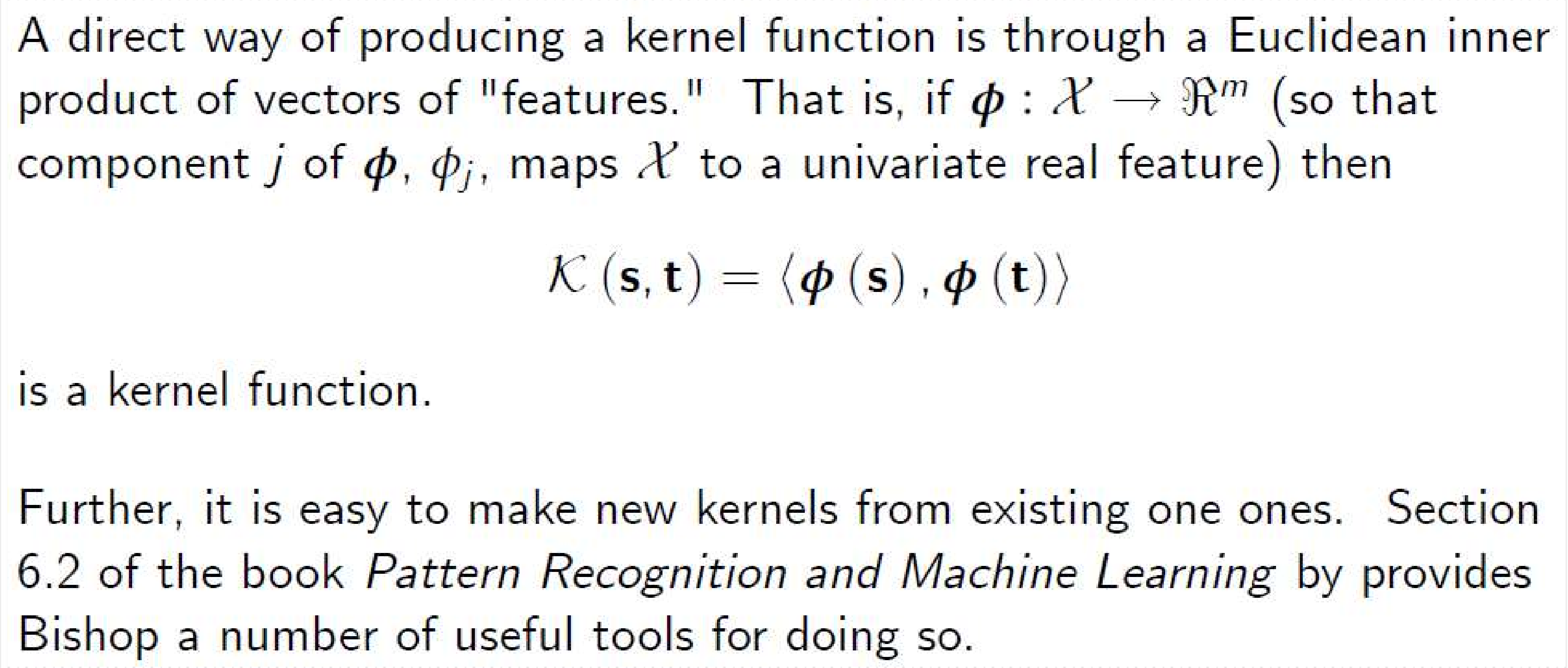
Kernel Mechanics (New Ones from Existing Ones)

