# Machine Learning 9: Regularization and Stability

Master 2 Computer Science

Aurélien Garivier

2018-2019

UNIVERSITÉ DE LYON

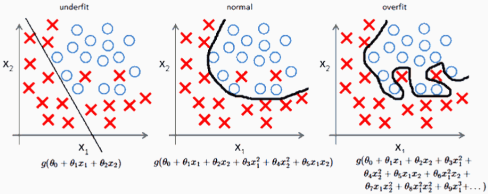
ENS DE LYON

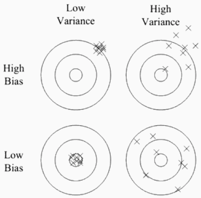
# Table of contents

# Regularization and Structural Risk Minimization

# Overfitting

Example: linear classification with polynomial features




Src: `http://mlwiki.org`

$\rightarrow$ how to get the best from several hypothesis classes?

# Nonuniform Learnability

**Definition**

A hypothesis class $\mathcal{H}$ is *nonuniformy learnable* if there exists a learning algorithm $A$ and a function $m_{\mathcal{H}}^{NUL} : (0,1)^2 \times \mathcal{H} \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every $h \in \mathcal{H}$, if $m \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$ then with probability at least $1 - \delta$ over the sample $S \sim D^{\otimes m}$,

$$L_D\big(A(S)\big) \leq L_D(h) + \epsilon \ .$$

**Theorem**

A hypothesis calss $\mathcal{H}$ of binary classifiers is nonuniformly learnable if and only if fit is a countable union of agnostic PAC learnable hypothesis classes.

# Structural Risk Minimization 1/2

Let $\mathcal{H} = \cup_{d \in \mathbb{N}} \mathcal{H}_d$, where each hypothesis class $\mathcal{H}_d$ is PAC learnable with uniform convergence rate $m^{UC}_{\mathcal{H}_d}$, and let $\epsilon_d : \mathbb{N} \times (0,1) \to (0,1)$ be defined as

$$\epsilon_d(m,\delta) = \min\left\{\epsilon \in (0,1) : m^{UC}_{\mathcal{H}_d}(\epsilon,\delta) \leq m\right\} \ .$$

For every $h \in \mathcal{H}$ let $d(h) = \min\left\{d : h \in \mathcal{H}_d\right\}$. Let also $w : \mathbb{N} \to [0,1]$ be such that $\sum_{d=0}^{\infty} w(d) \leq 1$.

**Lemma**

For every $\delta \in (0,1)$ and for every distribution $D$, with probability at least $1-\delta$ over the sample $S \sim D^{\otimes m}$,

$$\forall h \in \mathcal{H}, \ \ L_D(h) \leq L_S(h) + \epsilon_{d(h)}\Big(m, w\big(d(h)\big)\delta\Big) \ .$$

# Structural Risk Minimization 2/2

**Structural Risk Minimization (SRM)**

$$A(S) \in \arg\min_{h\in\mathcal{H}} L_s(h) + \epsilon_{d(h)}\Big(m, w\big(d(h)\big)\delta\Big) .$$

Typical choice: $w(d) = \frac{6}{\pi^2(d+1)^2}$ gives for SRM the nonuniform learning rate

$$m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h) \leq m_{\mathcal{H}_{d(h)}}^{UC}\left(\frac{\epsilon}{2}, \frac{6\delta}{\pi^2 d(h)^2}\right) .$$

If $\mathrm{VCdim}(\mathcal{H}_d) = d$, $m_{\mathcal{H}_d}^{UC}(\epsilon/2, \delta) = C\frac{d+\log(1/\delta)}{\epsilon^2}$ and hence

$$m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h) - m_{\mathcal{H}_d}^{UC}(\epsilon/2, \delta) \leq \frac{8C\log(2d)}{\epsilon^2} .$$

Remark: other strategy = *aggregation*, cf PAC-Bayes learning.

# Minimum Description Length and Occam's razor

Entiae non sunt multiplicanda praeter necessitatem
(Entities are not to be multiplied without necessity)
Here: A short explanation tends to be more valid
(generalize better) than a long explanation

Suggests a choice for $w(d)$: should penalize complexity.

