IFT-7002 Fondements de l'apprentissage machine

Apprentissage PAC-Bayésien
et stabilité algorithmique via l'information mutuelle

Mario Marchand
Université Laval

Hiver 2024

Sommaire

1. Introduction et définitions
2. La KL divergence et le changement de mesure Donsker-Varadhan
3. Théorème PAC-Bayes général et bornes PAC-Bayes
4. Algorithme d'apprentissage PAC-Bayésien et sa garantie PAC
   - Apprentissage PAC-Bayésien en pratique et Gaussiennes isotropes
5. Garantie PAC-Bayes pour algorithmes retournant distribution Q sur $\mathcal{W}$
6. Garantie PAC-Bayes pour alg. stochastiques et information mutuelle
7. Descente de gradient stochastique avec dynamique de Langevin

Qu'est-ce que l'approche PAC-Bayes ?

- L'approche PAC-Bayes, initiée par McAllester (1999), visait à obtenir des garanties PAC pour des algorithmes d'apprentissage Bayésiens.
- Cependant il s'agit d'une approche "fréquentiste" où le but est de trouver un prédicteur de risque minimal par rapport à une fonction de perte donnée.
- Cette approche a permis d'obtenir d'excellentes garanties sur le risque des SVM et des réseaux de neurones en plus d'établir un lien entre les approches Bayésiennes et fréquentistes et la stabilité algorithmique.

Définitions

- Soit $\mathcal{X}$ l'espace des instances, $\mathcal{Y}$, l'espace des étiquettes et $\mathcal{Z} \stackrel{\text{def}}{=} \mathcal{X} \times \mathcal{Y}$ l'espace des exemples.
- Chaque exemple $z = (x, y) \in \mathcal{Z}$ est la réalisation (ou l'observation) de la variable aléatoire $Z = (X, Z)$ est distribuée selon $\mathcal{D}$.
- Chaque échantillon s de m exemples est la réalisation de la variable aléatoire S distribuée selon $\mathcal{D}^m$.
- Chaque modèle prédictif est représenté par un vecteur $w \in \mathcal{W} \subseteq \mathbb{R}^d$
  - Comme les poids d'un classificateur linéaire ou d'un réseau de neurones.
- Ayant un échantillon s d'exemples, l'approche PAC-Bayes s'intéresse aux algorithmes produisant une "bonne" distribution Q sur $\mathcal{W}$
- On supposera que ces distributions Q possèdent une densité q telle que $q(w)$ désigne le densité de probabilité de Q au point $w \in \mathcal{W}$.

Définitions (suite)

- Chaque prédicteur w est la réalisation de la variable aléatoire W distribué selon une distribution Q choisie.
- Pour tout $w \in \mathcal{W}$,

$$L_{\mathcal{D}}(w) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{Z\sim\mathcal{D}} \ell(w, Z) \quad ; \quad L_S(w) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \ell(w, Z_i)\,.$$

- Étant donné que S est distribué selon $\mathcal{D}^m$, dénoté par $S \sim \mathcal{D}^m$, nous avons

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} L_S(w) \;=\; L_{\mathcal{D}}(w)\,.$$

- On utilisera aussi

$$\ell(Q, z) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{W\sim Q} \ell(W, z), \quad \forall z \in \mathcal{Z}$$

$$L_{\mathcal{D}}(Q) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{W\sim Q} L_{\mathcal{D}}(W) \quad ; \quad L_S(Q) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{W\sim Q} L_S(W)\,.$$

Distributions Q et P

- Considérons une fonction $f : \mathcal{W} \times \mathcal{Z}^m \to \mathbb{R}$.
- Par exemple, cela pourrait être $f(W, S) = L_{\mathcal{D}}(W) - L_S(W)$.
- Dans ce cas, une borne supérieure sur $\mathbb{E}_{W \sim Q}\, f(W, S)$ nous donnera une borne supérieure sur $L_{\mathcal{D}}(Q) - L_S(Q)$.
- Ayant une distribution P sur $\mathcal{W}$ fixée *a priori*, nous chercherons à borner $\mathbb{E}_{W \sim Q}\, f(W, S)$ simultanément pour toute distribution Q qui est absolument continue par rapport à P, dénoté par $Q \ll P$.
  - Lorsque les densités p et q existent, cela signifie que

$$\forall w \in \mathcal{W} : p(w) = 0 \implies q(w) = 0\,.$$

- Pour toute distribution Q sur un ensemble $\mathcal{W}$ ayant une fonction de densité q, le support de Q désigne l'ensemble

$$\operatorname{supp}(Q) \stackrel{\text{def}}{=} \{w \in \mathcal{W} : q(w) > 0\}\,.$$

- Donc, on a $\operatorname{supp}(Q) \subseteq \operatorname{supp}(P)$ lorsque $Q \ll P$.

Sommaire

1. Introduction et définitions
2. La KL divergence et le changement de mesure Donsker-Varadhan
3. Théorème PAC-Bayes général et bornes PAC-Bayes
4. Algorithme d'apprentissage PAC-Bayésien et sa garantie PAC
   - Apprentissage PAC-Bayésien en pratique et Gaussiennes isotropes
5. Garantie PAC-Bayes pour algorithmes retournant distribution Q sur $\mathcal{W}$
6. Garantie PAC-Bayes pour alg. stochastiques et information mutuelle
7. Descente de gradient stochastique avec dynamique de Langevin

Divergence Kullback-Leibler

- Cette borne sur $\mathbb{E}_{W\sim Q}\, f(W,S)$ dépends de la KL-divergence $D(Q\|P)$ entre Q et le "prior" P qui encode notre connaissance a priori sur les bonnes régions de $\mathcal{W}$.
- Pour toutes distributions Q et P sur $\mathcal{W}$ telles que $Q \ll P$, on a

$$D(Q\|P) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{W\sim Q} \ln\left(\frac{q(W)}{p(W)}\right) = \int_{\mathcal{W}} q(w) \ln\left(\frac{q(w)}{p(w)}\right) dw\,.$$

  (On a $D(Q\|P) = \infty$ si $\exists w : p(w) = 0 \wedge q(w) > 0$).
- Propriétés (voir Cover et Thomas, 1991) :
  - $D(Q\|P) \geq 0\ \forall (Q,P)$.
  - $D(Q\|P) = 0$ ssi $Q = P$.
  - En général on a $D(Q\|P) \neq D(P\|Q)$.

Changement de mesure de Donsker-Varadhan

Toutes les bornes PAC-Bayes sont une conséquence du théorème suivant.

Théorème (changement de mesure de Donsker-Varadhan)

Soit Q et P deux distributions sur $\mathcal{W}$ et soit $g : \mathcal{W} \to \mathbb{R}$ une fonction mesurable. Nous avons

$$\mathop{\mathbb{E}}_{W\sim Q} g(W) \leq \; D(Q\|P) + \ln \left[\mathop{\mathbb{E}}_{W\sim P} e^{g(W)} \right] .$$

En fait, pour tout Q et P, il existe toujours g où l'égalité est atteinte (corollaire 4.15 de Boucheron, Lugosi et Massart, 2013). On obtient donc une forme variationnelle pour la KL divergence :

$$D(Q\|P) \;=\; \sup_{g\in\mathcal{G}_P} \left\{ \mathop{\mathbb{E}}_{W\sim Q} g(W) - \ln \left[\mathop{\mathbb{E}}_{W\sim P} e^{g(W)} \right] \right\} ,$$

où $\mathcal{G}_P \stackrel{\text{def}}{=} \{g : \mathcal{W} \to \mathbb{R} \quad \text{t.q. } \mathbb{E}_{W\sim P}\, e^{g(W)} < \infty\}$.

- **Preuve:** Puisque $e^{g(w)} \geq 0$, on a

$$\begin{aligned}
\mathop{\mathbb{E}}_{W\sim P} e^{g(W)} &= \int_{\mathcal{W}} dw\, p(w) e^{g(w)} \\
&= \int_{\mathcal{W}:q(w)=0} dw\, p(w) e^{g(w)} + \int_{\mathcal{W}:q(w)>0} dw\, q(w) \frac{p(w)}{q(w)} e^{g(w)} \\
&\geq \int_{\mathcal{W}:q(w)>0} dw\, q(w) \frac{p(w)}{q(w)} e^{g(w)} \;=\; \mathop{\mathbb{E}}_{W\sim Q} \left[\frac{p(W)}{q(W)} e^{g(W)} \right].
\end{aligned}$$

- Alors, pour toute fonction mesurable g, on a

$$\mathop{\mathbb{E}}_{W\sim P} e^{g(W)} \;\geq\; \mathop{\mathbb{E}}_{W\sim Q} \left[\frac{p(W)}{q(W)} e^{g(W)} \right].$$

- En utilisant l'inégalité précédente et l'inégalité de Jensen sur la concavité de $\ln(\cdot)$, on a

$$\begin{aligned}\ln\left[\mathop{\mathbb{E}}_{W\sim P} e^{g(W)}\right] &\geq \ln\left[\mathop{\mathbb{E}}_{W\sim Q} \frac{p(W)}{q(W)} e^{g(W)}\right] \geq \mathop{\mathbb{E}}_{W\sim Q} \ln\left[\frac{p(W)}{q(W)} e^{g(W)}\right] \\ &= \mathop{\mathbb{E}}_{W\sim Q} \ln\left[\frac{p(W)}{q(W)}\right] + \mathop{\mathbb{E}}_{W\sim Q} g(W) \\ &= -D(Q\|P) + \mathop{\mathbb{E}}_{W\sim Q} g(W)\,.\end{aligned}$$

- Alors pour toute fonction mesurable g, nous avons

$$\mathop{\mathbb{E}}_{W\sim Q} g(W) \leq D(Q\|P) + \ln\left[\mathop{\mathbb{E}}_{W\sim P} e^{g(W)}\right]\,.$$

■

Sommaire

1. Introduction et définitions
2. La KL divergence et le changement de mesure Donsker-Varadhan
3. Théorème PAC-Bayes général et bornes PAC-Bayes
4. Algorithme d'apprentissage PAC-Bayésien et sa garantie PAC
   - Apprentissage PAC-Bayésien en pratique et Gaussiennes isotropes
5. Garantie PAC-Bayes pour algorithmes retournant distribution Q sur $\mathcal{W}$
6. Garantie PAC-Bayes pour alg. stochastiques et information mutuelle
7. Descente de gradient stochastique avec dynamique de Langevin

Théorème PAC-Bayes général

Le théorème suivant est une usine à produire des bornes PAC-Bayes.