Stephen Vardeman

Analytics Iowa LLC

ISU Statistics and IMSE

A way to make a kernel



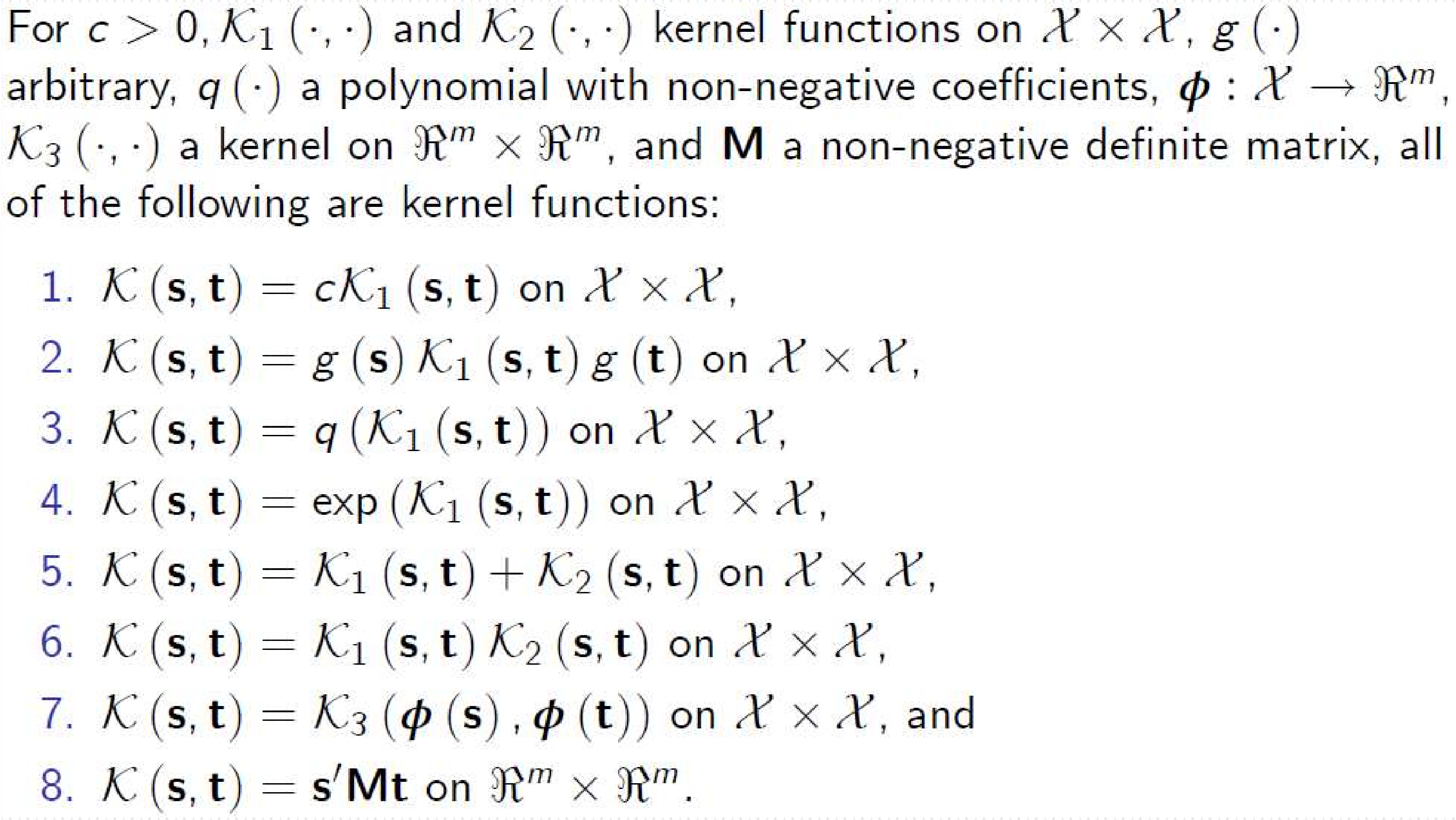
A direct way of producing a kernel function is through a Euclidean inner product of vectors of "features." That is, if $\boldsymbol{\phi} : \mathcal{X} \to \Re^m$ (so that component j of $\boldsymbol{\phi}$, ϕ_j, maps $\mathcal{X}$ to a univariate real feature) then

$$\mathcal{K}\left(\mathbf{s},\mathbf{t}\right) = \left\langle \boldsymbol{\phi}\left(\mathbf{s}\right), \boldsymbol{\phi}\left(\mathbf{t}\right)\right\rangle$$

is a kernel function.

Further, it is easy to make new kernels from existing one ones. Section 6.2 of the book *Pattern Recognition and Machine Learning* by provides Bishop a number of useful tools for doing so.

