A Table of notations

Symbol	**Name**	**Description**
Features and feature updates		
ψ_{X_i}	Implicit function of X_i	For example, the solution to an optimisation problem, differential equation or root finding problem depending on X_i.
$\psi_{X_i}^{(t+1)}$	Feature for input X_i	As in equation 1. The result of applying t iterations of a numerical procedure that compute the implicit function ψ_{X_i}.
$g^{(t)}$	Feature update rule	As in equation 1
$\boldsymbol{\psi}$	Feature space	Features are an element of this space.
Kernels and kernel updates (component-wise)		
Ψ_{ij}	ℓDEKER evaluated at X_i and X_j	As in equation 2.
$\Psi_{ij}^{(t+1)}$	DEKER evaluated at X_i and X_j	As in equation 2.
G_{ij}	DEKER update rule (component-wise)	As in Theorem 4.
$\mathbb{\Psi}$	DEKER evaluation space	Evaluations of the DEKER are an element of this space.
Kernels and kernel updates (matrix)		
$\mathsf{\Psi}$	ℓDEKER matrix evaluated at (X_1, X_2)	$\mathsf{\Psi} \in \mathbb{S}_+^2$ and the ijth element of $\mathsf{\Psi}$ is Ψ_{ij}.
$\mathsf{\Psi}^{(t+1)}$	DEKER matrix evaluated at (X_1, X_2)	$\mathsf{\Psi}^{(t+1)} \in \mathbb{S}_+^2$ and the ijth element of $\mathsf{\Psi}^{(t+1)}$ is $\Psi_{ij}^{(t+1)}$.
G	DEKER update rule (matrix version)	As in equation 3.
$\mathbb{S}_+^2$	Convex cone of PSD matrices	$\mathbb{S}_+^2 = \{M \in \mathbb{R}^{2\times 2} \mid M \succeq 0\}$.
Inputs and data		
$\mathbb{X}$	Input space	A space $\mathbb{X} \subseteq \mathbb{R}^l$ to which input belongs.
X_i	Input vector	An element of $\mathbb{X}$.
X	Input matrix	An $N \times l$ matrix, where each row represents a single element of $\mathbb{X}$.
Γ	Random mapping	$\Gamma : \mathbb{X} \to \mathbb{Y}$ is a random mapping which translates input X to data Y belonging to the support $\mathbb{Y}$ of an exponential family.
$Y = \Gamma(X)$	Data	The result of applying Γ to input X
$\gamma_i(X)$	Random mapping coordinate evaluation	The ith coordinate of $\Gamma(X)$.
Exponential families		
$\mathbb{Y}$	Exponential family support	As in § 2.3. The space over which the exponential family distribution has non-zero mass. We take $\mathbb{Y} \subseteq \mathbb{R}$.
T	Sufficient statistic	As in § 2.3.
η	Canonical parameter	As in § 2.3. The canonical parameter belongs to an open set $\mathcal{H} \subseteq \mathbb{R}$.
A	Log partition function	The function that returns logarithm of the normalising constant of the exponential family as a function of its canonical parameter. $A : \mathbb{H} \to \mathbb{R}$. As in § 2.3.
s^{-1}	Inverse link function	The function that maps a parameter of the exponential family to the expectation parameter of the exponential family. As in equation 8.

R	Canonical nonlinearity	As in Proposition 2. $R : \mathbb{R} \to \mathbb{H}$ maps the result of a linear transformation to a canonical parameter of the exponential family.
Activation functions		
ζ, ζ_1, ζ_2	General activation function	A generic activation function of a neural network.
ρ	factor activation	The derivative of the canonical nonlinearity. That is, $\rho(a) = R'(a)$.
σ	chain activation	The derivative of the composition of the log partition function and the canonical nonlinearity. That is, $\sigma(a) = (A \circ R)'(a)$
u	Heaviside step function	A special case of ζ. $\mathrm{u}(a)$ takes the value of 0 if $a < 0$, 1 if $a > 0$, and 0.5 if $a = 0$.
ReLU	Rectified linear unit	A special case of ζ. $\mathrm{ReLU}(a) = \mathrm{u}(a)a$.
erf	Error function	$\mathrm{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-z^2}\, dz$. As in § 3.3.

Table 3: Summary of notation used throughout this paper.

B What is the relationship between s^{-1}, R, σ and ρ?

Nonlinearly parameterised exponential families are densities Any member of a given exponential family is a density (mass) function. That is, for every $\nu \in \mathbb{H}$,

$$\int p(y \mid \nu)\, dy = 1 \quad \text{and} \quad p(y \mid \nu) \geq 0. \tag{22}$$

Given a nonlinearity $R : \mathbb{R} \to \mathbb{H}$, it is immediate that any member of a given nonlinearly parameterised exponential family is a density. That is, for any $a \in \mathbb{R}$, defining $\nu_0 = R(a) \in \mathbb{H}$, from equation 22 we have

$$\int p\big(y \mid R(a)\big)\, dy = \int p(y \mid \underbrace{\nu_0}_{\in \mathbb{H}})\, dy = 1 \quad \text{and} \quad p\big(y \mid R(a)\big) = p(y \mid \underbrace{\nu_0}_{\in \mathbb{H}}) \geq 0.$$

Identities relating s^{-1}, R, σ, and ρ The canonical link function is one which expresses the canonical parameter η in terms of the expectation parameter $\mathbb{E}[T(Y) \mid \eta]$. When R is the identity, we have that $A'(\eta) = \mathbb{E}[T(Y) \mid \eta]$ and so the canonical link function is $(A')^{-1}$. That $A'(\eta) = \mathbb{E}[T(Y) \mid \eta]$ follows from the fact that A is a cumulant generating function for the sufficient statistic (Wainwright et al., 2008, Proposition 3.1).

A (not necessarily canonical) link function is one which expresses a (not necessarily canonical) parameter a in terms of the expectation parameter $\mathbb{E}[T(Y)]$. We now discuss how given an exponential family and a link function can be related to a choice of R.

In the general setting, since A is a cumulant generating function, the inverse link function s^{-1} satisfies

$$A'\big(R(a)\big) = s^{-1}(a) \tag{23}$$

Noting that A' is invertible because A is strictly convex, equation 23 implies that for a desired link function s, we must choose

$$\begin{aligned} R(a) &= \big((A')^{-1} \circ s^{-1}\big)(a) \\ &= (A')^{-1}(\eta), \quad \eta = s^{-1}(a) \end{aligned} \tag{24}$$

Since $\rho(a) = R'(a)$, we have that

$$\begin{aligned} \rho(a) &= \frac{dR}{d\eta}\frac{d\eta}{da} \\ &= \big((A')^{-1}\big)'\big(s^{-1}(a)\big)\,(s^{-1})'(a) \\ &= \frac{(s^{-1})'(a)}{A'' \circ (A')^{-1} \circ s^{-1}(a)} \quad \text{(by the inverse function theorem).} \end{aligned} \tag{25}$$

Since $\sigma(a) = (A \circ R)'(a)$ and $\rho(a) = R'(a)$, we have that

$$\begin{aligned} \sigma(a) &= \underbrace{A'\big(R(a)\big)}_{s^{-1}(a)}\rho(a) \\ &= \frac{s^{-1}(a)\,(s^{-1})'(a)}{A'' \circ (A')^{-1} \circ s^{-1}(a)}. \end{aligned} \tag{26}$$

As expected, when s a canonical link function (which is to say that $s^{-1}(a) = A'(a)$), equation 24 implies that R is the identity, ρ takes a constant value of 1 and σ is A').

Some examples are given in Table 1.

C Detailed neural network kernel description

Let $\mathsf{W}^{(1)} \in \mathbb{R}^{d\times n}$ be the weights of a fully connected hidden layer with activation function σ. Suppose each entry of $\mathsf{W}^{(1)}$ is i.i.d. with distribution $\mathcal{N}(0,1)$. Given an input $\phi_1 \in \mathbb{R}^{n\times 1}$ (we take the convention that vectors are column vectors), the signal in the hidden layer is $h_1^{(1)} \triangleq \sigma(\mathsf{W}^{(1)}\phi_1)$. Given any two input features ϕ_1 and ϕ_2, a normalised inner product of the features in the hidden layer is

$$\frac{1}{d}{h_1^{(1)}}^\top h_2^{(1)} = \frac{1}{d}\sigma(\mathsf{W}^{(1)}\phi_1)^\top\sigma(\mathsf{W}^{(1)}\phi_2) = \frac{1}{d}\sum_{i=1}^{d}\sigma(W_i^\top\phi_1)\sigma(W_i^\top\phi_2), \tag{27}$$

where $W_i^\top$ is the ith row of $\mathsf{W}^{(1)}$. Note that since each row of $\mathsf{W}^{(1)}$ is i.i.d., equation 27 is an average of i.i.d. random variables. A strong law of large numbers says that the average of a sequence of i.i.d. random variables converges almost surely to the expectation if the expectation is finite. We therefore have that equation 27 converges almost surely to

$$k_\sigma(\phi_1,\phi_2) \triangleq \mathbb{E}_W\big[\sigma(W^\top\phi_1)\sigma(W^\top\phi_2)\big], \tag{28}$$

as $d\to\infty$. Here $W^\top \in \mathbb{R}^{1\times n}$ is a vector with i.i.d. entries drawn from $\mathcal{N}(0,1)$. We call k_σ a single hidden layer neural network kernel (NNK) with activation function σ.

Note that $W^\top$ is a row vector, and therefore $W^\top\phi_1$ is a scalar. This means that while equation 28 is written as an expectation over n-variate random vector W, it is actually only an expectation over the bivariate random vector $(\chi,\chi') = \big(W^\top\phi_1, W^\top\phi_2\big)$. Since Gaussian random vectors are closed under affine transformations, (χ,χ') is a Gaussian random vector. The mean of each component is zero. The 2-by-2 covariance matrix $\boldsymbol{\Sigma}^{(1)}$ has entries

$$\Sigma_{12}^{(1)} = \mathbb{E}\big[(W^\top\phi_1)(W^\top\phi_2)\big] = \mathbb{E}\left[\sum_{p=1}^{n}\sum_{q=1}^{n}W_p\phi_{1p}W_q\phi_q\right] = \sum_{p=1}^{n}\sum_{q=1}^{n}\mathbb{E}\big[W_p\phi_{1p}W_q\phi_q'\big] = \sum_{p=1}^{n}\mathbb{E}\big[W_p^2\big]\phi_{1p}\phi_{2p},$$

where the last equality is due to the fact that W_p and W_q are independent when $p\neq q$, and ϕ_{1p} and ϕ_{2q} are not random variables. Since $\mathbb{E}W_p^2 = 1$, the right most term is $\phi_1^\top\phi_2$. We may repeat a similar procedure for Σ_{11} and Σ_{22}, giving us an expression for the covariance

$$\boldsymbol{\Sigma}^{(1)} \triangleq \begin{pmatrix}\Sigma_{11}^{(1)} & \Sigma_{12}^{(1)}\\ \Sigma_{12}^{(1)} & \Sigma_{22}^{(1)}\end{pmatrix} = \begin{pmatrix}\phi_1^\top\phi_1 & \phi_1^\top\phi_2\\ \phi_2^\top\phi_1 & \phi_2^\top\phi_2\end{pmatrix}. \tag{29}$$

It is instructive to rewrite equation 28 in two other forms. The first form explicitly shows the expectation with respect to the bivariate Gaussian which has covariance given by equation 29,

$$k_\sigma(\phi_1,\phi_2) = \mathbb{E}_{(\chi,\chi')^\top\sim\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}^{(1)})}\big[\sigma(\chi)\sigma(\chi')\big], \tag{30}$$

For the second form, we use notation to remind us that the kernel k_σ actually depends only on three scalar values. From equation 29, we observe that $k_\sigma(\phi_1,\phi_2)$ depends on ϕ_1 and ϕ_2 *only* through the pairwise inner products $\Sigma_{12}^{(1)} = \phi_1^\top\phi_2$. Observe that by symmetry $\Sigma_{12}^{(1)} = \Sigma_{21}^{(1)}$. In other words, there exists a function κ_σ such that

$$k_\sigma(\phi_1,\phi_2) = \kappa_\sigma(\Sigma_{11}^{(1)},\Sigma_{22}^{(1)},\Sigma_{12}^{(1)}). \tag{31}$$

In summary, there are three equivalent ways to write an NNK, $k_\sigma(\phi_1,\phi_2)$:

- As an expectation over random vectors W corresponding to neural network weights equation 28,
- As an expectation over a bivariate Gaussian with covariance $\boldsymbol{\Sigma}^{(1)}$ equation 29,
- As a function of three arguments, explicitly showing the three parameters in the covariance of the bivariate Gaussian equation 31.