Théorème (Théorème PAC-Bayes général (en probabilité))

Soit $\mathcal{D}$ une distribution sur $\mathcal{Z}$ et P une distribution sur $\mathcal{W}$. Soit $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ et $f : \mathcal{W} \times \mathcal{Z}^m \to \mathbb{R}$ mesurable satisfaisant

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} e^{\lambda f(w,S)} \leq \psi(\lambda), \ \forall \lambda > 0 \text{ et } \forall w \in \mathcal{W} \tag{1}$$

Alors, pour tout $\lambda > 0$ et pour tout $\delta \in (0,1)$, on a

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left(\forall Q : \mathop{\mathbb{E}}_{W\sim Q} f(W,S) \ \leq \ \frac{1}{\lambda}\left[D(Q\|P) + \ln\frac{\psi(\lambda)}{\delta}\right]\right) \ \geq \ 1-\delta\,.$$

- **Preuve:** Puisque P ne dépend pas de S et que f satisfait (1), on a

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{W\sim P} e^{\lambda f(W,S)} \;=\; \mathop{\mathbb{E}}_{W\sim P}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} e^{\lambda f(W,S)} \;\leq\; \psi(\lambda)\,.$$

- L'inégalité de Markov appliquée à $\mathbb{E}_{W\sim P}\, e^{\lambda f(W,S)}$ nous donne

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left(\mathop{\mathbb{E}}_{W\sim P} e^{\lambda f(W,S)} > \frac{\psi(\lambda)}{\delta}\right) \;\leq\; \delta\,.$$

- De manière équivalente, nous avons

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^m}\left(\ln\left[\mathop{\mathbb{E}}_{W\sim P} e^{\lambda f(W,S)}\right] \leq \ln\frac{\psi(\lambda)}{\delta}\right) \;\geq\; 1-\delta\,.$$

- En appliquant sur $\lambda f(W,S)$ le thm sur le changement de mesure, nous obtenons qu'avec prob. $\geq 1-\delta$, nous avons

$$\forall Q \ll P : \lambda \mathop{\mathbb{E}}_{W\sim Q} f(W,S) - D(Q||P) \;\leq\; \ln\left[\mathop{\mathbb{E}}_{W\sim P} e^{\lambda f(W,S)}\right]\,.$$

Pertes sous Gaussiennes

- Afin d'obtenir des résultats généraux, nous supposons que la perte $\ell(w, Z)$ est une variable aléatoire σ-sous Gaussienne pour tout $w \in \mathcal{W}$, i.e., nous supposons que

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{D}} e^{\lambda(\ell(w,Z)-L_{\mathcal{D}}(w))} \leq e^{\frac{\lambda^2\sigma^2}{2}}, \ \forall w \in \mathcal{W}, \ \forall \lambda \in \mathbb{R}.$$

- Les fonctions de pertes bornées sont automatiquement sous Gaussiennes (mais pas l'inverse) : si $a \leq \ell(w, Z) \leq b$ pour tout $w \in \mathcal{W}$, le théorème de Hoeffding nous dit que

$$\mathop{\mathbb{E}}_{Z\sim\mathcal{D}} e^{\lambda(\ell(w,Z)-L_{\mathcal{D}}(w))} \leq e^{\frac{\lambda^2(b-a)^2}{8}}, \ \forall w \in \mathcal{W}, \ \forall \lambda \in \mathbb{R}.$$

  Donc, dans ce cas, $\ell(w, Z)$ est $(b-a)/2$-sous Gaussienne.
- Propriété : lorsque $\ell(w, Z)$ est σ-sous Gaussienne, pour tout $t > 0$, nous avons

$$\max\left(\mathbb{P}[\ell(w, Z) - L_{\mathcal{D}}(w) \geq t], \mathbb{P}[\ell(w, Z) - L_{\mathcal{D}}(w) \leq -t]\right) \leq e^{\frac{-t^2}{2\sigma^2}}.$$

Pertes sous gaussiennes (suite)

Lemme

Si $\ell(w, Z)$ est σ-sous Gaussienne pour tout $w \in \mathcal{W}$, alors $L_S(w)$ est $\sigma/\sqrt{m}$-sous Gaussienne.

Preuve:

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{S} e^{\lambda(L_S(w) - L_{\mathcal{D}}(w))} \\
&\quad = \mathop{\mathbb{E}}_{Z_1} \cdots \mathop{\mathbb{E}}_{Z_m} e^{\frac{\lambda}{m} \sum_{i=1}^{m} (\ell(w, Z_i) - L_{\mathcal{D}}(w))} \\
&\quad = \prod_{i=1}^{m} \mathop{\mathbb{E}}_{Z_i} e^{\frac{\lambda}{m}(\ell(w, Z_i) - L_{\mathcal{D}}(w))} \\
&\quad = \left[\mathop{\mathbb{E}}_{Z} e^{\frac{\lambda}{m}(\ell(w, Z) - L_{\mathcal{D}}(w))} \right]^m \\
&\quad \leq \left[e^{\frac{\lambda^2 \sigma^2}{2m^2}} \right]^m = e^{\frac{\lambda^2 \sigma^2}{2m}} , \ \forall \lambda \in \mathbb{R}, \ \forall w \in \mathcal{W} .
\end{aligned}
$$

Borner $L_{\mathcal{D}}(Q) - L_S(Q)$

Corollaire (adaptation de Germain et al. 2017)

Soit $\mathcal{D}$ une distribution sur $\mathcal{Z}$ et P une distribution sur $\mathcal{W}$. Soit ℓ une fonction de perte telle que $\ell(w, Z)$ est σ-sous Gaussienne $\forall w \in \mathcal{W}$. Alors $\forall \lambda > 0$ et $\forall \delta \in (0, 1)$, avec probabilité $\geq 1 - \delta$ sur les tirages de $S \sim \mathcal{D}^m$, on a simultanément pour tout Q :

$$L_{\mathcal{D}}(Q) - L_S(Q) \stackrel{def}{=} \mathop{\mathbb{E}}_{W \sim Q} L_{\mathcal{D}}(W) - L_S(W) \;\leq\; \frac{1}{\lambda}\left[D(Q\|P) + \ln\frac{1}{\delta}\right] + \frac{\lambda\sigma^2}{2m}$$

Preuve: Corollaire obtenu en utilisant $f(W, S) = L_{\mathcal{D}}(W) - L_S(W)$ dans le théorème général avec $\mathbb{E}_S\, e^{\lambda(L_D(w) - L_S(w))} \leq e^{\frac{\lambda^2\sigma^2}{2m}} = \psi(\lambda)$. ■

Remarque : en choisissant $\lambda = \sqrt{m}$, la borne sur $L_{\mathcal{D}}(Q) - L_S(Q)$ devient

$$\frac{1}{\sqrt{m}}\left[D(Q\|P) + \ln\left(\frac{1}{\delta}\right) + \frac{\sigma^2}{2}\right].$$

Borner $|L_{\mathcal{D}}(Q) - L_S(Q)|$

Corollaire

Soit $\mathcal{D}$ une distribution sur $\mathcal{Z}$. Soit ℓ tel que $\ell(w, Z)$ est σ-sous Gaussienne $\forall w \in \mathcal{W}$. Soit P une distribution sur $\mathcal{W}$. Alors $\forall \lambda > 0$ et $\forall \delta \in (0, 1)$, avec probabilité $\geq 1 - \delta$ sur les tirages de $S \sim \mathcal{D}^m$, on a simultanément pour tout Q :

$$|L_{\mathcal{D}}(Q) - L_S(Q)| \leq \frac{1}{\lambda}\left[D(Q\|P) + \ln\frac{2}{\delta}\right] + \frac{\lambda\sigma^2}{2m}\,.$$

Preuve: En utilisant $f(W, S) = L_S(W) - L_{\mathcal{D}}(W)$ dans le théorème général avec $\mathbb{E}_S\, e^{\lambda(L_S(w) - L_{\mathcal{D}}(w))} \leq e^{\frac{\lambda^2\sigma^2}{2m}} = \psi(\lambda)$, on obtient un corollaire identique au précédent, mais avec

$$L_S(Q) - L_{\mathcal{D}}(Q) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{W\sim Q} L_S(W) - L_{\mathcal{D}}(W) \;\leq\; \frac{1}{\lambda}\left[D(Q\|P) + \ln\frac{1}{\delta}\right] + \frac{\lambda\sigma^2}{2m}$$

- Pour les deux bornes précédentes, nous avons alors

$$\mathop{\mathbb{P}}_{S}\left(\exists Q : L_S(Q) - L_{\mathcal{D}}(Q) \ > \ \frac{1}{\lambda}\left[D(Q\|P) + \ln\frac{2}{\delta}\right] + \frac{\lambda\sigma^2}{2m}\right) \leq \delta/2$$

$$\mathop{\mathbb{P}}_{S}\left(\exists Q : L_{\mathcal{D}}(Q) - L_S(Q) \ > \ \frac{1}{\lambda}\left[D(Q\|P) + \ln\frac{2}{\delta}\right] + \frac{\lambda\sigma^2}{2m}\right) \leq \delta/2$$

