CPSC 532S: Assignment 4 – due Friday, 8 Apr 2022, 11:59pm

This late policy is extra-extra-generous because of end-of-term: -1 point on the assignment per day you're late, **not including Saturday/Sunday** so that handing in Monday is only -1, up to a hard deadline of Friday April 15th (and not accepted after that).

This assignment is split into four questions (the parts with big section headers; most have sub-parts). You can solve each question in groups of up to three. Groups don't need to be consistent between problems; you can do Q1 and Q2 alone, Q3 with Alice, and Q4 with Bob and Carlos if you want.

Please **do not** just split the questions up and do the parts separately. **If your name is on a solution, you are pledging that you contributed significantly to the solution and understand it fully.**

There is a separate Gradescope assignment for each problem; use the Gradescope groups feature to submit once and associate with each of you, but also put all of your names on the first page as a backup.

Prepare your answers to these questions **using LaTeX**. Hopefully you're reasonably familiar with it, but if not, try using Overleaf and looking around for tutorials online. (Note that free Overleaf accounts can only share with one "named collaborator," but you can collaborate with more people by sending them an edit link. Make sure you only share the parts of the homework you're handing in together!)

Feel free to ask questions if you get stuck on things on Piazza (but remove any details about the actual answers... feel free to make a private post if that's tough). If you look stuff up anywhere other than in the slides, one of the two course textbooks, or the Telgarsky notes, please **cite your sources**: just say in the answer to that question where you looked. (A link is fine, no need for a formal citation.) Please do not look at solution manuals or so on. If you accidentally come across a solution while looking for something related, still write the argument up in your own words, link to wherever you found it, and be clear about what happened.

If you like, the `.tex` source for this file is available on the course website, and you can put your answers in `\begin{answer} My answer here... \end{answer}` environments to make them stand out if so; feel free to delete whatever boilerplate you want (or not, I'm not printing them out). Or answer in a fresh document; just make it clear which question you're answering where.

If you're using a consistent group and want to write your answers in one document, you could split the PDF with e.g. `qpdf a2.pdf --pages . 2-3 -- q1.pdf` or through the GUI of a PDF viewer. Or you can upload the full file four times and just make sure you assign pages appropriately.

Submit your answers as a PDF on Gradescope: instructions on Piazza. You'll be prompted to mark where each sub-part is in your PDF; make sure you mark all relevant pages for each part. (This saves me a surprising amount of time in grading.) If something goes wrong, you can also email your assignment to me directly (`dsuth@cs.ubc.ca`).

1 Proving kerneldom [25 points + 5 bonus points]

Prove that the following functions are kernels, i.e. that they are positive semi-definite functions.

Hint: Recall that you can do so by directly proving all kernel matrices are psd, by writing an explicit feature mapping $k(x, x') = \langle \phi(x), \phi(x') \rangle$ where ϕ maps into any Hilbert space (including $\mathbb{R}^d$), or by using steps known to produce new kernels out of old ones as in lecture 15 slides "Building kernels from other kernels" through "Some more ways to build kernels." You could also use Bochner's theorem, which we did not cover in class, if you're a Fourier buff: a kernel $k(x, y) = \psi(x - y)$ with $\psi(0) = 1$ is psd iff it is the Fourier transform of a probability measure.

Hint: Here are two Hilbert spaces that might be useful to you. First, the space ℓ^2 of square-summable sequences $(a_k)_{k=1}^{\infty}$ with inner product $\langle (a_k), (b_k) \rangle_{\ell^2} = \sum_k a_k b_k$. Second, the space $L^2(\mathcal{X})$ of square-integrable functions[1] on $\mathcal{X}$, with inner product $\langle f, g \rangle_{L^2} = \int_{\mathcal{X}} f(x) g(x) \mathrm{d}x$.

(a) [5 points] $k(x, y) = \cos(x - y)$ on $\mathbb{R}$.

Hint: The list of trigonometric identities makes for good bedtime reading.

(b) [5 points] $k_n(x, y) = \frac{1}{2\pi} \left[1 + 2 \sum_{k=1}^{n} \cos(k(x - y))\right] = \frac{\sin\left(\left(n + \frac{1}{2}\right)(x-y)\right)}{2\pi \sin((x-y)/2)}$ on $\mathbb{R}$ for any $n \geq 0$.

(This is called the *Dirichlet kernel*, because it is a continuous kernel which converges to the Dirichlet delta function $\delta(x - y)$ as $n \to \infty$.)

(c) [5 points] $k(x, y) = \min(x, y)$ on $[0, 1]$.

Hint: You could consider the integral $\int_{\mathbb{R}} \mathbb{1}(t \in [0, x]) \mathbb{1}(t \in [0, y]) \mathrm{d}t$.

(d) [5 points] $k(X, Y) = \sum_{x \in X} \sum_{y \in Y} k_0(x, y)$ on finite sets with elements in $\mathcal{X}$, where k_0 is a kernel on $\mathcal{X}$.

(e) [5 points] $k(x, y) = 1/\sqrt{1 - xy}$ on $(-1, 1)$.

