

# La régression linéaire en grande dimension

## Cours#1 - M2 MIAGE Unité "Données Massives"

Clément Lejeune

Institut de Recherche en Informatique de Toulouse,  
Université Toulouse III Paul Sabatier

30 Novembre 2021



Institut de Recherche  
en Informatique de Toulouse



UNIVERSITÉ  
TOULOUSE III  
PAUL SABATIER



Université  
de Toulouse

## Questions 🙋

- Qu'est-ce qu'une régression linéaire (RL) ?

## Questions 🙋

- Qu'est-ce qu'une régression linéaire (RL) ?
- A quoi sert une RL ?

## Questions 🙋

- Qu'est-ce qu'une régression linéaire (RL) ?
- A quoi sert une RL ?
- Comment ajuster une RL ?

## Questions 🙋

- Qu'est-ce qu'une régression linéaire (RL) ?
- A quoi sert une RL ?
- Comment ajuster une RL ?
- Qu'entend-on par "grande dimension" ? Données massives ?

## 1 Introduction

- Qu'est-ce qu'une RL ? 🧑
- A quoi sert une RL ? 🧑
- En pratique 🖐️
- Résumé

## 2 Concepts généraux en apprentissage supervisé

- La minimisation de l'erreur empirique
- Le sur/sous-apprentissage

## 3 Apprentissage d'un modèle linéaire

- Formalisation en un problème d'optimisation
- Résolution du problème d'optimisation
- Évaluation du modèle

## 4 Exemple de "A à Z" 🧐

# Qu'est-ce qu'une RL ?

La régression linéaire (RL) est une méthode statistique permettant d'estimer les paramètres d'un modèle linéaire entre une *variable cible*  $Y$  (à prédire), aussi appelée "variable dépendante", et plusieurs *variables prédictrices*  $X_1, \dots, X_d$  aussi appelées des "prédicteurs".


# Qu'est-ce qu'une RL ?

La régression linéaire (RL) est une méthode statistique permettant d'estimer les paramètres d'un modèle linéaire entre une *variable cible*  $Y$  (à prédire), aussi appelée "variable dépendante", et plusieurs *variables prédictrices*  $X_1, \dots, X_d$  aussi appelées des "prédicteurs".

## Définition (modèle linéaire multivariée)

Soit  $Y \in \mathbb{R}$  une variable continue,  $X_1, \dots, X_d$  des variables continues et/ou discrètes ( $\mathbb{R}$  ou  $\mathbb{N}$ ), et  $w_1, \dots, w_d$  les paramètres (continues) du modèles:  $Y$  est dite linéaire en  $X_1, \dots, X_d$  si:

$$Y = \sum_{j=1}^d w_j X_j = w_1 X_1 + \dots + w_d X_d \quad (1)$$

 Chaque paramètre  $w_j$  pondère la variable  $X_j$  est alors aussi appelé "poids" (weight).



# A quoi sert une RL ? 🙋

Supposons que l'on connaisse la valeurs poids  $w_1, \dots, w_d$ , plusieurs objectifs de la RL:

- Analyser l'importance de chaque prédicteur  $X_j$  sur la variable dépendante  $Y$  au moyen des poids (statistique *descriptive*),  
Exemples:
  - (en socio-économie)  $Y$ : prix de vente d'un bien immobilier,  $\{X_j\}$ :  $\{m^2, \text{âge, nb de pièces, ...}\}$ ,
  - (en biologie)  $Y$ : niveau d'expression d'un gène,  $\{X_j\}$ :  $\{\text{sexe, âge, origine, ...}\}$ ,
- Prédire  $Y$  par une combinaison "simple" des prédicteurs (statistique *prédictive*).  
Exemples:
  - (en marketing):  $Y$ : demande d'un produit,  $\{X_j\}$ :  $\{\text{prix, saison, ...}\}$

# A quoi sert une RL ? 🙄

## Remarque

👉 La statistique prédictive, aussi appelée apprentissage statistique, est une branche du *machine learning*. Ici, on veut "apprendre" (=estimer les paramètres), un modèle à partir d'**observations** des variables **cibles** et **prédictives**, on parle d'apprentissage *supervisé*. Dans le cas où seules les variables prédictives sont observées, on parle d'apprentissage non-supervisé (ex: clustering).