- En utilisant la borne de l'union sur ces deux évènements, nous obtenons

$$\mathop{\mathbb{P}}_{S}\left(\exists Q : |L_{\mathcal{D}}(Q) - L_S(Q)| \ > \ \frac{1}{\lambda}\left[D(Q\|P) + \ln\frac{2}{\delta}\right] + \frac{\lambda\sigma^2}{2m}\right) \leq \delta$$

Ce qui est logiquement équivalent au corollaire. ■

Sommaire

1. Introduction et définitions
2. La KL divergence et le changement de mesure Donsker-Varadhan
3. Théorème PAC-Bayes général et bornes PAC-Bayes
4. Algorithme d'apprentissage PAC-Bayésien et sa garantie PAC
   - Apprentissage PAC-Bayésien en pratique et Gaussiennes isotropes
5. Garantie PAC-Bayes pour algorithmes retournant distribution Q sur $\mathcal{W}$
6. Garantie PAC-Bayes pour alg. stochastiques et information mutuelle
7. Descente de gradient stochastique avec dynamique de Langevin

Algorithme d'apprentissage PAC-Bayésien

- Ayant un échantillon d'apprentissage S et un prior P sur $\mathcal{W}$, le corollaire précédent suggère l'algorithme d'apprentissage suivant qui optimise une garantie PAC-Bayésienne sur $L_{\mathcal{D}}(Q)$:

$$\mathrm{PB}^{\lambda}_{\mathcal{W}}(S) \stackrel{\text{def}}{=} \operatorname*{argmin}_{Q \ll P} \left[L_S(Q) + \frac{1}{\lambda} D(Q \| P) \right] .$$

- Notez que $\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)$ retourne une distribution Q sur $\mathcal{W}$.
- Quelles sont les classes de distributions apprenables par $\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)$?
  - Et pour quelle valeur de λ ?

Garantie PAC agnostique pour $\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)$

- Par la définition de $\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)$ et le corollaire précédent, avec probabilité $\geq 1-\delta$, on a pour tout Q :

$$
\begin{aligned}
L_{\mathcal{D}}(\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)) &\leq L_S(\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)) + \frac{1}{\lambda}\left[D(\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)\|P) + \ln\left(\frac{2}{\delta}\right)\right] + \frac{\lambda\sigma^2}{2m} \\
&\leq L_S(Q) + \frac{1}{\lambda}\left[D(Q\|P) + \ln\left(\frac{2}{\delta}\right)\right] + \frac{\lambda\sigma^2}{2m} \\
&\leq L_{\mathcal{D}}(Q) + \frac{2}{\lambda}\left[D(Q\|P) + \ln\left(\frac{2}{\delta}\right)\right] + \frac{\lambda\sigma^2}{m}\,.
\end{aligned}
$$

- Considérons les Q t.q. $D(Q\|P) \leq D^*$ pour un choix de P et $D^* > 0$.
- Donc, avec probabilité $\geq 1-\delta$, on a

$$
L_{\mathcal{D}}(\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)) \;\leq\; \min_{Q:D(Q\|P)\leq D^*} L_{\mathcal{D}}(Q) + \frac{2}{\lambda}\left[D^* + \ln\left(\frac{2}{\delta}\right)\right] + \frac{\lambda\sigma^2}{m}\,.
$$

Garantie PAC agnostique pour $\mathrm{PB}_{\mathcal{W}}^{\lambda}(S)$

- Considérons maintenant $\lambda = c\sqrt{m}$ pour $c > 0$. Alors, avec probabilité $\geq 1 - \delta$, on a

$$L_{\mathcal{D}}(\mathrm{PB}_{\mathcal{W}}^{\lambda}(S)) \leq \min_{Q:D(Q\|P)\leq D^*} L_{\mathcal{D}}(Q) + \frac{1}{\sqrt{m}} \left\{ \frac{2}{c} \left[D^* + \ln\left(\frac{2}{\delta}\right) \right] + c\sigma^2 \right\}.$$

- Or, le terme à droite est minimal lorsque

$$c = \sqrt{2 \frac{D^* + \ln\left(\frac{2}{\delta}\right)}{\sigma^2}}. \tag{2}$$

- Donc, avec probabilité $\geq 1 - \delta$, on a

$$L_{\mathcal{D}}(\mathrm{PB}_{\mathcal{W}}^{\lambda}(S)) \leq \min_{Q:D(Q\|P)\leq D^*} L_{\mathcal{D}}(Q) + \sqrt{\frac{8\sigma^2}{m} \left[D^* + \ln\left(\frac{2}{\delta}\right) \right]}.$$

- Nous venons alors de démontrer le résultat suivant :

Garantie PAC agnostique pour $\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)$

Théorème

Soit $\mathcal{D}$ une distribution sur $\mathcal{Z}$ et ℓ tel que $\ell(w, Z)$ est σ-sous Gaussienne $\forall w \in \mathcal{W}$. Soit P une distribution sur $\mathcal{W}$. Soit $\lambda = c\sqrt{m}$, avec c donné par l'eq.2. Alors $\forall \delta \in (0,1)$, avec prob. $\geq 1-\delta$ sur $S \sim \mathcal{D}^m$, on a

$$L_{\mathcal{D}}(\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)) \leq \min_{Q: D(Q\|P) \leq D^*} L_{\mathcal{D}}(Q) + \sqrt{\frac{8\sigma^2}{m}\left[D^* + \ln\left(\frac{2}{\delta}\right)\right]}\,.$$

Donc, si $m \geq 8\sigma^2[D^* + \ln(2/\delta)]/\epsilon^2$, avec probabilité $\geq 1-\delta$, on a

$$L_{\mathcal{D}}(\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)) \leq \min_{Q: D(Q\|P) \leq D^*} L_{\mathcal{D}}(Q) + \epsilon\,.$$

En d'autres mots, $\mathrm{PB}^{\lambda}_{\mathcal{W}}(S)$ est un algorithme d'apprentissage au sens PAC agnostique pour la classe de distributions à KL-divergence bornée

$$\mathcal{M}_{\mathcal{W}}(P, D^*) \stackrel{\text{def}}{=} \{Q \text{ sur } \mathcal{W} : D(Q\|P) \leq D^*\}\,.$$

Algorithme d'apprentissage PAC-Bayésien en pratique

- Étant donné l'équation 2, choisir la valeur de λ pour l'algorithme

$$\mathrm{PB}_{\mathcal{W}}^{\lambda}(S) \stackrel{\text{def}}{=} \underset{Q \ll P}{\operatorname{argmin}} \left[L_S(Q) + \frac{1}{\lambda} D(Q\|P) \right],$$

  revient à choisir la valeur de D^*.
  - En pratique on utilise souvent $\lambda \approx m/\text{cste}$.
- Il faut aussi se restreindre à une classe de distributions sur $\mathcal{W}$ de manière à pouvoir calculer efficacement
  - $D(Q\|P)$
  - $L_S(Q) = (1/m) \sum_i \ell(Q, z_i)$.
    - Il faut donc pouvoir calculer efficacement $\ell(Q, z)$, $\forall z \in \mathcal{Z}$.
- Cela est possible lorsque l'on utilise les Gaussiennes isotropes pour notre classe de distributions.

Apprentissage PAC-Bayésien avec Gaussiennes isotropes

- Utilisons alors la classe $\mathcal{G}$ des Gaussiennes isotropes $\mathcal{G}$ sur $\mathcal{W} = \mathbb{R}^d$ où chaque densité $q_{\mathbf{w}}$, paramétrisée par $\mathbf{w} \in \mathbb{R}^d$, possède une valeur $q_{\mathbf{w}}(\mathbf{w}')$ au point $\mathbf{w}'$ donnée par

$$q_{\mathbf{w}}(\mathbf{w}') \;=\; \left(\frac{1}{\sqrt{2\pi}}\right)^d e^{-\frac{1}{2}\|\mathbf{w}-\mathbf{w}'\|^2} = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(w_i - w'_i)^2} \,.$$

Apprentissage PAC-Bayésien avec Gaussiennes isotropes

- Pour le calcul de $D(Q_{\mathbf{w}} \| P)$, choisissons notre système de coordonnées pour que la première coordonnée soit dans la direction de $\mathbf{w}$ (et que les autres coordonnées soient dans les directions $\perp$ à $\mathbf{w}$). Nous avons (pour $\mu \stackrel{\text{def}}{=} \|\mathbf{w}\|$)

$$q_{\mathbf{w}}(\mathbf{w}') = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(w'_1 - \mu)^2} \prod_{i=2}^{d} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(w'_i)^2} \,.$$

- Si $\mathcal{N}(\mu, \sigma^2)$ dénote une Gaussienne (distribution normale) de variance σ^2 et d'espérance μ, nous avons

$$Q_{\mathbf{w}} = \mathcal{N}(\mu, 1) \times \mathcal{N}(0, 1)^{d-1} \,.$$

- Pour notre distribution a priori P, choisissons

$$P = \mathcal{N}(0, 1)^d \,.$$

Apprentissage PAC-Bayésien avec Gaussiennes isotropes

- Nous obtenons alors

$$
\begin{aligned}
D(Q_{\mathbf{w}} \| P) &= \int d\mathbf{w}' q_{\mathbf{w}}(\mathbf{w}') \ln \left(\frac{q_{\mathbf{w}}(\mathbf{w}')}{p(\mathbf{w}')} \right) \\
&= \int d\mathbf{w}' q_{\mathbf{w}}(\mathbf{w}') \ln \left(\frac{e^{-\frac{1}{2}(w_1'-\mu)^2} \prod_{i=2}^{d} e^{-\frac{1}{2}(w_i')^2}}{\prod_{i=1}^{d} e^{-\frac{1}{2}(w_i')^2}} \right) \\
&= \int d\mathbf{w}' q_{\mathbf{w}}(\mathbf{w}') \ln \left(\frac{e^{-\frac{1}{2}(w_1'-\mu)^2}}{e^{-\frac{1}{2}(w_1')^2}} \right) \\
&= \int \frac{dw_1'}{\sqrt{2\pi}} e^{-\frac{1}{2}(w_1'-\mu)^2} \left[-\frac{1}{2}(w_1'-\mu)^2 + \frac{1}{2}(w_1')^2 \right] \\
&= \int \frac{dw_1'}{\sqrt{2\pi}} e^{-\frac{1}{2}(w_1'-\mu)^2} \left[\mu w_1' - \frac{\mu^2}{2} \right] = \frac{\mu^2}{2} = \frac{\|\mathbf{w}\|^2}{2} .
\end{aligned}
$$

Apprentissage PAC-Bayésien avec Gaussiennes isotropes

- Ayant choisi une fonction de projection $\boldsymbol{\phi} : \mathcal{X} \to \mathbb{R}^d$, chaque prédicteur $\mathbf{v} \in \mathcal{W} = \mathbb{R}^d$ sera un classificateur $h_{\mathbf{v}}$ linéaire t.q pour tout $\mathbf{x} \in \mathcal{X}$, on a

$$h_{\mathbf{v}}(\mathbf{x}) = \text{sign}(\langle \mathbf{v}, \boldsymbol{\phi}(\mathbf{x}) \rangle) \,,$$

avec $\text{sign}(x) = +1$ si $x > 0$ et $\text{sign}(x) = -1$ si $x \leq 0$.

