# ECE 434: Problem Set 3

Due on November 6, 2023

*Professor Chi Jin*

**Evan Dogariu**
Collaborators: Alex Zhang

# Problem 1

Determine whether the following functions are kernels (i.e. prove that function $k$ is a kernel or provide a counter-example showing that $k$ is not a kernel):

(a) $k(x, y) = \cos(x - y)$ for all $x, y \in \mathbb{R}$

(b) $k(x, y) = \cos \angle(x, y)$ for all $x, y \in \mathbb{R}^d$

## Solution

**Proof.** (a) We expand

$$k(x, y) = \cos(x - y) = \cos(x)\cos(y) + \sin(x)\sin(y)$$

So, letting $f : \mathbb{R} \to \mathbb{R}^2$ be the map sending $z \mapsto [\cos(z),\ \sin(z)]^T$, then we have that

$$k(x, y) = f(x) \cdot f(y),$$

where $\cdot$ denotes the usual Euclidean dot product on $\mathbb{R}^2$, which we know to be a valid kernel. So, by Theorem 1(3) from Lecture 6, we see that $k$ is a kernel.

(b) We expand

$$k(x, y) = \cos \angle(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$$

Letting $f : \mathbb{R}^d \to \mathbb{R}^d$ be the map sending $z \mapsto \frac{z}{\|z\|}$, we therefore have that

$$k(x, y) = f(x) \cdot f(y),$$

where $\cdot$ denotes the usual Euclidean dot product on $\mathbb{R}^d$, which we know to be a valid kernel. Again by Theorem 1(3) from Lecture 6, we get that $k$ is a kernel. ■

# Problem 2

Linear Support Vector Machine (SVM) finds the maximum margin linear classifier for binary classification problems. In this question, we will kernelize this algorithm. We first formally introduce linear SVM:

Consider a binary classification problem with label set $Y = \{-1, 1\}$, linear function class $\mathcal{F} := \{f(x) = w^\top x \mid w \in \mathbb{R}^d\}$, and the classification rule $y = \text{sgn}(f(x))$ which assigns value 1 if $f(x) \geq 0$, and assigns value $-1$ otherwise. Linear SVM can be formulated as optimizing the hinge loss with $\ell_2$ regularization:

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} [1 - y_i w^\top x_i]_+ + \lambda \|w\|_2^2$$

where $[z]_+ := \max\{0, z\}$ for any $z$.

(a) For a given kernel $k : X \times X \to \mathbb{R}$, write the kernelized version of linear SVM as an optimization problem over functions in the Reproducing Kernel Hilbert Space (RKHS).

(b) Argue why the RKHS optimization problem obtained in (a) can be reduced to an optimization problem over vectors in $\mathbb{R}^n$, and write down the latter optimization problem.

## Solution

**Proof.** (a) Let $k$ be the given kernel. By Theorem 2 from Lecture 4, there is a Hilbert space $\mathcal{H}$ and a feature map $\phi : X \to \mathcal{H}$ such that $k(x_1, x_2) = \langle \phi(x_1), \phi(x_2) \rangle_{\mathcal{H}}$ for all $x_1, x_2 \in X$. Kernelizing the linear SVM objective, we seek

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} [1 - y_i k(w, x)]_+ + \lambda k(w, w)$$

where we have replaced standard Euclidean dot products with the kernel. By the RKHS property, we therefore seek

$$\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} [1 - y_i \langle \phi(w), \phi(x_i) \rangle_{\mathcal{H}}]_+ + \lambda \|\phi(w)\|_{\mathcal{H}}^2$$

As was done in Section 1.2 of Lecture 6, we may relate this to an optimization over $\mathcal{H}$. In particular, we seek

$$\min_{f \in \mathcal{H}} \sum_{i=1}^{n} [1 - y_i \langle f, \phi(x_i) \rangle_{\mathcal{H}}]_+ + \lambda \|f\|_{\mathcal{H}}^2$$

Lastly, we observe that since $\phi(x_i) = k(\cdot, x_i)$ by construction, the reproducing property of $\mathcal{H}$ gives that $\langle f, \phi(x_i) \rangle_{\mathcal{H}} = \langle f, k(\cdot, x_i) \rangle_{\mathcal{H}} = f(x_i)$. In total, we seek

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

(b) We will reduce the above optimization problem to one over vectors in $\mathbb{R}^n$ using the Representer Theorem. To do so, define $h : \mathbb{R} \to \mathbb{R}$ via $h(z) := \lambda z$. Then, since $\lambda > 0$, $h$ is strictly increasing. If we define $L : \mathbb{R}^n \to \mathbb{R}$ as

$$L(z_1, \ldots, z_n) := \sum_{i=1}^{n} [1 - y_i z_i]_+,$$

then the above objective becomes trying to minimize

$$\min_{f \in \mathcal{H}} L(f(x_1), \ldots, f(x_n)) + h(\|f\|_{\mathcal{H}}^2)$$

Problem 2 continued on next page... 