En pratique:

- Les valeurs théoriques des paramètres  $w_1, \dots, w_j$  dans l'Eq-1 sont inconnues !

En pratique:

- Les valeurs théoriques des paramètres  $w_1, \dots, w_j$  dans l'Eq-1 sont inconnues !
- $Y$  et  $\{X_j\}_{1 \leq j \leq d}$  connues par un échantillon *finie* de  $n$  couples d'observations (prédictions, prédicteurs)  $= \{y_i, x_j\}_{i \leq n, j \leq d}$ .

On note  $x_i = (x_{i1}, \dots, x_{id})$  la  $i$ -ème mesure des  $X_1, \dots, X_d$

En pratique:

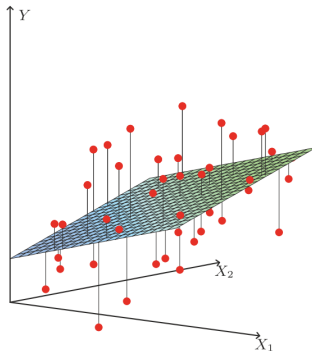
- Les valeurs théoriques des paramètres  $w_1, \dots, w_j$  dans l'Eq-1 sont inconnues !
- $Y$  et  $\{X_j\}_{1 \leq j \leq d}$  connues par un échantillon *finie* de  $n$  couples d'observations (prédictions, prédicteurs)  $= \{y_i, x_j\}_{i \leq n, j \leq d}$ .

On note  $x_i = (x_{i1}, \dots, x_{id})$  la  $i$ -ème mesure des  $X_1, \dots, X_d$

## Précision

👉  $w_1, \dots, w_j$  représentent les "vrais" paramètres du modèle (inconnus) que l'on remplace par une *estimation* notée  $\hat{w}_1, \dots, \hat{w}_j$  calculées à partir des  $n$  observations.


Exemple: deux variables prédictrices  $X_1, X_2$  et une variable cible  $Y$  observées  $\{y_i, x_{ij}\}_{i \leq n=30, j \leq d=2}$  aléatoirement  $n = 30$  fois.



**Figure:** Le plan (maillage bleu/vert/jaune) est l'espace engendré par la variable cible  $Y$  et une estimation  $(\hat{w}_1, \hat{w}_2, Y) \in \mathbb{R}^2$  de  $(w_1, w_2)$ . Chaque point du plan représente un tuple  $(x_1, x_2, y)$ . Source: (Hastie, Tibshirani, and Friedman 2009, pp.45)



# Introduction

Raisonnement général:

- On dispose de  $n$  mesures  $(y_i, x_i)$  avec  $i = 1 \dots n$ , des variables  $Y, X_1, \dots, X_d$ , 

# Introduction



## Raisonnement général:

- On dispose de  $n$  mesures  $(y_i, x_i)$  avec  $i = 1 \dots n$ , des variables  $Y, X_1, \dots, X_d$ , 
- 🙌 On fait l'hypothèse que le modèle linéaire  $Y = \sum_j w_j X_j$  est un bon candidat pour prédire  $y_i$  à partir de  $x_i$  pour tout  $i$ . Et on espère que les mesures sont représentatives de la distribution de  $Y, X_1, \dots, X_d$  !
- Problème 😱 on ne connaît pas les  $w_1, \dots, w_j$  





# Introduction

## Raisonnement général:

- On dispose de  $n$  mesures  $(y_i, x_i)$  avec  $i = 1 \dots n$ , des variables  $Y, X_1, \dots, X_d$ , 
- 🙌 On fait l'hypothèse que le modèle linéaire  $Y = \sum_j w_j X_j$  est un bon candidat pour prédire  $y_i$  à partir de  $x_i$  pour tout  $i$ . Et on espère que les mesures sont représentatives de la distribution de  $Y, X_1, \dots, X_d$  !
- Problème 😱 on ne connaît pas les  $w_1, \dots, w_d$  
- On souhaite estimer  $w_1, \dots, w_d$  par  $\hat{w}_1, \dots, \hat{w}_d$  de façon à ce que la prédiction  $\hat{y}_i = \sum_j \hat{w}_j x_{ij}$  soit la plus proche possible de  $y_i$  pour tout  $i = 1 \dots n$  🤔.

