Discussion sur la classification binaire
A retouver: l'hypothèse $f_+(x) = 1/(1+e^{-\omega^T x})$ revient à l'hypothèse suivante

$$Y = 2\mathbb{1}_{\{\omega^T x+\varepsilon\geq 0\}} - 1$$

avec $\varepsilon \perp\!\!\!\perp X$ et $\varepsilon \sim \mathcal{L}ogistique$.

Dans un réseau de neurones pour de la classification avec une dernière couche logistique, alors on pourrait le voir de la manière suivante

$$Y = 2\mathbb{1}_{\{\omega^T h(x)+\varepsilon\geq 0\}} - 1$$

avec $h(x)$ la fonction qui encode toutes les couches intermédiaires (fonctions d'activation ReLu par exemple) à travers lesquelles passe x.

Discussion sur les SVM A retenir sur les SVM: le SVM est un algo qui cherche à séparer par un hyperplan les observations selon leurs labels. Sur des données réelles, il est souvent difficile voire impossible de séparer linéairement les observations $\implies$ le SVM s'associe de manière très efficace, à travers l'utilisation des mystérieux noyaux, avec une étape de transformation (nonlinéaire) des données de manière à ce que les données transformées soient elles linéairement séparables. Le choix du noyau dans l'algorithme dépend entre autres des données initiales. Attention, le noyau effectue "lui-même" la transformation nonlinéaire des données.

Les SVM vus comme un problème de minimisation du risque empirique régularisé. La question nous montre que le programme (2) associé à un SVM peut se réécrire de la manière suivante

$$\underset{w,w_0}{\operatorname{argmin}} \left\{ \frac{||w||^2}{2} + C\sum_{i=1}^{n} f\big(y_i(\langle\Phi(x_i), w\rangle + w_0)\big) \right\},$$

avec $f : x \mapsto \max(0, 1-x)$ qui est la perte dite hinge (*cf.* les premiers TPs et le cours sur les pertes pour les problèmes de classification). On peut encore réécrire ce problème comme

$$\underset{w,w_0}{\operatorname{argmin}} \left\{ \underbrace{\sum_{i=1}^{n} f\big(y_i(\langle\Phi(x_i), w\rangle + w_0)\big)}_{\text{=: perte de classif empirique}} + \underbrace{\frac{||w||^2}{2C}}_{\text{=: terme de regularisation}} \right\}.$$

Le SVM peut se voir comme un classifieur standard où les features sont $\Phi(x)$ au lieu de x et la perte de classification utilisée est la perte hinge.

Preuve de la question 3: D'après les contraintes du programme (2) du SVM, nous voyons que

$$\begin{aligned} &\begin{cases} \xi_i \geq 0 & \forall i = 1, ..., n \\ y_i(\langle\Phi(x_i), w\rangle + w_0) \geq 1 - \xi_i & \forall i = 1, ..., n \end{cases} \\ \iff &\begin{cases} \xi_i \geq 0 & \forall i = 1, ..., n \\ \xi_i \geq 1 - y_i(\langle\Phi(x_i), w\rangle + w_0) & \forall i = 1, ..., n \end{cases} \end{aligned}$$

Par ailleurs, nous voulons minimiser $\sum_{i=1}^n \xi_i$. Nous remarquons que le choix $\xi_i = f\big(y_i(\langle\Phi(x_i), w\rangle + w_0)\big)$ satisfait les contraintes et minimise ξ_i pour chaque i et donc $\sum_{i=1}^n \xi_i$ car les ξ_i sont positifs par construction. En remplaçant ξ_i par $f\big(y_i(\langle\Phi(x_i), w\rangle + w_0)\big)$ dans le programme (2), nous obtenons le résultat.