IFT-7002 Fondements de l'apprentissage machine

Régularisation et stabilité

Shai Shalev-Shwartz
The Hebrew University of Jerusalem

Traduit et adapté par Mario Marchand
Université Laval

Hiver 2024

Sommaire

1. Minimisation du risque empirique régularisé
   - Régularisation de Tikhonov
   - Cas particulier : régression de ridge
2. Stabilité d'un algorithme d'apprentissage
3. Stabilité de la régularisation de Tikhonov
4. Apprendre avec la régularisation de Tikhonov
5. La DGS pour fonctions fortement convexes
6. La DGS pour la régularisation de Tikhonov
7. Garanties sur le risque : norme ou dimension ?
   - Exemple d'application : classification de texte

Problème d'apprentissage convexe-Lipschitzien-borné

Rappel :

Définition (problème d'apprentissage convexe-Lipschitzien-borné)

Un problème d'apprentissage $(\mathcal{H}, Z, \ell)$ est convexe-Lipschitzien-borné avec paramètres ρ, B si les propriétés suivantes sont satisfaites :

- $\mathcal{H}$ est un ensemble convexe et pour tout $\mathbf{w} \in \mathcal{H}$, nous avons $\|\mathbf{w}\| \leq B$.
- Pour tout $z \in Z$, la fonction de perte $\ell(\cdot, z)$ est convexe et ρ-Lipschitzienne.

- Nous avons vu que la DGS permet d'apprendre les problèmes convexes-Lipchitziens-bornés.
- Nous verrons ici que la minimisation du risque empirique régularisé permet également d'apprendre ces problèmes.

Sommaire

1. Minimisation du risque empirique régularisé
   - Régularisation de Tikhonov
   - Cas particulier : régression de ridge
2. Stabilité d'un algorithme d'apprentissage
3. Stabilité de la régularisation de Tikhonov
4. Apprendre avec la régularisation de Tikhonov
5. La DGS pour fonctions fortement convexes
6. La DGS pour la régularisation de Tikhonov
7. Garanties sur le risque : norme ou dimension ?
   - Exemple d'application : classification de texte

Régularisation de Tikhonov

Ayant une fonction de régularisation $R : \mathbb{R}^d \to \mathbb{R}$, une fonction de perte ℓ et un échantillon d'apprentissage S, la minimisation du risque empirique régularisé est l'algorithme A défini par

$$A(S) = \operatorname*{argmin}_{\mathbf{w}} \left(L_S(\mathbf{w}) + R(\mathbf{w}) \right) \ .$$

Examinons le cas particulier de la régularisation de Tikhonov défini par

$$A(S) = \operatorname*{argmin}_{\mathbf{w}} \left(L_S(\mathbf{w}) + \lambda \|\mathbf{w}\|^2 \right) \ .$$

Un sous-cas important est celui de la régression de ridge, défini par

$$A(S) = \operatorname*{argmin}_{\mathbf{w}} \left(\lambda \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 \right) \ .$$

Sommaire

1. Minimisation du risque empirique régularisé
   - Régularisation de Tikhonov
   - Cas particulier : régression de ridge
2. Stabilité d'un algorithme d'apprentissage
3. Stabilité de la régularisation de Tikhonov
4. Apprendre avec la régularisation de Tikhonov
5. La DGS pour fonctions fortement convexes
6. La DGS pour la régularisation de Tikhonov
7. Garanties sur le risque : norme ou dimension ?
   - Exemple d'application : classification de texte

Régression de ridge : solution

- Tout d'abord écrivons la fonction à minimiser sous forme matricielle.
  - Soit X la matrice $d \times m$ formée des m vecteurs colonnes $\mathbf{x}_1, \ldots, \mathbf{x}_m$.
  - soit $\mathbf{y} \stackrel{\text{def}}{=} (y_1, \ldots, y_m)^\top$.
  - Nous avons donc

$$\frac{1}{m} \sum_{i=1}^{m} (\langle \mathbf{w}, \mathbf{x}_i \rangle - y_i)^2 = \frac{1}{m} ||X^\top \mathbf{w} - \mathbf{y}||^2 .$$

  - Notre fonction f à minimiser est donc

$$f(\mathbf{w}) \stackrel{\text{def}}{=} \lambda \|\mathbf{w}\|^2 + \frac{1}{m} ||X^\top \mathbf{w} - \mathbf{y}||^2 .$$

- f est convexe et différentiable en $\mathbf{w}$. Son minimum global est donc donné par la valeur de $\mathbf{w}$ tel que $\nabla f(\mathbf{w}) = \mathbf{0}$.
- De manière équivalente, on cherche $\mathbf{w}$ t.q. le Jacobien $J_{\mathbf{w}}(f) = \mathbf{0}^\top$.

Rappel sur le Jacobien

- Le Jacobien d'une fonction $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ évalué à $\mathbf{x} \in \mathbb{R}^d$, dénoté $J_{\mathbf{x}}(\mathbf{f})$, est une matrice $m \times d$ telle que sa $i$ème ligne est $\nabla f_i(\mathbf{x})$.
- Si $m = 1$ alors $J_{\mathbf{x}}(f) = [\nabla f(\mathbf{x})]^\top$ (un vecteur ligne).
- Si $\mathbf{f}(\mathbf{w}) = A\mathbf{w}$ pour $A \in \mathbb{R}^{m,d}$ alors $J_{\mathbf{w}}(\mathbf{f}) = A$.
- Règle d'enchainement : Soit $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^m$ et $\mathbf{g} : \mathbb{R}^k \to \mathbb{R}^d$, le Jacobien de la composition $(\mathbf{f} \circ \mathbf{g}) : \mathbb{R}^k \to \mathbb{R}^m$, évalué à $\mathbf{x}$, est donné par

$$J_{\mathbf{x}}(\mathbf{f} \circ \mathbf{g}) = J_{\mathbf{g}(\mathbf{x})}(\mathbf{f}) J_{\mathbf{x}}(\mathbf{g}) \ .$$

Retour à la régression de ridge

- Il faut trouver le Jacobien de $\lambda||\mathbf{w}||^2$ et de $(1/m)||X^\top \mathbf{w} - \mathbf{y}||^2$.
- $J_{\mathbf{w}}(\lambda||\mathbf{w}||^2) = 2\lambda\mathbf{w}^\top$.
- Soit $\mathbf{g}(\mathbf{w}) = X^\top \mathbf{w} - \mathbf{y}$ et $\mathbf{f}(\mathbf{v}) = \frac{1}{m}\|\mathbf{v}\|^2 = \frac{1}{m}\sum_{i=1}^m v_i^2$.
- Règle d'enchaînement : $J_{\mathbf{w}}(\mathbf{f} \circ \mathbf{g}) = J_{\mathbf{g}(\mathbf{w})}(\mathbf{f}) J_{\mathbf{w}}(\mathbf{g})$.
- Or, $J_{\mathbf{w}}(\mathbf{g}) = X^\top$ et $J_{\mathbf{v}}(\mathbf{f}) = (2/m)(v_1, \ldots, v_m)$.
- Donc $J_{\mathbf{g}(\mathbf{w})}(\mathbf{f}) J_{\mathbf{w}}(\mathbf{g}) = (2/m)\mathbf{g}(\mathbf{w})^\top X^\top = (2/m)(X^\top \mathbf{w} - \mathbf{y})^\top X^\top$.
- En imposant que la somme de ces deux Jacobiens $= \mathbf{0}^\top$, on a

$$2\lambda\mathbf{w}^\top + \frac{2}{m}(X^\top \mathbf{w} - \mathbf{y})^\top X^\top = \mathbf{0}^\top \ .$$