# Introduction

## Raisonnement général:

- On dispose de  $n$  mesures  $(y_i, x_i)$  avec  $i = 1 \dots n$ , des variables  $Y, X_1, \dots, X_d$ , 
- 🙌 On fait l'hypothèse que le modèle linéaire  $Y = \sum_j w_j X_j$  est un bon candidat pour prédire  $y_i$  à partir de  $x_i$  pour tout  $i$ . Et on espère que les mesures sont représentatives de la distribution de  $Y, X_1, \dots, X_d$  !
- Problème 😱 on ne connaît pas les  $w_1, \dots, w_d$  
- On souhaite estimer  $w_1, \dots, w_d$  par  $\hat{w}_1, \dots, \hat{w}_d$  de façon à ce que la prédiction  $\hat{y}_i = \sum_j \hat{w}_j x_{ij}$  soit la plus proche possible de  $y_i$  pour tout  $i = 1 \dots n$  🤔.

## Remarque 🙌

Estimer les paramètres  $w_1, \dots, w_d$  revient à "apprendre au modèle à prédire  $Y$  à partir des  $X_1, \dots, X_d$ " (vocabulaire *machine learning*).

# La minimisation de l'erreur empirique

La théorie de l'apprentissage statistique supervisé est fondée sur la minimisation d'un critère appelé *risque empirique*. Cette théorie statistique a été formalisée en 1995 par Vladimir Vapnik (récent!). L'idée générale est la suivante:

# La minimisation de l'erreur empirique

La théorie de l'apprentissage statistique supervisé est fondée sur la minimisation d'un critère appelé *risque empirique*. Cette théorie statistique a été formalisée en 1995 par Vladimir Vapnik (récent!). L'idée générale est la suivante:

- Faire une hypothèse sur un modèle théorique  $h$  prédictif de  $Y$  à partir de  $X_1, \dots, X_d$  c-à-d on veut que  $h$  satisfasse  $Y = h(X)$   
(👉 ici, on a choisi  $h(X_1, \dots, X_d) = \sum_j w_j X_j$ ),

# La minimisation de l'erreur empirique

La théorie de l'apprentissage statistique supervisé est fondée sur la minimisation d'un critère appelé *risque empirique*. Cette théorie statistique a été formalisée en 1995 par Vladimir Vapnik (récent!). L'idée générale est la suivante:

- Faire une hypothèse sur un modèle théorique  $h$  prédictif de  $Y$  à partir de  $X_1, \dots, X_d$  c-à-d on veut que  $h$  satisfasse  $Y = h(X)$   
(👉 ici, on a choisi  $h(X_1, \dots, X_d) = \sum_j w_j X_j$ ),
- Choisir un critère d'erreur de prédiction  $\ell(y_i, h(x_i))$ ,

## Remarque

La fonction  $h$  dépend à la fois des prédicteurs et des paramètres  $w_j$ . Donc la fonction  $\ell$  dépend aussi des paramètres  $w_1, \dots, w_d$  ! 🐪

# La minimisation de l'erreur empirique

La théorie de l'apprentissage statistique supervisé est fondée sur la minimisation d'un critère appelé *risque empirique*. Cette théorie statistique a été formalisée en 1995 par Vladimir Vapnik (récent!). L'idée générale est la suivante:

- Faire une hypothèse sur un modèle théorique  $h$  prédictif de  $Y$  à partir de  $X_1, \dots, X_d$  c-à-d on veut que  $h$  satisfasse  $Y = h(X)$   
(👉 ici, on a choisi  $h(X_1, \dots, X_d) = \sum_j w_j X_j$ ),
- Choisir un critère d'erreur de prédiction  $\ell(y_i, h(x_i))$ ,

## Remarque

La fonction  $h$  dépend à la fois des prédicteurs et des paramètres  $w_j$ . Donc la fonction  $\ell$  dépend aussi des paramètres  $w_1, \dots, w_d$  ! 🐪

- Calculer  $\hat{w}_1, \dots, \hat{w}_d$  comme les minimiseur de la fonction 
$$g(w_1, \dots, w_d) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, h(x_i)).$$

# La minimisation de l'erreur empirique

🤔 Moralité: on fixe un modèle prédictif  $h$ , on fixe une mesure d'erreur entre les observations  $(y_i, x_i)$  et le modèle, et enfin on calcule les paramètres de  $h$  en minimisant la moyenne des mesures d'erreur.