Closed-form expressions of k_σ for different σ are available (Williams, 1997; Le Roux & Bengio, 2007; Cho & Saul, 2009; Tsuchida et al., 2018; Pearce et al., 2019; Tsuchida, 2020; Meronen et al., 2020; Tsuchida et al., 2021; Han et al., 2022). For example, when σ is the ReLU function, the resulting kernel is known as the arc-cosine kernel of order 1 and is given by (Cho & Saul, 2009)

$$k_{\mathrm{ReLU}}(\phi_1, \phi_2) = \frac{\|\phi_1\|\|\phi_2\|}{2\pi}\big(\sin\theta - (\pi - \theta)\cos\theta\big), \quad \text{where} \quad \cos\theta = \frac{\phi^\top \phi_2}{\|\phi_1\|\|\phi_2\|}.$$

D Tools for concentration inequalities

The main purpose of this appendix is to introduce Bernstein's inequality and associated tools to apply to our problem at hand. We first need to introduce sub-Gaussian and sub-exponential random variables, and discuss special cases of how we may construct such random variables.

Definition 8. *A centered random variable Y is sub-Gaussian if there exists an $S > 0$ such that*

$$\mathbb{E}\exp\left(Y^2/S^2\right) \le 2.$$

The sub-Gaussian norm of Y,

$$s \triangleq \inf\Big\{v > 0 : \mathbb{E}\exp\Big(Y^2/v^2\Big) \le 2\Big\},$$

is the smallest S.

Bounded random variables are sub-Gaussian, and as an immediate consequence, so are constant random variables.

Lemma 9 (Vershynin (2018) Example 2.5.8)**.** *A bounded random variable Y is sub-Gaussian with sub-Gaussian norm s satisfying*

$$s \le (\log 2)^{-1}\|Y\|_\infty,$$

where $\|Y\|_\infty$ is the essential supremum of Y.

Lipschitz functions of Gaussian random variables are also sub-Gaussian.

Lemma 10 (Vershynin (2018) Theorem 5.2.2)**.** *Let Y be a Gaussian random variable with variance a^2. Let $f : \mathbb{R} \to \mathbb{R}$ be L-Lipschitz. Then $f(Y) - \mathbb{E}f(Y)$ is sub-Gaussian with sub-Gaussian norm s_0 satisfying*

$$s_0 \le C|a|L,$$

for some absolute constant $C > 0$. Furthermore, by the triangle inequality and Lemma 9 $f(Y)$ is sub-Gaussian and the sub-Gaussian norm s of $f(Y)$ satisfies

$$s \le C|a|L + (\log 2)^{-1}\big|\mathbb{E}f(Y)\big|.$$

A class of random variables which includes sub-Gaussian random variables is the class of sub-exponential random variables.

Definition 11. *A random variable Y is sub-exponential if there exists an $A > 0$ such that*

$$\mathbb{E}\exp\left(|Y|/A\right) \le 2$$

The sub-exponential norm of Y,

$$a \triangleq \inf\Big\{v > 0 : \mathbb{E}\exp\Big(\big(|Y|\big)/v\Big) \le 2\Big\},$$

is the smallest A.

Centering a sub-exponential random variable results in another sub-exponential random variable.

Lemma 12 (Vershynin (2018) Exercise 2.7.10)**.** *If Y is sub-exponential with sub-exponential norm a_0 then $Y - \mathbb{E}Y$ is also sub-exponential, with sub-exponential norm a satisfying*

$$a \le Ca_0$$

for some absolute constant $C > 0$.

A useful fact is that a product of sub-Gaussian random variables is sub-exponential.

Lemma 13 (Vershynin (2018) Lemma 2.7.7)**.** *Let Y_1 and Y_2 be sub-Gaussian random variables with sub-Gaussian norms s_1 and s_2 respectively. Then their product $Y_1 Y_2$ is sub-exponential with sub-exponential norm a satisfying $a \leq s_1 s_2$.*

Finally, sub-exponential random variables obey a useful quantitative form of a law or large numbers, which is a form of a Bernstein inequality.

Theorem 14 (Bernstein's inequality, Corollary 2.8.3 of Vershynin (2018))**.** *Let $Y_1, \ldots, Y_d$ be a collection of random variables and write $\mu_i = \mathbb{E} Y_i$ for $i = 1, \ldots, d$. Suppose $Y_1 - \mu_i, \ldots, Y_d - \mu_d$ are independent sub-exponential random variables with sub-exponential norms $a_1, \ldots a_d$. Then, for every $r \geq 0$,*

$$\mathbb{P}\Big(\big| \frac{1}{d} \sum_{i=1}^{d} (Y_i - \mu_i) \big| \geq r \Big) \leq 2 \exp\Big(-cdM \Big),$$

where $M = \min\left\{ \frac{r^2}{\max_i a_i^2}, \frac{r}{\max_i a_i} \right\}$ and $c > 0$ is an absolute constant.

E Analysis

The stochastic gradient $\frac{\partial}{\partial \phi} L\big(\phi; X, \mathsf{V}\big)$ evaluated at an arbitrary point $\phi \in \boldsymbol{\psi}$ for input X and random V is after scaling the sum of the gradient of the negative log prior and the stochastic gradient of the negative log likelihood,

$$\frac{\partial}{\partial \phi} L\big(\phi; X, \mathsf{V}\big) = \underbrace{\sqrt{\frac{m}{d}}\lambda\phi}_{\text{Gradient of negative log prior}} \underbrace{- \frac{1}{d}\mathsf{V}^\top\big(T\big(\Gamma(X)\big) \odot \rho(\mathsf{V}\phi) - \sigma(\mathsf{V}\phi)\big)}_{\text{Stochastic gradient of log likelihood}}. \tag{20}$$

In order to prove Theorem 4, we will need to prove a series of lemmas. The intuition behind these lemmas is as follows. Assumption 2 means that if the limit were allowed to be applied, the gradient of the negative log prior term in equation 20 multiplied by the step size would look like ϕ. This means that the update of SGD would just be the stochastic gradient of the log likelihood. We then examine the inner product of the stochastic gradient of the log likelihood, which would be the kernel update rule. The series of Lemmas is then as follows. We first convert the inner product of the stochastic gradient of the log likelihood to an approximate form that is easier to deal with (Lemma 15). We then confirm that the kernel update only involves the inner product of the stochastic gradients of the log likelihood (Lemma 16). Finally, we show that the inner products of the approximate form converges to a closed form update rule G (Lemma 17). Assembling these lemmas together yields Theorem 4.

To this end, define the kernel

$$k_d(X_1, X_2; \phi_1, \phi_2) \triangleq \frac{1}{dm\lambda^2}\big(T\big(\Gamma(X_1)\big) \odot \rho\big(\mathsf{V}\phi_1\big) - \sigma(\mathsf{V}\phi_1)\big)^\top \mathsf{V}\mathsf{V}^\top\big(T\big(\Gamma(X_2)\big) \odot \rho\big(\mathsf{V}\phi_2\big) - \sigma(\mathsf{V}\phi_2)\big),$$

which is a scaled inner product of the gradient of the negative log likelihood evaluated at inputs X_1 and X_2 and arbitrary points ϕ_1 and ϕ_2. The factor $\frac{1}{m}\mathsf{V}\mathsf{V}^\top \in \mathbb{R}^{d\times d}$ is approximately the identity matrix for large m under Assumption 1, leading to an easier to deal with approximation $\widetilde{k}_d(X_1, X_2; \phi_1, \phi_2)$ for $k_d(X_1, X_2; \phi_1, \phi_2)$,

$$\widetilde{k}_d(X_1, X_2; \phi_1, \phi_2) \triangleq \frac{1}{d\lambda^2}\big(T\big(\Gamma(X_1)\big) \odot \rho\big(\mathsf{V}\phi_1\big) - \sigma(\mathsf{V}\phi_1)\big)^\top\big(T\big(\Gamma(X_2)\big) \odot \rho\big(\mathsf{V}\phi_2\big) - \sigma(\mathsf{V}\phi_2)\big).$$

Lemma 15 says that this approximation is exact in the infinite d limit, and quantifies the quality of this approximation when d is finite.

Lemma 15. *Let $\phi_1 \in \mathbb{R}^m$ and $\phi_2 \in \mathbb{R}^m$ be arbitrary and suppose Assumption 3 holds.*

(15A) *Under Assumption 1, $k_d(X_1, X_2; \phi_1, \phi_2)$ converges in probability to $\widetilde{k}_d(X_1, X_2; \phi_1, \phi_2)$.*

(15B) *Under Assumption 5, there exist constants $Q > 0$ and $c > 0$ such that for all $\delta > 0$ and $\epsilon_2 > 0$,*

$$\mathbb{P}\Big(\big|k_d(X_1, X_2; \phi_1, \phi_2) - \widetilde{k}_d(X_1, X_2; \phi_1, \phi_2)\big| \geq \frac{(K + \epsilon_2)}{\lambda^2}\big(2\epsilon_1 + \epsilon_1^2\big)\Big) \leq 2\exp\Big(-cdM\Big) + e^{-m\delta^2/2},$$

where $\epsilon_1 = \sqrt{\frac{d}{m}} + \delta$ and $M = \min\Big\{\frac{\epsilon_2^2}{Q^2}, \frac{\epsilon_2}{Q}\Big\}$.

