1. (Problem 8.15 Haykin) Let $\bar{k}_{ij}$ denote the centered counterpart of the ij-th element k_{ij} of the Gram **K**. Derive the following formula (Schölkopf, 1977):

$$\bar{k}_{ij}=k_{ij}-\frac{1}{N}\sum_{m=1}^{N}\mathbf{\Phi}^T(\mathbf{x}_m)\mathbf{\Phi}(\mathbf{x}_j)-\frac{1}{N}\sum_{n=1}^{N}\mathbf{\Phi}^T(\mathbf{x}_i)\mathbf{\Phi}(\mathbf{x}_n)+\frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}\mathbf{\Phi}^T(\mathbf{x}_m)\mathbf{\Phi}(\mathbf{x}_n) \quad .$$

Suggest a compact representation of this relation in matrix form.

I'll answer the second question first. Consider this compact representation of the relation:

$$\bar{k}_{ij}=\left(\Phi^T(\mathbf{x}_i)-\bar{\Phi}^T\right)\left(\Phi(\mathbf{x}_j)-\bar{\Phi}\right)=\left(\Phi^T(\mathbf{x}_i)-\frac{1}{N}\sum_{m=1}^{N}\Phi^T(\mathrm{x}_m)\right)\left(\Phi(\mathbf{x}_j)-\frac{1}{N}\sum_{n=1}^{N}\Phi(\mathbf{x}_n)\right)$$

$$=\Phi^T(\mathbf{x}_i)\Phi(\mathbf{x}_j)-\frac{1}{N}\sum_{m=1}^{N}\Phi^T(\mathrm{x}_m)\Phi(\mathbf{x}_j)-\frac{1}{N}\Phi^T(\mathbf{x}_i)\sum_{n=1}^{N}\Phi(\mathbf{x}_n)-\frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}\Phi^T(\mathrm{x}_m)\Phi(\mathbf{x}_n)$$

$$=k_{ij}-\frac{1}{N}\sum_{m=1}^{N}\Phi^T(\mathrm{x}_m)\Phi(\mathbf{x}_j)-\frac{1}{N}\sum_{n=1}^{N}\Phi^T(\mathbf{x}_i)\Phi(\mathbf{x}_n)+\frac{1}{N^2}\sum_{m=1}^{N}\sum_{n=1}^{N}\Phi^T(\mathrm{x}_m)\Phi(\mathbf{x}_n).$$

2. (Problem 8.16 Haykin) Show that the normalization of eigenvector $\boldsymbol{\alpha}$ of the Gram **K** is equivalent to the requirement that Eq. (8.109) be satisfied.

This problem is a relatively straightforward algebraic task. Equation (8.100) is the key. From this we get

$$1=\tilde{\mathbf{q}}^T\mathbf{q}=\left(\sum_{i=1}^{N}\alpha_i\boldsymbol{\phi}(\mathbf{x}_i)\right)^T\left(\sum_{j=1}^{N}\alpha_i\boldsymbol{\phi}(\mathbf{x}_j)\right)=\sum_{i=1}^{N}\sum_{j=1}^{N}\boldsymbol{\phi}^T(\mathbf{x}_i)\alpha_i\alpha_j\boldsymbol{\phi}(\mathbf{x}_j)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j\boldsymbol{\phi}^T(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_j)$$

$$=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j k(\mathbf{x}_i,\mathbf{x}_j)=\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j N\tilde{\lambda}[8.101]=\boldsymbol{\alpha}^T\boldsymbol{\alpha}\lambda$$

. So, since $\boldsymbol{\alpha}_r^T\boldsymbol{\alpha}_r\lambda_r=1$, $\boldsymbol{\alpha}_r^T\boldsymbol{\alpha}_r=\frac{1}{\lambda_r}$.

3. (Problem 9.8 Haykin) Using the transformation formula of Eq. (9.40) applied to Eq. (9.37), derive the probability density function of Eq. (9.41). [Note: I believe Haykin's Eq. 9.41 is missing a $\frac{1}{\sigma}$ term. Please verify or disprove this.]

   The X^2 probability density function is defined by

$$p_{X^2}(x)=\frac{x^{\frac{m}{2}-1}\exp(-\frac{x}{2})}{2^{\frac{m}{2}}\Gamma(\frac{m}{2})} \quad .$$

   To convert this to a density function with standard deviation σ , Haykin makes the substitution $\frac{v}{\sigma^2}=x$ (although I would think that substitution $\frac{v}{\sigma}=x$ is more appropriate, that is, dividing by the standard deviation). The substitution $v=\sigma^2 x$ gives us $\frac{\partial v}{\partial x}=\sigma^2$ and

$$p_V(v)=\frac{p_X\left(\frac{v}{\sigma^2}\right)}{\left|\frac{\partial v}{\partial x}\right|}=\frac{\left(\frac{v}{\sigma^2}\right)^{\frac{m}{2}-1}\exp\left(\frac{-v}{2\sigma^2}\right)}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}\frac{1}{\sigma^2}=\frac{v^{\frac{m}{2}-1}\exp\left(\frac{-v}{2\sigma^2}\right)}{(\sigma^2)^{\frac{m}{2}-1}\sigma^2 2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}$$

$$=\frac{v^{\frac{m}{2}-1}\exp\left(-\frac{v}{2\sigma^2}\right)}{\sigma^m 2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}$$ which is exactly Eqn. (9.37).