# La minimisation de l'erreur empirique

🤔 Moralité: on fixe un modèle prédictif  $h$ , on fixe une mesure d'erreur entre les observations  $(y_i, x_i)$  et le modèle, et enfin on calcule les paramètres de  $h$  en minimisant la moyenne des mesures d'erreur.

Facile ?! 🐱



# Le sur/sous-apprentissage

- En minimisant l'erreur empirique moyenne, il y a un risque que le modèle n'apprenne à "bien prédire" que sur les observations  $(y_i, x_i)$  utilisées minimiser l'erreur empirique...
- On parle alors de sur/sous-apprentissage 😞,

# Le sur/sous-apprentissage

- En minimisant l'erreur empirique moyenne, il y a un risque que le modèle n'apprenne à "bien prédire" que sur les observations  $(y_i, x_i)$  utilisées minimiser l'erreur empirique...
- On parle alors de sur/sous-apprentissage 😞,
- On veut un modèle qui sache prédire  $\hat{y}_{i'} = h(x_{i'}) = \sum_j \hat{w}_j x_{i'j} \approx y_{i'}$  sachant que  $(y_{i'}, x_{i'})$  n'a pas été utilisé pour calculer les paramètres,

# Le sur/sous-apprentissage

- En minimisant l'erreur empirique moyenne, il y a un risque que le modèle n'apprenne à "bien prédire" que sur les observations  $(y_i, x_i)$  utilisées minimiser l'erreur empirique...
- On parle alors de sur/sous-apprentissage 😞,
- On veut un modèle qui sache prédire  $\hat{y}_{i'} = h(x_{i'}) = \sum_j \hat{w}_j x_{i'j} \approx y_{i'}$  sachant que  $(y_{i'}, x_{i'})$  n'a pas été utilisé pour calculer les paramètres,
- En évaluant l'erreur empirique moyenne sur les données d'apprentissage, on ne sait pas si le modèle a sur-appris ou sous-appris,

# Le sur/sous-apprentissage

- En minimisant l'erreur empirique moyenne, il y a un risque que le modèle n'apprenne à "bien prédire" que sur les observations  $(y_i, x_i)$  utilisées minimiser l'erreur empirique...
- On parle alors de sur/sous-apprentissage 😞,
- On veut un modèle qui sache prédire  $\hat{y}_{i'} = h(x_{i'}) = \sum_j \hat{w}_j x_{i'j} \approx y_{i'}$  sachant que  $(y_{i'}, x_{i'})$  n'a pas été utilisé pour calculer les paramètres,
- En évaluant l'erreur empirique moyenne sur les données d'apprentissage, on ne sait pas si le modèle a sur-appris ou sous-appris,
- Pour le savoir, il faut évaluer l'erreur sur des observations non-utilisées dites *données de test*.

# Concepts généraux en apprentissage supervisé

Plus formellement,

## Définition: Erreur d'entraînement/généralisation

Soit  $\{y_i, x_i\}_{i \leq n} \in \Omega$  et  $\{y_{i'}, x_{i'}\}_{i' \leq n'} \in \Omega'$  deux ensembles de  $n$  et  $n'$  observations indépendantes de  $(Y, X_1 \dots, X_d)$ ,  $\Omega \cap \Omega' = \emptyset$ ,  $h$  une fonction de prédiction, telle que  $Y = h_{w_1, \dots, w_d}(X) = \sum_j w_j X_j$ , ayant pour paramètres  $w_1, \dots, w_d$  et  $\ell(\cdot, \cdot)$  un critère d'erreur de prédiction.

$$\hat{w}_1, \dots, \hat{w}_d = \arg \min_{w_1, \dots, w_d} \frac{1}{n} \sum_{\{y_i, x_i\}_{i \leq n} \in \Omega} \ell(y_i, h_{w_1, \dots, w_d}(x_i)) \quad (2)$$

sont les paramètres "entraînés" (ou "estimés"), et

$$err_{tr} = \frac{1}{n} \sum_{\{y_i, x_i\}_{i \leq n} \in \Omega} \ell(y_i, h_{\hat{w}_1, \dots, \hat{w}_d}(x_i)) \quad (3)$$

est l'erreur d'*entraînement*.