By the Representer Theorem (Theorem 2 in Lecture 6), we see that all solutions of this optimization problem are of the form $f^*(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ for some coefficients $\alpha_i$. Thus, to minimize this expression we simply must find the $\alpha \equiv (\alpha_1, \dots, \alpha_n) \in \mathbb{R}^n$ that minimizes it. To be precise, letting $f_\alpha := \sum_{i=1}^n \alpha_i k(\cdot, x_i) = \sum_{i=1}^n \alpha_i \phi(x_i)$ for each $\alpha$, we seek

$$\min_{\alpha \in \mathbb{R}^n} \sum_{i=1}^n [1 - y_i f_\alpha(x_i)]_+ + \lambda \|f_\alpha\|_{\mathcal{H}}^2,$$

which is an optimization problem over $\mathbb{R}^n$ as desired. ■

# Problem 3

Consider again the binary classification setting with label set $Y = \{-1, 1\}$, and arbitrary function class $\mathcal{F} \subseteq \{f : X \to \mathbb{R}\}$. Let the margin loss $\ell_\rho(f(x), y) = \Phi_\rho(yf(x))$ where $\rho > 0$, and

$$\Phi_\rho(z) = \begin{cases} 0 & \text{if } \rho \leq z \\ 1 - \frac{z}{\rho} & \text{if } 0 \leq z \leq \rho \\ 1 & \text{if } z \leq 0 \end{cases}$$

We note this margin loss can be viewed as an upper bound (and a smoothed version) of 0-1 loss. Let $r(f)$ denote the population risk of classifier $\text{sgn}(f(\cdot))$ under 0-1 loss. Let $r_\rho(f)$ and $\hat{r}_\rho(f)$ denote the population risk and the empirical risk of $f$ under margin loss $\ell_\rho$.

(a) Prove that for any function $f \in \mathcal{F}$, $r(f) \leq r_\rho(f)$.

(b) Prove that with probability at least $1 - \delta$, for any $f \in \mathcal{F}$ we have

$$r_\rho(f) \leq \hat{r}_\rho(f) + \frac{4}{\rho}\mathcal{R}_n(\mathcal{F}) + c\sqrt{\frac{1 + \log(1/\delta)}{n}}$$

(c) Suppose $\hat{f}$ is the solution of linear SVM problem from Problem 2 with an additional hard constraint $\|w\|_2 \leq R$. Suppose domain $X$ satisfies the condition that $\sup_{x \in X} \|x\|_2 \leq D$. Provide a margin-based generalization bound for $\hat{f}$ (i.e. a bound of the form $r(\hat{f}) \leq \hat{r}_\rho(\hat{f}) + \text{error terms}$).

## Solution

**Proof.** (a) Let $f \in \mathcal{F}$. Then, we have that for all $(x, y) \in X \times Y$,

$$\ell(f(x), y) \leq \ell_\rho(f(x), y)$$

since the 0-1 loss $\ell$ is upper bounded by $\ell_\rho$ (they agree for $z \leq 0$ and for $z \geq \rho$ but $0 \equiv \ell \leq \ell_\rho$ over $[0, \rho]$). Thus, in expectation over $(x, y) \sim \mathcal{D}$,

$$r(f) \equiv \mathbb{E}_{(x,y)\sim\mathcal{D}}\ell(f(x), y) \leq \mathbb{E}_{(x,y)\sim\mathcal{D}}\ell_\rho(f(x), y) \equiv r_\rho(f)$$

as desired.