If we now apply the change of variable $r=v^{\frac{1}{2}}$ we'll see where things go a little bit wrong.

Since $\frac{\partial r}{\partial v}=\frac{1}{2}v^{\frac{-1}{2}}=\frac{1}{2r}$, changing variables gives us

$$p_R(r)=\frac{p_V(r^2)}{\left|\frac{\partial r}{\partial v}\right|}=\frac{(r^2)^{\frac{m}{2}-1}\exp\left(-\frac{r^2}{2\sigma^2}\right)}{\sigma^m 2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}2\mathrm{r}$$

$$=\frac{2r^{m-2}r\exp\left(\frac{-r^2}{2\sigma^2}\right)}{\sigma^m 2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}=\frac{2r^{m-1}\exp\left(\frac{-r^2}{2\sigma^2}\right)}{\sigma^m 2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}=\frac{1}{2^{(m/2-1)}\Gamma(m/2)}\left(\frac{r}{\sigma}\right)^{m-1}\exp\left(\frac{-r^2}{2\sigma^2}\right)\left(\frac{1}{\sigma}\right)$$. But for the extra $\frac{1}{\sigma}$ term, this is identical to Eqn. (9.41) .

Clearly there's a problem here.

If, as Van Hulle does in his original paper, we make a single substitution of $\frac{r}{\sigma}=x^{\frac{1}{2}}$ or $r=\sigma x^{\frac{1}{2}}$, yielding $\frac{\partial r}{\partial x}=\frac{\sigma x^{-\frac{1}{2}}}{2}=\frac{\sigma^2}{2r}$, then we get

$$p_R(r)=\frac{p_X\left(\left(\frac{r}{\sigma}\right)^2\right)}{\left|\frac{\partial r}{\partial x}\right|}=\frac{\left(\left(\frac{r}{\sigma}\right)^2\right)^{\frac{m}{2}-1}\exp\left(\frac{-\left(\frac{r}{\sigma}\right)^2}{2}\right)}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}\frac{2r}{\sigma^2}=\frac{2\left(\frac{r}{\sigma}\right)^{m-2}\frac{r}{\sigma}\exp\left(\frac{-\left(\frac{r}{\sigma}\right)^2}{2}\right)}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}\left(\frac{1}{\sigma}\right)$$

$$= \frac{2\left(\frac{r}{\sigma}\right)^{m-1} \exp\left(\frac{-\left(\frac{r}{\sigma}\right)^2}{2}\right)}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}\left(\frac{1}{\sigma}\right)$$. Rearranging a bit, we get

$$p_R(r) = \frac{1}{2^{(m/2-1)}\Gamma(m/2)}\left(\frac{r}{\sigma}\right)^{m-1}\exp\left(\frac{-r^2}{2\sigma^2}\right)\left(\frac{1}{\sigma}\right)$$ which is (once again) Eqn. (9.41) with that nasty extra $\frac{1}{\sigma}$ term and which is ***not*** (by the way) Van Hulle's Eqn. (2.2). So it looks like either Haykin and Van Hulle both missed this term or I did my substitutions incorrectly.

4. (Problem 9.9 Haykin) This problem is in two parts, addressing the issues involved in deriving a couple of equations that pertain to the kernel SOM algorithm:

(a) The *incomplete gamma distribution* of a random variable *X*, with sample value *x*, is defined by (Abramowitz and Stegun, 2965, p. 260):

$$P_X(x|\alpha) = \frac{1}{\Gamma(\alpha)}\int_0^x t^{\alpha-1}\exp(-t)\,dt$$

where $\Gamma(\alpha)$ is the gamma function. The *complement* of the incomplete gamma distribution is correspondingly defined by

$$\Gamma(\alpha, x) = \int_x^\infty t^{\alpha-1}\exp(-t)\,dt \quad .$$

Using these two formulas, derive the cumulative distribution function of random variable *R* that is defined in Eq. (9.43). This propagates the error found in 9.8.