- Nous utilisons une fonction de perte $\ell : \mathcal{W}, \mathcal{Z} \to \mathbb{R}_+$ t.q.

$$\ell(\mathbf{v}, (\mathbf{x}, y)) \stackrel{\text{def}}{=} \mathbb{1}\left(\langle \mathbf{v}, y\boldsymbol{\phi}(\mathbf{x}) \rangle \leq 0\right) \,.$$

avec $\mathbb{1}(a) = 1$ si a est vrai et $\mathbb{1}(a) = 0$ si a est faux.

- Soit $\boldsymbol{\psi} \stackrel{\text{def}}{=} y\boldsymbol{\phi}(\mathbf{x})$. Définissons

$$\ell(\mathbf{v}, \boldsymbol{\psi}) \stackrel{\text{def}}{=} \mathbb{1}\left(\langle \mathbf{v}, \boldsymbol{\psi} \rangle \leq 0\right)$$

$$\ell(Q_{\mathbf{w}}, \boldsymbol{\psi}) \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} d\mathbf{v}\, q_{\mathbf{w}}(\mathbf{v}) \mathbb{1}\left(\langle \mathbf{v}, \boldsymbol{\psi} \rangle \leq 0\right) \,.$$

Apprentissage PAC-Bayésien avec Gaussiennes isotropes

- Notez que $\ell(Q_{\mathbf{w}}, \boldsymbol{\psi}) = 1$ lorsque $\boldsymbol{\psi} = \mathbf{0}$.
- Considérons alors n'importe quel vecteur $\boldsymbol{\psi}$ non nul.
- Décomposons le vecteur $\mathbf{v}$ en sa partie $v_{\parallel}$ parallèle à $\boldsymbol{\psi}$ et sa partie $\mathbf{v}_{\perp}$ perpendiculaire à $\boldsymbol{\psi}$. Nous avons alors $\mathbf{v} = (v_{\parallel}, \mathbf{v}_{\perp})$.
- De plus, si nous définissons $\psi \stackrel{\text{def}}{=} \|\boldsymbol{\psi}\| > 0$, alors $\langle \mathbf{v}, \boldsymbol{\psi} \rangle = v_{\parallel} \psi$.
- Effectuons également la décomposition $\mathbf{w} = (w_{\parallel}, \mathbf{w}_{\perp})$ en les parties parallèles et perpendiculaires à $\boldsymbol{\psi}$. Nous avons alors

$$\|\mathbf{v} - \mathbf{w}\|^2 \;=\; (v_{\parallel} - w_{\parallel})^2 + \|\mathbf{v}_{\perp} - \mathbf{w}_{\perp}\|^2 \,.$$

- Alors

$$\begin{aligned} q_{\mathbf{w}}(\mathbf{v}) &= \left(\frac{1}{\sqrt{2\pi}}\right)^d e^{-\frac{1}{2}\|\mathbf{v}-\mathbf{w}\|^2} \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^{d-1} e^{-\frac{1}{2}\|\mathbf{v}_{\perp}-\mathbf{w}_{\perp}\|^2} \left(\frac{1}{\sqrt{2\pi}}\right) e^{-\frac{1}{2}(v_{\parallel}-w_{\parallel})^2} \,. \end{aligned}$$

Apprentissage PAC-Bayésien avec Gaussiennes isotropes

- Nous avons alors

$$
\begin{aligned}
\ell(Q_{\mathbf{w}}, \boldsymbol{\psi}) &= \int_{\mathbb{R}^d} d\mathbf{v}\, q_{\mathbf{w}}(\mathbf{v}) \mathbb{1}\left(\langle \mathbf{v}, \boldsymbol{\psi} \rangle \leq 0\right) \\
&= \int_{\mathbb{R}^d} d\mathbf{v} \left(\frac{1}{\sqrt{2\pi}}\right)^d e^{-\frac{1}{2}\|\mathbf{v}-\mathbf{w}\|^2} \mathbb{1}\left(v_{\|}\psi \leq 0\right) \\
&= \int_{-\infty}^{+\infty} dv_{\|} \left(\frac{1}{\sqrt{2\pi}}\right) e^{-\frac{1}{2}(v_{\|}-w_{\|})^2} \mathbb{1}(v_{\|}\psi \leq 0) \\
&\qquad \times \int_{\mathbb{R}^{d-1}} d\mathbf{v}_{\perp} \left(\frac{1}{\sqrt{2\pi}}\right)^{d-1} e^{-\frac{1}{2}\|\mathbf{v}_{\perp}-\mathbf{w}_{\perp}\|^2} \\
&= \int_{-\infty}^{+\infty} dv_{\|} \left(\frac{1}{\sqrt{2\pi}}\right) e^{-\frac{1}{2}(v_{\|}-w_{\|})^2} \mathbb{1}(v_{\|} \leq 0)\,,
\end{aligned}
$$

car $\psi > 0$.

Apprentissage PAC-Bayésien avec Gaussiennes isotropes

- En posant $x \stackrel{\text{def}}{=} v_{\|} - w_{\|}$. Nous obtenons

$$\begin{aligned}
\ell(Q_{\mathbf{w}}, \boldsymbol{\psi}) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} dx \; e^{-\frac{1}{2}x^2} I(x \leq -w_{\|}) \\
&= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-w_{\|}} dx \; e^{-\frac{1}{2}x^2} \\
&= \frac{1}{\sqrt{2\pi}} \int_{w_{\|}}^{+\infty} dx \; e^{-\frac{1}{2}x^2} \\
&= \mathop{\mathbb{P}}_{X \sim \mathcal{N}(0,1)} (X \geq w_{\|}) \\
&\stackrel{\text{def}}{=} \Phi(w_{\|}) \, .
\end{aligned}$$

Apprentissage PAC-Bayésien avec Gaussiennes isotropes

- Notez que

$$w_{\|} = \frac{\langle \mathbf{w}, \boldsymbol{\psi} \rangle}{\|\boldsymbol{\psi}\|} = \frac{\langle y\mathbf{w}, \boldsymbol{\phi} \rangle}{\|\boldsymbol{\phi}\|} = \|\mathbf{w}\| \frac{\langle y\mathbf{w}, \boldsymbol{\phi} \rangle}{\|\mathbf{w}\|\|\boldsymbol{\phi}\|} \stackrel{\text{def}}{=} \|\mathbf{w}\| \Gamma_{\mathbf{w}}(\boldsymbol{\phi}, y) \,,$$

  où $\Gamma_{\mathbf{w}}(\boldsymbol{\phi}, y)$ est la marge normalisée de $\mathbf{w}$ sur l'exemple $(\boldsymbol{\phi}(\mathbf{x}), y)$.

- Alors

$$\ell(Q_{\mathbf{w}}, (\mathbf{x}, y)) \;=\; \Phi\left(\|\mathbf{w}\| \Gamma_{\mathbf{w}}(\boldsymbol{\phi}(\mathbf{x}), y)\right) \,.$$