Proof. We use the shorthand $\Gamma_1 = \Gamma(X_1)$ and $\Gamma_2 = \Gamma(X_2)$. We have

$$\begin{aligned}
&\big|k_d(X_1, X_2; \phi_1, \phi_2) - \widetilde{k}_d(X_1, X_2; \phi_1, \phi_2)\big| \\
&= \frac{1}{d\lambda^2}\big|\big(T(\Gamma_1) \odot \rho\big(\mathsf{V}\phi_1\big) - \sigma(\mathsf{V}\phi_1)\big)^\top\big(\frac{1}{m}\mathsf{V}\mathsf{V}^\top - \mathsf{I}\big)\big(T(\Gamma_2) \odot \rho\big(\mathsf{V}\phi_2\big) - \sigma(\mathsf{V}\phi_2)\big)\big| \\
&\leq \frac{1}{d\lambda^2}\Big\|T(\Gamma_1) \odot \rho\big(\mathsf{V}\phi_1\big) - \sigma(\mathsf{V}\phi_1)\Big\|\Big\|\frac{1}{m}\mathsf{V}\mathsf{V}^\top - \mathsf{I}\Big\|\Big\|T(\Gamma_2) \odot \rho\big(\mathsf{V}\phi_2\big) - \sigma(\mathsf{V}\phi_2)\Big\| \\
&\leq \frac{1}{d\lambda^2}\max_{\phi_1, \phi_2}\Big\{\Big\|T(\Gamma_1) \odot \rho\big(\mathsf{V}\phi_1\big) - \sigma(\mathsf{V}\phi_1)\Big\|^2, \Big\|T(\Gamma_2) \odot \rho\big(\mathsf{V}\phi_2\big) - \sigma(\mathsf{V}\phi_2)\Big\|^2\Big\}\Big\|\frac{1}{m}\mathsf{V}\mathsf{V}^\top - \mathsf{I}\Big\| \\
&= \frac{1}{\lambda^2}\max_{\hat{\phi} \in \{\phi_1, \phi_2\}}\Big(K_{\hat{\phi}} - \mathbb{E}[K_{\hat{\phi}}]\Big)\Big\|\frac{1}{m}\mathsf{V}\mathsf{V}^\top - \mathsf{I}\Big\| + \mathbb{E}[K_{\hat{\phi}}]\Big\|\frac{1}{m}\mathsf{V}\mathsf{V}^\top - \mathsf{I}\Big\|,
\end{aligned} \tag{32}$$

where $K_{\hat{\phi}} = \frac{1}{d}\sum_{i=1}^{d}\Big(T\big(\gamma_i(X_1)\big) \odot \rho\big(V_i^\top \hat{\phi}\big) - \sigma(V_i^\top \hat{\phi})\Big)^2$ and $V_i^\top$ is the ith row of V. The quantity $\mathbb{E}[K_{\hat{\phi}}]$ is finite by Assumption 3.

Using a standard result (Wainwright, 2019, Example 6.2), we have that

$$\mathbb{P}\Bigg(\Big\|\mathsf{I} - \frac{1}{m}\mathsf{V}\mathsf{V}^\top\Big\| \geq 2\epsilon_1 + \epsilon_1^2\Bigg) \leq e^{-m\delta^2/2}, \quad \epsilon_1 = \sqrt{\frac{d}{m}} + \delta. \tag{33}$$

Combining equation 32 and equation 33 and taking $d \to \infty$ under Assumption 1, we have **(15A)**.

We may apply a Bernstein concentration inequality to $K_{\hat{\phi}} - \mathbb{E}[K_{\hat{\phi}}]$ as follows. The variables $T\big(\gamma_i(X_1)\big)\rho\big(V_i^\top \psi\big) - \sigma(V_i^\top \psi)$ for each i are mutually independent. The quantities $V_i^\top \hat{\psi}$ are zero-mean Gaussian (since Gaussian random variables are closed under linear combinations). By Assumption 5, each variable in the sum contains sub-Gaussian elements since bounded random variables are sub-Gaussian (Lemma 9), and Lipschitz functions of Gaussian random variables are sub-Gaussian (Lemma 10). The square of sub-Gaussian random variables is sub-exponential (Lemma 13). Sub-exponential random variables that are centered by subtracting their mean are also sub-exponential (Lemma 12). Therefore, by Bernstein's Theorem (Theorem 14), there exist constants $c, Q > 0$ (depending on ρ, σ, X_1 and X_2) such that for every $\epsilon_2 \geq 0$,

$$\mathbb{P}\Big(\max_{\hat{\phi}\in\{\phi_1,\phi_2\}} \big|K_{\hat{\phi}} - \mathbb{E}K_{\hat{\phi}}\big| \geq \epsilon_2\Big) \leq 2\exp\Big(-cdM\Big), \tag{34}$$

where $M = \min\Big\{\frac{\epsilon_2^2}{Q^2}, \frac{\epsilon_2}{Q}\Big\}$.

Finally, combining equation 32, equation 33 and equation 34 via a union bound, we have

$$\mathbb{P}\Big(\big|k_d(X_1, X_2; \phi_1, \phi_2) - \widetilde{k}_d(X_1, X_2; \phi_1, \phi_2)\big| \leq \frac{(K + \epsilon_2)}{\lambda^2}\big(2\epsilon_1 + \epsilon_1^2\big)\Big) \geq 1 - 2\exp\Big(-cdM\Big) - e^{-m\delta^2/2}.$$

□

We now confirm that the kernel update rule only involves the inner product of the stochastic gradient of the log likelihood, and not the gradient of the log prior.

Lemma 16. *Suppose Assumptions 1, 2 (a) and 3 hold. Then applying SGD to objective equation 17,*

$$\Psi_{ij}^{(t+1)} = \lim_{d\to\infty} k_d(X_i, X_j; \psi_{X_i}^{(t)}, \psi_{X_j}^{(t)}) = \operatorname*{plim}_{d\to\infty} \widetilde{k}_d(X_i, X_j; \psi_{X_i}^{(t)}, \psi_{X_j}^{(t)})$$

Proof. Combining the SGD update equation 15 and the stochastic gradient equation 20, we have that for input X, the $t+1$th iterate of SGD satisfies

$$\psi_X^{(t+1)} = \psi_X^{(t)}\big(1 - \alpha^{(t)}\sqrt{\frac{m}{d}}\lambda\big) + \alpha^{(t)}\frac{1}{d}\mathsf{V}^{(t)^\top}\big(T\big(\Gamma(X)\big) \odot \rho(\mathsf{V}^{(t)}\psi_X^{(t)}) - \sigma(\mathsf{V}^{(t)}\psi_X^{(t)})\big)$$

$$\psi_X^{(t+1)} - \psi_X^{(t)}\big(1 - \alpha^{(t)}\sqrt{\frac{m}{d}}\lambda\big) = \alpha^{(t)}\frac{1}{d}\mathsf{V}^{(t)^\top}\big(T\big(\Gamma(X)\big) \odot \rho(\mathsf{V}^{(t)}\psi_X^{(t)}) - \sigma(\mathsf{V}^{(t)}\psi_X^{(t)})\big) \tag{35}$$

Evaluating equation 35 at input X_i and X_j, and taking inner products, we find

$$\Big(\psi_{X_i}^{(t+1)} - \psi_{X_i}^{(t)}\big(1 - \alpha^{(t)}\sqrt{\frac{m}{d}}\lambda\big)\Big)^\top \Big(\psi_{X_j}^{(t+1)} - \psi_{X_j}^{(t)}\big(1 - \alpha^{(t)}\sqrt{\frac{m}{d}}\lambda\big)\Big)$$
$$= \alpha^{(t)^2}\frac{1}{d^2}\big(T\big(\Gamma(X_i)\big) \odot \rho(\mathsf{V}^{(t)}\psi_{X_i}^{(t)}) - \sigma(\mathsf{V}^{(t)}\psi_{X_i}^{(t)})\big)^\top \mathsf{V}^{(t)}\mathsf{V}^{(t)^\top}\big(T\big(\Gamma(X_j)\big) \odot \rho(\mathsf{V}^{(t)}\psi_{X_j}^{(t)}) - \sigma(\mathsf{V}^{(t)}\psi_{X_j}^{(t)})\big)$$

Invoking Assumption 2, we see that the left hand side satisfies

$$\begin{aligned}
&\lim_{d\to\infty}\Big(\psi_{X_i}^{(t+1)} - \psi_{X_i}^{(t)}\big(1-\alpha^{(t)}\sqrt{\frac{m}{d}}\lambda\big)\Big)^\top\Big(\psi_{X_j}^{(t+1)} - \psi_{X_j}^{(t)}\big(1-\alpha^{(t)}\sqrt{\frac{m}{d}}\lambda\big)\Big)\\
&= \lim_{d\to\infty}\psi_{X_i}^{(t+1)^\top}\psi_{X_j}^{(t+1)} + \psi_{X_i}^{(t)^\top}\psi_{X_j}^{(t)}\underbrace{\big(1-\alpha^{(t)}\sqrt{\frac{m}{d}}\lambda\big)^2}_{\to 0}\\
&\qquad - \underbrace{\big(1-\alpha^{(t)}\sqrt{\frac{m}{d}}\lambda\big)}_{\to 0}\Big(\|\psi_{X_i}^{(t+1)}\|\|\psi_{X_j}^{(t)}\|a_1 + \|\psi_{X_i}^{(t)}\|\|\psi_{X_j}^{(t+1)}\|a_2\Big)\\
&= \Psi_{ij}^{(t+1)}
\end{aligned}$$

where a_1 and a_2 are cosine angles belonging to $[-1,1]$. On the other hand, under Assumption 2 the right hand side satisfies

$$\begin{aligned}
&\lim_{d\to\infty}\alpha^{(t)^2}\frac{1}{d^2}\big(T\big(\Gamma(X_i)\big)\odot\rho(\mathsf{V}^{(t)}\psi_{X_i}^{(t)}) - \sigma(\mathsf{V}^{(t)}\psi_{X_i}^{(t)})\big)^\top\mathsf{V}^{(t)}\mathsf{V}^{(t)^\top}\big(T\big(\Gamma(X_j)\big)\odot\rho(\mathsf{V}^{(t)}\psi_{X_j}^{(t)}) - \sigma(\mathsf{V}^{(t)}\psi_{X_j}^{(t)})\big)\\
&= \lim_{d\to\infty}\frac{1}{dm\lambda^2}\big(T\big(\Gamma(X_i)\big)\odot\rho(\mathsf{V}^{(t)}\psi_{X_i}^{(t)}) - \sigma(\mathsf{V}^{(t)}\psi_{X_i}^{(t)})\big)^\top\mathsf{V}^{(t)}\mathsf{V}^{(t)^\top}\big(T\big(\Gamma(X_j)\big)\odot\rho(\mathsf{V}^{(t)}\psi_{X_j}^{(t)}) - \sigma(\mathsf{V}^{(t)}\psi_{X_j}^{(t)})\big)\\
&= \lim_{d\to\infty}k_d(X_i,X_j;\psi_{X_i}^{(t)},\psi_{X_j}^{(t)}).
\end{aligned}$$

By Lemma 15, this limit is well defined and is given by $\operatorname*{plim}_{d\to\infty}\widetilde{k}_d(X_i,X_j;\psi_{X_i}^{(t)},\psi_{X_j}^{(t)})$. □

Finally, we show that the approximate form of inner products $\widetilde{k}_d$ converges to a closed form update rule G.