- En prenant la transposée et en multipliant par $m/2$ on a

$$m\lambda\mathbf{w} + X(X^\top \mathbf{w} - \mathbf{y}) = \mathbf{0} \ .$$

Régression de ridge : la solution

- Si I dénote la matrice identité, la dernière équation s'écrit

$$\left(m\lambda I + XX^\top\right)\mathbf{w} = X\mathbf{y}\ .$$

- Or, $m\lambda I + XX^\top$ est toujours inversible pour $\lambda > 0$ car elle est symétrique définie positive lorsque $\lambda > 0$.
- Alors, le prédicteur $\mathbf{w}_{rr}$ minimisant la fonction objectif de la régularisation de ridge est donné par

$$\mathbf{w}_{rr} = \left(m\lambda I + XX^\top\right)^{-1} X\mathbf{y}\,.$$

- Examinons maintenant les garanties qu'il est possible d'établir sur les prédicteurs obtenus de la régularisation de Tikhonov lorsque l'on utilise une fonction de perte ℓ qui est convexe et Lipschitzienne.

Pourquoi régulariser ?

- Similaire à l'approche MDL : spécifier une préférence a priori sur les hypothèses. Ici nous avons une préférence envers les prédicteurs $\mathbf{w}$ de petites normes.
- Permet de stabiliser : nous allons démontrer que la régularisation de Tikhonov stabilise l'algorithme d'apprentissage par rapport aux petites perturbations effectuées sur l'échantillon d'apprentissage.
  - Cela aide l'algorithme à bien généraliser.

Sommaire

1. Minimisation du risque empirique régularisé
   - Régularisation de Tikhonov
   - Cas particulier : régression de ridge
2. Stabilité d'un algorithme d'apprentissage
3. Stabilité de la régularisation de Tikhonov
4. Apprendre avec la régularisation de Tikhonov
5. La DGS pour fonctions fortement convexes
6. La DGS pour la régularisation de Tikhonov
7. Garanties sur le risque : norme ou dimension ?
   - Exemple d'application : classification de texte

Stabilité

- Informellement : un algorithme A est stable si une petite modification de son entrée S provoque seulement une petit changement dans le prédicteur produit par A.
- Il faut préciser quantitativement ce qu'est une "petite modification de l'entrée" et ce qu'est un "petit changement dans le prédicteur produit".

Stabilité

- Remplacer un des exemples : soit $S = (z_1, \ldots, z_m)$ et un exemple additionnel z', soit $S^{(i)} \stackrel{\text{def}}{=} (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$.
- On s'intéresse à l'effet d'un tel remplacement en moyenne sur toutes les positions possibles dans S.
  - $U(m)$ dénote la distribution uniforme sur $\{1, \ldots, m\}$.

Définition (en-moyenne-remplacer-un-stable)

Soit $\epsilon : \mathbb{N} \to \mathbb{R}$ une fonction monotone décroissante. Nous disons qu'un algorithme d'apprentissage A (relativement à une fonction de perte ℓ) est en-moyenne-remplacer-un-stable avec taux $\epsilon(m)$ si pour toute distribution $\mathcal{D}$, on a

$$\mathop{\mathbb{E}}_{(S,z')\sim\mathcal{D}^{m+1},i\sim U(m)}[\ell(A(S^{(i)}), z_i)) - \ell(A(S), z_i)] \leq \epsilon(m) \ .$$

Les algorithmes stables n'"overfit" pas

Théorème

Si A est en-moyenne-remplacer-un-stable avec taux $\epsilon(m)$, alors

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \leq \epsilon(m) \ .$$

Preuve: Puisque S et z' sont tirés i.i.d. selon $\mathcal{D}$, pour tout i nous avons

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S))] = \mathop{\mathbb{E}}_{S,z'}[\ell(A(S), z')] = \mathop{\mathbb{E}}_{S,z'}[\ell(A(S^{(i)}), z_i)] \ .$$

De plus, par définition du risque empirique, nous avons

$$\mathop{\mathbb{E}}_{S}[L_S(A(S))] = \mathop{\mathbb{E}}_{S,i}[\ell(A(S), z_i)] \ .$$

Donc pour tout i, on a

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) - L_S(A(S))] = \mathop{\mathbb{E}}_{S,z'}[\ell(A(S^{(i)}), z_i)] - \mathop{\mathbb{E}}_{S,i}[\ell(A(S), z_i)] \ .$$

Le théorème est obtenu en prenant l'espérance $\mathbb{E}_{i\sim U(m)}$ des 2 côtés et en utilisant la définition précédente. ■

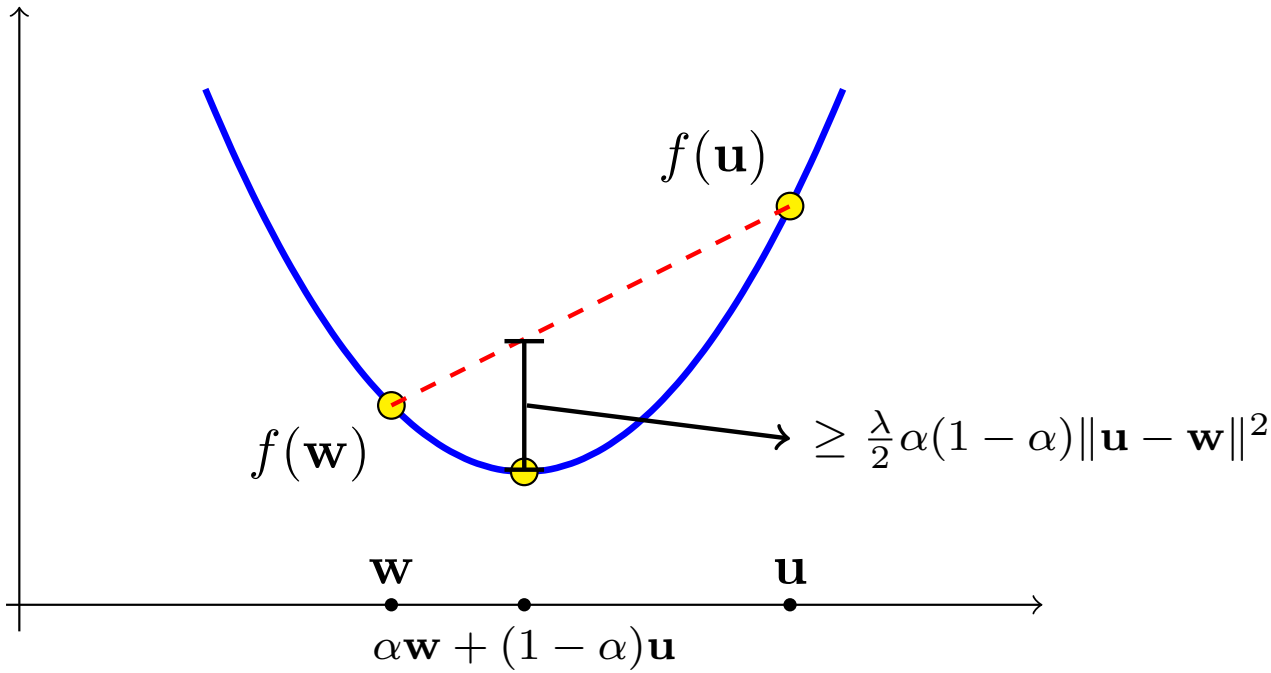
Sommaire

1. Minimisation du risque empirique régularisé
   - Régularisation de Tikhonov
   - Cas particulier : régression de ridge
2. Stabilité d'un algorithme d'apprentissage
3. Stabilité de la régularisation de Tikhonov
4. Apprendre avec la régularisation de Tikhonov
5. La DGS pour fonctions fortement convexes
6. La DGS pour la régularisation de Tikhonov
7. Garanties sur le risque : norme ou dimension ?
   - Exemple d'application : classification de texte