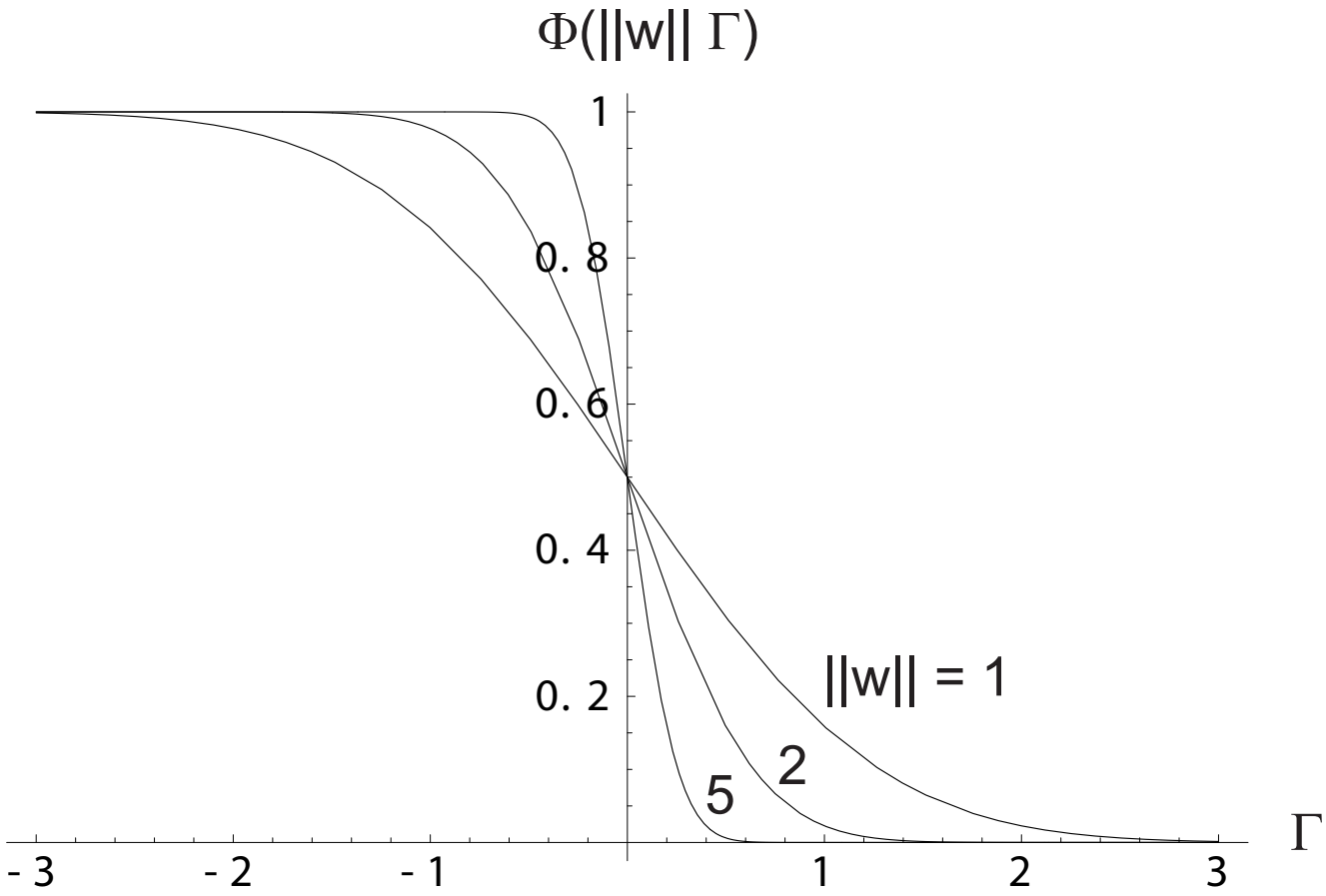
- Ainsi

$$L_S(Q_{\mathbf{w}}) \;=\; \frac{1}{m} \sum_{i=1}^{m} \Phi\left(\|\mathbf{w}\| \Gamma_{\mathbf{w}}(\boldsymbol{\phi}(\mathbf{x}_i), y_i)\right) \tag{3}$$

$$L_{\mathcal{D}}(Q_{\mathbf{w}}) \;=\; \mathop{\mathbb{E}}_{(\mathbf{X},Y)} \Phi\left(\|\mathbf{w}\| \Gamma_{\mathbf{w}}(\boldsymbol{\phi}(\mathbf{X}), Y)\right) \,. \tag{4}$$

- Notez que $\Phi(\|\mathbf{w}\|\Gamma) = I(\Gamma < 0)$ lorsque $\|\mathbf{w}\| \to \infty$.

Apprentissage PAC-Bayésien avec Gaussiennes isotropes



Apprentissage PAC-Bayésien avec Gaussiennes isotropes

- Avec ces expressions analytiques pour $\ell(Q_{\mathbf{w}}, z)$ et $D(Q_{\mathbf{w}} \| P)$, l'algorithme d'apprentissage PAC-Bayésien pour Gaussiennes isotropes s'écrit

$$\mathrm{PB}^{\lambda}_{\mathbb{R}^n}(S) \stackrel{\text{def}}{=} \operatorname*{argmin}_{\mathbf{w}} \left[\frac{1}{m} \sum_{i=1}^{m} \Phi\left(\|\mathbf{w}\| \Gamma_{\mathbf{w}}(\boldsymbol{\phi}(\mathbf{x}_i), y_i) \right) + \frac{1}{2\lambda} \|\mathbf{w}\|^2 \right] .$$

- Ceci est une régularisation de Tikhonov (similaire au SVM).
- Pour obtenir un risque minimal, on doit trouver $\mathbf{w}$ de faible norme et ayant une assez grande "marge douce" $\|\mathbf{w}\| \cdot \Gamma_{\mathbf{w}}(\boldsymbol{\phi}(\mathbf{x}_i), y_i)$ sur la majorité des exemples $(\boldsymbol{\phi}(\mathbf{x}_i), y_i)$.

Sommaire

1. Introduction et définitions
2. La KL divergence et le changement de mesure Donsker-Varadhan
3. Théorème PAC-Bayes général et bornes PAC-Bayes
4. Algorithme d'apprentissage PAC-Bayésien et sa garantie PAC
   - Apprentissage PAC-Bayésien en pratique et Gaussiennes isotropes
5. Garantie PAC-Bayes pour algorithmes retournant distribution Q sur $\mathcal{W}$
6. Garantie PAC-Bayes pour alg. stochastiques et information mutuelle
7. Descente de gradient stochastique avec dynamique de Langevin

Algorithmes retournant une distribution Q sur $\mathcal{W}$

- Considérons n'importe quel algorithme qui retourne une distribution Q sur $\mathcal{W}$ à partir d'un échantillon d'apprentissage $S \sim \mathcal{D}^m$.
  - Cela pourrait être

$$\mathrm{PB}^{\lambda}_{\mathcal{W}}(S) \stackrel{\text{def}}{=} \operatorname*{argmin}_{Q \in \mathcal{Q}_{\mathcal{W}}} \left[L_S(Q) + \frac{1}{\lambda} D(Q \| P) \right] .$$

    où $\mathcal{Q}_{\mathcal{W}}$ dénote un ensemble des distributions Q sur $\mathcal{W}$.
- Notons par $Q(S)$ la distribution retournée (par un algorithme) afin de souligner qu'il s'agit d'une distribution aléatoire dépendante de S.
  - Ce n'est que pour une réalisation s de S, que $Q(s)$ est une distribution.
  - Donc $Q(\cdot) : \mathcal{Z}^m \to \mathcal{Q}_{\mathcal{W}}$ caractérise l'algorithme d'apprentissage.
- Pour $\mathcal{Q}_{\mathcal{W}}$ donné, nous cherchons à borner $\mathbb{E}_{S \sim \mathcal{D}^m} \mathbb{E}_{W \sim Q(S)} f(W, S)$ quelque soit $Q(\cdot)$.

Théorème PAC-Bayes général (en espérance) pour $Q(S)$

Théorème (Théorème PAC-Bayes général (en espérance) pour $Q(S)$)

Soit $\mathcal{D}$ une distribution sur $\mathcal{Z}$ et soit P une distribution sur $\mathcal{W}$. Soit $\mathcal{Q}_{\mathcal{W}}$ l'ensemble de toutes les distributions sur $\mathcal{W}$. Soit $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ et $f : \mathcal{W} \times \mathcal{Z}^m \to \mathbb{R}$ mesurable satisfaisant

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} e^{\lambda f(w,S)} \leq \psi(\lambda), \ \forall \lambda > 0 \text{ et } \forall w \in \mathcal{W} \tag{5}$$

Alors, pour tout $\lambda > 0$ et pour tout $Q(\cdot) : \mathcal{Z}^m \to \mathcal{Q}_{\mathcal{W}}$, on a

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{W\sim Q(S)} f(W,S) \ \leq \ \frac{1}{\lambda}\left(\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \left[D(Q(S)\|P) \right] \ + \ \ln\left[\psi(\lambda)\right] \right) .$$

Notez que $\psi(\lambda)$ est normalement une fonction croissante en λ.

- **Preuve:** Puisque P ne dépend pas de S et que f satisfait (5), on a

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\mathop{\mathbb{E}}_{W\sim P} e^{\lambda f(W,S)} \;=\; \mathop{\mathbb{E}}_{W\sim P}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} e^{\lambda f(W,S)} \;\leq\; \psi(\lambda)\,. \qquad (6)$$

- En prenant le logarithme de chaque côté de l'équation (6) et en utilisant l'inégalité de Jensen sur la concavité de $\ln(\cdot)$, nous avons

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \ln\left[\mathop{\mathbb{E}}_{W\sim P} e^{\lambda f(W,S)}\right] \leq \ln\left[\psi(\lambda)\right]\,. \qquad (7)$$

- Or, selon le changement de mesure Donsker Varadhan, $\forall S \in \mathcal{Z}^m$ on a

$$\forall Q \in \mathcal{Q}_{\mathcal{W}}: \quad \mathop{\mathbb{E}}_{W\sim Q} \lambda f(W,S) - \; D(Q\|P) \leq \ln\left[\mathop{\mathbb{E}}_{W\sim P} e^{\lambda f(W,S)}\right]\,.$$

- Donc $\forall Q(\cdot): \mathcal{Z}^m \to \mathcal{Q}_{\mathcal{W}}$ et $\forall S \in \mathcal{Z}^m$ on a

$$\mathop{\mathbb{E}}_{W\sim Q(S)} \lambda f(W,S) - \; D(Q(S)\|P) \leq \ln\left[\mathop{\mathbb{E}}_{W\sim P} e^{\lambda f(W,S)}\right]\,.$$

- Alors, selon cette dernière équation et l'équation (7), $\forall \lambda > 0$ et $\forall Q(\cdot) : \mathcal{Z}^m \to \mathcal{Q}_{\mathcal{W}}$, on a

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{W\sim Q(S)} \lambda f(W,S) \leq \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} D(Q(S)\|P) + \ln\left[\psi(\lambda)\right] .$$

■

Le prior P minimisant $\mathbb{E}_S\, D(Q(S)\|P)$

Pour trouver le P indépendant de S qui minimise $\mathbb{E}_S\, D(Q(S)\|P)$, procédons comme Lever et al. (TCS, 2013). Nous avons alors

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} D(Q(S)\|P) &= \mathop{\mathbb{E}}_{S} \int_{\mathcal{W}} dw\, q(S)(w) \ln\left(\frac{q(S)(w)}{p(w)}\right) \\ &= \mathop{\mathbb{E}}_{S} \int_{\mathcal{W}} dw\, q(S)(w) \left[\ln(q(S)(w)) + \ln\left(\frac{1}{p(w)}\right)\right]\end{aligned}$$

Pour trouver P minimisant $\mathbb{E}_S\, D(Q(S)\|P)$, il faut donc minimiser l'entropie croisée entre $\mathbb{E}_S\, Q(S)$ et P :

$$\mathop{\mathbb{E}}_{S}\left[\int_{\mathcal{W}} dw\, q(S)(w) \ln\left(\frac{1}{p(w)}\right)\right] \;=\; \left[\int_{\mathcal{W}} dw\, \left[\mathop{\mathbb{E}}_{S} q(S)\right](w) \ln\left(\frac{1}{p(w)}\right)\right].$$

Puisque le minimum est obtenu pour $p(w) = [\mathbb{E}_S\, q(S)]\,(w)$, le prior P minimisant $\mathbb{E}_S\, D(Q(S)\|P)$ est donc $\mathbb{E}_S\, Q(S)$.