More precisely: if $|h|$ is the length of a prefix-free binary code for the hypothesis $h$, set

$$w(h) = 2^{-|h|} .$$

By Hoeffding's inequality, this typically yields the

**Minimum Description Length (MDL) estimator:**

$$A(S) \in \arg\min_{h \in \mathcal{H}} L_S(h) + \sqrt{\frac{|h| + \log \frac{2}{\delta}}{2m}} .$$

This heuristic needs to be justified statistically (often possible).

# Regularization and Stability

# Stable Rules do not overfit

**Theorem**

Let $D$ be a distribution on $\mathcal{X} \times \{\pm 1\}$, $S = (z_1, \ldots, z_m)$ be an iid sequence of examples, $z'$ be another independent sample of $D$, and let I be an independent sample of the uniform distribution on $\{1, \ldots, m\}$. For all $1 \leq i \leq m$, let $S^{(i)} = (z_1, \ldots, z_{i-1}, z', z_{i+1}, \ldots, z_m)$. Then, for any learning alogrithm $A$,

$$\mathbb{E}_S\left[L_D\big(A(S)\big) - L_S\big(A(S)\big)\right] = \mathbb{E}_{S,z',I}\left[\ell\big(A(S^{(I)}), z_I\big) - \ell\big(A(S), z_I\big)\right] .$$

Indeed, $\mathbb{E}_{S,z',I}\left[\ell\big(A(S^{(I)}), z_I\big)\right] = \mathbb{E}_S\left[L_D\big(A(S)\big)\right]$, and $\mathbb{E}_{S,I}\left[\ell\big(A(S), z_I\big)\right] = \mathbb{E}_S\left[L_S\big(A(S)\big)\right]$.

**Definition**

Algorithm $A$ is said to be *on-average-replace-one*-stable with rate $\epsilon : \mathbb{N} \to \mathbb{R}$ if for every distribution $D$ and every sample size $m \in \mathbb{N}$,

$$\mathbb{E}_{S,z',I}\left[\ell\big(A(S^{(I)}), z_I\big) - \ell\big(A(S), z_I\big)\right] \leq \epsilon_m .$$

# Tikhonov Regularization as a Stabilizer

We consider a class $\mathcal{H} = \left\{ h_w : w \in \bigcup_{d \geq 0} \mathbb{R}^d \right\}$.

**Definition**

Tikhonov's Regularized Loss Minimizer is defined as

$$A(S) \in \underset{h_w \in \mathcal{H}}{\arg\min}\, L_S(h) + \lambda \|w\|^2 ,$$

where $\lambda > 0$ is a parameter.

With square loss on $\mathbb{R}^d$, the resulting estimator is called *ridge regression*:

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\arg\min}\, \frac{1}{m} \sum_{i=1}^{m} \frac{1}{2} \left( \langle w, x_i \rangle - y_i \right)^2 + \lambda \|w\|^2 = \left( 2\lambda m I_d + X^T X \right)^{-1} X^T y ,$$

where $X = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{pmatrix}$ and $y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{pmatrix}$.

# Tikhonov's RLM for convex loss is stable

Denote $f_S(w) = L_S(w) + \lambda\|w\|^2$. If $\ell$ is convex, then $f$ is $2\lambda$-strongly convex, and thus

$$f_S\big(A(S^{(i)}) - f_S\big(A(S)\big) \geq \lambda\big\|A(S^{(i)}) - A(S)\big\|^2 ,$$

and

$$f_S\big(A(S^{(i)})\big) - f_S\big(A(S)\big) = \underbrace{L_{S^{(i)}}\big(A(S^{(i)})\big) + \lambda\big|A(S^{(i)}\big|^2 - L_{S^{(i)}}\big(A(S)\big) - \lambda\big|A(S)\big|^2}_{\leq 0}$$

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

and hence

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

# Lipschitz loss

When the loss $\ell(\cdot, z)$ is $\rho$-Lipschitz for every $z$, we obtain that

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

when entails $\left\| A(S^{(i)}) - A(S) \right\| \leq \dfrac{2\rho}{\lambda m}$.