Fonctions fortement convexes

- Pour statuer sur la stabilité de la régularisation de Tikhonov, exploitons la forte convexité de la fonction de régularisation $\lambda||\mathbf{w}||^2$.
- Une fonction f est dite λ-fortement convexe si pour tout $\mathbf{u}$ et $\mathbf{w}$ et pour tout $\alpha \in [0, 1]$, nous avons

$$\alpha f(\mathbf{w}) + (1-\alpha) f(\mathbf{u}) - f(\alpha \mathbf{w} + (1-\alpha)\mathbf{u}) \geq \frac{\lambda}{2} \alpha (1-\alpha) ||\mathbf{w} - \mathbf{u}||^2 .$$



Fonctions fortement convexes

Si f est λ-fortement convexe et que $\mathbf{v} \in \partial f(\mathbf{u})$, alors

$$f(\mathbf{w}) \geq f(\mathbf{u}) + \langle \mathbf{w} - \mathbf{u}, \mathbf{v} \rangle + \frac{\lambda}{2} ||\mathbf{w} - \mathbf{u}||^2 .$$

- **Preuve:** En utilisant la définition de fortement convexe et en divisant par α, on trouve

$$f(\mathbf{w}) - f(\mathbf{u}) \geq \frac{f(\mathbf{u} + \alpha(\mathbf{w} - \mathbf{u})) - f(\mathbf{u})}{\alpha} + \frac{\lambda}{2}(1 - \alpha)||\mathbf{w} - \mathbf{u}||^2 .$$

- Pour tout point $\mathbf{u}$, pour tout $\mathbf{v} \in \partial f(\mathbf{u})$ et pour tout point $\mathbf{w}'$, par la définition du sous-gradient, nous avons $f(\mathbf{w}') - f(\mathbf{u}) \geq \langle \mathbf{w}' - \mathbf{u}, \mathbf{v} \rangle$.
- En choisissant $\mathbf{w}' = \mathbf{u} + \alpha(\mathbf{w} - \mathbf{u})$ nous avons alors

$$f(\mathbf{u} + \alpha(\mathbf{w} - \mathbf{u})) - f(\mathbf{u}) \geq \alpha \langle \mathbf{w} - \mathbf{u}, \mathbf{v} \rangle .$$

- Ce qui donne le résultat prétendu lorsque $\alpha \to 0$. ■

Fonctions fortement convexes

- Le résultat précédent implique que si f est λ-fortement convexe et que $f(\mathbf{u})$ est minimal (*i.e.*, $\mathbf{0} \in \partial f(\mathbf{u})$), alors

$$f(\mathbf{w}) - f(\mathbf{u}) \geq \frac{\lambda}{2}||\mathbf{w} - \mathbf{u}||^2 .$$

- Donc, le point $\mathbf{u}$ minimisant $f(\mathbf{u})$ est unique !
- Notez qu'une fonction convexe est 0-fortement convexe.
- Exercice : si f est λ-fortement convexe et g est convexe, alors $f + g$ est λ-fortement convexe.
- Exercice : démontrez que $\lambda||\mathbf{w}||^2$ est 2λ-fortement convexe.
- Exercice : démontrez que $(\langle \mathbf{w}, \mathbf{x} \rangle - y)^2$ est 0-fortement convexe sans être λ-fortement convexe pour $\lambda > 0$.

Stabilité de la régularisation de Tikhonov

- Soit $f_S(\mathbf{w}) \stackrel{\text{def}}{=} L_S(\mathbf{w}) + \lambda||\mathbf{w}||^2$. Puisque f_S est 2λ-fortement convexe et que $A(S)$ minimise f_S, pour tout $\mathbf{v}$ on a

$$f_S(\mathbf{v}) - f_S(A(S)) \geq \lambda||\mathbf{v} - A(S)||^2 .$$

- De plus, pour tout $\mathbf{v}$ et $\mathbf{u}$ et pour tout i, on a

$$\begin{aligned} f_S(\mathbf{v}) - f_S(\mathbf{u}) &= L_S(\mathbf{v}) + \lambda||\mathbf{v}||^2 - (L_S(\mathbf{u}) + \lambda||\mathbf{u}||^2) \\ &= L_{S^{(i)}}(\mathbf{v}) + \lambda||\mathbf{v}||^2 - (L_{S^{(i)}}(\mathbf{u}) + \lambda||\mathbf{u}||^2) \\ &\quad + \frac{\ell(\mathbf{v}, z_i) - \ell(\mathbf{u}, z_i)}{m} + \frac{\ell(\mathbf{u}, z') - \ell(\mathbf{v}, z')}{m} . \end{aligned}$$

Stabilité de la régularisation de Tikhonov

- En choisissant $\mathbf{v} = A(S^{(i)})$ et $\mathbf{u} = A(S)$ et en exploitant le fait que $\mathbf{v}$ minimise $L_{S^{(i)}}(\mathbf{v}) + \lambda||\mathbf{v}||^2$, on a

$$f_S(A(S^{(i)}) - f_S(A(S)) \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{m} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{m}.$$

- En combinant avec la première équation de la page précédente, et en utilisant le fait que $\ell(\cdot, z_i)$ est ρ-Lipschitzienne, on a

$$\lambda||A(S^{(i)}) - A(S)||^2 \leq \frac{\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i)}{m} + \frac{\ell(A(S), z') - \ell(A(S^{(i)}), z')}{m} \leq \frac{\rho}{m}||A(S^{(i)}) - A(S)|| + \frac{\rho}{m}||A(S) - A(S^{(i)})||.$$

Stabilité de la régularisation de Tikhonov

- Ce qui implique

$$||A(S^{(i)}) - A(S)||^2 \leq \frac{2\rho}{\lambda m}||A(S^{(i)}) - A(S)||\,,$$

- et donc

$$||A(S^{(i)}) - A(S)|| \leq \frac{2\rho}{\lambda m}\,,$$

- Puisque $\ell(\cdot, z_i)$ est ρ-Lipschitzienne, nous avons finalement

$$\ell(A(S^{(i)}), z_i) - \ell(A(S), z_i) \leq \rho||A(S^{(i)}) - A(S)|| \leq \frac{2\rho^2}{\lambda m}\,.$$

- L'espérance $\mathbb{E}_{(S,z'),i}$ de cette équation implique que A est en-moyenne-remplacer-un-stable avec taux $\frac{2\rho^2}{\lambda m}$.