(b) Define the function classes $\mathcal{G} := \{(x, y) \mapsto \ell_\rho(f(x), y) : \ f \in \mathcal{F}\}$ and $\mathcal{H} := \{(x, y) \mapsto yf(x) : \ f \in \mathcal{F}\}$ for convenience. Then, by definition of $\ell_\rho$ we may represent $\mathcal{G}$ via

$$\mathcal{G} = \{(x, y) \mapsto (\Phi_\rho \circ h)(x, y) : \ h \in \mathcal{H}\} \implies \mathcal{G} = \Phi_\rho \circ \mathcal{H}$$

Note that, letting $\Pi_{[a,b]} : \mathbb{R} \to \mathbb{R}$ be the projection of a point onto the interval $[a, b]$ (i.e. clipping), then

$$\Phi_\rho(z) \equiv 1 - \frac{1}{\rho}\Pi_{[0,\rho]}(z)$$

So, by Theorem 1(5) and 1(3) from Lecture 4,

$$\mathcal{R}_n(\mathcal{G}) \equiv \mathcal{R}_n\left(1 - \frac{1}{\rho}\Pi_{0,\rho} \circ \mathcal{H}\right) \leq \frac{1}{\sqrt{n}} + \frac{1}{\rho}\mathcal{R}_n(\Pi_{[0,\rho]} \circ \mathcal{H})$$

Since projection operators are 1-Lipschitz and this projection operator maps 0 to 0, we may apply Theorem 1(4) to see that

$$\mathcal{R}_n(\mathcal{G}) \leq \frac{1}{\sqrt{n}} + \frac{2}{\rho}\mathcal{R}_n(\mathcal{H})$$

Problem 3 continued on next page...

By Equation 3 in Lecture 4, we know that $\mathcal{R}_n(\mathcal{H}) = \mathcal{R}_n(\mathcal{F})$ (since multiplying by labels $y \in \{-1, 1\}$ doesn't change the distribution of Rademacher variables). As such,

$$\mathcal{R}_n(\mathcal{G}) \leq \frac{1}{\sqrt{n}} + \frac{2}{\rho}\mathcal{R}_n(\mathcal{F})$$

Now, by combining Proposition 1 and Theorem 1 from Lecture 2, we see that with probability $\geq 1 - \delta$,

$$r_\rho(f) - \hat{r}_\rho(f) \leq 2\mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Thus,

$$r_\rho(f) - \hat{r}_\rho(f) \leq \frac{4}{\rho}\mathcal{R}_n(\mathcal{F}) + \frac{2}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{2n}}$$

Letting $a := \log(1/\delta) > 1$ for notation, we observe that

$$\frac{2}{\sqrt{n}} + \sqrt{\frac{a}{2n}} = \sqrt{\left(\frac{2}{\sqrt{n}} + \sqrt{\frac{a}{2n}}\right)^2} = \sqrt{\frac{4}{n} + \frac{a}{2n} + 2\sqrt{\frac{4a}{2n^2}}} = \sqrt{\frac{4 + 0.5a}{n} + \frac{2\sqrt{2}}{n}\sqrt{a}}$$

Now, we know that $a \geq 1 \implies \sqrt{a} \leq a \implies 4+0.5a+2\sqrt{2}\sqrt{a} \leq 4+0.5a+2\sqrt{2}a$. Since $0.5+2\sqrt{2} \approx 3.33 \leq 4$, we find that

$$\frac{2}{\sqrt{n}} + \sqrt{\frac{a}{2n}} \leq \sqrt{\frac{4 + 4a}{n}} = 2\sqrt{\frac{1 + a}{n}}$$

Plugging this algebraic detour back into our original expression,

$$r_\rho(f) - \hat{r}_\rho(f) \leq \frac{4}{\rho}\mathcal{R}_n(\mathcal{F}) + 2\sqrt{\frac{1 + \log(1/\delta)}{n}},$$

which is the desired inequality.

(c) Let $\mathcal{F}_{\text{lin}}$ be the function class given by linear SVMs with bounded $w$'s over bounded inputs. This is precisely the function class handled in Section 1.2.1 of Lecture 5, and so we may use the Rademacher complexity bound derived there (note that even though the problem isn't regression as it was in the notes, the function class $\mathcal{F}_{\text{lin}}$ is identical and therefore so is the complexity). Thus, by part (b) above,

$$\mathcal{R}_n(\mathcal{F}_{\text{lin}}) \leq \frac{RD}{\sqrt{n}} \implies r_\rho(\hat{f}) - \hat{r}_\rho(\hat{f}) \leq \frac{4RD}{\rho\sqrt{n}} + 2\sqrt{\frac{1 + \log(1/\delta)}{n}}$$