## Définition: Erreur d'entraînement/généralisation (suite)

$$err_{gen} = \frac{1}{n'} \sum_{\{y_{i'}, x_{i'}\}_{i' \leq n'} \in \Omega'} \ell(y_{i'}, h_{\hat{w}_1, \dots, \hat{w}_d}(x_{i'})) \quad (4)$$

est l'erreur de *généralisation*. Si:

- $err_{tr} \approx err_{gen}$ : le modèle a bien appris,
- si  $err_{tr} \ll err_{gen}$  ou  $err_{tr} \gg err_{gen}$ : le modèle a sur- ou sous-appris.

Moralité: le pouvoir de prédiction d'un modèle  $h$  (linéaire ou non), appris, se mesure sur des observations non utilisées pour l'apprentissage.

# Concepts généraux en apprentissage supervisé

La procédure décrite est très générale.

- Pour chaque type de données,
- pour chaque type de modèle  $h$ ,
- ...
- il faut choisir un critère d'erreur  $\ell(\cdot, \cdot)$ .

# Apprentissage d'un modèle linéaire

Très souvent, et encore plus dans le cas du modèle linéaire, le critère  $\ell(\cdot, \cdot)$  est l'erreur quadratique:

## Erreur quadratique

$$\ell_{quad}(y_i, \hat{y}_i = \sum_j w_j x_{ij}) = \frac{1}{2} (y_i - \sum_j w_j x_{ij})^2 \quad (5)$$

Avantages: les calculs théoriques sont faciles (identité remarquable -> décomposition statistique biais/variance), l'apprentissage revient à résoudre un problème d'optimisation facile (différentiable), etc. Voir Hastie, Tibshirani, and Friedman 2009 pour plus justifications.



# Apprentissage d'un modèle linéaire

On a défini:

- Un modèle prédictif  $h(x_i) = \sum_j w_j x_{ij}$ ,
- un critère d'erreur  $\ell_{quad}(y_i, h(x_i)) = \frac{1}{2}(y_i - h(x_i))^2$ ,

Il faut un algorithme pour calculer le minimiseur  $\hat{w}_1, \dots, \hat{w}_d$  de la fonction:

$$g(w_1, \dots, w_d) = \frac{1}{2n} \sum_i^n (y_i - \sum_{j=1}^d w_j x_{ij})^2$$

avec  $(y_i, x_i) \in \Omega$ .

# Formalisation en un problème d'optimisation

On note  $\mathbf{w} = (w_1, \dots, w_d) \in \mathbb{R}^d$  le vecteur des  $d$  paramètres.  
L'apprentissage des paramètres du modèle  $h(x_i) = \sum_j w_j x_{ij}$  revient à résoudre un problème d'optimisation:

$$\text{MINIMIZE } g(\mathbf{w}) = \frac{1}{2n} \sum_{i=1}^n \underbrace{(y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle)^2}_{\ell(y_i, h(\mathbf{x}_i))} \quad (6)$$

$\langle u, v \rangle = \sum_j u_j v_j$  est la notation du produit scalaire.

# Formalisation en un problème d'optimisation

Remarquons que la fonction  $g$  est *quadratique* en  $\mathbf{w}$ .

# Formalisation en un problème d'optimisation

Remarquons que la fonction  $g$  est *quadratique* en  $\mathbf{w}$ . Exemple d'une fonction quadratique à  $d = 2$  variables:

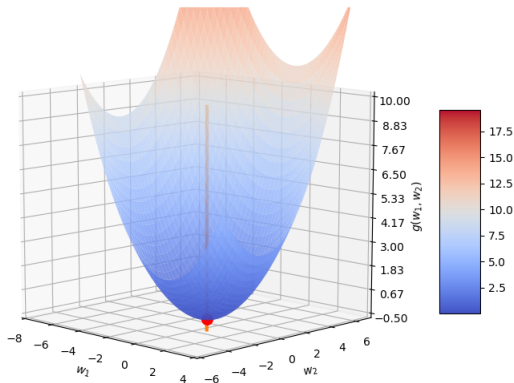


Figure: Le minimum de  $g$  est atteint en  $\hat{\mathbf{w}} = (1, -2)$ .

