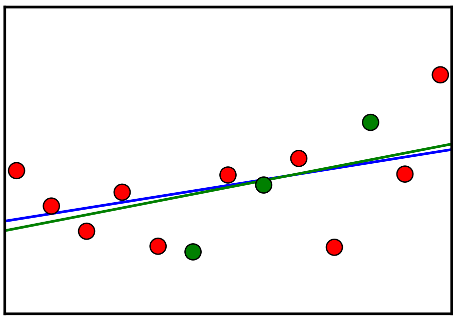
- dans un cas, on cherche une *forme approchée* qui résume globalement l'ensemble des données et on cherche à estimer *deux* paramètres (la pente et l'ordonnée à l'origine de la droite), quelle que soit la taille N du nuage de points ;
- dans l'autre, on cherche une expression algébrique qui *coïncide exactement* avec les données sans chercher à en donner une vision d'ensemble et si le nuage de points contient N points, le degré du polynôme interpolateur est en général égal à $(N-1)$: il faut donc déterminer N coefficients.

Conséquences pratiques

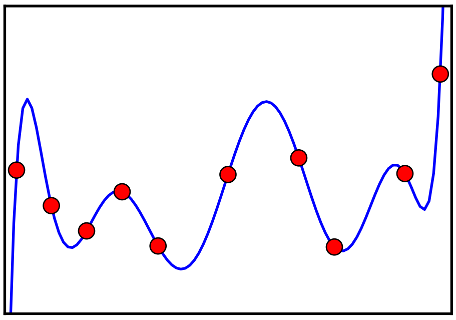
Ces approches différentes ont des conséquences pratiques qu'il convient d'avoir en tête lorsque la taille du nuage de points augmente.

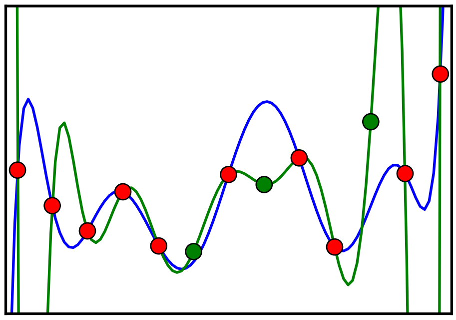
► L'ajout de quelques points supplémentaires au nuage (les points verts) ne modifie pas sensiblement l'allure globale du nuage de points. Puisque la droite de régression se borne à décrire l'allure globale de ce nuage de points, l'ajout de quelques points ne modifie pas beaucoup la droite de régression (la nouvelle droite, en vert, est proche de l'ancienne, en bleu).





► Au contraire, l'ajout de quelques points va modifier le polynôme interpolateur et pas seulement en augmentant son degré : le graphe peut également changer de manière très sensible.





Analogie avec le surapprentissage en intelligence artificielle

En termes d'intelligence artificielle, cette discussion illustre le problème du surapprentissage.

► On parle de **surapprentissage** lorsque le nombre de données dont on dispose pour entraîner une intelligence artificielle est trop faible relativement au nombre de paramètres à déterminer pour pouvoir utiliser l'intelligence artificielle.

► Dans le cas de la régression linéaire, on dispose de N données pour estimer deux paramètres et le résultat obtenu (= les valeurs des deux paramètres) est robuste au sens où l'introduction de nouvelles données modifiera peu le résultat obtenu.

► Dans le cas de l'interpolation polynomiale, on dispose de N données qui déterminent les N coefficients du polynôme interpolateur. Le polynôme trouvé est parfaitement adapté aux données (il permet de retrouver les coordonnées exactes de chaque points du nuage) mais change sensiblement dès qu'on augmente le nombre de données.