Lemma 17. *Suppose Assumptions 1, 3 and 4 hold. Then*

$$\operatorname*{plim}_{d\to\infty}\widetilde{k}_d(X_i,X_j;\psi_{X_i}^{(t)},\psi_{X_j}^{(t)}) = G(\Psi_{ii}^{(t)},\Psi_{jj}^{(t)},\Psi_{ij}^{(t)};X_i,X_j),$$

where

$$\begin{aligned}
&G(\Psi_{ii}^{(t)},\Psi_{jj}^{(t)},\Psi_{ij}^{(t)};X_i,X_j)\\
&= \frac{1}{\lambda^2}\Bigg(C_{ij}\kappa_\rho\big(\Psi_{ii}^{(t)},\Psi_{jj}^{(t)},\Psi_{ij}^{(t)}\big) - \kappa_{\sigma,\rho}\big(\Psi_{ii}^{(t)},\Psi_{jj}^{(t)},\Psi_{ij}^{(t)}\big)\mu_i - \kappa_{\rho,\sigma}\big(\Psi_{ii}^{(t)},\Psi_{jj}^{(t)},\Psi_{ij}^{(t)}\big)\mu_j + \kappa_\sigma\big(\Psi_{ii}^{(t)},\Psi_{jj}^{(t)},\Psi_{ij}^{(t)}\big)\Bigg).
\end{aligned}$$

Proof. We have

$$\begin{aligned}
&\operatorname*{plim}_{d\to\infty}\widetilde{k}_d(X_i,X_j;\psi_{X_i}^{(t)},\psi_{X_j}^{(t)})\\
&= \operatorname*{plim}_{d\to\infty}\frac{1}{d\lambda^2}\big(T\big(\Gamma(X_i)\big)\odot\rho\big(\mathsf{V}^{(t)}\psi_{X_i}^{(t)}\big) - \sigma(\mathsf{V}^{(t)}\psi_{X_i}^{(t)})\big)^\top\big(T\big(\Gamma(X_j)\big)\odot\rho\big(\mathsf{V}^{(t)}\psi_{X_j}^{(t)}\big) - \sigma(\mathsf{V}^{(t)}\psi_{X_j}^{(t)})\big)
\end{aligned}$$

Expanding the quadratic, we find

$$\begin{aligned}
\Psi_{ij}^{(t+1)} &= \operatorname*{plim}_{d\to\infty}\frac{1}{d\lambda^2}\big(T\big(\Gamma(X_i)\big)\odot\rho\big(\mathsf{V}^{(t)}\psi_{X_i}^{(t)}\big) - \sigma(\mathsf{V}^{(t)}\psi_{X_i}^{(t)})\big)^\top\big(T\big(\Gamma(X_j)\big)\odot\rho\big(\mathsf{V}^{(t)}\psi_{X_j}^{(t)}\big) - \sigma(\mathsf{V}^{(t)}\psi_{X_j}^{(t)})\big)\\
&= \operatorname*{plim}_{d\to\infty}\frac{1}{d\lambda^2}\Big(\big(T\big(\Gamma(X_i)\big)\odot\rho(\mathsf{V}^{(t)}\psi_{X_i}^{(t)})\big)^\top\big(T\big(\Gamma(X_j)\big)\odot\rho(\mathsf{V}^{(t)}\psi_{X_j}^{(t)})\big)-\\
&\big(T\big(\Gamma(X_i)\big)\odot\rho(\mathsf{V}^{(t)}\psi_{X_i}^{(t)})\big)^\top\sigma(\mathsf{V}^{(t)}\psi_{X_j}^{(t)})-\\
&\big(T\big(\Gamma(X_j)\big)\odot\rho(\mathsf{V}^{(t)}\psi_{X_j}^{(t)})\big)^\top\sigma(\mathsf{V}^{(t)}\psi_{X_i}^{(t)})+\\
&\sigma(\mathsf{V}^{(t)}\psi_{X_i}^{(t)})^\top\sigma(\mathsf{V}^{(t)}\psi_{X_j}^{(t)})\Big)
\end{aligned}$$

The collection of d pairs $\left\{\left(\mathsf{V}^{(t)}\psi^{(t)}_{X_i}, \mathsf{V}^{(t)}\psi^{(t)}_{X_j}\right)_p\right\}_{p=1}^d$ is mutually independent and Gaussian given $\psi^{(t)}_{X_i}$ and $\psi^{(t)}_{X_j}$. Letting $V^\top \in \mathbb{R}^m$ be equal in distribution to a row of $\mathsf{V}^{(t)}$, a law of large numbers says that

$$\Psi^{(t+1)}_{ij} = \frac{1}{\lambda^2}\Big(c(X_i,X_j)\mathbb{E}_{V^\top}\Big[\lim_{d\to\infty}\rho(V^\top\psi^{(t)}_{X_i})\rho(V^\top\psi^{(t)}_{X_j})\Big] -$$

$$\mu(X_i)\mathbb{E}_{V^\top}\Big[\lim_{d\to\infty}\rho(V^\top\psi^{(t)}_{X_i})\sigma(V^\top\psi^{(t)}_{X_j})\Big] -$$

$$\mu(X_j)\mathbb{E}_{V^\top}\Big[\lim_{d\to\infty}\rho(V^\top\psi^{(t)}_{X_j})\sigma(V^\top\psi^{(t)}_{X_i})\Big] +$$

$$\mathbb{E}_{V^\top}\Big[\lim_{d\to\infty}\sigma(V^\top\psi^{(t)}_{X_i})\sigma(V^\top\psi^{(t)}_{X_j})\Big]\Big).$$

Now observe that conditional on $\psi^{(t)}_{X_i}$ and $\psi^{(t)}_{X_j}$, the random vector $(\chi,\chi')^\top = \lim_{d\to\infty}\left(V^\top\psi^{(t)}_{X_i}, V^\top\psi^{(t)}_{X_j}\right)^\top$ is bivariate Gaussian with mean 0 and covariance matrix

$$\begin{pmatrix}\Psi^{(t)}_{ii} & \Psi^{(t)}_{ij}\\ \Psi^{(t)}_{ij} & \Psi^{(t)}_{jj}\end{pmatrix}.$$

Therefore, the terms on the right hand side involve evaluations of the functions $\kappa_{\rho,\rho}, \kappa_{\rho,\sigma}, \kappa_{\sigma,\rho}$ and $\kappa_{\sigma\sigma}$. □

Chaining Lemmas 16 and 17, we obtain our main theorem.

Theorem 4. *Suppose Assumptions 1, 2 (a), 3, and 4 hold. Let $C_{ij} = c(X_i,X_j)$ and $\mu_i = \mu(X_i)$ be as defined in Assumption 4. Then applying SGD to objective equation 17, the update rule* G *equation 3 exists and can be decomposed into G equation 5 satisfying*

$$G(\Phi_{ii},\Phi_{jj},\Phi_{ij};X_i,X_j)$$

$$= \frac{1}{\lambda^2}\Bigg(C_{ij}\kappa_\rho\big(\Phi_{ii},\Phi_{jj},\Phi_{ij}\big) - \kappa_{\sigma,\rho}\big(\Phi_{ii},\Phi_{jj},\Phi_{ij}\big)\mu_i - \kappa_{\rho,\sigma}\big(\Phi_{ii},\Phi_{jj},\Phi_{ij}\big)\mu_j + \kappa_\sigma\big(\Phi_{ii},\Phi_{jj},\Phi_{ij}\big)\Bigg).$$

Here κ_σ, κ_ρ, $\kappa_{\sigma,\rho}$ and $\kappa_{\rho,\sigma}$ are as defined by equation 11, equation 19 and Proposition 2.

Proof. By Lemma 16, under Assumptions 1, 2 (a) and 3, we have

$$\Psi^{(t+1)}_{12} = \lim_{d\to\infty} k_d(X_1,X_2;\psi^{(t)}_{X_1},\psi^{(t)}_{X_2}) = \operatorname*{plim}_{d\to\infty}\widetilde{k}_d(X_1,X_2;\psi^{(t)}_{X_1},\psi^{(t)}_{X_2}).$$

By Lemma 17, under Assumptions 1, 3 and 4 we have

$$\operatorname*{plim}_{d\to\infty}\widetilde{k}_d(X_1,X_2;\psi^{(t)}_{X_1},\psi^{(t)}_{X_2}) = G(\Psi^{(t)}_{ii},\Psi^{(t)}_{jj},\Psi^{(t)}_{ij};X_i,X_j).$$

□

The effect of the finite approximation is not described in Theorem 4. In order to describe the effect of finite approximations, we combine the previously proven Lemma 15B with the following Lemma 18, assembling them in Theorem 7.

We now turn to analysing the sensitivity of the ffDEKer when initialised around the ℓDEKer. For this, we combine Lemma 15 with the following lemma.

Lemma 18. *Suppose Assumptions 1,2 (a), 3, 4 and 5 hold. Then for all $\epsilon > 0$ there exists some $Q > 0$ (depending on ρ, σ, X_i, and X_j) such that*

$$\mathbb{P}\Big(\big|\widetilde{k}_d(X_1,X_2;r,r') - \Psi_{12}\big| \geq \epsilon\Big) \leq 2\exp\Big(-dcM\Big),$$

where $M = \min\left\{\frac{\epsilon^2}{Q^2}, \frac{\epsilon}{Q}\right\}$ and $c > 0$ is some absolute constant.

Proof. We first write $\widetilde{k}_d$ as a sum. Letting $V_i^\top \in \mathbb{R}^m$ denote the ith row of V and $\gamma_i(X)$ denote the ith coordinate of $\Gamma(X)$,

$$\widetilde{k}_d(X_1, X_2; r, r') = \frac{1}{d\lambda^2} \sum_{i=1}^{d} \big(T\big(\gamma_i(X_1)\big)\rho\big(V_i^\top r\big) - \sigma(V_i^\top r)\big)\big(T\big(\gamma_i(X_2)\big)\rho\big(V_i^\top r'\big) - \sigma(V_i^\top r')\big).$$

In this form, we observe that $\mathbb{E}\widetilde{k}_d(X_1, X_2; r, r') = G(\Psi_{11}, \Psi_{22}, \Psi_{12}; X_1, X_2) = \Psi_{12}$. We therefore seek to concentrate $\widetilde{k}_d(X_1, X_2; r, r')$ about its mean.

The bivariate pairs $\Big(T\big(\gamma_i(X_1)\big)\rho\big(V_i^\top r\big) - \sigma(V_i^\top r),\ T\big(\gamma_i(X_2)\big)\rho\big(V_i^\top r'\big) - \sigma(V_i^\top r')\Big)$ are independent from every other pair. The quantities $V_i^\top r$ and $V_i^\top r'$ are zero-mean Gaussian (since Gaussian random variables are closed under linear combinations). Each pair contains sub-Gaussian elements since bounded random variables are sub-Gaussian (Lemma 9), and Lipschitz functions of Gaussian random variables are sub-Gaussian (Lemma 10). The product of two sub-Gaussian random variables is sub-exponential (Lemma 13). Sub-exponential random variables that are centered by subtracting their mean are also sub-exponential (Lemma 12). Therefore, by Bernstein's Theorem (Theorem 14), there exist constants $c, Q > 0$ (depending on ρ, σ, X_i, and X_j) such that for every $\epsilon \geq 0$,

$$\mathbb{P}\Big(\big|\widetilde{k}_d(X_1, X_2; r, r') - \Psi_{12}\big| \geq \epsilon\Big) \leq 2\exp\Big(-cdM\Big),$$

where $M = \min\left\{\frac{\epsilon^2}{Q^2}, \frac{\epsilon}{Q}\right\}$. □

Theorem 7. *Suppose Assumptions 1, 2 (b), 3, 4 and 5 hold. Let initial guesses be $\psi_{X_1}^{(0)} = r_1$ and $\psi_{X_2}^{(0)} = r_2$ as in Definition 6. Then there exist constants $Q_2, Q_3, c_2, c_3 > 0$ such that for all $\delta > 0$, $\epsilon_2 > 0$ and ε_2,*

$$\mathbb{P}\Big(\big|\overline{\Psi}_{12}^{(1)} - \Psi_{12}\big| \leq \varepsilon_1 + \varepsilon_2\Big) \geq 1 - \delta_1 - \delta_2,$$

where

$$\varepsilon_1 = \frac{K + \epsilon_2}{\lambda^2}(2\epsilon_1 + \epsilon_1^2), \quad \delta_1 = 2\exp\Big(-c_2 d M_2\Big) + \exp\big(-m\delta^2/2\big) \quad \textit{and} \quad \delta_2 = 2\exp\big(-dc_3 M_3\big)$$

and $\epsilon_1 = \sqrt{\frac{d}{m}} + \delta$, $M_2 = \min\left\{\frac{\epsilon_2^2}{Q_2^2}, \frac{\epsilon_2}{Q_2}\right\}$ and $M_3 = \min\left\{\frac{\varepsilon_2^2}{Q_3^2}, \frac{\varepsilon_2}{Q_3}\right\}$ and $c_3 > 0$ is some absolute constant.