# Formalisation en un problème d'optimisation

Évidemment, on ne cherche jamais une solution visuellement (impossible pour  $d > 3$  !):

- on exploite les propriétés de la fonction  $g$  puis, on calcule la solution soit:
- soit analytiquement (calcul à la main),
- soit avec un algorithme: on part d'une solution initiale  $\hat{\mathbf{w}}_0$  et on applique une procédure itérative renvoyant des itérées  $\hat{\mathbf{w}}_k, \hat{\mathbf{w}}_{k+1}, \dots$  convergeant vers la solution optimale.

# Formalisation en un problème d'optimisation

Ici, on va exploiter la propriété quadratique. Pourquoi ?

- $g$  est quadratique  $\Rightarrow$  (ici) *convexe*  
( $g(t\mathbf{w} + (1-t)\mathbf{w}) \leq tg(\mathbf{w}) + (1-t)g(\mathbf{w}), t \in [0, 1]$ ),
- $g$  est quadratique  $\Rightarrow g$  est différentiable (les dérivées partielles de  $g$  par rapport à  $w_1, \dots, w_d$  existent),
- Et en considérant:  $n > d$  et les variables prédictives  $X_1, \dots, X_d$  linéairement indépendantes  $\Rightarrow g$  est "strictement convexe" (la raison est hors-scope du cours mais importante),
- La stricte convexité implique l'unicité des minima: il n'y a qu'un seul minimiseur, donc il est global.

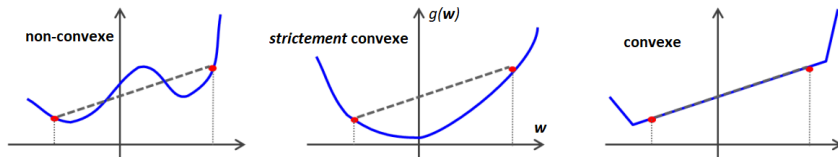


Figure: Non-convexité/convexité en une dimension.

# Résolution du problème d'optimisation

En optimisation, pour trouver un minimum local, on utilise souvent la propriété d'optimalité du premier ordre ("first-order optimality condition") qui se traduit de façon informelle par:

" $\mathbf{w}^*$  est un minimum local de  $g \Rightarrow$  les dérivées de  $g$  en  $\mathbf{w}^*$  sont nulles.

Donc puisqu'ici le minimum est unique, le  $\mathbf{w} = w_1, \dots, w_d$  qui annule les dérivées de  $g$  est le minimum global (le meilleur  $\hat{\mathbf{w}}$  selon le critère d'erreur  $\ell$ ). Il suffit donc de:

- calculer les dérivées partielles  $g'(w_1, \dots, w_d)$ ,
- résoudre  $\nabla g = g'(w_1, \dots, w_d) = \left( \frac{\partial g}{\partial w_1}, \dots, \frac{\partial g}{\partial w_d} \right) = \mathbf{0}$ .

🧐 Et c'est tout ! 🎉

👉 Pour les intéressé(e)s en optimisation: Boyd and Vandenberghe 2004 (excellente référence internationale).

# Résolution du problème d'optimisation

Solution (voir détails calcul au tableau): D'abord, on réécrit  $g$  sous forme vectorielle:

$$g(\mathbf{w}) = \frac{1}{2n} \|\mathbf{y} - X\mathbf{w}\|_2^2 = \frac{1}{2n} (\mathbf{y} - X\mathbf{w})^\top (\mathbf{y} - X\mathbf{w}) = \frac{1}{2n} \langle \mathbf{y} - X\mathbf{w}, \mathbf{y} - X\mathbf{w} \rangle \quad (7)$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1d} \\ \vdots & \ddots & x_{ij} & \ddots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nd} \end{pmatrix}$$