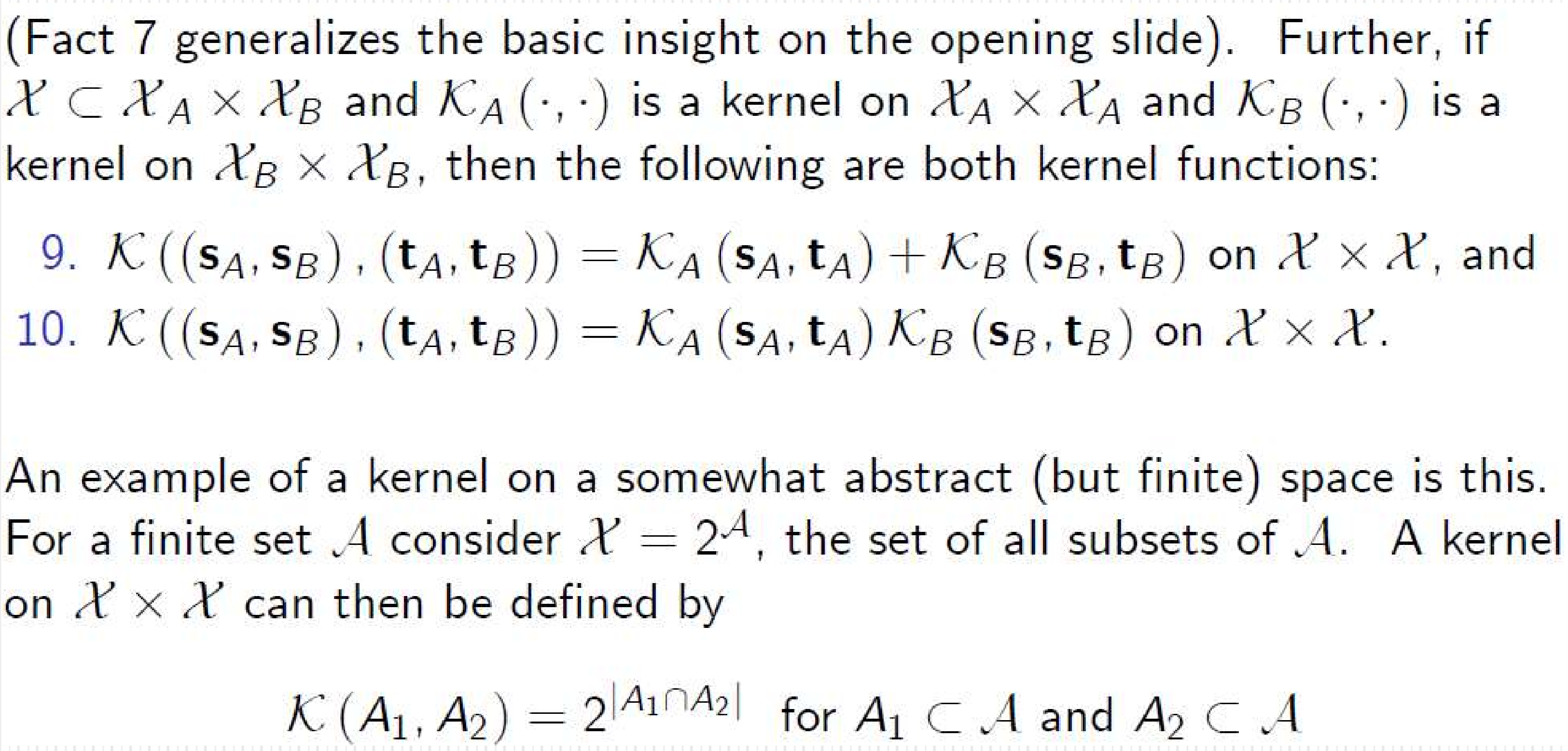
Facts from Bishop’s Section 6.2



For $c > 0$, $\mathcal{K}_1(\cdot,\cdot)$ and $\mathcal{K}_2(\cdot,\cdot)$ kernel functions on $\mathcal{X} \times \mathcal{X}$, $g(\cdot)$ arbitrary, $q(\cdot)$ a polynomial with non-negative coefficients, $\boldsymbol{\phi} : \mathcal{X} \to \Re^m$, $\mathcal{K}_3(\cdot,\cdot)$ a kernel on $\Re^m \times \Re^m$, and $\mathbf{M}$ a non-negative definite matrix, all of the following are kernel functions:

1. $\mathcal{K}(\mathbf{s},\mathbf{t}) = c\mathcal{K}_1(\mathbf{s},\mathbf{t})$ on $\mathcal{X} \times \mathcal{X}$,
2. $\mathcal{K}(\mathbf{s},\mathbf{t}) = g(\mathbf{s})\,\mathcal{K}_1(\mathbf{s},\mathbf{t})\,g(\mathbf{t})$ on $\mathcal{X} \times \mathcal{X}$,
3. $\mathcal{K}(\mathbf{s},\mathbf{t}) = q(\mathcal{K}_1(\mathbf{s},\mathbf{t}))$ on $\mathcal{X} \times \mathcal{X}$,
4. $\mathcal{K}(\mathbf{s},\mathbf{t}) = \exp(\mathcal{K}_1(\mathbf{s},\mathbf{t}))$ on $\mathcal{X} \times \mathcal{X}$,
5. $\mathcal{K}(\mathbf{s},\mathbf{t}) = \mathcal{K}_1(\mathbf{s},\mathbf{t}) + \mathcal{K}_2(\mathbf{s},\mathbf{t})$ on $\mathcal{X} \times \mathcal{X}$,
6. $\mathcal{K}(\mathbf{s},\mathbf{t}) = \mathcal{K}_1(\mathbf{s},\mathbf{t})\,\mathcal{K}_2(\mathbf{s},\mathbf{t})$ on $\mathcal{X} \times \mathcal{X}$,
7. $\mathcal{K}(\mathbf{s},\mathbf{t}) = \mathcal{K}_3(\boldsymbol{\phi}(\mathbf{s}), \boldsymbol{\phi}(\mathbf{t}))$ on $\mathcal{X} \times \mathcal{X}$, and
8. $\mathcal{K}(\mathbf{s},\mathbf{t}) = \mathbf{s}'\mathbf{M}\mathbf{t}$ on $\Re^m \times \Re^m$.