Defining $\Gamma(\alpha, x) = \int_0^x e^{-t} t^{\alpha-1}\,dt$, although Haykin may want to show something else, I want to show that the cumulative distribution of our modified Eqn. (9.41) is given by

$$P_R(r) = \frac{\Gamma\left(\frac{m}{2}, \frac{r^2}{2\sigma^2}\right)}{\Gamma\left(\frac{m}{2}\right)}$$. Rather than using variable *d* which is used in the problem (probably a typo), I'll use *m* to make this look more like the equations in the chapter. Since this represents a change of variables, $\frac{r^2}{2\sigma^2} = t$ we must calculate $\frac{\partial r}{\partial t}$ where $r = (2\sigma^2 t)^{\frac{1}{2}}$, or

$$\frac{\partial r}{\partial t} = \frac{\sqrt{2}\sigma}{2} t^{\frac{-1}{2}} = \frac{\sqrt{2}\sigma}{2}\frac{\sqrt{2}\sigma}{r} = \frac{\sigma^2}{r}$$

Substituting $\frac{r^2}{2\sigma^2}$ for t and $\frac{m}{2}$ for alpha, we get that the cumulative distribution is

$$P_R(r)=\frac{\int_0^x \exp\left(\frac{-r^2}{2\sigma^2}\right)\left(\frac{r^2}{2\sigma^2}\right)^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)}\frac{r}{\sigma^2}dr$$

and by the fundamental theorem of calculus, the derivative of this is

$$\frac{\exp\left(\frac{-r^2}{2\sigma^2}\right)\left(\frac{r^2}{2\sigma^2}\right)^{\frac{m}{2}-1}}{\Gamma\left(\frac{d}{2}\right)}\frac{r}{\sigma^2\Gamma\left(\frac{m}{2}\right)}=\frac{\left(\frac{r}{\sigma}\right)^{m-2}\exp\left(\frac{-r^2}{2\sigma^2}\right)}{2^{\frac{m}{2}-1}\Gamma\left(\frac{m}{2}\right)}\left(\frac{r}{\sigma^2}\right)=\frac{\left(\frac{r}{\sigma}\right)^{m-1}\exp\left(\frac{-r^2}{2\sigma^2}\right)}{2^{\frac{m}{2}-1}\Gamma\left(\frac{m}{2}\right)}\left(\frac{1}{\sigma}\right) .$$

So my conclusion is that the nasty extra $\frac{1}{\sigma}$ term really belongs in there. You'll note the extra σ term hanging out in front of Haykin's Eqn. (9.43). If I'm right, that's there to compensate for the missing term in Eqn. (9.41).

(b) Using the formula of the incomplete gamma distribution as the definition of the averaged neural output $\bar{y}_i$, derive Eq. (9.51) for the partial derivative $\partial\bar{y}_i(r)/\partial r$.

If you look at Eqn. (9.51) you can see plainly that $\frac{\partial\bar{y}(r)}{\partial r}$ must be negative if this equation is correct.

Recall that *r* is always positive and as is σ and the value of the gamma function as well as the exponential. Now if this is really negative, then why would we expect to take the log of it as is done in Eqn.s (9,50), (9.49), (9.48), and (9.47) and expect to get a real number? Once again, something is amiss.

Here's how I respond to this question:

Given that we've shown that this is the incomplete gamma distribution is the cumulative distribution function for *r*, and that it is given by

$$p_R(r)=\left(1-\frac{\Gamma\left(\frac{m}{2},\frac{r^2}{2\sigma^2}\right)}{\Gamma\left(\frac{m}{2}\right)}\right)$$

Then once again, using the fundamental theorem of calculus, $\frac{\partial\bar{y}(r)}{\partial r}$ is identical to our result in 9.9a, which can be rewritten

$$\frac{2}{\Gamma(m/2)(\sqrt{2}\sigma_i)^{m-1}} r^{m-1} \exp\left(\frac{-r^2}{2\sigma_i^2}\right)\left(\frac{1}{\sigma}\right)$$

However, this does not account for the averaging over the sample values of *r*, so we really should be summing this over all *r* in our sample and dividing by the number of samples.

5. (Problem 9.10 Haykin) In developing the approximate update formula of Eq. (9.55) for the weight vector of the kernel SOM algorithm, we justified ignoring the second term in Eq. (9.52). Yet, in deriving the update formula of Eq. (9.58) for the kernel width σ_i, no approximation was made. Justify this latter choice.

   What happened in Equation (9.52) was slightly more complicated that ignoring a single term. In fact, this expression: $\frac{\mathbf{x}-\mathbf{w}_i}{\sigma_i^2}-(m-1)\left(\frac{\mathbf{x}-\mathbf{w}_i}{\|\mathbf{x}-\mathbf{w}_i\|^2}\right)$ is shown to be approximately equal to $\frac{\mathbf{x}-\mathbf{w}_i}{m\sigma_i^2}$ by noting that $E[\![\|\mathbf{x}-\mathbf{w}_i\|^2]\!]=m\sigma_i^2$. If we try to exploit that same observation in Equation 9.58, we get $\Delta\sigma_i=\frac{\eta_\sigma}{\sigma_i}\left(\frac{\|\mathbf{x}-\mathbf{w}_i\|}{m\sigma_i^2}-1\right)=\frac{\eta_\sigma}{\sigma_i}\left(\frac{m\sigma_i^2}{m\sigma_i^2}-1\right)=\frac{\eta_\sigma}{\sigma_i}(1-1)=0$ which would yield an update rule in which σ never changes.