Stabilité de la régularisation de Tikhonov

Nous avons donc démontré le théorème suivant.

Théorème (Stabilité de la régularisation de Tikhonov)

Supposons une fonction de perte convexe et ρ-Lipschitzienne. Alors, la régularisation de Tikhonov avec la fonction de régularisation $\lambda\|\mathbf{w}\|^2$ est en-moyenne-remplacer-un-stable avec taux de $\frac{2\,\rho^2}{\lambda\,m}$. Conséquemment,

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \leq \frac{2\,\rho^2}{\lambda\,m}\ .$$

Notez que la stabilité augmente avec λ.

Sommaire

1. Minimisation du risque empirique régularisé
   - Régularisation de Tikhonov
   - Cas particulier : régression de ridge
2. Stabilité d'un algorithme d'apprentissage
3. Stabilité de la régularisation de Tikhonov
4. Apprendre avec la régularisation de Tikhonov
5. La DGS pour fonctions fortement convexes
6. La DGS pour la régularisation de Tikhonov
7. Garanties sur le risque : norme ou dimension ?
   - Exemple d'application : classification de texte

Le compromis entre ajustement aux données et stabilité

Écrivons :

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S))] = \mathop{\mathbb{E}}_{S}[L_S(A(S))] + \mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \ .$$

- Le premier terme est faible lorsque A s'ajuste bien à S.
- Le deuxième terme est faible lorsque A est stable.
  - Il y a "overfitting" lorsque ce terme est élevé.
- λ contrôle le compromis entre ces deux termes.

Garantie pour la régularisation de Tikhonov

- Lorsque $A(S)$ effectue la régularisation de Tikhonov avec une fonction de perte ρ-Lipschitzienne, nous avons

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S)) - L_S(A(S))] \leq \frac{2\,\rho^2}{\lambda\, m}\,.$$

- Pour tout vecteur arbitraire $\mathbf{w}^*$, nous avons

$$L_S(A(S)) \leq L_S(A(S)) + \lambda\|A(S)\|^2 \leq L_S(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2\,.$$

- En effectuant l'espérance sur S et en observant que $\mathbb{E}_S[L_S(\mathbf{w}^*)] = L_{\mathcal{D}}(\mathbf{w}^*)$, nous obtenons

$$\mathop{\mathbb{E}}_{S}[L_S(A(S))] \leq L_{\mathcal{D}}(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2\,.$$

- Alors, pour tout $\mathbf{w}^*$, nous avons

$$\mathop{\mathbb{E}}_{S}[L_{\mathcal{D}}(A(S))] \leq L_{\mathcal{D}}(\mathbf{w}^*) + \lambda\|\mathbf{w}^*\|^2 + \frac{2\,\rho^2}{\lambda\, m}\,.$$

Garantie pour la régularisation de Tikhonov

- Supposons alors que nos vecteurs $\mathbf{w}^*$ de comparaison possèdent une norme bornée *i.e.*, $||\mathbf{w}^*|| \leq B$.
- La somme des deux derniers termes devient $\lambda B^2 + (2\rho^2)/(\lambda m)$.
- Cette somme est minimale et vaut $\rho B\sqrt{8/m}$ lorsque $\lambda = (\rho/B)\sqrt{2/m}$. Nous avons donc le résultat suivant.

Théorème (Garantie pour la régularisation de Tikhonov)

Soit un algorithme d'apprentissage A effectuant la régularisation de Tikhonov avec une fonction de perte ρ-Lipschitzienne et avec la fonction de régularisation $\lambda||\mathbf{w}||^2$. Lorsque $\lambda = (\rho/B)\sqrt{2/m}$, nous avons

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \leq \min_{\mathbf{w}:||\mathbf{w}||\leq B} L_{\mathcal{D}}(\mathbf{w}) + \rho B\sqrt{\frac{8}{m}} \ .$$

Donc, pour tout $\epsilon > 0$, lorsque $m \geq 8\rho^2 B^2/\epsilon^2$, pour tout $\mathcal{D}$, on a $\mathbb{E}_S[L_{\mathcal{D}}(A(S))] \leq \min_{\mathbf{w}:||\mathbf{w}||\leq B} L_{\mathcal{D}}(\mathbf{w}) + \epsilon$.

Choisir λ

- Le choix de $\lambda = (\rho/B)\sqrt{2/m}$ est celui minimisant la borne supérieure sur $\mathbb{E}_S[L_{\mathcal{D}}(A(S))]$ et nous permet d'obtenir un risque à ϵ près de l'optimal parmi tous les prédicteurs ayant une norme d'au plus B lorsque $m \geq 8\rho^2 B^2/\epsilon^2$. Mais on ne connait pas B.
- Plus B est élevée, plus faible sera l'erreur d'approximation $\min_{\mathbf{w}:||\mathbf{w}||\leq B} L_{\mathcal{D}}(\mathbf{w})$.
- De $\lambda = (\rho/B)\sqrt{2/m}$, on déduit que B est élevée lorsque λ est faible.
- Donc diminuer λ contribue à diminuer l'erreur d'approximation.
- Par contre, $m \geq 8\rho^2 B^2/\epsilon^2$ implique $\epsilon^2 \geq 8\rho^2 B^2/m$. Donc si on augmente B, on augmente aussi l'erreur d'estimation ϵ.
- Donc diminuer λ contribue à augmenter l'erreur d'estimation.
- Donc λ est un paramètre permettant de trouver le bon compromis entre erreur d'approximation et erreur d'estimation. En pratique, λ est choisi à l'aide d'un ensemble de validation.

Sommaire

1. Minimisation du risque empirique régularisé
   - Régularisation de Tikhonov
   - Cas particulier : régression de ridge
2. Stabilité d'un algorithme d'apprentissage
3. Stabilité de la régularisation de Tikhonov
4. Apprendre avec la régularisation de Tikhonov
5. La DGS pour fonctions fortement convexes
6. La DGS pour la régularisation de Tikhonov
7. Garanties sur le risque : norme ou dimension ?
   - Exemple d'application : classification de texte

DGS pour fonctions fortement convexes

- Le théorème précédent implique que la régularisation de Tikhonov apprends, au sens PAC agnostique, les problèmes convexes Lipschitziens bornés.
- La complexité d'échantillon dépend de la constante de Lipschitz de la fonction de perte utilisée et de la norme maximale des prédicteurs de la classe $\mathcal{H}$.
- La régularisation de Tikhonov consiste à trouver $\mathbf{w}$ minimisant $L_S(\mathbf{w}) + \lambda||\mathbf{w}||^2$.
- Il est possible d'effectuer la DGS pour résoudre ce problème de minimisation.
- Puisqu'il s'agit d'une fonction fortement convexe, examinons d'abord la DGS pour le cas plus général de la minimisation de fonctions fortement convexes.