Sommaire

1. Introduction et définitions
2. La KL divergence et le changement de mesure Donsker-Varadhan
3. Théorème PAC-Bayes général et bornes PAC-Bayes
4. Algorithme d'apprentissage PAC-Bayésien et sa garantie PAC
   - Apprentissage PAC-Bayésien en pratique et Gaussiennes isotropes
5. Garantie PAC-Bayes pour algorithmes retournant distribution Q sur $\mathcal{W}$
6. Garantie PAC-Bayes pour alg. stochastiques et information mutuelle
7. Descente de gradient stochastique avec dynamique de Langevin

Changement de notation

- Pour éclaircir la signification du théorème général lorsque $P = \mathbb{E}_S\, Q(S)$, un changement de notation s'impose.
- On ajoutera aux distributions les indices spécifiant les variables aléatoires assujetties par ces distributions.
- On dénotera par P_S la distribution $\mathcal{D}^m$ sur $\mathcal{Z}^m$ pour la variable aléatoire S et $p_S(s)$ désignera sa densité au point $s \in \mathcal{Z}^m$.
- Nous dénoterons maintenant par $P_{W|S}$, la distribution aléatoire $Q(S)$.
- $P_{W|S=s}$ désignera la distribution $Q(s)$ suite à une réalisation s de S.
- Ainsi, $p_{W|S}$ sera la densité aléatoire de $P_{W|S}$ et $p_{W|S=s}$ sera la densité de $P_{W|S=s}$.
- Dénotons par $P_{W,S}$ la distribution jointe sur $\mathcal{W} \times \mathcal{Z}^m$ et par $p_{W,S}(w,s)$ sa densité de probabilité au point $(w,s) \in \mathcal{W} \times \mathcal{Z}^m$. On a

$$P_{W,S} = P_{W|S} P_S \quad ; \quad p_{W,S}(w,s) = p_{W|S=s}(w) p_S(s)$$

Algorithmes d'apprentissage stochastiques

- Notez que W est affecté par deux sources de stochasticité :
  - Une source venant de la stochasticité des données S produit par $P_S = \mathcal{D}^m$.
  - Une source venant de $Q(s) = P_{W|S=s}$ suite à une réalisation s de S. Cette partie est attribuable à la stochasticité de l'algorithme d'apprentissage sur l'échantillon s de S.
- Bien que les algorithmes PAC-Bayes produisent une distribution connue $Q(s)$, cela n'est pas le cas pour la majorité des algorithmes stochastiques (comme la descente de gradient stochastique).
- Suite à une réalisation s de S, la sortie $A(s)$ d'un algorithme stochastique est la variable aléatoire W dont la distribution est généralement inconnue (bien qu'elle existe).
- Nous allons démontrer que la nouvelle formulation (en espérance) du théorème PAC-Bayes général nous donne une garantie pour les algorithmes stochastiques.

Ré-écriture des espérances

- Avec ce changement de notation, nous avons

$$\mathop{\mathbb{E}}_{S} Q(S) = \mathop{\mathbb{E}}_{S} P_{W|S} = P_W \quad \text{(la marginale de } P_{W,S} \text{ sur } \mathcal{W})$$

- Nous avons également

$$\begin{aligned}\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \mathop{\mathbb{E}}_{W\sim Q(S)} f(W,S) &= \mathop{\mathbb{E}}_{S\sim P_S} \mathop{\mathbb{E}}_{W\sim P_{W|S}} f(W,S) = \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{W|S} f(W,S) \\ &= \mathop{\mathbb{E}}_{S,W\sim P_{W,S}} f(W,S) = \mathop{\mathbb{E}}_{W,S} f(W,S)\,.\end{aligned}$$

Ré-écriture des espérances

- En utilisant la définition de l'information mutuelle, nous obtenons

$$
\begin{aligned}
\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} D(Q(S)\|\mathop{\mathbb{E}}_{S} Q(S)) &= \mathop{\mathbb{E}}_{S} D(P_{W|S}\|P_W) \\
&= \mathop{\mathbb{E}}_{S} \int_{\mathcal{W}} dw\, p_{W|S}(w) \ln\left(\frac{p_{W|S}(w)}{p_W(w)}\right) \\
&= \int_{\mathcal{Z}^m} ds\, p_S(s) \int_{\mathcal{W}} dw\, p_{W|S=s}(w) \ln\left(\frac{p_S(s)p_{W|S=s}(w)}{p_S(s)p_W(w)}\right) \\
&= \int_{\mathcal{Z}^m} ds \int_{\mathcal{W}} dw\, p_{W,S}(w,s) \ln\left(\frac{p_{W,S}(w,s)}{p_S(s)p_W(w)}\right) \\
&= D(P_{W,S}\|P_W P_S) \stackrel{\text{def}}{=} I(W;S)\,.
\end{aligned}
$$

Théorème général pour algorithmes d'appr. stochastiques

Le dernier théorème se formule donc de la manière suivante.

Théorème (Théorème général pour algorithmes d'appr. stochastiques)

Soit $\mathcal{D}$, une distribution sur $\mathcal{Z}$ et $P_S = \mathcal{D}^m$. Soit A un algorithme d'apprentissage stochastique caractérisé par la distribution $P_{W|S}$ sur $\mathcal{W}$ tel que $A(S) = W \sim P_{W|S}$ lorsque $S \sim P_S$. Soit $\psi : \mathbb{R}_+ \to \mathbb{R}_+$ et $f : \mathcal{W} \times \mathcal{Z}^m \to \mathbb{R}$ satisfaisant

$$\mathop{\mathbb{E}}_{S} e^{\lambda f(w,S)} \leq \psi(\lambda), \ \forall \lambda > 0 \text{ et } \forall w \in \mathrm{supp}(P_W) \tag{8}$$

Alors pour tout $\lambda > 0$, on a

$$\mathop{\mathbb{E}}_{W,S} f(W,S) \ \leq \ \frac{1}{\lambda}\left[I(W;S) \ + \ \ln\left(\psi(\lambda)\right)\right].$$

Information mutuelle $I(W;S)$

- L'information mutuelle $I(W;S)$ mesure la réduction de l'incertitude sur W que nous procure l'observation de S. Nous avons

$$I(W;S) \;=\; H(W) - H(W|S)\,,$$

avec

$$H(W) \stackrel{\text{def}}{=} \int_{\mathcal{W}} dw\, p_W(w) \ln\left(\frac{1}{p_W(w)}\right) \quad \text{(entropie de } W\text{)}$$

$$H(W|S) \stackrel{\text{def}}{=} \int_{\mathcal{Z}^m} ds\, p_S(s) \int_{\mathcal{W}} dw\, p_{W|S=s}(w) \ln\left(\frac{1}{p_{W|S=s}(w)}\right)\,.$$

- Notez que, directement de la définition, on a $I(W;S) = I(S;W)$.
- Nous avons $I(W;S) = 0$ lorsque $P_{W,S} = P_W P_S$ (*i.e.*, lorsque W et S sont indépendantes).
- Référence de base : T.M. Cover et J.A. Thomas, *Elements of information theory (1991).*

Spécialisation du théorème général

- Utilisons maintenant $f(W,S) = L_{\mathcal{D}}(W) - L_S(W)$ dans le théorème général avec une fonction de perte sous Gaussienne, *i.e.*,

$$\mathop{\mathbb{E}}_{Z} e^{\lambda(\ell(w,Z)-L_{\mathcal{D}}(w))} \leq e^{\frac{\lambda^2\sigma^2}{2}}, \ \forall\lambda\in\mathbb{R}, \ \forall w\in\mathrm{supp}(P_W)\,.$$

- Nous avons vu que cela implique que $L_S(w)$ est $\sigma/\sqrt{m}$-sous Gaussien. Alors

$$\mathop{\mathbb{E}}_{S} e^{\lambda(L_{\mathcal{D}}(w)-L_S(w))} \leq e^{\frac{\lambda^2\sigma^2}{2m}}, \ \forall\lambda\in\mathbb{R}, \ \forall w\in\mathrm{supp}(P_W)\,.$$

- Dans ce cas, la condition (8) du thm général est satisfaite pour

$$\psi(\lambda) \ = \ e^{\frac{\lambda^2\sigma^2}{2m}}$$

- Le thm général implique donc que nous avons pour tout $\lambda > 0$,

$$\mathop{\mathbb{E}}_{W,S}\left[L_{\mathcal{D}}(W) - L_S(W)\right] \ \leq \ \frac{1}{\lambda}\left[I(W;S) \ + \ \frac{\lambda^2\sigma^2}{2m}\right].$$