Proof. Under Assumption 2 (b), plugging the stochastic gradient equation 20 into the SGD update rule equation 35, we have

$$\begin{aligned}
\overline{\Psi}_{ij}^{(t+1)} &= \frac{1}{\lambda^2 dm}\big(T\big(\Gamma(X_i)\big) \odot \rho(\mathsf{V}^{(t)}\psi_{X_i}^{(t)}) - \sigma(\mathsf{V}^{(t)}\psi_{X_i}^{(t)})\big)^\top \mathsf{V}^{(t)}\mathsf{V}^{(t)\top}\big(T\big(\Gamma(X_i)\big) \odot \rho(\mathsf{V}^{(t)}\psi_{X_j}^{(t)}) - \sigma(\mathsf{V}^{(t)}\psi_{X_j}^{(t)})\big) \\
&= k_d(X_i, X_j; \psi_{X_i}^{(t)}, \psi_{X_j}^{(t)}) \\
\overline{\Psi}_{ij}^{(1)} &= k_d(X_i, X_j; \psi_{X_i}^{(0)}, \psi_{X_j}^{(0)}).
\end{aligned}$$

By Lemma 15, we may approximate $k_d(X_i, X_j; \psi_{X_i}^{(0)}, \psi_{X_j}^{(0)})$ by $\widetilde{k}_d(X_i, X_j; \psi_{X_i}^{(0)}, \psi_{X_j}^{(0)})$ with high probability. By Lemma 18, we may approximate $\widetilde{k}_d(X_i, X_j; \psi_{X_i}^{(0)}, \psi_{X_j}^{(0)})$ by Ψ_{ij} with high probability. The proof then follows by applying a triangle inequality and a union bound.

In more detail, the triangle inequality says

$$\big|\overline{\Psi}_{12}^{(1)} - \Psi_{12}\big| \leq \big|k_d(X_i, X_j; \psi_{X_i}^{(0)}, \psi_{X_j}^{(0)}) - \widetilde{k}_d(X_i, X_j; \psi_{X_i}^{(0)}, \psi_{X_j}^{(0)})\big| + \big|\widetilde{k}_d(X_i, X_j; \psi_{X_i}^{(0)}, \psi_{X_j}^{(0)}) - \Psi_{ij}\big|. \quad (36)$$

Let $\delta > 0$ and $\epsilon_2 > 0$ be arbitrary. Define $\epsilon_1 = \sqrt{\frac{d}{m}} + \delta$. Define $\varepsilon_1 = \frac{K+\epsilon_2}{\lambda^2}(2\epsilon_1 + \epsilon_1^2)$. Let A_1 denote the event that $\big|k_d(X_i, X_j; \psi_{X_i}^{(0)}, \psi_{X_j}^{(0)}) - \widetilde{k}_d(X_i, X_j; \psi_{X_i}^{(0)}, \psi_{X_j}^{(0)})\big| \geq \varepsilon_1$. Then by Lemma 15 there exists some constants $Q_2 > 0$ and $c_2 > 0$ such that

$$\mathbb{P}(A_1) \leq \delta_1, \tag{37}$$

where $\delta_1 = 2\exp\Big(-c_2 d M_2\Big) + \exp\big(-m\delta^2/2\big)$ and $M_2 = \min\big\{\frac{\epsilon_2^2}{Q_2^2}, \frac{\epsilon_2}{Q_2}\big\}$.

Let $\varepsilon_2 > 0$ be arbitrary. Let A_2 denote the event that $\big|\widetilde{k}_d(X_i, X_j; \psi_{X_i}^{(0)}, \psi_{X_j}^{(0)}) - \Psi_{ij}\big| \geq \varepsilon_2$. Then by Lemma 18, there exists some $Q_3 > 0$ such that

$$\mathbb{P}(A_2) \leq \delta_2, \tag{38}$$

where $\delta_2 = 2\exp\big(-dc_3M_3\big)$, $M_3 = \min\big\{\frac{\varepsilon_2^2}{Q_3^2}, \frac{\varepsilon_2}{Q_3}\big\}$ and $c_3 > 0$ is an absolute constant.

Combining equation 37 and equation 38 by a union bound, we obtain

$$\mathbb{P}(A_1 \cup A_2) \leq \delta_1 + \delta_2$$
$$\mathbb{P}(A_1^{\complement} \cap A_2^{\complement}) \geq 1 - \delta_1 - \delta_2$$

Finally, if $A_1^{\complement} \cap A_2^{\complement}$ then by equation 36, $\big|\overline{\Psi}_{12}^{(1)} - \Psi_{12}\big| \leq \varepsilon_1 + \varepsilon_2$. □

F Random finite forms for the explicit kernel c and mean function μ

Suppose the sufficient statistic T is the identity.

Linear kernel. Let $\Gamma(X_1) = \mathsf{Q}X$, where each entry of $\mathsf{Q} \in \mathbb{R}^{d \times l}$ is sampled i.i.d. from $\mathcal{N}(0, v^2)$. Then we obtain the linear kernel,

$$\begin{aligned} c(X_1, X_2) &= v^2 \lim_{d \to \infty} \frac{1}{d} X_1^\top \mathsf{Q}^\top \mathsf{Q} X_2 \\ &= v^2 X_1^\top X_2. \end{aligned}$$

In this case, $\mu(X_1) = 0$. since $\frac{1}{d}\sum_{i=1}^d \mathbb{E}[Q_i^\top X_1] = \frac{1}{d}\mathbb{E}[Q_i]^\top X_1 = 0$, where $Q_i^\top$ is the ith row of Q.

Squared exponential kernel. We may obtain stationary nonlinear kernels via a random Fourier feature type construction (Rahimi & Recht, 2007). Suppose d is even and $\mathsf{Q} \in \mathbb{R}^{d/2 \times l}$ is sampled i.i.d. from $\mathcal{N}(0, 1)$. Define

$$\Gamma(X_1) = A \odot \begin{pmatrix} \cos(\mathsf{Q}X_1) \\ \sin(\mathsf{Q}X_1) \end{pmatrix} \in \mathbb{R}^{d \times 1},$$

where elements of $A = (a_1, \ldots, a_{d/2}, b_1, \ldots, b_{d/2})^\top \in \mathbb{R}^{d \times 1}$ are sampled i.i.d. from $\mathcal{N}(\mu_v, v^2)$. Then $\mathbb{E}[a_i^2] = v^2 + \mu_v^2$ and

$$\begin{aligned} c(X_1, X_2) &= \lim_{d \to \infty} \frac{1}{d} \sum_{i=1}^{d/2} a_i^2 \cos(Q_i^\top X_1)\cos(Q_i^\top X_2) + \frac{1}{d}\sum_{j=1}^{d/2} b_j^2 \sin(Q_j^\top X_1)\sin(Q_j^\top X_2) \\ &= \frac{v^2 + \mu_v^2}{2} \mathbb{E}\big[\cos(Q^\top X_1)\cos(Q^\top X_2) + \sin(Q^\top X_1)\sin(Q^\top X_2)\big], \quad Q \sim \mathcal{N}(0, \mathsf{I}) \\ &= \frac{v^2 + \mu_v^2}{2} \mathbb{E}\big[\cos\big(Q^\top (X_1 - X_2)\big)\big] \\ &= \frac{v^2 + \mu_v^2}{2} \exp\big(-\frac{1}{2}\|X_1 - X_2\|_2^2\big). \end{aligned}$$

An extension to arbitrary stationary kernels follows using Bochner's theorem to define the probability measure of Q via a Fourier transform (Rahimi & Recht, 2007). An extension to arbitrary covariance structures can be obtained by introducing a dependency structure among elements of rows of Q.

The reason that A is introduced is to allow the mean to converge to zero, so that $\mu(X) = 0$ can be realised. That is,

$$\begin{aligned} \mu(X_1) &= \lim_{d \to \infty} \frac{1}{d} \sum_{i=1}^{d/2} a_i \cos(Q_i^\top X_1) + \frac{1}{d} \sum_{j=1}^{d/2} b_j \sin(Q_j^\top X_1) \\ &= \frac{\mu_v}{2} \mathbb{E}\big[\cos(Q^\top X_1)\big] \\ &= \frac{\mu_v}{2} \exp\big(-\frac{1}{2}\|X_1\|^2\big) \end{aligned}$$

is zero whenever $\mu_v = 0$.

F.1 A model we found to be practically useful

Let $\Gamma(X_1) = A \cdot \mathrm{ReLU}(\mathsf{Q}X_1)$, where elements of A are sampled i.i.d. from $\mathcal{N}(\mu_v, v^2)$. Then c is the arc-cosine kernel of degree 1 (Cho & Saul, 2009) (see equation 47),

$$c(X_1, X_2) = (v^2 + \mu_v^2) \frac{\|X_1\|\|X_2\|}{2\pi} \big(\sin\theta + (\pi - \theta)\cos\theta\big), \quad \theta = \arccos \frac{X_1^\top X_2}{\|X_1\|\|X_2\|}.$$

The mean function is given by

$$\begin{aligned}\mu(X_1) &= \mu_v \mathbb{E}\big[\operatorname{ReLU}(\|X_1\| Z)\big], \quad Z \sim \mathcal{N}(0,1) \\ &= \frac{\mu_v}{2}\|X_1\|\sqrt{\frac{2}{\pi}}. \end{aligned} \tag{39}$$