More facts from Bishop and an example



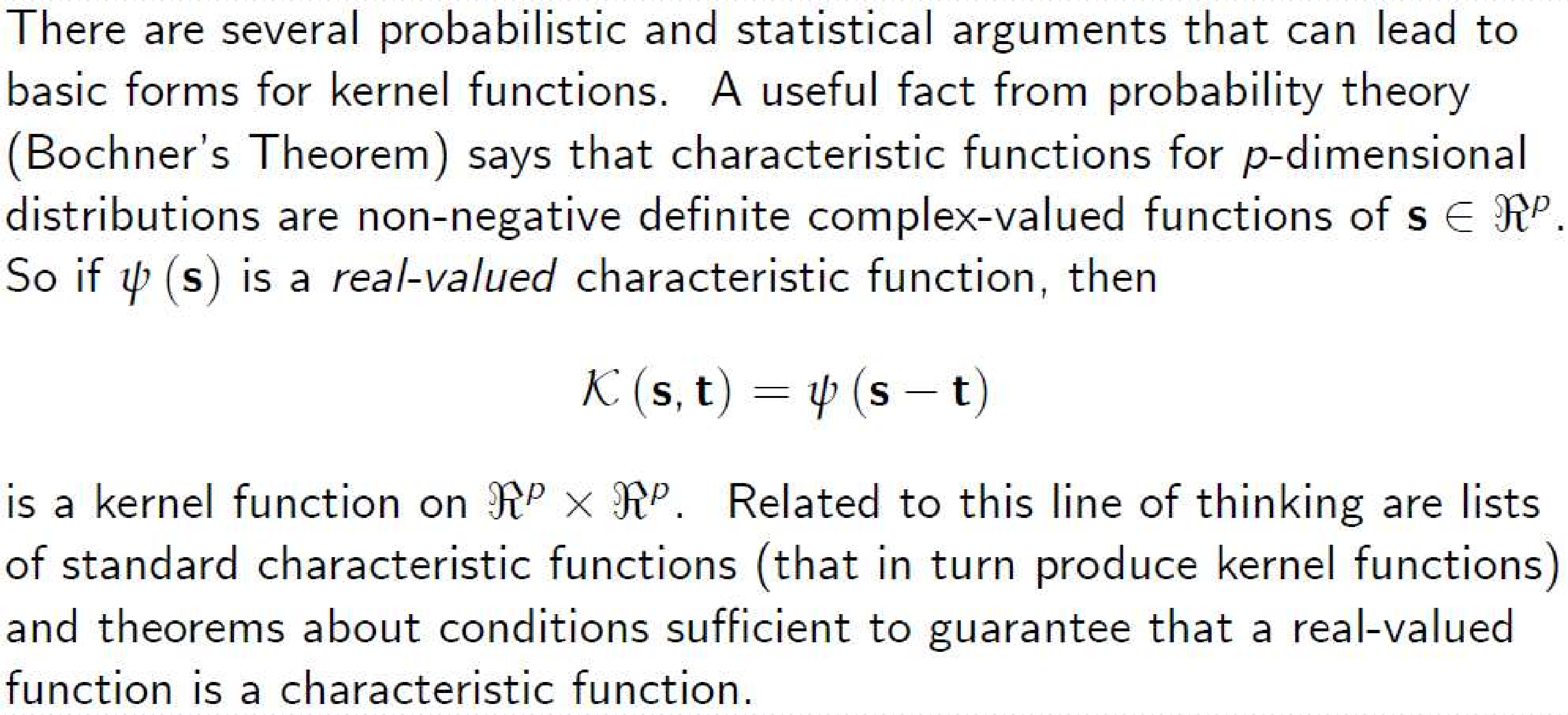
(Fact 7 generalizes the basic insight on the opening slide). Further, if $\mathcal{X} \subset \mathcal{X}_A \times \mathcal{X}_B$ and $\mathcal{K}_A(\cdot,\cdot)$ is a kernel on $\mathcal{X}_A \times \mathcal{X}_A$ and $\mathcal{K}_B(\cdot,\cdot)$ is a kernel on $\mathcal{X}_B \times \mathcal{X}_B$, then the following are both kernel functions:

9. $\mathcal{K}((\mathbf{s}_A, \mathbf{s}_B), (\mathbf{t}_A, \mathbf{t}_B)) = \mathcal{K}_A(\mathbf{s}_A, \mathbf{t}_A) + \mathcal{K}_B(\mathbf{s}_B, \mathbf{t}_B)$ on $\mathcal{X} \times \mathcal{X}$, and
10. $\mathcal{K}((\mathbf{s}_A, \mathbf{s}_B), (\mathbf{t}_A, \mathbf{t}_B)) = \mathcal{K}_A(\mathbf{s}_A, \mathbf{t}_A)\,\mathcal{K}_B(\mathbf{s}_B, \mathbf{t}_B)$ on $\mathcal{X} \times \mathcal{X}$.

An example of a kernel on a somewhat abstract (but finite) space is this. For a finite set $\mathcal{A}$ consider $\mathcal{X} = 2^{\mathcal{A}}$, the set of all subsets of $\mathcal{A}$. A kernel on $\mathcal{X} \times \mathcal{X}$ can then be defined by

$$\mathcal{K}(A_1, A_2) = 2^{|A_1 \cap A_2|} \quad \text{for } A_1 \subset \mathcal{A} \text{ and } A_2 \subset \mathcal{A}$$