Spécialisation du théorème général

- Notez que la valeur de λ minimisant

$$\frac{1}{\lambda} I(W;S) \; + \; \frac{\lambda \sigma^2}{2m} \,,$$

est donnée par

$$\frac{-1}{\lambda^2} I(W;S) + \frac{\sigma^2}{2m} \; = 0 \quad \Longrightarrow \quad \lambda = \sqrt{\frac{2m I(W;S)}{\sigma^2}} \,.$$

Ce qui donne

$$\frac{1}{\lambda} I(W;S) \; + \; \frac{\lambda \sigma^2}{2m} \; = \; 2\sqrt{\frac{\sigma^2}{2m} I(W;S)} \,.$$

Donc, le thm général implique que nous avons

$$\mathop{\mathbb{E}}_{W,S} \left[L_{\mathcal{D}}(W) - L_S(W) \right] \; \leq \; \sqrt{\frac{2\sigma^2}{m} I(W;S)} \,.$$

Spécialisation du théorème général

- Similairement, si nous utilisons $f(W,S) = L_S(W) - L_{\mathcal{D}}(W)$ et exploitons le fait que

$$\mathop{\mathbb{E}}_{S} e^{\lambda(L_S(w) - L_{\mathcal{D}}(w))} \leq e^{\frac{\lambda^2\sigma^2}{2m}}, \ \forall \lambda \in \mathbb{R}, \ \forall w \in \mathrm{supp}(P_W)\,.$$

- Nous avons pour tout $\lambda > 0$,

$$\mathop{\mathbb{E}}_{W,S}\left[L_S(W) - L_{\mathcal{D}}(W)\right] \ \leq \ \sqrt{\frac{2\sigma^2}{m} I(W;S)}\,.$$

- En combinant ces 2 derniers résultats nous obtenons, de manière différente, le théorème de Xu et Raginsky (NeurIPS 2017) :

Apprentissage stochastique et information mutuelle

Théorème (Xu et Raginsky (NeurIPS 2017))

Soit $\mathcal{D}$, une distribution sur $\mathcal{Z}$, $P_S = \mathcal{D}^m$ et ℓ tel que $\ell(w, Z)$ est σ-sous Gaussienne $\forall w \in \mathcal{W}$. Soit A un algorithme d'apprentissage stochastique caractérisé par la distribution $P_{W|S}$ sur $\mathcal{W}$ tel que $A(S) = W \sim P_{W|S}$ lorsque $S \sim P_S$. Alors on a

$$\left| \mathop{\mathbb{E}}_{W,S} \left[L_{\mathcal{D}}(W) - L_S(W) \right] \right| \leq \sqrt{\frac{2\sigma^2}{m} I(S;W)}\,.$$

- Il n'y a pas d'"overfitting" lorsque $I(S;W) \ll m$.
- La stabilité de l'algorithme d'apprentissage A (caractérisant $P_{W|S}$) augmente en diminuant $I(S;W)$.

Comment s'approcher du risque minimal ?

- Soit $w^* \in \operatorname{argmin}_{w \in \mathcal{W}} L_{\mathcal{D}}(w)$ le prédicteur optimal dans $\mathcal{W}$.
- Soit $w_S \in \operatorname{argmin}_{w \in \mathcal{W}} L_S(w)$.
- Nous avons alors : $\mathbb{E}_S\, L_S(w_S) \;\leq\; \mathbb{E}_S\, L_S(w^*) = L_{\mathcal{D}}(w^*)$.
- Alors on a

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{S}\ \mathop{\mathbb{E}}_{W\sim P_{W|S}} L_{\mathcal{D}}(W) - L_{\mathcal{D}}(w^*) \\
&\quad\leq \mathop{\mathbb{E}}_{S}\ \mathop{\mathbb{E}}_{W\sim P_{W|S}} L_S(W) - L_{\mathcal{D}}(w^*) + \sqrt{\frac{2\sigma^2}{m} I(S;W)} \\
&\quad\leq \mathop{\mathbb{E}}_{S}\ \mathop{\mathbb{E}}_{W\sim P_{W|S}} L_S(W) - L_S(w_S) + \sqrt{\frac{2\sigma^2}{m} I(S;W)} \\
&\quad= \mathop{\mathbb{E}}_{S} \underbrace{\left[\mathop{\mathbb{E}}_{W\sim P_{W|S}} L_S(W) - L_S(w_S)\right]}_{\text{erreur d'optimisation } \epsilon_S^{\text{opt}}(P_{W|S})} + \sqrt{\frac{2\sigma^2}{m} I(S;W)}\,.
\end{aligned}
$$

- $P_{W|S}$ doit donc minimiser $\mathbb{E}_S\, \epsilon_S^{\text{opt}}(P_{W|S})$ avec $I(S;W) \ll m$.

Sommaire

1. Introduction et définitions
2. La KL divergence et le changement de mesure Donsker-Varadhan
3. Théorème PAC-Bayes général et bornes PAC-Bayes
4. Algorithme d'apprentissage PAC-Bayésien et sa garantie PAC
   - Apprentissage PAC-Bayésien en pratique et Gaussiennes isotropes
5. Garantie PAC-Bayes pour algorithmes retournant distribution Q sur $\mathcal{W}$
6. Garantie PAC-Bayes pour alg. stochastiques et information mutuelle
7. Descente de gradient stochastique avec dynamique de Langevin

Analyse de la SGLD de Pensia et al. (ISIT 2018)

- Il est possible de borner $I(W;S)$ pour la descente de gradient stochastique avec dynamique de Langevin (SGLD).
  - C'est une DGS avec une injection explicite de bruit.
- Soit $S = (Z_1, \ldots, Z_m)$ la variable aléatoire associée à l'échantillon d'apprentissage.
- À chaque itération t de la SGLD, on tire $i \sim \mathcal{U}([m])$ et on utilise $Z^t = Z_i$. Ensuite on tire $\xi_t \sim \mathcal{N}(0, \sigma_t^2 I_d)$.
- La variable aléatoire W associée au prédicteur final est donnée par

$$W = \frac{1}{T} \sum_{t=1}^{T} W^t \,,$$

avec $W^0 = \mathbf{0}$. En utilisant $\xi_t \sim \mathcal{N}(0, \sigma_t^2 I_d)$, la règle de mise à jour s'écrit

$$W^t = W^{t-1} - \eta_t \nabla \ell(W^{t-1}, Z^t) + \xi_t \,.$$

Analyse de la SGLD de Pensia et al. (ISIT 2018) - suite

- Pour tout $t \in [T]$, $W^{(t)} \stackrel{\text{def}}{=} (W^1, \ldots, W^t)$ et $Z^{(t)} \stackrel{\text{def}}{=} (Z^1, \ldots, Z^t)$.
- **Définition :** X, Y, Z forment la chaîne de Markov $X \to Y \to Z$ si Z est conditionnellement indépendant de X sachant Y.
- "Data processing inequality" : $I(X;Z) \leq \left\{ \begin{array}{l} I(X;Y) \\ I(Y;Z) \end{array} \right.$.
  - Aucun traitement Y ne peut contribuer à augmenter $I(X;Z)$.
- Dans notre cas, nous avons $S \to Z^{(T)} \to W^{(T)} \to W$. Alors

$$\begin{aligned} I(S;W) \leq \; & I(Z^{(T)};W^{(T)}) \\ & = I(Z^{(T)};W^1) + I(Z^{(T)};W^2|W^1) + I(Z^{(T)};W^3|W^2,W^1) + \\ & \quad \ldots + I(Z^{(T)};W^T|W^{(T-1)}) \,, \end{aligned}$$

où l'égalité est due à la règle d'enchainement de l'information mutuelle (Cover et Thomas, THM 2.5.2).

Analyse de la SGLD de Pensia et al. (ISIT 2018) - suite

- De plus, en utilisant la relation entre information mutuelle et entropie et en exploitant la dépendance de W^t sur $W^{(t-1)}$, on a

$$\begin{aligned} I(Z^{(t)}; W^t | W^{(t-1)}) &= H(W^t | W^{(t-1)}) - H(W^t | W^{(t-1)}, Z^{(t)}) \\ &= H(W^t | W^{t-1}) - H(W^t | W^{t-1}, Z^t) \\ &\stackrel{\text{def}}{=} I(W^t; Z^t | W^{t-1}) \end{aligned}$$

- En combinant ce résultat avec la page précédente, on a

$$I(S; W) \leq \sum_{t=1}^{T} I(W^t; Z^t | W^{t-1}) \, .$$

- Pour tout w^{t-1}, cherchons à borner

$$I(W^t; Z^t | W^{t-1} = w^{t-1}) = H(W^t | W^{t-1} = w^{t-1}) - H(W^t | Z^t, W^{t-1} = w^{t-1})$$

indépendamment de la valeur de w^{t-1}.