Again we may take $\mu_v = 0$ to realise $\mu(X) = 0$. However, in order to construct a model that is statistically not mis-specified, when using $\sigma = \operatorname{ReLU}$ and $\rho = \mathrm{u}$ it is useful to consider the case where μ_v is non-zero (say 1). Otherwise, the model tries to describe symmetric observations $\Gamma(X_1) = A \cdot \operatorname{ReLU}(\mathsf{Q}X_1)$ that are equally likely negative or positive as a Gaussian distribution with a skewed non-negative mean. In order to handle non-zero μ_v, we require evaluating additional cross-terms, given by

$$\begin{aligned}\kappa_{\sigma,\rho}\big(\Psi_{11}, \Psi_{22}, \Psi_{12}\big) &= \mathbb{E}_{(\chi,\chi')^\top \sim \mathcal{N}(\mathbf{0},\Psi)}\big[\operatorname{ReLU}\big(\chi\big)\,\mathrm{u}\big(\chi'\big)\big] \\ &= \mathbb{E}_{(\chi,\chi')^\top \sim \mathcal{N}(\mathbf{0},\Psi)}\big[\chi\,\mathrm{u}(\chi)\,\mathrm{u}\big(\chi'\big)\big] \\ &= \Psi_{11}\mathbb{E}\big[\delta(\chi)\,\mathrm{u}(\chi')\big] + \Psi_{12}\mathbb{E}\big[\mathrm{u}(\chi)\delta(\chi')\big] \quad \text{(multivariate Stein's lemma)} \\ &= \Big(\Psi_{11}\mathbb{E}\big[\delta(\sqrt{\Psi_{11}}Z_1)\,\mathrm{u}\big(\sqrt{\Psi_{22}}(Z_1\cos\theta + Z_2\sin\theta)\big)\big] \\ &\qquad + \Psi_{12}\mathbb{E}\big[\mathrm{u}\big(\sqrt{\Psi_{11}}(Z_2\cos\theta + Z_1\sin\theta)\big)\delta(\sqrt{\Psi_{22}}Z_2)\big]\Big), \end{aligned} \tag{40}$$

where $(Z_1, Z_2)^\top \sim \mathcal{N}(0, \mathsf{I})$ and $\cos\theta = \frac{\Psi_{12}}{\sqrt{\Psi_{22}\Psi_{11}}}$. We have that

$$\begin{aligned}&\mathbb{E}\big[\delta(\sqrt{\Psi_{11}}Z_1)\,\mathrm{u}\big(\sqrt{\Psi_{22}}(Z_1\cos\theta + Z_2\sin\theta)\big)\big] \\ &= \frac{1}{2\pi}\int \exp\big(-\frac{1}{2}(z_1^2 + z_2^2)\big)\delta(\sqrt{\Psi_{11}}z_1)\,\mathrm{u}\big(\sqrt{\Psi_{22}}(z_1\cos\theta + z_2\sin\theta)\big)\,dz_1\,dz_2 \\ &= \frac{1}{2\pi\sqrt{\Psi_{11}}}\int \exp\big(-\frac{1}{2}z_2^2\big)\,\mathrm{u}\big(\sqrt{\Psi_{22}}z_2\sin\theta\big)\,dz_2 \\ &= \frac{1}{\sqrt{2\pi}\sqrt{\Psi_{11}}}\int \frac{1}{\sqrt{2\pi}}\exp\big(-\frac{1}{2}z_2^2\big)\,\mathrm{u}\big(z_2\big)\,dz_2 \\ &= \frac{1}{2\sqrt{2\pi}\sqrt{\Psi_{11}}}. \end{aligned} \tag{41}$$

Combining equation 39, equation 40 and equation 41, we have

$$\mu(X_1)\kappa_{\sigma,\rho}\big(\Psi_{11}, \Psi_{22}, \Psi_{12}\big) = \mu_v\frac{\sqrt{\Sigma_{11}}}{4\pi}\Big(\sqrt{\Psi_{11}} + \frac{\Psi_{12}}{\sqrt{\Psi_{22}}}\Big). \tag{42}$$

The terms involving κ_ρ and κ_σ are arc-cosine kernels and are given in equation 46 and equation 47.

G Examples

G.1 Error function example

We consider a special case where inputs are mapped to a Gaussian with a conditional expectation between -1 and 1 through the random mapping Γ. We then use a Gaussian likelihood with a choice of R that maps to values between -1 and 1. Equivalently, we use an inverse link function that maps to values between -1 and 1.

Let p the pdf of a univariate standard Gaussian. Suppose input X is mapped to data $Y = \operatorname{erf}(\mathsf{W}X/\sqrt{2}) + Q$ for some linear mapping $\mathsf{W} \in \mathbb{R}^{d\times l}$ and noise $Q \sim \mathcal{N}(0, \mathsf{I})$ each with elements drawn i.i.d. from a standard Gaussian. An appropriate model is then to let $A = \eta^2/2$ and $R(a) = \operatorname{erf}(a/\sqrt{2})$. Then $\rho(a) = 2p(a)$ and $\sigma(a) = 2p(a)\operatorname{erf}(a/\sqrt{2})$.

We now invoke the general Theorem 4. In the following, we compute the individual terms in the update rule. Recall from equation 11, that for a particular activation function ζ, a neural network kernel (NNK) is computed by taking the bivariate Gaussian expectation,

$$\kappa_\zeta\big(\Phi_{11}, \Phi_{22}, \Phi_{12}\big) = \mathbb{E}_{(\chi_1,\chi_2)^\top \sim \mathcal{N}(\mathbf{0},\mathsf{\Phi})}\big[\zeta\big(\chi_1\big)\zeta\big(\chi_2\big)\big], \quad \mathsf{\Phi} \triangleq \begin{pmatrix} \phi_1^\top \phi_1 & \phi_1^\top \phi_2 \\ \phi_2^\top \phi_1 & \phi_2^\top \phi_2 \end{pmatrix}.$$

It is helpful to write covariance matrices in terms of variances $\phi_1^\top \phi_1 = \Phi_{11}$, $\phi_2^\top \phi_2 = \Phi_{22}$ and covariances $\phi_1^\top \phi_2 = \sqrt{\Phi_{11}\Phi_{22}}\cos\theta$, where θ is the angle between ϕ_1 and ϕ_2. That is,

$$\mathsf{\Phi} = \begin{pmatrix} \Phi_{11} & \sqrt{\Phi_{11}\Phi_{22}}\cos\theta \\ \sqrt{\Phi_{11}\Phi_{22}}\cos\theta & \Phi_{22} \end{pmatrix}.$$

The resulting determinant and inverse then satisfy

$$\begin{aligned} \det \mathsf{\Phi} &= \Phi_{11}\Phi_{22}\sin^2\theta \\ \mathsf{\Phi}^{-1} &= \frac{1}{\Phi_{11}\Phi_{22}\sin^2\theta}\begin{pmatrix} \Phi_{22} & -\sqrt{\Phi_{11}\Phi_{22}}\cos\theta \\ -\sqrt{\Phi_{11}\Phi_{22}}\cos\theta & \Phi_{11} \end{pmatrix}. \end{aligned}$$

The NNK for factor activations A more general result is given in Tsuchida (2020, Proposition 20). For completeness, we reproduce the result here. For activation function $\rho(a) = 2p(a)$, we expand the 2D integral corresponding to the expectation for the NNK κ_ρ,

$$\begin{aligned} \kappa_\rho(\Phi_{11}, \Phi_{22}, \Phi_{12}) &= \frac{4}{2\pi}\int \exp\big(-\frac{1}{2}(a_1^2 + a_2^2)\big)\frac{1}{2\pi\sqrt{\Phi_{11}\Phi_{22}}\sin\theta}\exp\big(-\frac{1}{2}(a_1, a_2)\mathsf{\Phi}^{-1}(a_1, a_2)^\top\big)\, da_1\, da_2 \\ &= \frac{2}{\pi}\int \frac{1}{2\pi\sqrt{\Phi_{11}\Phi_{22}}\sin\theta}\exp\big(-\frac{1}{2}(a_1, a_2)(\mathsf{\Phi}^{-1} + \mathsf{I})(a_1, a_2)^\top\big)\, da_1\, da_2. \end{aligned} \tag{43}$$

We now complete the square inside the argument of exp, so that we may express the integrand of equation 43 as a product of a bivariate Gaussian pdf and a constant.

Letting $\mathsf{F}^{-1} = \mathsf{\Phi}^{-1} + \mathsf{I}$, we compute F as

$$\begin{aligned} \mathsf{F}^{-1} &= \frac{1}{\Phi_{11}\Phi_{22}\sin^2\theta}\begin{pmatrix} \Phi_{22}\big(1 + \Phi_{11}\sin^2\theta\big) & -\sqrt{\Phi_{11}\Phi_{22}}\cos\theta \\ -\sqrt{\Phi_{11}\Phi_{22}}\cos\theta & \Phi_{11}\big(1 + \Phi_{22}\sin^2\theta\big) \end{pmatrix} \\ \det \mathsf{F}^{-1} &= 1 + \det \mathsf{\Phi}^{-1} + \mathrm{Trace}\mathsf{\Phi}^{-1} \\ &= \frac{1 + \Phi_{11} + \Phi_{22} + \Phi_{11}\Phi_{22}\sin^2\theta}{\Phi_{11}\Phi_{22}\sin^2\theta} \\ \mathsf{F} &= \frac{1}{1 + \Phi_{11} + \Phi_{22} + \Phi_{11}\Phi_{22}\sin^2\theta}\begin{pmatrix} \Phi_{11}\big(1 + \Phi_{22}\sin^2\theta\big) & \sqrt{\Phi_{11}\Phi_{22}}\cos\theta \\ \sqrt{\Phi_{11}\Phi_{22}}\cos\theta & \Phi_{22}\big(1 + \Phi_{11}\sin^2\theta\big) \end{pmatrix}. \end{aligned}$$

We then rewrite equation 43 as

$$\begin{aligned}\kappa_\rho(\Phi_{11}, \Phi_{22}, \Phi_{12}) &= \frac{2\sqrt{\det \mathsf{F}}}{\pi\sqrt{\Phi_{11}\Phi_{22}}\sin\theta} \underbrace{\int \frac{1}{2\pi\sqrt{\det \mathsf{F}}} \exp\big(-\frac{1}{2}(a_1,a_2)\mathsf{F}^{-1}(a_1,a_2)^\top\big)\, da_1\, da_2}_{=1} \\ &= \frac{2}{\pi\sqrt{1+\Phi_{11}+\Phi_{22}+\Phi_{11}\Phi_{22}\sin^2\theta}} \\ &= \frac{2}{\pi\sqrt{(1+\Phi_{11})(1+\Phi_{22})-\Phi_{12}^2}}\end{aligned}$$

The NNK for chain activations This result follows by a similar completing the square type derivation, but instead of the resulting integrand being a bivariate Gaussian density, the resulting integrand is a product of a bivariate Gaussian density with probit activations. The result then follows from Williams (1997). Concretely, the NNK κ_σ satisfies

$$\kappa_\sigma(\Phi_{11}, \Phi_{22}, \Phi_{12}) = \frac{4}{2\pi} \int \frac{\operatorname{erf}(a_1/\sqrt{2})\operatorname{erf}(a_2/\sqrt{2})}{2\pi\sqrt{\Phi_{11}\Phi_{22}}\sin\theta} \exp\big(-\frac{1}{2}(a_1,a_2)(\mathsf{\Phi}^{-1}+\mathsf{I})(a_1,a_2)^\top\big)\, da_1\, da_2.$$

Completing the square, we have

$$\begin{aligned}\kappa_\sigma(\Phi_{11}, \Phi_{22}, \Phi_{12}) &= \frac{2\sqrt{\det \mathsf{F}}}{\pi\sqrt{\Phi_{11}\Phi_{22}}\sin\theta} \underbrace{\int \frac{\operatorname{erf}(a_1/\sqrt{2})\operatorname{erf}(a_2/\sqrt{2})}{2\pi\sqrt{\det \mathsf{F}}} \exp\big(-\frac{1}{2}(a_1,a_2)\mathsf{F}^{-1}(a_1,a_2)^\top\big)\, da_1\, da_2}_{=\kappa_{\operatorname{erf}(\cdot/\sqrt{2})}(F_{11},F_{22},F_{12})} \\ &= \frac{2}{\pi\sqrt{(1+\Phi_{11})(1+\Phi_{22})-\Phi_{12}^2}} \left(\frac{2}{\pi}\sin^{-1}\frac{F_{12}}{\sqrt{\big(1+F_{11}\big)\big(1+F_{22}\big)}}\right)\end{aligned}$$

where the last line follows from equation (11) of Williams (1997).