As $\frac{1}{\sqrt{n}} \leq \sqrt{\frac{1+\log(1/\delta)}{n}}$ clearly (since $\delta < 1 \implies 1/\delta > 1 \implies \log(1/\delta) > 0$), we see that

$$r_\rho(\hat{f}) - \hat{r}_\rho(\hat{f}) \leq \left(\frac{4RD}{\rho} + 2\right)\sqrt{\frac{1 + \log(1/\delta)}{n}}$$

Lastly, since $r(\hat{f}) \leq r_\rho(\hat{f})$ by part (a), we combine everything to get that

$$r(\hat{f}) \leq \hat{r}_\rho(\hat{f}) + \left(\frac{4RD}{\rho} + 2\right)\sqrt{\frac{1 + \log(1/\delta)}{n}},$$

which is a generalization bound of the desired form. ■

# Problem 4

Prove that the following operations preserve the convexity of the functions. For simplicity, you can always assume the domain of the function is $\mathbb{R}^d$.

(a) If $f_1, \ldots, f_n$ are convex functions, $\alpha_1, \ldots, \alpha_n$ are nonnegative scalars, show that $f := \sum_{i=1}^n \alpha_i f_i$ is also a convex function.

(b) If $f_\theta$ is a convex function for all $\theta \in \Theta$, show that $f := \sup_{\theta \in \Theta} f_\theta$ is also a convex function.

(c) If $g$ is a convex function, show that for an arbitrary matrix $A \in \mathbb{R}^{d \times d}$, and vector $b \in \mathbb{R}^d$, the function $f(x) := g(Ax + b)$ is also convex.

## Solution

**Proof.** (a) Let $t \in (0, 1)$ and $x, y \in \mathbb{R}^d$ be arbitrary. We wish to show that

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

To this end, we compute

$$f(tx + (1-t)y) = \sum_{i=1}^n \alpha_i f_i(tx + (1-t)y)$$

Since each $f_i$ is convex, we know that $f_i(tx + (1-t)y) \leq tf_i(x) + (1-t)f_i(y)$, and so

$$f(tx+(1-t)y) \leq \sum_{i=1}^n \alpha_i(tf_i(x)+(1-t)f_i(y)) = t\left(\sum_{i=1}^n \alpha_i f_i(x)\right)+(1-t)\left(\sum_{i=1}^n \alpha_i f_i(y)\right) = tf(x)+(1-t)f(y)$$

So, $f$ is convex since this holds for all choices of $t, x, y$.

(b) Let $t \in (0, 1)$ and $x, y \in \mathbb{R}^d$ be arbitrary. Note that for all $\theta \in \Theta$, $f_\theta \leq f$ by definition. So, for every $\theta \in \Theta$ we have

$$f_\theta(tx + (1-t)y) \leq tf_\theta(x) + (1-t)f_\theta(y) \leq tf(x) + (1-t)f(y),$$

where the first inequality follows by convexity of $f_\theta$. Since the above bound holds for all $\theta$, it will hold in supremum. So,

$$f(tx + (1-t)y) \equiv \sup_{\theta \in \Theta} \{f_\theta(tx + (1-t)y)\} \leq tf(x) + (1-t)f(y)$$

Therefore, $f$ is convex since this holds for all choices of $t, x, y$.

(c) Let $A \in \mathbb{R}^{d \times d}$ and $b \in \mathbb{R}^d$ be arbitrary, and define $f(x) := g(Ax + b)$. Let $t \in (0, 1)$ and $x, y \in \mathbb{R}^d$ be arbitrary. Then,

$$f(tx + (1-t)y) = g(A(tx + (1-t)y) + b) = g(tAx + (1-t)Ay + b)$$

Since we may write $b = tb + (1-t)b$ for free, this equals

$$= g(tAx + tb + (1-t)Ay + (1-t)b) = g(t(Ax + b) + (1-t)(Ay + b))$$

By convexity of $g$ and the definition of $f$,

$$g(t(Ax + b) + (1-t)(Ay + b)) \leq tg(Ax + b) + (1-t)g(Ay + b) = tf(x) + (1-t)f(y)$$

Combining everything,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$$

So, $f$ is convex since this holds for all choices of $t, x, y$. ■