Basic kernels from characteristic functions

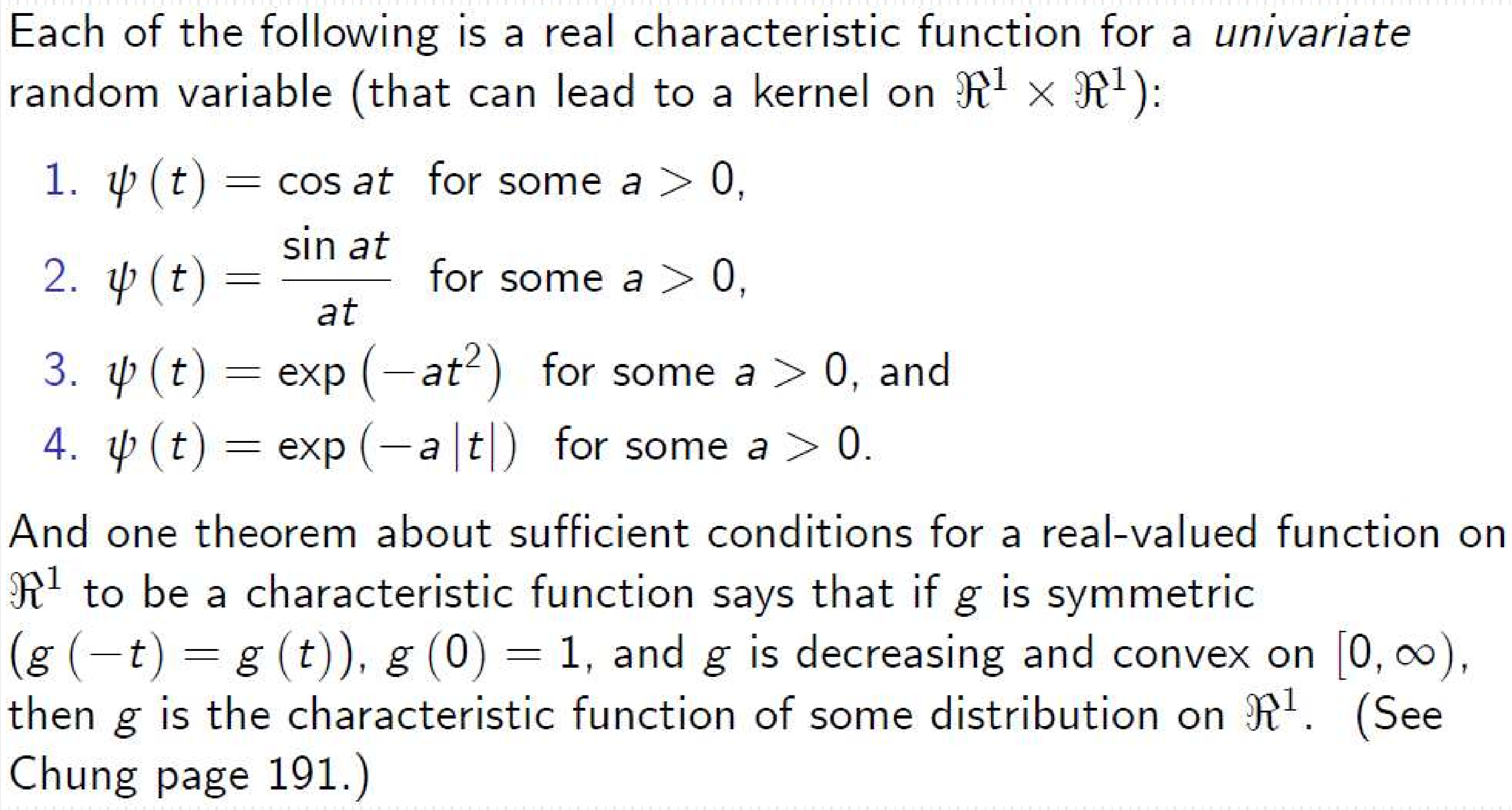


There are several probabilistic and statistical arguments that can lead to basic forms for kernel functions. A useful fact from probability theory (Bochner's Theorem) says that characteristic functions for *p*-dimensional distributions are non-negative definite complex-valued functions of $\mathbf{s} \in \Re^p$. So if $\psi(\mathbf{s})$ is a *real-valued* characteristic function, then

$$\mathcal{K}(\mathbf{s}, \mathbf{t}) = \psi(\mathbf{s} - \mathbf{t})$$

is a kernel function on $\Re^p \times \Re^p$. Related to this line of thinking are lists of standard characteristic functions (that in turn produce kernel functions) and theorems about conditions sufficient to guarantee that a real-valued function is a characteristic function.