**RLM generalizes well Lispchitz Losses**

When the loss function $\ell(\cdot, z)$ is convex and $\rho$-Lipschitz for all $z$, Tihkohnov's RLM is on-average-one-stable with rate $\frac{2\rho^2}{\lambda m}$, and hence

$$\mathbb{E}_S \left[ L_D(A(S)) - L_S(A(S)) \right] \leq \frac{2\rho^2}{\lambda m} .$$

Remark: when $\ell$ is $\beta$-smooth and non-negative, and when $\ell(0, z) \leq C$ for all $z$, one can prove that for $\lambda \geq \frac{2\beta}{m}$ Tikhonov's RLM satisfies

$$\mathbb{E}_S \left[ L_D(A(S)) - L_S(A(S)) \right] \leq \frac{48\beta}{\lambda m} \mathbb{E}\left[ L_S(A(S)) \right] \leq \frac{48\beta C}{\lambda m} .$$

# Controlling Fitting-Stability Tradeoff

Fitting-stability tradeoff:

$$\mathbb{E}_S\Big[L_D\big(A(S)\big)\Big] = \underbrace{\mathbb{E}_S\Big[L_S\big(A(S)\big)\Big]}_{\text{fitting error}} + \underbrace{\mathbb{E}_S\Big[L_D\big(A(S)\big) - L_S\big(A(S)\big)\Big]}_{\text{generalization error} = \text{stability}} .$$

The stronger the regularization (the larger $\lambda$), the better the stability BUT the higher the bias.

But for every $h_w \in \mathcal{H}$,

$$\mathbb{E}_S\Big[L_S\big(A(S)\big)\Big] \leq \mathbb{E}_S\Big[L_S(h_w) + \lambda\|w\|^2\Big] = L_D(h_w) + \lambda\|w\|^2 .$$

**Oracle inequality**

If the loss function $\ell(\cdot, z)$ is convex and $\rho$-Lipschitz for all $z$, Tikhonov's RLM satisfies

$$\mathbb{E}_S\Big[L_D\big(A(S)\big)\Big] \leq \inf_{h_w \in \mathcal{H}} L_D(h_w) + \lambda\|w\|^2 + \frac{2\rho^2}{\lambda m}$$

**Corollary**

If $\forall h_w \in \mathcal{H}, \|w\| \leq B$ and if the loss function $\ell(\cdot, z)$ is convex and $\rho$-Lipschitz for all $z$, Tikhonov's RLM with $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$ satisfies:

$$\mathbb{E}_S\left[L_D\big(A(S)\big)\right] \leq \inf_{h_w \in \mathcal{H}} L_D(h_w) + \rho B \sqrt{\frac{8}{m}} \,.$$

Hence, for every $\epsilon > 0$, if $m \geq \frac{8\rho^2 B^2}{\epsilon^2}$ then for every distribution $D$ $\mathbb{E}_S\left[L_D\big(A(S)\big)\right] \leq \inf_{h_w \in \mathcal{H}} L_D(h_w) + \epsilon$.

The same kind of result can be obtained for $\beta$-smooth, non-negative losses: with $\lambda = \epsilon/(3B^2)$, for every $m \geq \frac{150\beta B^2}{\epsilon^2}$, whatever the distribution $D$, $\mathbb{E}_S\left[L_D\big(A(S)\big)\right] \leq \inf_{h_w \in \mathcal{H}} L_D(h_w) + \epsilon$.

In practice, $\lambda$ is most often chosen by cross-validation.

# Example: Ridge regression generalizes well

**Theorem**

Let $D$ be a distribution over $\mathcal{X} \times [-1, 1]$, where $\mathcal{X} = \left\{x \in \mathbb{R}^d : \|x\| \leq 1\right\}$. Let $\mathcal{H} = \left\{w \in \mathbb{R}^d : \|w\| \leq B\right\}$. For any $\epsilon \in (0, 1)$, let $m \geq m_{\mathcal{H}}(\epsilon) = 150B^2/\epsilon^2$. Then ridge regression with parameter $\lambda = \epsilon/(3B^2)$ satisfies:

$$\mathbb{E}_S\Big[L_D\big(A(S)\big)\Big] \leq \min_{w \in \mathcal{H}} L_D(w) + \epsilon \ .$$

Furthermore, for every $\delta \in (0, 1)$ and every $m \geq m_{\mathcal{H}}(\epsilon, \delta) = m_{\mathcal{H}}(\epsilon\delta)$,

$$\mathbb{P}_S\Big(L_D\big(A(S)\big) \leq \min_{w \in \mathcal{H}} L_D(w) + \epsilon\Big) \geq 1 - \delta \ .$$

Expectation to high-probability PAC learning: the sample complexity can be reduced to $m_{\mathcal{H}}(\epsilon, \delta) = m_{\mathcal{H}}(\epsilon/2)\left\lceil \log_2(1/\delta)\right\rceil + \left\lceil \frac{\log(4/\delta) + \log(\lceil \log_2(1/\delta)\rceil)}{\epsilon^2}\right\rceil$ when the loss function is bounded by 1.