For [5 bonus points], you can instead show $1/\sqrt{1 - x^\mathsf{T} y}$ is psd on $\{x \in \mathbb{R}^d : \|x\| < 1\}$.

Hint: It might help to use the following expansion (see e.g. here), which converges for $|z| < 1$:

$$\frac{1}{\sqrt{1 - z}} = \sum_{k=0}^{\infty} c_k z^k \quad \text{for} \quad c_k := \frac{1}{2^{2k}} \binom{2k}{k}.$$

[1] Really, this should be a space of equivalence classes of functions, since a function that's zero only almost everywhere will have norm zero. That won't matter for this question.

2 Maximizing differences [25 points]

Let's consider learning a kernel classifier with the somewhat unusual *linear loss*, $\ell(h,(x,y)) = -yh(x)$, where $y \in \{-1,1\}$. Take the kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with associated RKHS $\mathcal{H}_k$.

(a) [10 points] Find the regularized loss minimizer

$$\hat{h}_\lambda = \operatorname*{arg\,min}_{h \in \mathcal{H}_k} L_S(h) + \tfrac{1}{2}\lambda \|h\|^2_{\mathcal{H}_k}, \tag{RLM}$$

for a training sample $S = ((x_1, y_1), \ldots, (x_n, y_n))$.

(b) [5 points] Show that $L_S(\hat{h}_\lambda) = -\frac{1}{\lambda}\left\| \frac{1}{n}\sum_{i:y_i=1} k(x_i,\cdot) - \frac{1}{n}\sum_{i:y_i=-1} k(x_i,\cdot)\right\|^2_{\mathcal{H}_k}$.

(c) [5 points] Find a (data-dependent) value of λ, call it $\hat{\lambda}$, such that $\|\hat{h}_{\hat{\lambda}}\|_{\mathcal{H}_k} = 1$, and simplify the expression for $L_S(\hat{h}_{\hat{\lambda}})$.

(d) [5 points] Argue that $\hat{h}_{\hat{\lambda}}$ is a solution to

$$\min_{h \in \mathcal{H}_k : \|h\| \le 1} L_S(h). \tag{ERM}$$

Further argue that solving (ERM) is equivalent to solving

$$\max_{h \in \mathcal{H}_k : \|h\| \le 1} \sum_{i:y_i=1} h(x_i) - \sum_{i:y_i=-1} h(x_i), \tag{MAX}$$

i.e. finding a function high on the positively-labeled points and low on the negatively-labeled ones.

3 One way to do semi-supervised learning [25 points]

Semi-supervised learning is when you're given not only a training set of (x, y) pairs, but also a set of unlabeled x samples from the marginal distribution of x. (For instance, maybe you have a really big dataset scraped from the web, and have only paid for human annotation of a small, random selection from it.) Even though there aren't any labels, this can be useful for determining the optimal decision function under some reasonable assumptions: for instance, if you have a clear cluster structure, it's perhaps more likely that the labeling function is constant on that cluster.

One way to try to implement this is to penalize the *gradient norm* of the decision function, evaluated at the data points – the decision function should be smooth where there's data. (You might be familiar with this type of gradient penalty from GANs.) It turns out that the special structure of RKHSes will allow for this. Specifically, let's let $\mathcal{H}_k$ be an RKHS for some twice-differentiable kernel k on $\mathbb{R}$, i.e. $f \in \mathcal{H}_k$ maps $\mathbb{R}$ to $\mathbb{R}$. (Everything here will work for $\mathbb{R}^d$, the notation just gets a little messier.)

Our goal will be to minimize the following regularized loss over all of $\mathcal{H}_k$, where both ν and λ are positive scalars:

$$J(h) = \frac{1}{n}\sum_{i=1}^{n}\big(h(x_i) - y_i\big)^2 + \frac{\nu}{m}\sum_{i=1}^{m}\big(h'(x_i)\big)^2 + \lambda\|h\|_{\mathcal{H}_k}^2. \tag{J}$$

Here we assume that we have our usual sample set $\big((x_1, y_1), \ldots, (x_n, y_n)\big)$, but we *also* have an unlabeled sequence $\big(x_{n+1}, \ldots, x_m\big)$, so that we have m total samples for x (of which the first n are labeled).

Kernels are two-argument functions, so differentiation notation can be slightly awkward. For brevity, we will use ∂_1 to refer to differentiating with respect to the first argument and ∂_2 the second, so that $\partial_1 k(x, y)$ means $\frac{\partial}{\partial x}k(x, y)$, $\partial_2^2 k(x, y)$ means $\frac{\partial^2}{\partial y^2}k(x, y)$, and $\partial_1\partial_2 k(x, x)$ means $\frac{\partial^2}{\partial z_1 \partial z_2}k(z_1, z_2)\big|_{\substack{z_1=x\\z_2=x}}$ – note that differentiation happens "before" passing the arguments in (this is *not* $\frac{\partial^2}{\partial x^2}k(x, x)$).