DGS pour fonctions fortement convexes

DGS pour fonctions fortement convexes :

- **Objectif :** Minimiser $f(\mathbf{w})$ lorsque f est λ-fortement convexe.
- **initialiser :** $\mathbf{w}^{(1)} = \mathbf{0}$
- **pour** $t = 1, 2, \ldots, T$
  - tirer un vecteur $\mathbf{v}_t$ tel que $\mathbb{E}[\mathbf{v}_t | \mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$.
  - choisir $\eta_t = 1/(\lambda t)$.
  - mise à jour : $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta_t \mathbf{v}_t$
- **sortie :** $\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)}$

Remarques :

- On choisi η_t dépendant de t pour exploiter la forte convexité de f.
- On tire $\mathbf{v}_t$, indépendamment de $\mathbf{v}_1, \ldots, \mathbf{v}_{t-1}$ lorsque $\mathbf{w}^{(t)}$ est fixe, selon une distribution ayant la propriété que $\mathbb{E}[\mathbf{v}_t | \mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$.
- Cela nous donnera une garantie pour $\mathbb{E}\, f(\bar{\mathbf{w}})$.

Convergence de la DGS pour fonctions fortement convexes

Théorème

Soit f une fonction λ-fortement convexe et $\mathbb{E}[||\mathbf{v}_t||^2] \leq \rho^2$ pour tous les vecteurs aléatoires $\mathbf{v}_t$ échantillonnés dans la DGS pour fonctions fortement convexes. Soit $\mathbf{w}^\star \in \operatorname{argmin}_{\mathbf{w}\in\mathbb{R}^d} f(\mathbf{w})$. Alors

$$\mathbb{E}[f(\bar{\mathbf{w}})] - f(\mathbf{w}^\star) \leq \frac{\rho^2}{2\lambda T}(1+\log(T))\,.$$

Preuve: Puisque f est λ-fortement convexe et que $\mathbb{E}[\mathbf{v}_t|\mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$, pour tout t on a

$$\langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbb{E}[\mathbf{v}_t|\mathbf{w}^{(t)}]\rangle \geq f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) + \frac{\lambda}{2}||\mathbf{w}^{(t)} - \mathbf{w}^\star||^2\,.$$

Notez que $\mathbf{v}_t$ est tiré indépendamment de $\mathbf{v}_1, \ldots, \mathbf{v}_{t-1}$ lorsque $\mathbf{w}^{(t)}$ est fixe. En prenant $\mathbb{E}_{\mathbf{v}_1 \ldots \mathbf{v}_T}$ de chaque côté, on a pour tout t

$$\mathop{\mathbb{E}}_{\mathbf{v}_1\ldots\mathbf{v}_T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t\rangle \geq \mathop{\mathbb{E}}_{\mathbf{v}_1\ldots\mathbf{v}_T} \left(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) + \frac{\lambda}{2}||\mathbf{w}^{(t)} - \mathbf{w}^\star||^2 \right)\,.$$

Convergence de la DGS pour fonctions fortement convexes

En utilisant la règle de mise à jour $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta_t \mathbf{v}_t$, pour tout t on a

$$||\mathbf{w}^{(t)} - \mathbf{w}^\star||^2 - ||\mathbf{w}^{(t+1)} - \mathbf{w}^\star||^2 = 2\eta_t \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle - \eta_t^2 ||\mathbf{v}_t||^2 .$$

En prenant l'espérance des deux côtés et en utilisant $\mathbb{E}[||\mathbf{v}_t||^2] \leq \rho^2$, on a

$$\mathop{\mathbb{E}}_{\mathbf{v}_1 \dots \mathbf{v}_T} \langle \mathbf{w}^{(t)} - \mathbf{w}^\star, \mathbf{v}_t \rangle \leq \frac{\mathbb{E}_{\mathbf{v}_1 \dots \mathbf{v}_T}[||\mathbf{w}^{(t)} - \mathbf{w}^\star||^2 - ||\mathbf{w}^{(t+1)} - \mathbf{w}^\star||^2]}{2\eta_t} + \frac{\eta_t}{2}\rho^2 .$$

En combinant cette équation avec celle de la page précédente et en sommant sur t, on a

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\mathbf{v}_1 \dots \mathbf{v}_T} \left(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) + \frac{\lambda}{2} ||\mathbf{w}^{(t)} - \mathbf{w}^\star||^2 \right)$$

$$\leq \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\mathbf{v}_1 \dots \mathbf{v}_T} \left(\frac{[||\mathbf{w}^{(t)} - \mathbf{w}^\star||^2 - ||\mathbf{w}^{(t+1)} - \mathbf{w}^\star||^2]}{2\eta_t} + \frac{\eta_t}{2}\rho^2 \right) .$$

Convergence de la DGS pour fonctions fortement convexes

Alors, puisque la somme des espérances est l'espérance de la somme, on a

$$\mathop{\mathbb{E}}_{\mathbf{v}_1 \dots \mathbf{v}_T} \sum_{t=1}^{T} \Big(f(\mathbf{w}^{(t)}) - f(\mathbf{w}^\star) \Big) \;\leq$$

$$\mathop{\mathbb{E}}_{\mathbf{v}_1 \dots \mathbf{v}_T} \sum_{t=1}^{T} \left(\frac{||\mathbf{w}^{(t)} - \mathbf{w}^\star||^2 - ||\mathbf{w}^{(t+1)} - \mathbf{w}^\star||^2}{2\eta_t} - \frac{\lambda}{2}||\mathbf{w}^{(t)} - \mathbf{w}^\star||^2 + \frac{\eta_t}{2}\rho^2 \right) .$$

En utilisant $\eta_t = 1/(\lambda t)$, et en utilisant $g(t) \stackrel{\text{def}}{=} (t-1)||\mathbf{w}^{(t)} - \mathbf{w}^\star||^2$, on a

$$\sum_{t=1}^{T} \left(\frac{||\mathbf{w}^{(t)} - \mathbf{w}^\star||^2 - ||\mathbf{w}^{(t+1)} - \mathbf{w}^\star||^2}{2\eta_t} - \frac{\lambda}{2}||\mathbf{w}^{(t)} - \mathbf{w}^\star||^2 \right)$$

$$= \sum_{t=1}^{T} \left(\frac{\lambda t}{2} \left[\frac{g(t)}{t-1} - \frac{g(t+1)}{t} \right] - \frac{\lambda}{2} \frac{g(t)}{t-1} \right) = \sum_{t=1}^{T} \frac{\lambda}{2} \left(g(t) - g(t+1) \right)$$

$$= -\frac{\lambda}{2} g(T+1) = -\frac{\lambda}{2} T ||\mathbf{w}^{(T+1)} - \mathbf{w}^\star||^2 \leq 0 \, .$$

Convergence de la DGS pour fonctions fortement convexes

Alors, en bornant la somme par l'intégrale, nous avons

$$\left(\mathop{\mathbb{E}}_{\mathbf{v}_1 \dots \mathbf{v}_T} \sum_{t=1}^{T} f(\mathbf{w}^{(t)}) \right) - T f(\mathbf{w}^\star) \;\leq\; \frac{\rho^2}{2\lambda} \sum_{t=1}^{T} \frac{1}{t} \;\leq\; \frac{\rho^2}{2\lambda}(1 + \log(T)) \,.$$

En divisant pat T et en utilisant l'inégalité de Jensen pour f convexe

$$f\left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)} \right) \leq \frac{1}{T} \sum_{t=1}^{T} f(\mathbf{w}^{(t)}) \,,$$

nous obtenons finalement que

$$\mathop{\mathbb{E}}_{\mathbf{v}_1 \dots \mathbf{v}_T} f(\bar{\mathbf{w}}) \;=\; \mathop{\mathbb{E}}_{\mathbf{v}_1 \dots \mathbf{v}_T} f\left(\frac{1}{T} \sum_{t=1}^{T} \mathbf{w}^{(t)} \right) \leq f(\mathbf{w}^\star) + \frac{\rho^2}{2\lambda T}(1 + \log(T)) \,.$$