# Support Vector Machines

# Margin for linear separation

- Training sample $S = \big\{(x_1, y_1), \ldots, (x_m, y_m)\big\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$.
- Linearly separable if there exists a halfspace $h = (w, b)$ such that $\forall i, y_i = \text{sign}\big(\langle w, x_i \rangle + b\big)$.
- What is the best separating hyperplane for generalization?

**Distance to hyperplane**

If $\|w\| = 1$, then the distance from $x$ to the hyperplane $h = (w, b)$ is $d(x, \mathcal{H}) = \big|\langle w, x \rangle + b\big|$.

**Proof:** Check that $\min \big\{\|x - v\|^2 : v \in h\big\}$ is reached at $v = x - \big(\langle w, x \rangle + b\big) w$.

# Hard-SVM

Formulation 1:

$$\underset{(w,b):\|w\|=1}{\arg\max} \min_{1\le i\le m} \left|\langle w, x_i\rangle + b\right| \quad \text{such that } \forall i, y_i\big(\langle w, x_i\rangle + b\big) > 0 \ .$$

Formulation 2:

$$\min_{w,b} \|w\|^2 \quad \text{such that } \forall i, y_i\big(\langle w, x_i\rangle + b\big) \ge 1 \ .$$

Remark: $b$ is not penalized.

**Proposition**

The two formulations are equivalent.

**Proof:** if $(w_0, b_0)$ is the solution of Formulation 2, then $\hat{w} = \frac{w_0}{\|w_0\|}$ , $\hat{b} = \frac{b_0}{|w|}$ is a solution of Formulation 1: if $(w^*, b^*)$ is another solution, then letting $\gamma^* = \min_{1\le i\le m} y_i\big(\langle w, x_i\rangle + b\big)$ we see that $\left(\frac{w^*}{\gamma^*}, \frac{b^*}{\gamma^*}\right)$ satisfies the constraint of Formulation 2, hence $\|w_0\| \le \frac{\|w^*\|}{\gamma^*} = \frac{1}{\gamma^*}$ and thus $\min_{1\le i\le m} \left|\langle \hat{w}, x_i\rangle + \hat{b}\right| = \frac{1}{\|w_0\|} \ge \gamma^*$.

# Sample Complexity

**Definition**

A distribution $D$ over $\mathbb{R}^d \times \{\pm 1\}$ is *separable with a* $(\gamma, \rho)$*-margin* if there exists $(w^*, b^*)$ such that $\|w^*\| = 1$ and with probability 1 on a pair $(X, Y) \sim D$, it holds that $\|X\| \leq \rho$ and $Y\big(\langle w^*, X\rangle + b\big) \geq \gamma$.

Remark: by multiplying the $x_i$ by $\alpha$, the margin is mutlplied by $\alpha$.

**Theorem**

For any distribution $D$ over $\mathbb{R}^d \times \{\pm 1\}$ that satisfies the $(\gamma, \rho)$-separability with margin assumption using a homogenous halfspace, with probability at least $1 - \delta$ over the training set of size $m$ the $0 - 1$ loss of the output of Hard-SVM is at most

$$\sqrt{\frac{4(\rho/\gamma)^2}{m}} + \sqrt{\frac{2\log(2/\delta)]}{m}} .$$

Remark: depends on dimension $d$ only thru $\rho$ and $\gamma$.

# Soft-SVM

When the data is not linearly separable, allow *slack variables* $\xi_i$:

$$\min_{w,b,\xi} \lambda\|w\|^2 + \frac{1}{m}\sum_{i=1}^{m} \xi_i \quad \text{such that } \forall i, y_i\big(\langle w, x_i\rangle + b\big) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

$$= \min_{w,b} \lambda\|w\|^2 + L_S^{\mathrm{hinge}}(w, b) \quad \text{where } \ell^{\mathrm{hinge}}(u) = \max(0, 1 - u)\ .$$