1-D characteristic functions

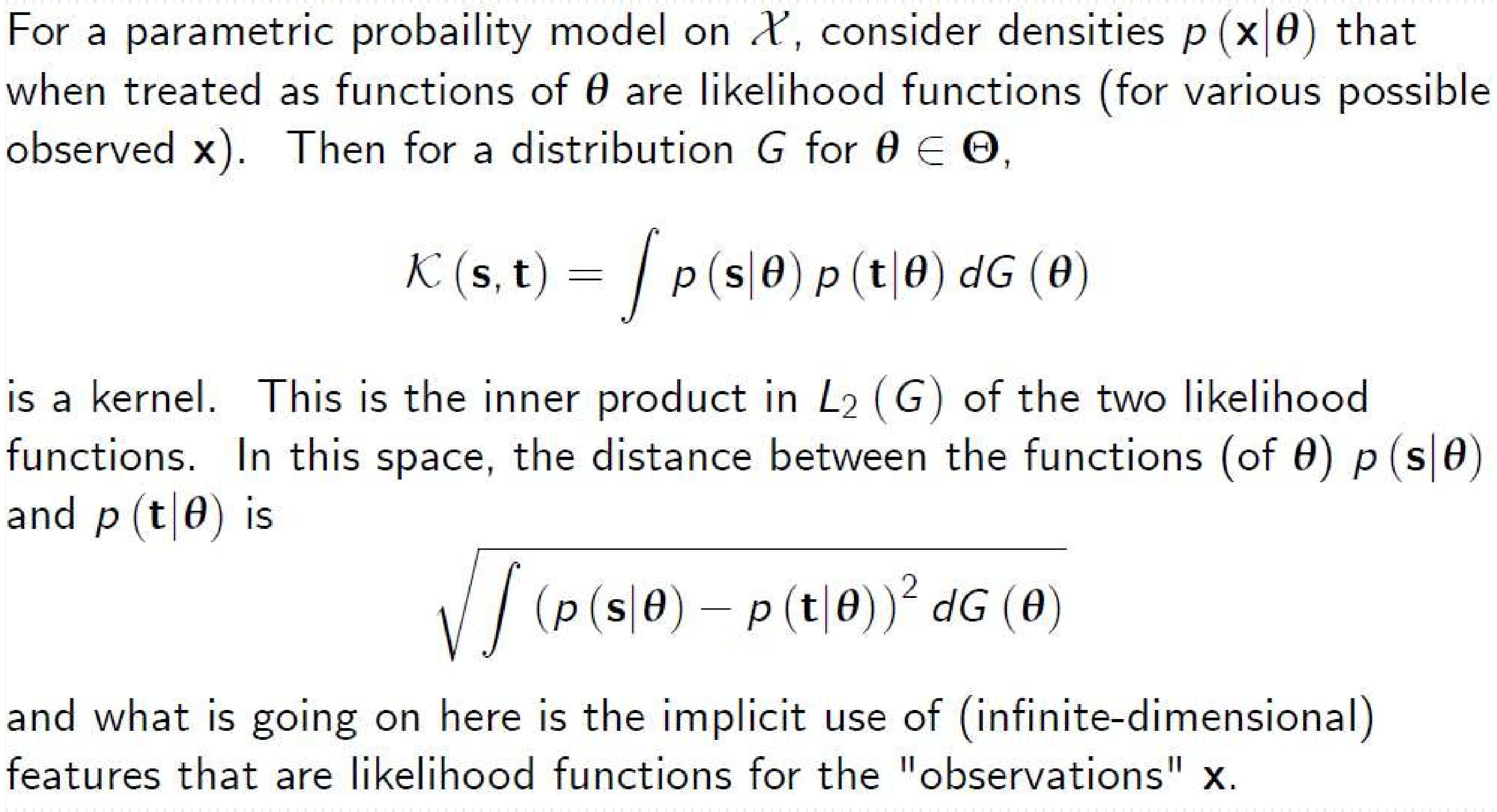


Each of the following is a real characteristic function for a *univariate* random variable (that can lead to a kernel on $\Re^1 \times \Re^1$):

1. $\psi(t) = \cos at$ for some $a > 0$,
2. $\psi(t) = \dfrac{\sin at}{at}$ for some $a > 0$,
3. $\psi(t) = \exp\left(-at^2\right)$ for some $a > 0$, and
4. $\psi(t) = \exp\left(-a\left|t\right|\right)$ for some $a > 0$.

And one theorem about sufficient conditions for a real-valued function on $\Re^1$ to be a characteristic function says that if g is symmetric $(g(-t) = g(t))$, $g(0) = 1$, and g is decreasing and convex on $[0, \infty)$, then g is the characteristic function of some distribution on $\Re^1$. (See Chung page 191.)

Kernels from average products of likelihoods



For a parametric probaility model on $\mathcal{X}$, consider densities $p(\mathbf{x}|\boldsymbol{\theta})$ that when treated as functions of $\boldsymbol{\theta}$ are likelihood functions (for various possible observed $\mathbf{x}$). Then for a distribution G for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

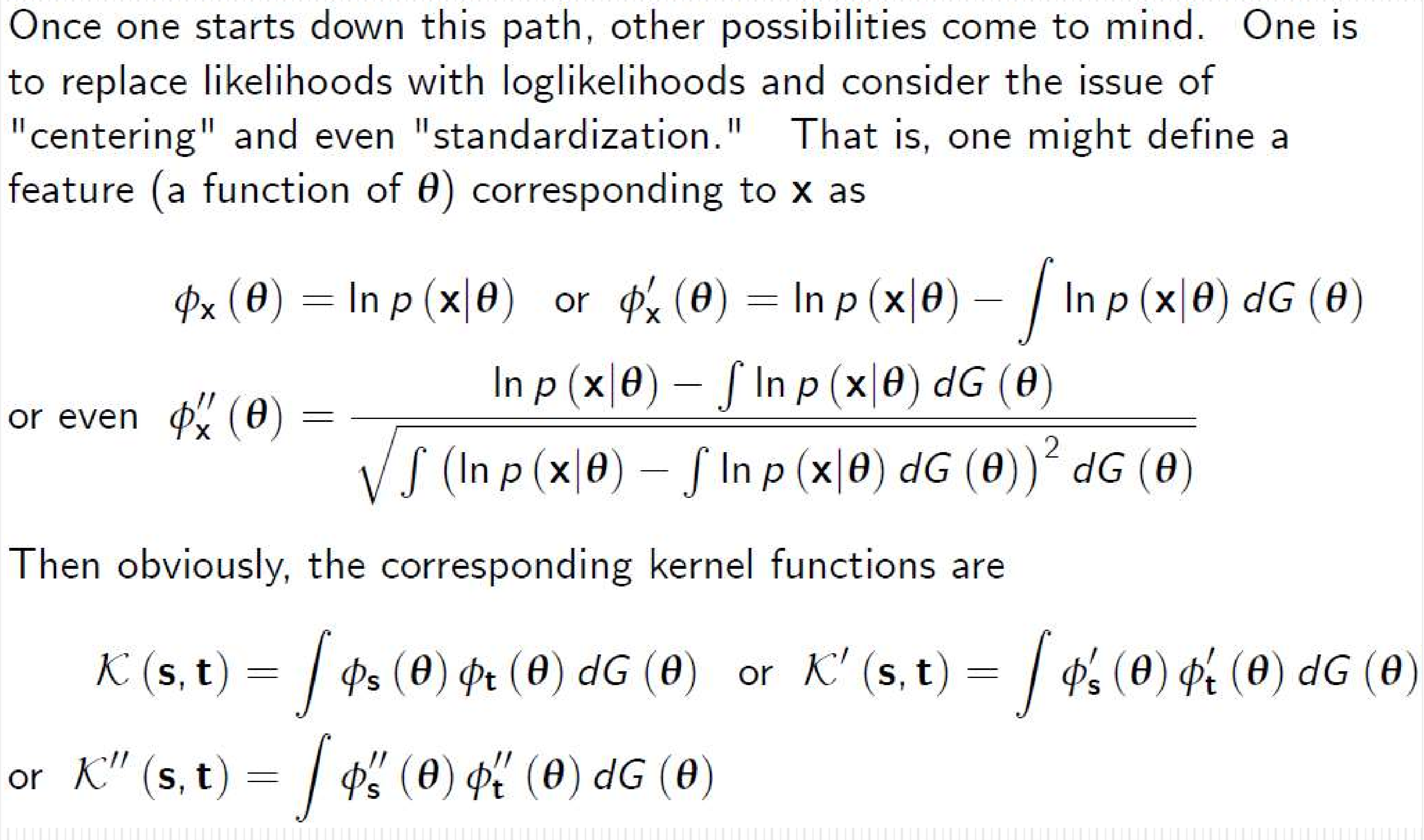
$$\mathcal{K}(\mathbf{s}, \mathbf{t}) = \int p(\mathbf{s}|\boldsymbol{\theta})\, p(\mathbf{t}|\boldsymbol{\theta})\, dG(\boldsymbol{\theta})$$