■

Sommaire

1. Minimisation du risque empirique régularisé
   - Régularisation de Tikhonov
   - Cas particulier : régression de ridge
2. Stabilité d'un algorithme d'apprentissage
3. Stabilité de la régularisation de Tikhonov
4. Apprendre avec la régularisation de Tikhonov
5. La DGS pour fonctions fortement convexes
6. La DGS pour la régularisation de Tikhonov
7. Garanties sur le risque : norme ou dimension ?
   - Exemple d'application : classification de texte

DGS pour la régularisation de Tikhonov

- Examinons maintenant le cas $f(\mathbf{w}) = L_S(\mathbf{w}) + \lambda||\mathbf{w}||^2$. Donc,

$$f(\mathbf{w}) = \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{w}, z_i) + \lambda||\mathbf{w}||^2 = \mathop{\mathbb{E}}_{i\sim U(m)} \ell(\mathbf{w}, z_i) + \lambda||\mathbf{w}||^2 \,.$$

- Chaque $\mathbf{v}_t$ doit satisfaire $\mathbb{E}[\mathbf{v}_t|\mathbf{w}^{(t)}] \in \partial f(\mathbf{w}^{(t)})$.
- Soit, $\mathbf{u}_i^{(t)} \in \partial \ell(\mathbf{w}^{(t)}, z_i)$. Alors on a

$$\mathop{\mathbb{E}}_{i\sim U(m)} \mathbf{u}_i^{(t)} \;=\; \frac{1}{m}\sum_{i=1}^{m} \mathbf{u}_i^{(t)} \;\in\; \partial L_S(\mathbf{w}^{(t)}) \,.$$

- Puisque $2\lambda\mathbf{w}$ est le gradient de $\lambda||\mathbf{w}||^2$, on a

$$2\lambda\mathbf{w}^{(t)} + \mathop{\mathbb{E}}_{i\sim U(m)} \mathbf{u}_i^{(t)} \;\in\; \partial f(\mathbf{w}^{(t)}) \,.$$

- Soit $\mathbf{v}_t \overset{\text{def}}{=} 2\lambda\mathbf{w}^{(t)} + \mathbf{u}_i^{(t)}$, avec $i \sim U(m)$. Alors

$$\mathbb{E}[\mathbf{v}_t|\mathbf{w}^{(t)}] \;=\; 2\lambda\mathbf{w}^{(t)} + \mathop{\mathbb{E}}_{i\sim U(m)} \mathbf{u}_i^{(t)} \;\in\; \partial f(\mathbf{w}^{(t)}) \,.$$

DGS pour la régularisation de Tikhonov

Puisque $f(\mathbf{w})$ est 2λ-fortement convexe, notre algorithme est le suivant :

DGS pour la régularisation de Tikhonov (version préliminaire) :

- **Objectif :** Minimiser $L_S(\mathbf{w}) + \lambda||\mathbf{w}||^2$.
- **initialiser :** $\mathbf{w}^{(1)} = \mathbf{0}$
- **pour** $t = 1, 2, \ldots, T$
  - tirer $i \sim U(m)$.
  - soit $\mathbf{u}_i^{(t)} \in \partial\ell(\mathbf{w}^{(t)}, z_i)$.
  - soit $\mathbf{v}_t = 2\lambda\mathbf{w}^{(t)} + \mathbf{u}_i^{(t)}$.
  - soit $\eta_t = 1/(2\lambda t)$.
  - mise à jour : $\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \eta_t\mathbf{v}_t \quad = \left(1 - \frac{1}{t}\right)\mathbf{w}^{(t)} - \frac{1}{2\lambda t}\mathbf{u}_i^{(t)}$.
- **sortie :** $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$

Étant donné l'égalité pour la mise à jour de $\mathbf{w}$, nous pouvons simplifier cet algorithme en la version suivante.

DGS pour la régularisation de Tikhonov

DGS pour la régularisation de Tikhonov (version définitive) :

- **Objectif :** Minimiser $L_S(\mathbf{w}) + \lambda||\mathbf{w}||^2$.
- **initialiser :** $\mathbf{w}^{(1)} = \mathbf{0}$
- **pour** $t = 1, 2, \ldots, T$
  - tirer $i \sim U(m)$.
  - soit $\mathbf{u}_i^{(t)} \in \partial\ell(\mathbf{w}^{(t)}, z_i)$.
  - mise à jour : $\mathbf{w}^{(t+1)} = \left(1 - \frac{1}{t}\right)\mathbf{w}^{(t)} - \frac{1}{2\lambda t}\mathbf{u}_i^{(t)}$.
- **sortie :** $\bar{\mathbf{w}} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}^{(t)}$

Garantie de la DGS pour la régularisation de Tikhonov

- Notez que le théorème de convergence s'applique seulement lorsque $\mathbb{E}[||\mathbf{v}_t||^2] \leq \rho^2$ à chaque itération de la DGS.
- Ici, on a $\mathbf{v}_t = 2\lambda \mathbf{w}^{(t)} + \mathbf{u}_i^{(t)}$.
- L'équation de mise à jour implique

$$
\begin{aligned}
\mathbf{w}^{(t+1)} &= \frac{t-1}{t}\mathbf{w}^{(t)} - \frac{1}{2\lambda t}\mathbf{u}_i^{(t)} \\
&= \frac{t-1}{t}\left(\frac{t-2}{t-1}\mathbf{w}^{(t-1)} - \frac{1}{2\lambda(t-1)}\mathbf{u}_i^{(t-1)}\right) - \frac{1}{2\lambda t}\mathbf{u}_i^{(t)} \\
&= \frac{t-1}{t}\frac{t-2}{t-1}\mathbf{w}^{(t-1)} - \frac{1}{2\lambda t}\mathbf{u}_i^{(t-1)} - \frac{1}{2\lambda t}\mathbf{u}_i^{(t)} \\
&\cdots \quad \cdots\cdots\cdots \\
&= -\frac{1}{2\lambda t}\sum_{t'=1}^{t}\mathbf{u}_i^{(t')}
\end{aligned}
$$

Garantie de la DGS pour la régularisation de Tikhonov

- Nous supposons une fonction de perte ρ-Lipschitzienne.
- Puisque $\mathbf{u}_i^{(t)} \in \partial \ell(\mathbf{w}^{(t)}, z_i)$, alors $||\mathbf{u}_i^{(t)}|| \leq \rho$.
- Par l'inégalité du triangle, on a

$$||\mathbf{w}^{(t)}|| \leq \frac{1}{2\lambda(t-1)} \sum_{t'=1}^{t-1} ||\mathbf{u}_i^{(t')}|| \leq \frac{\rho}{2\lambda} \,.$$

- En conséquence, pour tout t, on a

$$||\mathbf{v}_t|| = ||2\lambda \mathbf{w}^{(t)} + \mathbf{u}_i^{(t)}|| \leq ||2\lambda \mathbf{w}^{(t)}|| + ||\mathbf{u}_i^{(t)}|| \leq \rho + \rho = 2\rho \,.$$

- Donc, puisque $\mathbb{E}[||\mathbf{v}_t||^2] \leq 4\rho^2$ et que $f_S(\mathbf{w}) \stackrel{\text{def}}{=} L_S(\mathbf{w}) + \lambda ||\mathbf{w}||^2$ est 2λ-fortement convexe, on a pour tout S :

$$\mathbb{E}\, f_S(\bar{\mathbf{w}}) \leq f_S(\mathbf{w}_S) + \frac{\rho^2}{\lambda T}(1 + \log(T)) \;;\; \text{pour } \mathbf{w}_S \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{w} \in \mathbb{R}^d} f_S(\mathbf{w}) \,.$$

Note : l'espérance est sur $\mathbf{i} \stackrel{\text{def}}{=} (i_1, \ldots, i_T) \sim U(m)^T$.