Analyse de la SGLD de Pensia et al. (ISIT 2018) - suite

- La translation d'une variable ne modifie pas son entropie. Donc

$$\begin{aligned} H(W^t|W^{t-1}=w^{t-1}) &= H(W^t - w^{t-1}|W^{t-1}=w^{t-1}) \\ &= H(-\eta_t \nabla \ell(w^{t-1}, Z^t) + \xi_t)\,. \end{aligned}$$

- Or, $\eta_t \nabla \ell(w^{t-1}, Z^t)$ et ξ_t sont deux vecteurs aléatoires indépendants. Donc, pour une fonction de perte ℓ qui est ρ-Lipshitzienne, nous avons

$$\begin{aligned} \mathbb{E}\,\| - \eta_t \nabla \ell(w^{t-1}, Z^t) + \xi_t\|^2 &= \mathbb{E}\,\|\eta_t \nabla \ell(w^{t-1}, Z^t)\|^2 + \mathbb{E}\,\|\xi_t\|^2 \\ &\leq \eta_t^2 \rho^2 + d\sigma_t^2\,. \end{aligned}$$

- Or, parmi toutes les variables aléatoires X dont $\mathbb{E}\,\|X\|^2 < C$, $Y \sim \mathcal{N}(0, \frac{C}{d} I_d)$ possède la plus grande entropie, donnée par

$$H(Y) = \frac{d}{2} \ln\left(\frac{2\pi e C}{d}\right)\,.$$

- Alors on a

$$H(W^t|W^{t-1}=w^{t-1}) \leq \frac{d}{2} \ln\left(2\pi e \frac{\eta_t^2 \rho^2 + d\sigma_t^2}{d}\right)\,.$$

Analyse de la SGLD de Pensia et al. (ISIT 2018) - suite

- Similairement,

$$H(W^t|Z^t = z^t, W^{t-1} = w^{t-1}) \;=\; H(\xi_t)\,.$$

- Puisque ce résultat est valide indépendamment de z^t et w^{t-1}, on a

$$H(W^t|Z^t, W^{t-1}) \;=\; \frac{d}{2}\ln\left(2\pi e\sigma_t^2\right)\,.$$

- Conséquemment,

$$\begin{aligned} I(W^t; Z^t|W^{t-1}) &= H(W^t|W^{t-1}) - H(W^t|W^{t-1}, Z^t) \\ &\leq \frac{d}{2}\ln\left(2\pi e\frac{\eta_t^2\rho^2 + d\sigma_t^2}{d}\right) - \frac{d}{2}\ln\left(2\pi e\sigma_t^2\right) \\ &= \frac{d}{2}\ln\left(\frac{\eta_t^2\rho^2 + d\sigma_t^2}{d\sigma_t^2}\right) = \frac{d}{2}\ln\left(1 + \frac{\eta_t^2\rho^2}{d\sigma_t^2}\right) \quad \leq \frac{\eta_t^2\rho^2}{2\sigma_t^2} \end{aligned}$$

en utilisant $\ln(1+x) \leq x$.

Analyse de la SGLD de Pensia et al. (ISIT 2018) - suite

- En combinant les résultats des 3 dernières pages et en utilisant $\sigma_t^2 = \eta_t$ et $\eta_t = c/t$, nous obtenons

$$I(S;W) \leq \sum_{t=1}^{T} \frac{\eta_t^2 \rho^2}{2\sigma_t^2} = \frac{c\rho^2}{2} \sum_{t=1}^{T} \frac{1}{t} \leq \frac{c\rho^2}{2}[1 + \ln(T)]\,.$$

- Nous obtenons donc le corollaire suivant.

Corollaire (Pensia et al. (ISIT 2018))

Soit $\mathcal{D}$, une distribution sur $\mathcal{Z}$, $P_S = \mathcal{D}^m$ et ℓ tel que $\ell(w, Z)$ est σ-sous Gaussienne $\forall w \in \mathcal{W}$ et que $\ell(w, z)$ soit ρ-Lipschitzienne pour tout z. Soit A l'algorithme d'apprentissage SGLD (avec $\sigma_t^2 = \eta_t$ et $\eta_t = c/t$) donnant une distribution $P_{W|S}$ sur $\mathcal{W}$ tel que $A(S) = W \sim P_{W|S}$ lorsque $S \sim P_S$. Alors on a

$$\left| \mathop{\mathbb{E}}_{W,S} [L_{\mathcal{D}}(W) - L_S(W)] \right| \leq \sqrt{\frac{2\sigma^2}{m} I(S;W)} \leq \rho\sigma \sqrt{\frac{c[1 + \ln(T)]}{m}}\,.$$

Analyse de la SGLD de Pensia et al. (ISIT 2018) - suite

- Rappel : on cherche $P_{W|S}$ minimisant le membre de droite de

$$\mathop{\mathbb{E}}_{S}\mathop{\mathbb{E}}_{W\sim P_{W|S}} L_{\mathcal{D}}(W) \;-\; L_{\mathcal{D}}(w^*)$$

$$\leq \mathop{\mathbb{E}}_{S}\underbrace{\left[\mathop{\mathbb{E}}_{W\sim P_{W|S}} L_S(W) \;-\; L_S(w_S)\right]}_{\text{erreur d'optimisation } \epsilon_S^{\mathsf{opt}}(P_{W|S})} + \sqrt{\frac{2\sigma^2}{m} I(S;W)}\,.$$

- Cependant les choix $\sigma_t^2 = \eta_t$ et $\eta_t = c/t$ ne garantissent pas nécessairement une faible valeur pour $\mathbb{E}_S\, \epsilon_S^{\mathsf{opt}}(P_{W|S})$.
- Par contre, ce sera le cas si la fonction de perte ℓ est convexe et si l'on choisi $\sigma_t = \sigma_\xi$ et $\eta_t = \eta$. Dans ce cas

$$W^t \;=\; W^{t-1} - \eta\left(\nabla\ell(W^{t-1}, Z^t) + \frac{\xi}{\eta}\right) \;\stackrel{\mathsf{def}}{=}\; W^{t-1} - \eta V^{t-1}\,,$$

- Étant donné que $\mathbb{E}\,\xi = \mathbf{0}$, on a $\mathbb{E}[V^t|W^t, S] = \nabla L_S(W^t)$.

Analyse de la SGLD de Pensia et al. (ISIT 2018) - suite

- Puisque V^t est un estimateur non biaisé de $\nabla L_S(W^t)$, le théorème de convergence de la DGS du chap.14 du manuel s'applique aussi pour la minimisation de $L_S(W)$ avec la SGLD, mais pour

$$\mathbb{E}\,\|V^t\|^2 \;\leq\; \mathbb{E}\,\|\nabla \ell(W^{t-1}, Z^t)\|^2 + \mathbb{E}\left\|\frac{\xi}{\eta}\right\|^2 \;\leq\; \rho^2 + d\frac{\sigma_\xi^2}{\eta^2}$$

- Donc si

$$w_S \stackrel{\text{def}}{=} \operatorname*{argmin}_{w:\|w\|\leq B} L_S(w)\,,$$

- Nous avons (S est fixe, la randomisation est sur les Z^t et ξ)

$$\epsilon_S^{\text{opt}}(P_{W|S}) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{W|S} L_S(W) - L_S(w_S) \;=\; \mathop{\mathbb{E}}_{W|S} L_S\left(\frac{1}{T}\sum_{t=1}^T W^t\right) - L_S(w_S)$$

$$\leq \mathop{\mathbb{E}}_{W|S}\left(\frac{1}{T}\sum_{t=1}^T L_S(W^t)\right) - L_S(w_S) \leq \frac{B^2}{2\eta T} + \frac{\eta}{2}\left[\rho^2 + d\frac{\sigma_\xi^2}{\eta^2}\right]$$

Analyse de la SGLD de Pensia et al. (ISIT 2018) - suite

- En combinant ce dernier résultat avec

$$I(S;W) \;\leq\; \sum_{t=1}^{T} \frac{\eta_t^2 \rho^2}{2\sigma_t^2} \;=\; T\frac{\eta^2\rho^2}{2\sigma_\xi^2}\,,$$

- Nous obtenons

$$\begin{aligned}
\mathop{\mathbb{E}}_{S,W} L_{\mathcal{D}}(W) \;-\; L_{\mathcal{D}}(w^*) &\leq \mathop{\mathbb{E}}_{S} \epsilon_S^{\mathsf{opt}}(P_{W|S}) + \sqrt{\frac{2\sigma^2}{m} I(S;W)} \\
&\leq \frac{B^2}{2\eta T} + \frac{\eta}{2}\left[\rho^2 + d\frac{\sigma_\xi^2}{\eta^2}\right] + \sqrt{T\frac{\eta^2\rho^2\sigma^2}{m\sigma_\xi^2}} \\
&= \frac{1}{\eta}\left[\frac{B^2}{2T} + \frac{d\sigma_\xi^2}{2}\right] + \eta\left[\frac{\rho^2}{2} + \frac{\rho\sigma}{\sigma_\xi}\sqrt{\frac{T}{m}}\right]
\end{aligned}$$

Analyse de la SGLD de Pensia et al. (ISIT 2018) - suite

- Si nous choisissons

$$\sigma_\xi = \frac{B}{\sqrt{dT}} \quad ; \quad \eta = \sqrt{\frac{B^2}{\rho T\left(\frac{\rho}{2} + \sigma T\sqrt{\frac{d}{m}}\right)}},$$

- Nous obtenons finalement

$$\mathop{\mathbb{E}}_{S,W} L_{\mathcal{D}}(W) \; - \; L_{\mathcal{D}}(w^*) \; \leq \; 2B\rho\sqrt{\frac{1}{2T} + \frac{\sigma}{\rho}\sqrt{\frac{d}{m}}},$$

ce qui est une convergence très lente : en $O\left(\left[\frac{d}{m}\right]^{1/4}\right)$.

Résumé

- Le changement de mesure Donsker-Varadhan permet de borner la différence entre le risque et le risque empirique simultanément pour toute distribution Q sur $\mathcal{W}$ en fonction de $D(Q\|P)$ pour un P choisi.
- Ceci permet de borner le risque des algorithmes PAC-Bayésiens.
- $\mathrm{PB}_{\mathcal{W}}(S)$ permet d'apprendre au sens PAC-agnostique la classe

$$\mathcal{M}_{\mathcal{W}}(P, D^*) \stackrel{\text{def}}{=} \{Q \text{ sur } \mathcal{W} : D(Q\|P) \leq D^*\}\,.$$

- Pour un algorithme retournant un postérieur $Q(S)$, le prior P optimal est $\mathbb{E}_S\, Q(S)$.
- Pour tout algorithme d'apprentissage stochastique, on a

$$\left| \mathop{\mathbb{E}}_{S,W} \left[L_{\mathcal{D}}(W) - L_S(W) \right] \right| \leq \sqrt{\frac{2\sigma^2}{m} I(S;W)}\,.$$

- Il est possible de borner $I(S;W)$ pour SGLD.