# Résolution du problème d'optimisation

Solution (voir détails calcul au tableau): D'abord, on réécrit  $g$  sous forme vectorielle:

$$g(\mathbf{w}) = \frac{1}{2n} \|\mathbf{y} - X\mathbf{w}\|_2^2 = \frac{1}{2n} (\mathbf{y} - X\mathbf{w})^\top (\mathbf{y} - X\mathbf{w}) = \frac{1}{2n} \langle \mathbf{y} - X\mathbf{w}, \mathbf{y} - X\mathbf{w} \rangle \quad (7)$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1d} \\ \vdots & \ddots & x_{ij} & \ddots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nd} \end{pmatrix}$$

$$\nabla g = \left( \frac{\partial g}{\partial w_1}, \dots, \frac{\partial g}{\partial w_d} \right)^\top = \frac{1}{n} X^\top (X\mathbf{w} - \mathbf{y})$$

On résout  $\nabla g = \mathbf{0}$

# Résolution du problème d'optimisation

Solution (voir détails calcul au tableau): D'abord, on réécrit  $g$  sous forme vectorielle:

$$g(\mathbf{w}) = \frac{1}{2n} \|\mathbf{y} - X\mathbf{w}\|_2^2 = \frac{1}{2n} (\mathbf{y} - X\mathbf{w})^\top (\mathbf{y} - X\mathbf{w}) = \frac{1}{2n} \langle \mathbf{y} - X\mathbf{w}, \mathbf{y} - X\mathbf{w} \rangle \quad (7)$$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1d} \\ \vdots & \ddots & x_{ij} & \ddots & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nd} \end{pmatrix}$$

$$\nabla g = \left( \frac{\partial g}{\partial w_1}, \dots, \frac{\partial g}{\partial w_d} \right)^\top = \frac{1}{n} X^\top (X\mathbf{w} - \mathbf{y})$$

On résout  $\nabla g = \mathbf{0}$

$$\hat{\mathbf{w}} = \arg \min_{\mathbf{w}} g(\mathbf{w}) = (X^\top X)^{-1} X^\top \mathbf{y} \quad (8)$$

# Évaluation du modèle

Rappel: on évalue TOUJOURS le modèle appris sur un ensemble de données de test  $\{y_{i'}, x_{i'}\}_{i' \leq n'} \in \Omega'$ :

$$err_{gen} = \frac{1}{n'} \sum_{\{y_{i'}, x_{i'}\}_{i' \leq n'} \in \Omega'} (y_{i'} - \langle \hat{\mathbf{w}}, x_{i'} \rangle)^2 \quad (9)$$

On peut utiliser d'autres types d'erreur pour mesurer la généralisation comme le  $R^2$ , la somme des erreurs absolues, les erreurs relatives, etc. (voir TP)

# Exemple de "A à Z" 🧐

Prenons le jeu de données "boston house-prices" (données des 70's)  
Harrison Jr and Rubinfeld 1978, p. 96–97:

- $Y$ , variable cible "MEDV": prix médian d'un bien immobilier à Boston,
- $d = 13$  variables prédictives:

## Data Set Characteristics:

### Number of Instances:

506

### Number of Attributes:

13 numeric/categorical predictive. Median Value (attribute 14) is usually the target.

### Attribute Information (in order):

- CRIM per capita crime rate by town
- ZN proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS proportion of non-retail business acres per town
- CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX nitric oxides concentration (parts per 10 million)
- RM average number of rooms per dwelling
- AGE proportion of owner-occupied units built prior to 1940
- DIS weighted distances to five Boston employment centres
- RAD index of accessibility to radial highways
- TAX full-value property-tax rate per \$10,000
- PTRATIO pupil-teacher ratio by town
- B  $1000(Bk - 0.63)^2$  where Bk is the proportion of black people by town
- LSTAT % lower status of the population
- MEDV Median value of owner-occupied homes in \$1000's

# Exemple de "A à Z" 🧐

Questions ?

# Références I

- [1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [2] David Harrison Jr and Daniel L Rubinfeld. “Hedonic housing prices and the demand for clean air”. In: *Journal of environmental economics and management* 5.1 (1978), pp. 81–102.
- [3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. 2009. DOI: 10.1007/978-0-387-84858-7. URL: <https://link.springer.com/content/pdf/10.1007%2F978-0-387-84858-7.pdf%0Ahttp://link.springer.com/10.1007/978-0-387-84858-7>.