The following result will be useful for us:

Lemma 3.1 (Special case of Steinwart/Christmann Lemma 4.34)**.** *Let $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a kernel such that both $\frac{\partial}{\partial x}\frac{\partial}{\partial y}k(x, y)$ and $\frac{\partial^2}{\partial x^2}\frac{\partial^2}{\partial y^2}k(x, y)$ exist and are continuous. Then, for all $x \in \mathbb{R}$, $\partial_1 k(x, \cdot)$ and $\partial_1^2 k(x, \cdot)$ are functions in $\mathcal{H}_k$ such that for all $f \in \mathcal{H}_k$ we have*

$$\langle \partial_1 k(x, \cdot), f\rangle_{\mathcal{H}_k} = f'(x) \quad \text{and} \quad \langle \partial_1^2 k(x, \cdot), f\rangle_{\mathcal{H}_k} = f''(x).$$

For example, this also means that

$$\langle \partial_1 k(x, \cdot), \partial_1 k(x', \cdot)\rangle_{\mathcal{H}_k} = \partial_1\partial_2 k(x, x') \quad \text{and} \quad \langle \partial_1^2 k(x, \cdot), k(x', \cdot)\rangle_{\mathcal{H}_k} = \partial_1^2 k(x, x').$$

(a) [10 points] Show a representer theorem for $\arg\min_{h\in\mathcal{H}} J(h)$, i.e. that you can write the optimal h as a linear combination of some set of vectors in $\mathcal{H}$.

Hint: The representer theorem we showed in class (lecture 16, starting around page 18 – sorry that slides aren't numbered in that lecture...) will not *directly apply, because J depends on the derivatives of h. You'll need to make an analogous argument, taking advantage of the lemma above.*

Define the following matrices:

$$K = \begin{bmatrix} k(x_1, x_1) & \dots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \dots & k(x_n, x_n) \end{bmatrix} \in \mathbb{R}^{n \times n}$$

$$G = \begin{bmatrix} \partial_1 k(x_1, x_1) & \dots & \partial_1 k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \partial_1 k(x_m, x_1) & \dots & \partial_1 k(x_m, x_n) \end{bmatrix} \in \mathbb{R}^{m \times n}$$

$$H = \begin{bmatrix} \partial_1 \partial_2 k(x_1, x_1) & \dots & \partial_1 \partial_2 k(x_1, x_m) \\ \vdots & \ddots & \vdots \\ \partial_1 \partial_2 k(x_m, x_1) & \dots & \partial_1 \partial_2 k(x_m, x_m) \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

(b) [15 points] Write an explicit form for $J(h)$ in terms of usual matrix and vector operations on the K, G, and H matrices and the vector $y \in \mathbb{R}^n$ of labels, as well as the parameters of your linear combination.

Hint: It will probably help to start by writing out $h(x_i)$, $h'(x_i)$, and $\|h\|^2_{\mathcal{H}_k}$, then plugging those together. It's might be helpful in intermediate steps to use the standard basis vectors e_i, which have a one in the ith entry and zero in all others. Be careful about shapes matching.

If you did it right, the final form for J should be a quadratic form of your coefficients in terms of the matrices K, G, and H. Thus, setting the gradient to zero will give an analytical solution written as the solution to a certain linear system, although I don't need you to write out that system since it's a little messy.

4 Tangent kernels [25 points]

Recall that the empirical tangent kernel for a general model $h(x; w)$ is given by

$$k_w(x, x') = \left[\nabla_w h(x; w)\right]^\mathsf{T} \left[\nabla_w h(x'; w)\right],$$

where here we're thinking of $\nabla_w h(x; w)$ as a *column* vector, so that $k_w(x, x') \in \mathbb{R}$.

For randomly initialized $w \sim \mathbb{Q}$, in certain regimes k_w will converge as the model gets bigger to its expectation

$$k_{\mathbb{Q}}(x, x') = \mathop{\mathbb{E}}_{w \sim \mathbb{Q}} k_w(x, x').$$

(a) [10 points] Explicitly write out the empirical tangent kernel k_w for a linear model $h(x; w) = w^\mathsf{T} x$, and its expected kernel $k_{\mathbb{Q}}$ when $w \sim \mathcal{N}(0, I)$, the standard normal distribution. Does kernel regression with $k_{\mathbb{Q}}$ agree with ERM for h under the square loss (linear regression)?

(b) [15 points] Explicitly write out the empirical tangent kernel k_w for a (not-very-)deep linear model $f(x; V, W) = v^\mathsf{T} W x$, where $v \in \mathbb{R}^m$ and $W \in \mathbb{R}^{m \times d}$; you can think of a vector $w \in \mathbb{R}^{m(d+1)}$ as stacking up all the entries of W and v. Assume that $\mathbb{Q}$ has $w \sim \mathcal{N}(0, I)$, and also write out $k_{\mathbb{Q}}$. Does kernel regression with $k_{\mathbb{Q}}$ agree with the ERM of f with square loss?

Hint: `http://matrixcalculus.org/` is handy to check that you're taking your derivatives of linear algebra expressions correctly, or The Matrix Cookbook is what I used in grad school, back before computers could just solve all our problems for us. But make sure that you handle the shapes correctly – in this case, k_w should be scalar-valued.

Hint: Gee, that Wishart distribution sure is neat, huh? Nothing to do with anything in particular, just thought I'd mention.