is a kernel. This is the inner product in $L_2(G)$ of the two likelihood functions. In this space, the distance between the functions (of $\boldsymbol{\theta}$) $p(\mathbf{s}|\boldsymbol{\theta})$ and $p(\mathbf{t}|\boldsymbol{\theta})$ is

$$\sqrt{\int \left(p(\mathbf{s}|\boldsymbol{\theta}) - p(\mathbf{t}|\boldsymbol{\theta})\right)^2 dG(\boldsymbol{\theta})}$$

and what is going on here is the implicit use of (infinite-dimensional) features that are likelihood functions for the "observations" $\mathbf{x}$.

Kernels from average log-likelihoods



Once one starts down this path, other possibilities come to mind. One is to replace likelihoods with loglikelihoods and consider the issue of "centering" and even "standardization." That is, one might define a feature (a function of θ) corresponding to $\mathbf{x}$ as

$$\phi_{\mathbf{x}}(\boldsymbol{\theta}) = \ln p(\mathbf{x}|\boldsymbol{\theta}) \quad \text{or} \quad \phi'_{\mathbf{x}}(\boldsymbol{\theta}) = \ln p(\mathbf{x}|\boldsymbol{\theta}) - \int \ln p(\mathbf{x}|\boldsymbol{\theta})\, dG(\boldsymbol{\theta})$$

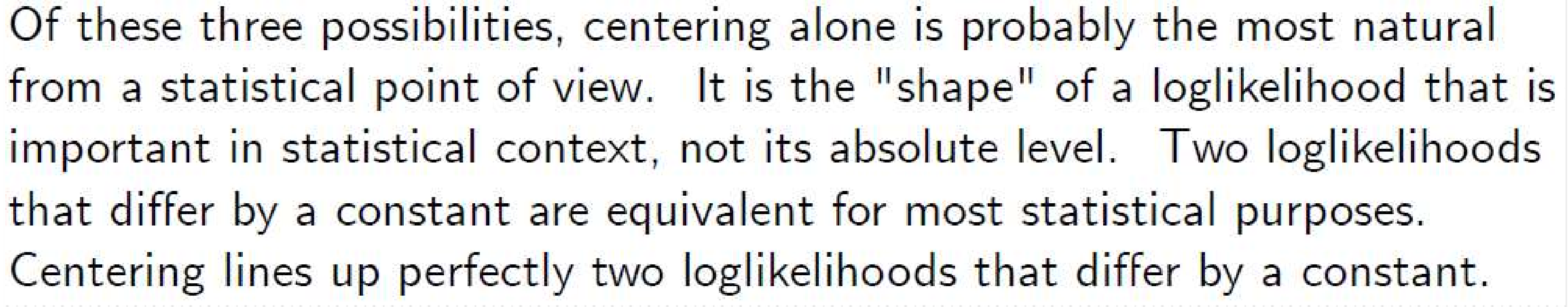
or even $\phi''_{\mathbf{x}}(\boldsymbol{\theta}) = \dfrac{\ln p(\mathbf{x}|\boldsymbol{\theta}) - \int \ln p(\mathbf{x}|\boldsymbol{\theta})\, dG(\boldsymbol{\theta})}{\sqrt{\int \left(\ln p(\mathbf{x}|\boldsymbol{\theta}) - \int \ln p(\mathbf{x}|\boldsymbol{\theta})\, dG(\boldsymbol{\theta})\right)^2 dG(\boldsymbol{\theta})}}$

Then obviously, the corresponding kernel functions are

$$\mathcal{K}(\mathbf{s},\mathbf{t}) = \int \phi_{\mathbf{s}}(\boldsymbol{\theta})\, \phi_{\mathbf{t}}(\boldsymbol{\theta})\, dG(\boldsymbol{\theta}) \quad \text{or} \quad \mathcal{K}'(\mathbf{s},\mathbf{t}) = \int \phi'_{\mathbf{s}}(\boldsymbol{\theta})\, \phi'_{\mathbf{t}}(\boldsymbol{\theta})\, dG(\boldsymbol{\theta})$$