**Theorem**

Let $D$ be a distribution over $B(0,\rho) \times \{\pm 1\}$. If $A(S)$ is the output of the soft-SVM algorithm on the sample $S$ of $D$ of size $m$,

$$\mathbb{E}\Big[L_D^{0-1}\big(A(S)\big)\Big] \leq \mathbb{E}\Big[L_D^{\mathrm{hinge}}\big(A(S)\big)\Big] \leq \inf_u L_D^{\mathrm{hinge}}(u) + \lambda\|u\|^2 + \frac{2\rho^2}{\lambda m}\ .$$

For every $B > 0$, setting $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$ yields:

$$\mathbb{E}\Big[L_D^{0-1}\big(A(S)\big)\Big] \leq \mathbb{E}\Big[L_D^{\mathrm{hinge}}\big(A(S)\big)\Big] \leq \inf_{w:\|w\|\leq B} L_D^{\mathrm{hinge}}(w) + \sqrt{\frac{8\rho^2 B^2}{m}}\ .$$

# Dual Form of the SVM Optimization Problem

To simplify, we consider only the homogeneous case of hard-SVM. Let

$$g(w) = \max_{\alpha \in [0,+\infty)^m} \sum_{i=1}^{m} \alpha_i \big(1 - y_i \langle w, x_i \rangle\big) = \begin{cases} 0 & \text{if } \forall i, y_i \langle w, x_i \rangle \geq 1, \\ +\infty & \text{otherwise .} \end{cases}$$

Then the hard-SVM problem is equivalent to

$$\begin{aligned}
\min_{w:\forall i, y_i \langle w, x_i \rangle \geq 1} \frac{1}{2}\|w\|^2 &= \min_{w} \frac{1}{2}\|w\|^2 + g(w) \\
&= \min_{w} \max_{\alpha \in [0,+\infty)^m} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i \big(1 - y_i \langle w, x_i \rangle\big) \\
&\overset{\text{min}-\text{max thm}}{=} \max_{\alpha \in [0,+\infty)^m} \min_{w} \frac{1}{2}\|w\|^2 + \sum_{i=1}^{m} \alpha_i \big(1 - y_i \langle w, x_i \rangle\big) \; .
\end{aligned}$$

The inner min is reached at $w = \sum_{i=1}^{m} \alpha_i y_i x_i$ and can thus be written as

$$\max_{\alpha \in \mathbb{R}^m, \alpha \geq 0} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{1 \leq i,j \leq m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle \; .$$

# Support vectors

Still for the homogeneous case of hard-SVM:

**Property**

Let $w_0$ be a solution of and let $I = \{i : \left|\langle w_0, x_i\rangle\right| = 1\}$. There exist $\alpha_1, \ldots, \alpha_m$ such that

$$w_0 = \sum_{i \in I} \alpha_i x_i \ .$$

The dual problem involves the $x_i$ only thru scalar products $\langle x_i, x_j \rangle$.

It is of size $m$ (independent of the dimension $d$).

These computations can be extended to the non-homogeneous soft-SVM

$\rightarrow$ **Kernel trick**.

# Numerically solving Soft-SVM

$f(w) = \frac{\lambda}{2}\|w\|^2 + L_S^{\text{hinge}}(w)$ is $\lambda$-strongly convex.
$\rightarrow$ Stochastic Gradient Descent with learning rate $1/(\lambda t)$. Stochastic subgradient of $L_S^{\text{hinge}}(w)$ : $v_t = -y_{I_t} x_{I_t} \mathbb{1}\{y_{I_t}\langle w, x_{I_t}\rangle < 1\}$.

$$w_{t+1} = w_t - \frac{1}{\lambda t}(\lambda w_t + v_t) = \frac{t-1}{t} w_t - \frac{1}{\lambda t} v_t = -\frac{1}{\lambda t}\sum_{i=1}^{t} v_t \ .$$

**Algorithm:** SGD for Soft-SVM

Set $\theta_0 = 0$
**for** $t = 0 \dots T-1$ **do**
  Let $w_t = \frac{1}{\lambda t}\theta_t$
  Pick $I_t \sim \mathcal{U}\big(\{1, \dots, m\}\big)$
  **if** $y_{I_t}\langle w_t, x_{I_t}\rangle < 1$ **then**
    $\theta_{t+1} \leftarrow \theta_t + y_{I_t} x_{I_t}$
  **else**
    $\theta_{t+1} \leftarrow \theta_t$
**return** $\bar{w}_T = \frac{1}{T}\sum_{t=0}^{T-1} w_t$