# Homework 2

Mathilde Bateson
Kernel methods in machine learning
Master Vision Apprentissage

## 1 Dual coordinate ascent algorithms for SVMs

We recall the primal formulation of SVMs seen in class :

$$\min_{f\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i f(x_i)) + \lambda \|f\|_{\mathcal{H}}^2$$

and its dual formulation:

$$\begin{array}{ll} \max_{\alpha\in\mathbb{R}^n} & 2\alpha^\intercal y - \alpha^\intercal K\alpha \\ \text{s.t.} & 0 \leq y_i\alpha_i \leq \frac{1}{2\lambda n} \end{array} \tag{1}$$

### 1.1 One variable update rule

The coordinate ascent method consists of iteratively optimizing with respect to the j-th variable, while fixing the other ones. The objective function $g$ simplifies to a function of one variable:

$$\begin{aligned} g(\alpha) &= 2\alpha^\intercal y - \alpha^\intercal K\alpha \\ &= 2\sum_{\substack{i=1\\i\neq j}}^{n} \alpha_i y_i - \sum_{\substack{i=1\\i\neq j}}^{n}\sum_{\substack{k=1\\k\neq j}}^{n} \alpha_i\alpha_k K(x_i,x_k) + 2\alpha_j\left(y_j - \sum_{\substack{i=1\\i\neq j}}^{n}\alpha_i K(x_i,x_j)\right) - \alpha_j^2 K(x_j,x_j) \end{aligned}$$

g is a simple quadratic function of $\alpha_j$, so it is convex differentiable. The gradient is of the following form:

$$\nabla_{\alpha_j} g(\alpha) = 2\left(y_j - \sum_{\substack{i=1\\i\neq j}}^{n}\alpha_i K(x_i,x_j)\right) - 2\alpha_j K(x_j,x_j)$$

g attains an optimum solution in $\alpha_j$ if and only if :

$$\nabla_{\alpha_j} g(\alpha_j) = 0$$

We thus move to the index j+ 1 without updating $\alpha_j$. Otherwise, we update $\alpha_j$ with the optimal solution of (1):

$$\alpha_j^* = max(min(\frac{y_j - \sum_{\substack{i=1\\i\neq j}}^{n}\alpha_i K(x_i,x_j)}{K(x_j,x_j)}, 0), \frac{1}{2\lambda n})$$

### 1.2 SVM with intercept dual formulation

We now consider now the primal formulation of SVMs with intercept:

$$\min_{f\in\mathcal{H}, b\in\mathcal{R}} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i(f(x_i)+b)) + \lambda\|f\|_{\mathcal{H}}^2 \tag{2}$$

Denoting $\tilde{f} = f + b$, an expansion of the representer theorem is necessary (we don't want to regularize the bias term, so we don't want to regularize on $\tilde{f}$ but on $f$. The so-called Semiparametric Representer Theorem says (in our simple case) that any solution to the optimization problem of minimizing the regularized risk (2) admits a representation of the form:

$$\tilde{f}(\cdot) = \sum_{i=1}^{n} \alpha_i K(\cdot, x_i) + c$$

We introduce slack variables $\xi_i$ to overcome the non-differentiability in zero problem, replacing $\max(0, 1-y_i(f(x_i)+b))$ by $\xi_i$. The minimization problem becomes:

$$\begin{aligned} &\min_{\substack{\alpha\in\mathbb{R}^n \\ \xi\in\mathbb{R}^n}} \frac{1}{n}\sum_{i=1}^{n}\xi_i + \lambda\alpha^\intercal K\alpha \\ \text{s.t.} \quad &\forall i, \quad \xi_i \geq 0 \\ &\forall i, \quad \xi_i \geq 1 - y_i\left(\textstyle\sum_{j=1}^{n}\alpha_j K(x_i,x_j) + b\right) \end{aligned}$$

We introduce the Lagrangian multipliers $\eta \in \mathbb{R}^n$ and $\nu \in \mathbb{R}^n$ corresponding to the Lagrangian:

$$\mathcal{L}(\alpha, b, \xi, \eta, \nu) = \frac{1}{n}\sum_{i=1}^{n}\xi_i + \lambda\alpha^\intercal K\alpha - \sum_{i=1}^{n}\nu_i\xi_i - \sum_{i=1}^{n}\eta_i\left(y_i\left(\sum_{j=1}^{n}\alpha_j K(x_i,x_j)+b\right) - 1 + \xi_i\right) \tag{3}$$

The dual problem writes as follows:

$$\max_{\eta,\nu\in\mathbb{R}^n} \inf_{\alpha,\xi\in\mathbb{R}^n} \mathcal{L}(\alpha, b, \xi, \eta, \nu)$$

We first need to determine the optimal $\alpha$ and $\xi$ in terms of the dual variables. L being convex and differentiable with respect to the primal variables, L is minimized when the gradient is null:

$$\begin{aligned} \frac{\partial L}{\partial \xi_i} = 0 &\Rightarrow \frac{1}{2\lambda n} - \eta_i - \nu_i = 0 \\ &\Rightarrow 0 \leq \eta_i \leq \frac{1}{2\lambda n} \\ \frac{\partial L}{\partial \alpha_i} = 0 &\Rightarrow \alpha_i = y_i\xi_i \end{aligned}$$