Garantie de la DGS pour la régularisation de Tikhonov

- Pour tout $\lambda > 0$ et pour tout S, nous avons alors

$$\mathop{\mathbb{E}}_{\mathbf{i}\sim U(m)^T} f_S(\bar{\mathbf{w}}) - f_S(\mathbf{w}_S) \leq \epsilon' \,, \text{ avec } \epsilon' \stackrel{\text{def}}{=} \frac{\rho^2}{\lambda}\left[\frac{1+\log(T)}{T}\right].$$

- Pour borner supérieurement

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{\mathbf{i}\sim U(m)^T} L_{\mathcal{D}}(\bar{\mathbf{w}}) \;-\; \min_{\mathbf{w}:||\mathbf{w}||\leq B} L_{\mathcal{D}}(\mathbf{w})\,,$$

utilisons le fait que $f_S(\mathbf{w})$ est 2λ-fortement convexe. Ce qui implique

$$f_S(\bar{\mathbf{w}}) - f_S(\mathbf{w}_S) \;\geq\; \lambda||\bar{\mathbf{w}} - \mathbf{w}_S||^2\,,\; \forall S, \forall \mathbf{i}, \forall \lambda > 0\,.$$

- Lorsque T satisfait la condition ci-haut, on a donc

$$\mathop{\mathbb{E}}_{\mathbf{i}\sim U(m)^T} ||\bar{\mathbf{w}} - \mathbf{w}_S||^2 \;\leq\; \frac{\epsilon'}{\lambda}\,,\; \forall S, \forall \lambda > 0\,.$$

Garantie de la DGS pour la régularisation de Tikhonov

- En utilisant l'inégalité de Jensen pour $\sqrt{\cdot}$ et le fait que $L_{\mathcal{D}}(\mathbf{w})$ est ρ-Lipschitzienne, nous avons pour tout S

$$\sqrt{\frac{\epsilon'}{\lambda}} \geq \sqrt{\mathop{\mathbb{E}}_{\mathbf{i}} ||\bar{\mathbf{w}} - \mathbf{w}_S||^2} \geq \mathop{\mathbb{E}}_{\mathbf{i}} ||\bar{\mathbf{w}} - \mathbf{w}_S|| \geq \mathop{\mathbb{E}}_{\mathbf{i}} \frac{1}{\rho} \left[L_{\mathcal{D}}(\bar{\mathbf{w}}) - L_{\mathcal{D}}(\mathbf{w}_S) \right] .$$

- En prenant l'espérance sur S, nous avons donc

$$\mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{\mathbf{i}} L_{\mathcal{D}}(\bar{\mathbf{w}}) \leq \mathop{\mathbb{E}}_{S} L_{\mathcal{D}}(\mathbf{w}_S) + \rho \sqrt{\frac{\epsilon'}{\lambda}} .$$

- En utilisant $\mathbf{w}^* \stackrel{\text{def}}{=} \operatorname{argmin}_{\mathbf{w}:||\mathbf{w}|| \leq B} L_{\mathcal{D}}(\mathbf{w})$, et que

$$\mathop{\mathbb{E}}_{S} L_{\mathcal{D}}(\mathbf{w}_S) \leq L_{\mathcal{D}}(\mathbf{w}^*) + \epsilon$$

lorsque $\lambda = (\rho/B)\sqrt{2/m}$ et $m \geq 8\rho^2 B^2/\epsilon^2$, nous avons

$$\mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{\mathbf{i}} L_{\mathcal{D}}(\bar{\mathbf{w}}) \leq L_{\mathcal{D}}(\mathbf{w}^*) + \epsilon + \rho \sqrt{\frac{\epsilon'}{\lambda}} = L_{\mathcal{D}}(\mathbf{w}^*) + \epsilon + \frac{\rho^2}{\lambda} \sqrt{\frac{1 + \log(T)}{T}} .$$

Garantie de la DGS pour la régularisation de Tikhonov

- Donc, lorsque $\lambda = (\rho/B)\sqrt{2/m}$ et $m \geq 8\rho^2 B^2/\epsilon^2$, nous avons

$$\begin{aligned}\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{\mathbf{i}} L_{\mathcal{D}}(\bar{\mathbf{w}}) &\leq L_{\mathcal{D}}(\mathbf{w}^*) + \epsilon + \frac{\rho^2}{\lambda}\sqrt{\frac{1+\log(T)}{T}} \\ &= L_{\mathcal{D}}(\mathbf{w}^*) + \epsilon + \rho B\sqrt{\frac{m}{2}\left[\frac{1+\log(T)}{T}\right]}.\end{aligned}$$

- Le dernier terme devient $\leq \epsilon''$ dès que

$$\frac{T}{1+\log(T)} \geq \frac{m\rho^2 B^2}{2(\epsilon'')^2}.$$

- Donc, dans ce cas

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{\mathbf{i}\sim U(m)^T} L_{\mathcal{D}}(\bar{\mathbf{w}}) \leq \min_{\mathbf{w}:||\mathbf{w}||\leq B} L_{\mathcal{D}}(\mathbf{w}) + \varepsilon\,,$$

pour $\epsilon = \epsilon'' = \varepsilon/2$.

Garantie de la DGS pour la régularisation de Tikhonov

Nous avons donc le résultat suivant

Théorème

Considérez la DGS pour minimiser $L_S(\mathbf{w}) + \lambda||\mathbf{w}||^2$ *avec* $|S| = m$ *et une fonction de perte* $\ell(\cdot, z)$ *convexe et* ρ*-Lipschitzienne pour tout* $z \in Z$*. Lorsque* $\lambda = (\rho/B)\sqrt{2/m}$ *et* $m \geq 32\rho^2 B^2/\varepsilon^2$ *pour* $B > 0$*, nous avons*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{\mathbf{i}\sim U(m)^T} L_{\mathcal{D}}(\bar{\mathbf{w}}) \;\leq\; \min_{\mathbf{w}:||\mathbf{w}||\leq B} L_{\mathcal{D}}(\mathbf{w}) + \varepsilon\,,$$

lorsque le nombre T *d'itérations satisfait*

$$\frac{T}{1+\log(T)} \;\geq \frac{2m\rho^2 B^2}{\varepsilon^2}\,.$$

Sommaire

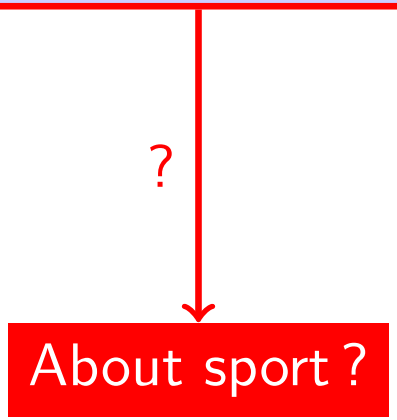
1. Minimisation du risque empirique régularisé
   - Régularisation de Tikhonov
   - Cas particulier : régression de ridge
2. Stabilité d'un algorithme d'apprentissage
3. Stabilité de la régularisation de Tikhonov
4. Apprendre avec la régularisation de Tikhonov
5. La DGS pour fonctions fortement convexes
6. La DGS pour la régularisation de Tikhonov
7. Garanties sur le risque : norme ou dimension ?
   - Exemple d'application : classification de texte