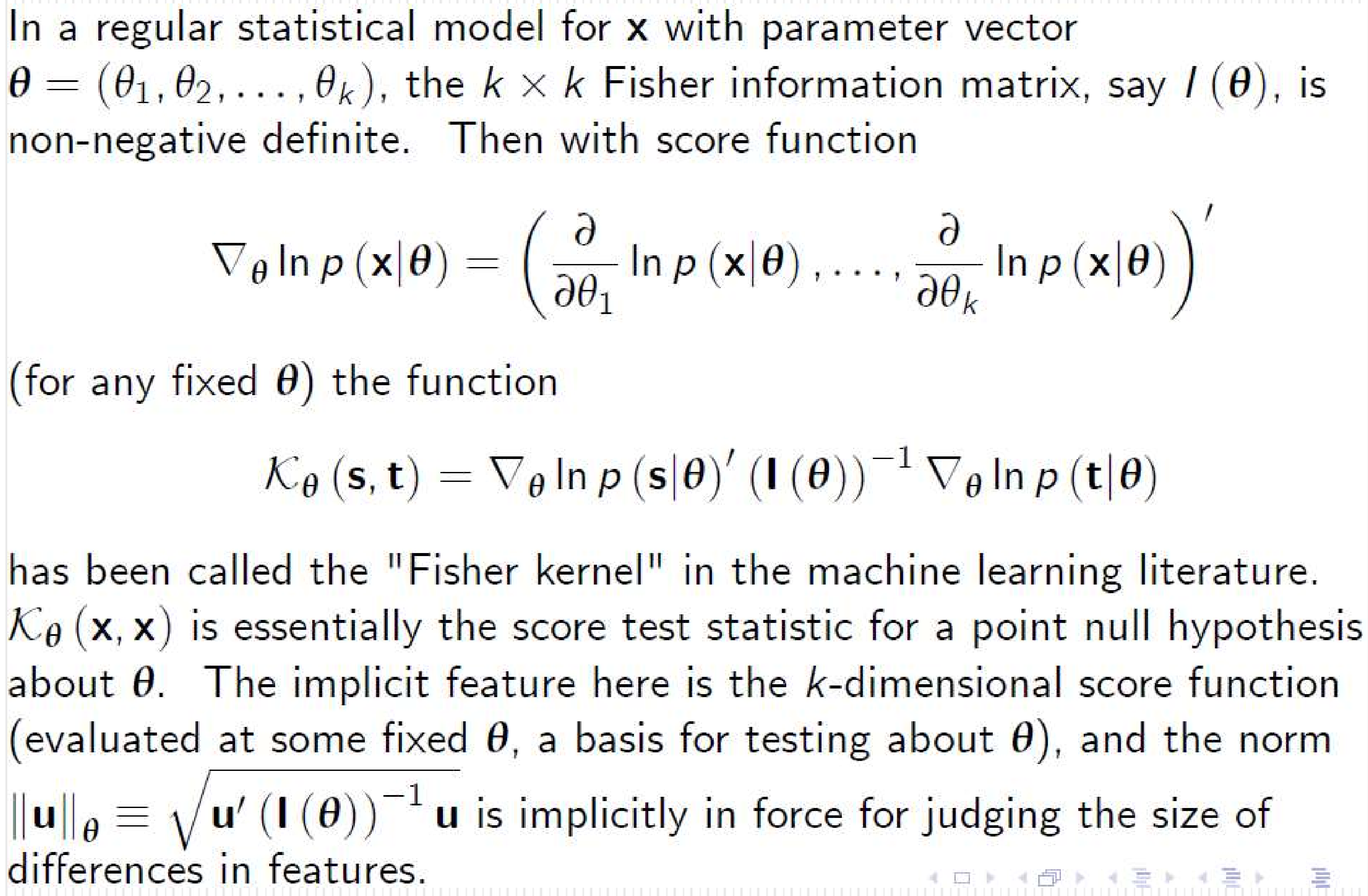
or $\mathcal{K}''(\mathbf{s},\mathbf{t}) = \int \phi''_{\mathbf{s}}(\boldsymbol{\theta})\, \phi''_{\mathbf{t}}(\boldsymbol{\theta})\, dG(\boldsymbol{\theta})$

Kernels from average log-likelihoods (cont.)



Of these three possibilities, centering alone is probably the most natural from a statistical point of view. It is the "shape" of a loglikelihood that is important in statistical context, not its absolute level. Two loglikelihoods that differ by a constant are equivalent for most statistical purposes. Centering lines up perfectly two loglikelihoods that differ by a constant.

“Fisher Kernels” from Score Functions



In a regular statistical model for $\mathbf{x}$ with parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$, the $k \times k$ Fisher information matrix, say $I(\boldsymbol{\theta})$, is non-negative definite. Then with score function

$$\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{x}|\boldsymbol{\theta}) = \left(\frac{\partial}{\partial \theta_1} \ln p(\mathbf{x}|\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial \theta_k} \ln p(\mathbf{x}|\boldsymbol{\theta}) \right)'$$

(for any fixed $\boldsymbol{\theta}$) the function

$$\mathcal{K}_{\boldsymbol{\theta}}(\mathbf{s}, \mathbf{t}) = \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{s}|\boldsymbol{\theta})' \left(\mathbf{I}(\boldsymbol{\theta})\right)^{-1} \nabla_{\boldsymbol{\theta}} \ln p(\mathbf{t}|\boldsymbol{\theta})$$

has been called the "Fisher kernel" in the machine learning literature. $\mathcal{K}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x})$ is essentially the score test statistic for a point null hypothesis about $\boldsymbol{\theta}$. The implicit feature here is the k-dimensional score function (evaluated at some fixed $\boldsymbol{\theta}$, a basis for testing about $\boldsymbol{\theta}$), and the norm $\|\mathbf{u}\|_{\boldsymbol{\theta}} \equiv \sqrt{\mathbf{u}' \left(\mathbf{I}(\boldsymbol{\theta})\right)^{-1} \mathbf{u}}$ is implicitly in force for judging the size of differences in features.