Plugging these equations back into (3), the dual problem becomes (denoting $Q = y^T K y$):

$$\begin{aligned} &\max_{\alpha\in\mathbb{R}^n} \alpha^\intercal y - \frac{1}{2}\alpha^\intercal Q\alpha \\ \text{s.t.} \quad &0 \leq \alpha_i \leq \tfrac{1}{2\lambda n} \\ &\textstyle\sum_{i=1}^{n} y_i\alpha_i = 0 \end{aligned} \tag{4}$$

The dual is easier to solve than the primal problem, but there is a new constraint $\sum_{i=1}^{n} y_i\alpha_i = 0$ , hence $\alpha_i$ is exactly determined by other $\alpha_j$, so we can't change only $\alpha_i$ without violating the constraint: we can't use the classical coordinate ascent.

### 1.3 Two variables update rule

Without loss of generality, we suppose that at a given iteration, $\alpha_3 ... \alpha_n$ are fixed while we optimize with respect to $\alpha_1$ and $\alpha_2$. From (4) we require that :

$$y_1\alpha_1 + y_1\alpha_2 = -\sum_{i=3}^{n} y_i\alpha_i = \zeta$$

$$\text{ie} \quad \alpha_1 = (\zeta - y_2\alpha_2)y_1$$

Hence the objective function $g(\alpha_1, ..., \alpha_n) = g((\zeta - y_2\alpha_2)y_1, \alpha_2..., \alpha_n)$ is a quadratic function in $\alpha_2$.
If we ignore the box constraint, we can maximize the quadratic function by setting its gradient to zero. With similar calculus as with question 1.2, we get the following update rule:

$$\alpha_j := \alpha_j - \frac{y_j(E_i - E_j)}{\eta}$$
$$\text{where} \quad \begin{array}{l} E_k = \sum_{i=1}^n \alpha_k K(x_i, x_k) - y_k \\ \eta = 2K(x_i, x_j) - K(x_i, x_i) - K(x_j, x_j) \end{array}$$

Now, considering the box constraint $0 \leq \alpha_j \leq C$ is verified at a certain iteration, we want to find bounds L and H such that $L \leq \alpha_j \leq H$ still holds when updating $\alpha_j$. We easily show that L and H are given by:

$$\text{if } y_i = y_j, L = \max(0, \alpha_j + \alpha_i - C), H = \min(C, \alpha_j + \alpha_i)$$
$$\text{if } y_i \neq y_j, L = \max(0, \alpha_j - \alpha_i), H = \min(C, C + \alpha_j - \alpha_i)$$

Finally we clip $\alpha_j$ to lie within the box constraint:

$$\text{if } \alpha_j \geq H, \alpha_j := H$$
$$\text{if } \alpha_j \leq L, \alpha_j := L$$
$$\text{else } \alpha_j := \alpha_j$$

And we solve for $\alpha_i$ :

$$\alpha_i := \alpha_i + y_i y_j(\alpha_j^{old} - \alpha_j)$$

# 2 Kernel mean embedding

Let us consider a Borel probability measure P of some random variable X on a compact set X . Let $K : X \times X \to \mathbb{R}$ be a continuous, bounded, p.d. kernel and H be its RKHS. The kernel mean embedding of P is defined as the function:

$$\mu(P) : y \to \mathbb{E}_{X\sim P}\left[k(X, y)\right]$$

1.Let $L_P$ be a linear operator defined as $L_P f := \mathbb{E}_{X\sim P}[k(X, y)]$. K is continuous, bounded kernel so using Jensen's inequality, we have that for all f in H:

$$\begin{aligned} |L_P f| &= |\mathbb{E}_{X\sim P}\left[f(x)\right]| \leq \mathbb{E}_{X\sim P}\left[|f(x)|\right] \\ &= \mathbb{E}_{X\sim P}\left[|\langle f, k(X, \cdot)\rangle|\right] \\ &\leq \mathbb{E}_{X\sim P}\left[\|f\|\sqrt{k(X, X)}\right] \end{aligned}$$

Using the Riesz representation theorem there exists $h \in H$ such that $L_P f = \langle f, h\rangle$. If we take $f = K(X, \cdot)$, we have that:

$$h(x) = L_P K(X, \cdot) = \int K(x, x'))dP(x') \Rightarrow h = \mu(P) \in H$$

2.Thus for $f \in H$, $\mathbb{E}_{X\sim P}\left[f(X)\right] = \langle f, \mu_P\rangle_H$. So we immediately get that if P and Q are two Borel probability measures such that $\mu_P = \mu_Q$, then $\mathbb{E}_{X\sim P}\left[f(X)\right] = \langle f, \mu_P\rangle_H = \langle f, \mu_Q\rangle_H = \mathbb{E}_{X\sim Q}\left[f(X)\right]$