Dimension vs. norme

- Rappel : $\mathrm{VCdim}(\mathcal{H})$ augmente souvent avec le nombre d de paramètres modifiables.
  - Dans ce cas, la complexité d'échantillon $m_{\mathcal{H}}(\epsilon, \delta)$ augmente avec d.
- Pour la régularisation de Tikhonov et la DGS, nous utilisons d paramètres, mais la complexité d'échantillon dépend de la norme maximale B d'une classe de prédicteurs.
  - Ça dépends aussi de la constante de Lipschitz ρ de la fonction de perte.
  - Mais ça ne dépends pas de d.
- L'approche qui nous donnera le meilleur prédicteur ($\mathrm{ERM}_{\mathcal{H}}(S)$ ou régularisation de Tikhonov) dépends de certaines propriétés de la distribution $\mathcal{D}$.

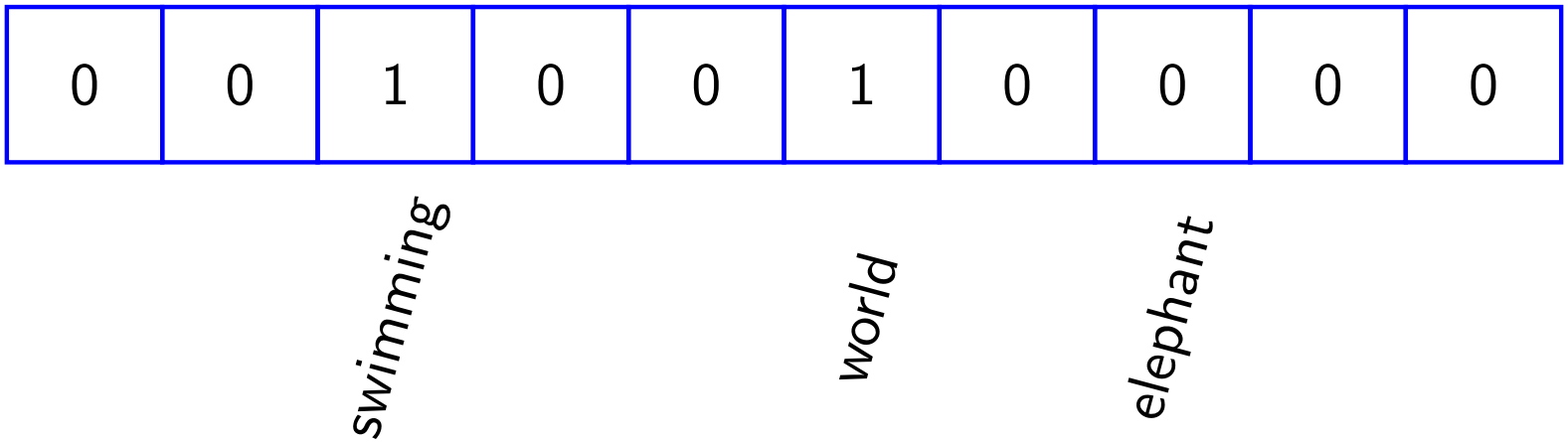
Exemple : classification de documents

Signs all encouraging for Phelps in comeback. He did not win any gold medals or set any world records but Michael Phelps ticked all the boxes he needed in his comeback to competitive swimming.



Représentation "bag-of-words"

Signs all encouraging for Phelps in comeback. He did not win any gold medals or set any world records but Michael Phelps ticked all the boxes he needed in his comeback to competitive swimming.



Classification de documents

- Ici, $\mathcal{X} = \{\mathbf{x} \in \{0,1\}^{d+1} : \|\mathbf{x}\|^2 \leq R^2, x_0 = 1\}$.
- Chaque document $\mathbf{x} \in \mathcal{X}$ est représenté par un "bag of words" :
  - La composante x_0 de chaque document $\mathbf{x}$ est fixée à 1 pour utiliser un biais.
  - Le dictionnaire possède d mots.
  - On a au plus R^2 mots dans chaque document $\mathbf{x}$.
- Soit $\mathcal{Y} = \{\pm 1\}$ (e.g., le sport est, oui ou non, un sujet du document).
- Les prédicteurs sont les classificateurs linéaires : $\mathbf{x} \mapsto \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle)$
- On s'attends à ce que w_i soit grand (positif) pour les mots indicatifs du sujet sport et que w_i soit petit (negatif) pour les mots indicatif que sport n'est pas un sujet.
- Utilisation du "hinge-loss" : $\ell(\mathbf{w}, (\mathbf{x}, y)) = [1 - y\langle \mathbf{w}, \mathbf{x} \rangle]_+$

Dimension vs. norme

- Ici, $\mathrm{VCdim}(\mathcal{H}) = d + 1$, mais d est typiquement très grand (le nombre de mots dans le dictionnaire).
- La fonction de perte est convexe et R-Lipschitzienne.
- Propriété de la distribution $\mathcal{D}$ (donc, de la tâche d'apprentissage) : Si le nombre de mots pertinents est petit, c'est qu'il existe un $\mathbf{w}^\star$ de petite norme ayant un faible risque $L_{\mathcal{D}}(\mathbf{w}^\star)$.
- Dans ce cas, la complexité d'échantillon dépends de $R^2\|\mathbf{w}^\star\|^2$, mais ne de dépends pas d.
- Mais, il existe des situations opposées (*i.e.*, d'autres tâches d'apprentissage) où d est beaucoup plus petit que $R^2\|\mathbf{w}^\star\|^2$.

Résumé

- La régularisation de Tikhonov.
  - La régression de ridge comme un cas particulier important.
- La définition de stabilité d'un algorithme d'apprentissage.
  - Les algorithmes stables n'"overfit" pas.
- La régularisation de Tikhonov est un algorithme stable (lorsqu'utilisé avec une fonction de perte convexe et Lipschitzienne).
  - Car sa fonction de régularisation est fortement convexe.
- La régularisation de Thikhonov est un algorithme d'apprentissage pour les problèmes convexes-Lipschitziens-bornés.
  - La complexité d'échantillon dépends de la constante de Lipschitz de la fonction de perte utilisée et de la norme maximale de la classe de prédicteurs.
  - La complexité d'échantillon ne dépends pas du nombre de paramètres modifiables (*e.g.*, la dimension des vecteurs de caractéristiques).
  - Cela peut s'avérer très utile pour l'apprentissage de certaines tâches, *e.g.*, la classification de documents.