The explicit kernel c By a law of large numbers, we have that the explicit kernel is an NNK,

$$\begin{aligned}c(X_1, X_2) &= \mathbb{E}\big[\operatorname{erf}(Z_1/\sqrt{2})\operatorname{erf}(Z_2/\sqrt{2})\big] + 1 \qquad (Z_1, Z_2) \sim \mathcal{N}\Big(0, \begin{pmatrix} X_1^\top X_1 & X_1^\top X_2 \\ X_2^\top X_1 & X_2^\top X_2 \end{pmatrix}\Big), \\ &= \frac{2}{\pi}\sin^{-1}\frac{X_1^\top X_2}{\sqrt{\big(1+X_1^\top X_1\big)\big(1+X_2^\top X_2\big)}} + 1,\end{aligned}$$

again invoking the result of Williams (1997).

The explicit mean μ By a law of large numbers, the average $\frac{1}{d}1^\top \operatorname{erf}(\mathsf{W}X/\sqrt{2})$ converges to zero as $d \to \infty$. We therefore have that $\mu(X) = 0$.

G.2 Other examples

We now investigate some other important examples. In each example, the central question is whether or not a unique fixed point exists. By Theorem 1, G admits a unique fixed point if it is a contraction. It is a contraction whenever its Jacobian determinant is less than 1. The Jacobian is lower triangular, since $0 = \frac{\partial G}{\partial \Phi_{22}} = \frac{\partial G}{\partial \Phi_{12}} = \frac{\partial G}{\partial \Phi_{12}} = \frac{\partial G}{\partial \Phi_{11}}$, so in order to compute the Jacobian determinant, it suffices to compute the diagonal entries. These can be computed with the following identity.

Theorem 19 (Theorem 6 of Tsuchida et al. (2021). See also Theorem 3 of Han et al. (2022).)**.** *Suppose the absolute value of $\zeta : \mathbb{R} \to \mathbb{R}$ is bounded by a polynomial. Let $\dot{\zeta}$ denote the distributional (Schwartz) derivative of ζ. Then $\frac{\partial k_\zeta(\Phi_{11},\Phi_{22},\Phi_{12})}{\partial \Phi_{12}} = k_{\dot{\zeta}}(\Phi_{11}, \Phi_{22}, \Phi_{12})$ and $\frac{\partial k_\zeta(\Phi_{11},\Phi_{11},\Phi_{11})}{\partial \Phi_{11}} = \mathbb{E}[(Z^2 - 1)\zeta^2\big(\sqrt{\Phi_{11}}Z\big)]/\big(2\Phi_{11}\big)$, where $Z \sim \mathcal{N}(0, 1)$.*

Theorem 19 allows one to compute the kernel $k_{\sigma'}$, where σ' is the derivative of σ, by differentiating the kernel k_σ. This is easier than computing $k_{\sigma'}$ from scratch. There are two immediate uses for such a result. Firstly, the quantity $k_{\sigma'}$ is needed to compute the neural tangent kernel. Secondly, and the reason the theorem is useful in our current context, is that it lets us have sufficient conditions for the update G to be a contraction.

G.2.1 Gaussian $A(\eta) = \eta^2/2$, identity $R(a) = a$, general C, zero μ

This important special case yields an ℓDEKer that may be computed in closed form. Setting $\sigma(z) = z$, Theorem 4 and Corollary 5 say that the DEKer converges to an ℓDEKer with a closed-form,

$$\Psi_{ij}^{(t+1)} = \frac{1}{\lambda^2}\big(C_{ij} + \Psi_{ij}^{(t)}\big) \qquad \text{and} \qquad \Psi_{ij} = \frac{1}{\lambda^2}\big(C_{ij} + \Psi_{ij}\big) \implies \Psi_{ij} = \frac{C_{ij}}{\lambda^2 - 1},$$

whenever $\lambda > 1$, since $\lambda > 1$ implies G is a contraction. In this particular case, the ℓDEKer is simply a rescaling of the kernel c.

G.2.2 General A, identity $R(a) = a$, general C, zero μ

In the general setting of § 3.3, Theorem 4 and Corollary 5 yield fixed point equations for the ℓDEKer that do not in general admit a closed-form,

$$\Psi_{ij}^{(t+1)} = \frac{1}{\lambda^2}\Big(C_{ij} + k_\sigma\big(\Psi_{ii}^{(t)}, \Psi_{jj}^{(t)}, \Psi_{ij}^{(t)}\big)\Big) \qquad \text{and} \qquad \Psi_{ij} = \frac{1}{\lambda^2}\Big(C_{ij} + k_\sigma\big(\Psi_{ii}, \Psi_{jj}, \Psi_{ij}\big)\Big).$$

By Theorem 19, the ℓDEKer is the fixed point of a contraction whenever $\kappa_{\dot\sigma}/\lambda^2 < 1$, for which it is sufficient that A'' is less than λ. Statistically speaking, since A acts as a cumulant generating function, this is equivalent to the largest variance of the exponential family being less than λ.

G.2.3 General A, general R, zero C, zero μ

A pathological but informative example is obtained when the PSD kernel c and the cross terms μ are chosen to be the constant zero function. In this case, from Theorem 4 and Corollary 5 we obtain

$$\Psi_{ij}^{(t+1)} = \frac{1}{\lambda^2}\kappa_\sigma\big(\Psi_{ii}^{(t)}, \Psi_{jj}^{(t)}, \Psi_{ij}^{(t)}\big) \qquad \text{and} \qquad \Psi_{ij} = \frac{1}{\lambda^2}\kappa_\sigma\big(\Psi_{ii}, \Psi_{jj}, \Psi_{ij}\big).$$

The ℓDEKer is the fixed point of a contraction whenever $\kappa_{\dot\sigma}/\lambda^2 < 1$, by Theorem 19.

Note that the (infinite τ) ℓDEKer does not depend on the input X_1, X_2, but the (finite τ) DEKer depends on the initial guess. For a given initial guess of $\Psi_{11}^{(1)} = \|X_1\|^2, \Psi_{22}^{(1)} = \|X_2\|^2, \Psi_{12}^{(1)} = X_1^\top X_2$, solving for the ℓDEKer using τ iterations of naive fixed point iteration is exactly the same as a τ-layer NNK equation 12. Therefore, the DEKer is an NNK if for an arbitrary activation σ there exist corresponding configurations of A and R.

G.2.4 Gaussian $A(\eta) = \eta^2/2$, ReLU $R(a) = a\,\mathrm{u}(a)$, general C, zero μ

Let u denote the Heaviside step function, which takes values 0, 1/2 and 1 when evaluated at < 0, 0 and > 0 respectively. The rectified linear unit may be written $\mathrm{ReLU}(a) = a\,\mathrm{u}(a)$. Choosing $A(\eta) = \eta^2/2$, ReLU $R(a) = a\,\mathrm{u}(a)$, we find that ρ is the Heaviside step function and σ is the ReLU. The corresponding kernels k_ρ and k_σ are known as the arc-cosine kernels of order 0 and 1, and have closed-form expressions (see Appendix I),

$$\begin{aligned}\kappa_{\mathrm{u}}\big(\Sigma_{11}, \Sigma_{22}, \Sigma_{12}\big) &= \frac{1}{2\pi}(\pi - \theta),\\ \kappa_{\mathrm{ReLU}}\big(\Sigma_{11}, \Sigma_{22}, \Sigma_{12}\big) &= \frac{\sqrt{\Sigma_{11}\Sigma_{22}}}{2\pi}\big(\sin\theta + (\pi - \theta)\cos\theta\big),\end{aligned}$$

where $\theta = \arccos \frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}$.

Fixed points for Ψ_{11} and Ψ_{22} can be computed in closed-form provided $\lambda^2 < 1/2$,

$$\Psi_{ii}^{(t+1)} = \frac{1}{2\lambda^2}\Big(C_{ii} + \Psi_{ii}^{(t)}\Big) \qquad \text{and} \qquad \Psi_{ii} = \frac{1}{2\lambda^2}\Big(C_{ii} + \Psi_{ii}\Big) \implies \Psi_{ii} = \frac{C_{ii}}{2\lambda^2 - 1}.$$

However, Ψ_{12} cannot be computed in closed-form. We leave the analysis for determining whether G results in a contraction for future work. Nevertheless, we may still compute using the result in Theorem 4 without violating any assumptions.

Interestingly, in this setting, $\rho = \dot{\sigma}$ almost everywhere and the DEKer iterates very closely resemble NTK iterates. There are three differences in calculating the DEKer and the NTK. Firstly, the DEKer uses c where the NTK uses $\Theta^{(t)}$. Secondly, the DEKer uses $\Psi^{(t)}$ as an input to $\Sigma^{(t+1)}$ and $\dot{\Sigma}^{(t+1)}$, whereas the NTK uses $\Sigma^{(t)}$. Finally, the DEKer may be initialised at any guess $\Psi^{(1)}$, whereas the NTK must be initialised at $X_1^\top X_2$.

G.2.5 Gaussian $A(\eta) = \eta^2/2$, ReLU $R(a) = a\,\mathrm{u}(a)$, arc-cosine c and corresponding μ

We now describe a setting that we found practically useful in our experiments (see § 4.2.). We use the setting described in § G.2.4, but without the assumption that $\mu(X) = 0$. For the features $\Gamma(X)$ of the kernel $c(X_1, X_2)$, we choose $\Gamma(X) = \mu_v\,\mathrm{ReLU}(\mathsf{Q}X)$, where $\mu_v \in \mathbb{R}$ is a hyperparameter and Q is a $d \times l$ matrix with entries drawn independently from the standard Gaussian distribution, resulting in an arc-cosine kernel for c. The mean function μ and the cross terms $\kappa_{\sigma,\rho}$ admit closed-form expressions, as given in Appendix F.1. The resulting DEKer can represent deep arc-cosine kernels when μ_v is zero, and resembles (but is not the same as) an NTK with extra cross-terms otherwise.

H Other considerations

H.1 Why the expected negative log posterior?

We may frame our optimisation objective in terms of exponential family PCA (Collins et al., 2001; Mohamed et al., 2008). Given a dataset $\{Y_s\}_{s=1}^N$ of N examples, exponential family PCA models observation $Y_s \in \mathbb{Y}^d$ as following a factored exponential family with canonical parameter $\mathsf{V}\phi_s$, for some basis V and latent $\phi_s \in \mathbb{R}^m$. The resulting graphical model is shown in Figure 3a. A maximum a posteriori estimate is

$$\phi_s^* \triangleq \arg\min -\log p\big(\phi_s \mid Y_s\big) = \arg\min -\log \int p\big(Y_s \mid \mathsf{V}, \phi_s)p(\mathsf{V})p(\phi_s)\, d\mathsf{V}, \tag{44}$$

in which the basis V is marginalised before the evaluation of the logarithm.

Our objective equation 17 differs from equation 44 in two respects. Firstly, we generalise the canonical parameter $\mathsf{V}\phi_s$ so that a nonlinearly parameterised canonical parameter $R(\mathsf{V}\phi_s)$ is used. Secondly and more critically, the order of the logarithm and the expectation is swapped. This may be understood by examining a variational lower bound (VLB) of the posterior. Note that the VLB has been used for MAP estimation in similar contexts (Kingma & Welling, 2014), and can be seen as a regularised or penalised variant of the ELBO. Let $\mathsf{\Phi} = (\phi_1, \dots, \phi_N)$. For any density $q(\mathsf{V})$ which ostensibly approximates $p(\mathsf{V} \mid \mathsf{Y}, \mathsf{\Phi})$, the log model evidence decomposes into a sum of a KL divergence and an ELBO,

$$\log p(\mathsf{Y} \mid \mathsf{\Phi}) = \mathrm{KL}\big(q(\mathsf{V}) \| p(\mathsf{V} \mid \mathsf{Y}, \mathsf{\Phi})\big) + \mathbb{E}_{\mathsf{V}\sim q} \log \frac{p(\mathsf{V}, \mathsf{Y} \mid \mathsf{\Phi})}{q(\mathsf{V})}, \quad \text{so that}$$
$$-\mathbb{E}_{\mathsf{V}\sim q} \log p(\mathsf{\Phi} \mid \mathsf{V}, \mathsf{Y}) = \mathrm{KL}\big(q(\mathsf{V}) \| p(\mathsf{V} \mid \mathsf{Y}, \mathsf{\Phi})\big) - \log p(\mathsf{\Phi} \mid \mathsf{Y}) + \mathbb{E}_{\mathsf{V}\sim q} \log p(\mathsf{V}, \mathsf{Y}) - \mathbb{E}_{\mathsf{V}\sim q} \log q(\mathsf{V}).$$

By minimising the left hand side with respect to $\mathsf{\Phi}$, we are maximising the log model evidence minus the KL divergence. By selecting hyperparameters π of the variational density q over V, we alter our approximate posterior.

Note a clashing nomenclature between EM-algorithm and variational inference — where the marginalised variable V is called a latent variable — against an unsupervised dimensionality reduction setting — where the low dimensional representation ϕ_s is called a latent variable.

H.2 Scaling and parameterisation of weight distributions

It is widely appreciated that the prior over W in the Bayesian setting (MacKay, 1998, §11.1) and the initialisation of W in the gradient-flow setting play an role in directing the limiting behaviour of the neural network (Sohl-Dickstein et al., 2020). On the one hand, convenient parameterisations and choices of prior and initial distributions lead to tractable large width limits. On the other hand, while limiting models can outperform their finite width counterparts in small data regimes (Arora et al., 2020), GPs in general are most often outperformed by deep learning models for many problems of interest. This might suggest that the tractable limits are the "wrong" ones to analyse if one seeks to explain the success stories of deep learning (Chizat et al., 2019; Woodworth et al., 2020). Other works consider more general heavy-tailed (Der & Lee, 2005; Peluchetti et al., 2020; Favaro et al., 2021; 2022) or differently scaled priors, but it is not yet clear whether these models can more accurately emulate deep learning models.

In our work in particular, the scaling of the prior with precision $\sqrt{md}$ (less than the m that might often be expected, since $d < m$) in equation 17 was crucial for finding a tractable limit. Independently of whether this limiting regime represents any meaningful feature representation, our analysis is valuable because (1) DEKERs are better than or competitive with other neural network kernel models in the settings that we tried, (2) we are the first to place deep neural network related kernels in a more fundamental footing of statistical estimation and optimisation, and (3) our analysis describes a limiting invariant of SGD.

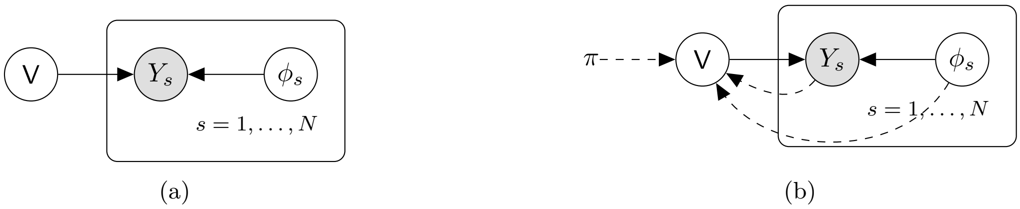

Figure 3: (a) Exponential family PCA, which may be viewed as an unsupervised problem, in which observed data Y_s is a realisation from an exponential family with canonical parameter $\mathsf{V}\phi_s$ for some basis V and latent ϕ_s. (b) Nonlinearly parameterised exponential family PCA, in which Y_s is a realisation from an exponential family with canonical parameter $R(\mathsf{V}\phi_s)$. We additionally choose $Y_s = \Gamma(X_s)$, and employ a variational approximation (indicated by the dashed lines) for the distribution $p(\mathsf{V} \mid \mathsf{Y}, \Phi) \approx q(\mathsf{V} \mid \pi)$, and take an infinite width limit.

H.3 Implicit differentiation

From Corollary 5, we have that the ℓDEKer Ψ satisfies $\Psi = \mathsf{G}(\Psi)$. Suppose Ψ depends on v-dimensional hyperparameter $\zeta \in \mathbb{R}^v$, such as the weight and bias variance (see Footnotes 2 and 3), or a hyperparameter of R. If G is continuously differentiable, the implicit function theorem says

$$\underbrace{\frac{d\Psi}{d\zeta}}_{3\times v} = \underbrace{\frac{\partial \mathsf{G}(\Psi)}{\partial \zeta}}_{3\times v} + \underbrace{\frac{\partial \mathsf{G}(\Psi)}{\partial \Psi}}_{3\times 3} \underbrace{\frac{d\Psi}{d\zeta}}_{3\times v} \implies \Big(\mathsf{I} - \frac{\partial \mathsf{G}(\Psi)}{\partial \Psi}\Big)\frac{d\Psi}{d\zeta} = \frac{\partial \mathsf{G}(\Psi)}{\partial \zeta},$$

which may be solved for $\frac{d\Psi}{d\zeta}$ using a backslash operator. This derivative may be used for gradient-based hyperparameter selection. For example, if the ℓDEKer were to be used as the covariance function of a Gaussian process, one could perform type II maximum marginal likelihood to compute point estimates for ζ. This implicit differentiation mirrors the finite-width counterpart, the DEQ (Bai et al., 2019). We leave its empirical investigation for future work.

I Arc-cosine kernels via derivatives

While the Dirac distribution is not a function and therefore cannot be used as an activation function in finite-width networks, it does arise as the derivative of NNKs with Heaviside activations, by Theorem 19. With an abuse of notation that extends the usual operation of integrating against a Dirca delta distribution, we may understand an expectation involving Dirac delta distributions as a limiting expectation involving nascent delta functions. We may evaluate the corresponding NNK as follows.

$$\begin{aligned}
&\kappa_\delta\big(\Sigma_{11}, \Sigma_{22}, \Sigma_{12}\big) \\
&= \mathbb{E}\big[\delta(\chi)\delta(\chi')\big] \\
&= \mathbb{E}\big[\delta\big(\sqrt{\Sigma_{11}}Z_1\big)\delta\big(\sqrt{\Sigma_{22}}(Z_1\rho + Z_2\sqrt{1-\rho^2})\big)\big], \quad (Z_1, Z_2)^\top \sim \mathcal{N}(0, I), \quad \rho = \Sigma_{12}/\sqrt{\Sigma_{11}/\Sigma_{22}} \\
&= \frac{1}{\sqrt{\Sigma_{11}}}\frac{1}{\sqrt{2\pi}}\int \delta\big(\sqrt{\Sigma_{22}}z_2\sqrt{1-\rho^2}\big)p(z_2)dz_2, \quad p \text{ is pdf of standard Gaussian} \\
&= \frac{1}{\sqrt{\Sigma_{11}\Sigma_{22}(1-\rho^2)}}\frac{1}{2\pi} \\
&= \frac{1}{2\pi\sqrt{\Sigma_{11}\Sigma_{22} - \Sigma_{12}^2}}.
\end{aligned} \tag{45}$$

Note the singularity whenever the Gaussian distribution is degenerate, i.e. $\Sigma_{11} = \Sigma_{22} = \Sigma_{12}$, which is an instance of the more general undefinedness of a product of Dirac delta distributions.

The NNK corresponding with Heaviside activations u was first evaluated using a geometric argument by Sheppard (1899), and is given by

$$\begin{aligned}
&\kappa_{\mathrm{u}}\big(\Sigma_{11}, \Sigma_{22}, \Sigma_{12}\big) \\
&= \mathbb{E}[\mathrm{u}(\chi)\,\mathrm{u}(\chi')] \\
&= \frac{1}{2\pi}(\pi - \theta), \quad \theta = \arccos\frac{\Sigma_{12}}{\sqrt{\Sigma_{11}\Sigma_{22}}}.
\end{aligned} \tag{46}$$

The NNK equation 46 was generalised to activations of the form $\mathrm{u}(z)z^p$ for positive integers p by Cho & Saul (2009). Of particular relevance is the case $p = 1$, in which case the activation function is ReLU and

$$\begin{aligned}
&\kappa_{\mathrm{ReLU}}\big(\Sigma_{11}, \Sigma_{22}, \Sigma_{12}\big) \\
&= \mathbb{E}[\mathrm{ReLU}(\chi)\,\mathrm{ReLU}(\chi')] \\
&= \frac{\sqrt{\Sigma_{11}\Sigma_{22}}}{2\pi}\big(\sin\theta + (\pi - \theta)\cos\theta\big).
\end{aligned} \tag{47}$$

Note that equation 45– equation 47 represent a sequence of derivatives, since the Dirac delta distribution, Heaviside function and ReLU represent a sequence of distributional derivatives. More concretely, by Theorem 19,

$$\frac{\partial^2\kappa_{\mathrm{ReLU}}}{\partial{\Sigma_{12}}^2} = \frac{\partial\kappa_{\mathrm{u}}}{\partial\Sigma_{12}} = \kappa_\delta,$$

as can be otherwise verified.

J Experiments

J.1 Measuring finite-width effects

We consider elements of an input space $\mathbb{X}$ which are 100 evenly spaced points over $[-5, 5]^2$. This results in an input matrix of size 100 × 2. We compute two 100 × 100 kernel matrices with ijth element: Ψ_{ij} (calculated to high tolerance using a fixed point solver) and $k_d^{(t)}(X_i, X_j)$ (calculated using SGD). Finite features Γ are chosen to be $\Gamma = \mathsf{T}X$, where $\mathsf{T} \in \mathbb{R}^{d \times m}$ is a zero-mean Gaussian random matrix. This results in a linear kernel $c(X_1, X_2) = X_1^\top X_2$. We set $t = 400$, $\lambda = 6$ and use a step length of $\alpha^{(t)} = \frac{1}{\lambda}\sqrt{d/m}$. We vary d between 5 and 500 in steps of 5 and choose $m = d^{3/2}$. We also provide the CKA between the (finite-d, finite-τ) ffDEKer and a squared exponential kernel $A \exp\left(-\|X_1 - X_2\|_2^2/2\right)$ for control, where the scaling parameter A is the largest value in the (infinite-d, infinite-τ) ℓDEKer matrix.

J.2 Inference using the DEKer

The hyperparameter grid over which `GridSearchCV` operates is given in table 4.

Hyperparameter	Present in	Values
Data scale (see footnote 2)	NTK, NNK, DEKer, SEK	$\{0.5, 1, 2, 4\}$
KRR regularisation strength	NTK, NNK, DEKer, SEK	$\{0.05, 0.1, 0.5\}$
Input augmented bias (see footnote 3)	NTK, NNK, DEKer	$\{-1.0, -0.1, 0.0, 0.1, 1.0\}$
Number of iterations / layers T	NTK, NNK	$\{2, 3, 4, 5\}$
Number of iterations / layers T	DEKer	$\{2, 3, 4, 5, \infty\}$
Inner regularisation strength λ	DEKer	$\{1, 2, 4\}$
Cross-term strength μ_v	DEKer	$\{0, 0.1, 0.5, 1, 2\}$
Lengthscale	SEK	$\{0.5, 1, 2, 4, 8, 16\}$

Table 4: Search space for `GridSearchCV`. We use Anderson acceleration to compute the DEKer when $T = \infty$ and there are less than 500 points in the dataset.