Kernel Multimodal Continuous Attention

Anonymous Author(s)
Affiliation
Address
email

Abstract

Attention mechanisms average a data representation with respect to probability weights. Recently, Martins et al. [2020, 2021] proposed continuous attention mechanisms, focusing on unimodal exponential and deformed exponential family attention densities: the latter can have sparse support. Farinhas et al. [2021] extended to multimodality via Gaussian mixture attention densities. In this paper, we propose using kernel exponential families [Canu and Smola, 2006] and our new sparse counterpart, kernel *deformed* exponential families. Theoretically, we show new existence results for both families, and approximation capabilities for the deformed case. Lacking closed form expressions for the context vector, we use numerical integration: we prove exponential convergence for both families. Experiments show that kernel continuous attention often outperforms unimodal continuous attention, and the sparse variant tends to highlight peaks of time series.

1 Introduction

Attention mechanisms are weighted averages of data representations [Bahdanau et al., 2015] used to make predictions. Discrete attention 1) cannot easily handle irregularly spaced observations, and 2) attention maps may be scattered, lacking focus. Martins et al. [2020, 2021] proposed continuous attention, showing that attention densities maximize the regularized expectation of a function of the data location (i.e. time). Special cases lead to exponential and deformed exponential families: the latter has sparse support. They form a continuous data representation and take expectations with respect to attention densities. In Martins et al. [2022] they apply this to a transformer architecture.

Martins et al. [2020, 2021, 2022] used unimodal attention densities, giving importance to *one* data region. Farinhas et al. [2021] extended this to multimodal Gaussian mixture attention densities. However 1) Gaussian mixtures lie in neither the exponential nor deformed exponential families, and are difficult to study in the context of Martins et al. [2020, 2021]; and 2) they have dense support. Sparse support can say that certain regions of data do not matter: a region of time has *no* effect on class probabilities, or a region of an image is *not* some object. We would like to use multimodal exponential and deformed exponential family attention densities, and understand how Farinhas et al. [2021] relates to the framework of Martins et al. [2020, 2021].

This paper makes three contributions: 1) we introduce kernel *deformed* exponential families, a sparse multimodal density class, and apply it along with the multimodal kernel exponential families [Canu and Smola, 2006] as attention densities. The latter have been used for density estimation, but not weighting data importance; 2) we theoretically analyze kernel exponential and deformed exponential family i) normalization, ii) approximation and iii) context vector numerical integration properties; 3) we apply them to real world datasets, showing that multimodal continuous attention outperforms unimodal, and that kernel deformed exponential family densities often highlight the peaks of time series. Approximation properties for the kernel deformed case are challenging: similar kernel exponential family results [Sriperumbudur et al., 2017] relied on exponential and logarithm

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

properties to bound the difference of the log-partition functional at two functions: these do not hold for deformed analogues. We provide similar bounds by using a functional mean value inequality.

We first review unimodal continuous attention [Martins et al., 2020, 2021]. We motivate multimodal continuous attention via time warping. We next describe kernel exponential families and give a novel normalization condition relating kernel growth to the base density's tail decay. We then propose kernel deformed exponential families, new densities which can have support over disjoint regions. We describe normalization and approximation capabilities. Next we describe using these densities for continuous attention, including numerical integration convergence analysis. We show experiments comparing unimodal and multimodal attention, and conclude with limitations and future work.

2 Related Work

Attention Mechanisms closely related are Martins et al. [2020, 2021, 2022], Farinhas et al. [2021]. Martins et al. [2020, 2021] frame continuous attention as an expectation of a value function with respect to a density, where the density solves an optimization problem. They only used unimodal (deformed) exponential exponential family densities: we extend this to the multimodal setting by leveraging kernel exponential families and proposing a deformed counterpart. Farinhas et al. [2021] proposed a multimodal continuous attention mechanism via a Gaussian mixture. We show in Appendix A that this solves a slightly different optimization problem from Martins et al. [2020, 2021]. A limitation of Gaussian mixtures is lack of flexible tail decay. Finally, Martins et al. [2022] apply continuous attention within a transformer architecture to model long context. This is a new application of continuous attention rather than an extension of specific continuous attention mechanisms.

Also relevant are Tsai et al. [2019], Shukla and Marlin [2021, 2022]. Shukla and Marlin [2021] provide an attention mechanism for irregularly sampled time series by use of a continuous-time kernel regression framework, but do not take an expectation of a data representation over time with respect to a continuous pdf. Instead they evaluate the kernel regression model at fixed time points. This describes importance of data at a set of points rather than over continuous regions. Shukla and Marlin [2022] extend this to incorporate uncertainty quantification. Other papers connect attention and kernels, but focus on discrete attention [Tsai et al., 2019, Choromanski et al., 2020]. Also relevant are temporal transformer papers, including Xu et al. [2019], Li et al. [2019, 2020], Song et al. [2018]. However, none have continuous attention densities.

Kernel Exponential Families Canu and Smola [2006] proposed kernel exponential families: Sriperumbudur et al. [2017] analyzed theory for density estimation. Wenliang et al. [2019] parametrized the kernel with a deep neural network. Other density estimation papers include Arbel and Gretton [2018], Dai et al. [2019], Sutherland et al. [2018]. We apply kernel exponential families as attention densities to *weight* a value function which represents the data, rather than for density estimation. Further, Wenliang et al. [2019] showed a condition for an unnormalized kernel exponential family density to have a finite normalizer. However, they used exponential power base densities. We instead relate kernel growth rates to the base density tail decay, allowing non-symmetric base densities.

To summarize our theoretical contributions: 1) showing that multimodal continuous attention is required to represent time warping 2) introducing kernel *deformed* exponential families with approximation and normalization analysis 3) improved kernel exponential family normalization results 4) convergence analysis of numerical integration for kernel-based attention 5) characterizing Farinhas et al. [2021] in terms of the framework of Martins et al. [2020, 2021].

3 Continuous Attention Mechanisms

An attention mechanism involves: 1) the value function approximates a data representation. 2) the attention density is chosen to be 'similar' to another data representation, encoding it into a density 3) the context c takes an expectation of the value function with respect to the attention density. It is

$$c = \mathbb{E}_{T \sim p}[V(T)]. \tag{1}$$

The value function $V(t)$ approximates a data representation, $T \sim p(t)$ is the random variable or vector for locations (temporal, spatial, etc), and $p(t)$ is the attention density (potentially with respect to a discrete measure). For discrete attention, one could have $V(t)$ be a time series where $t \in S$ a finite set of time points. One then weights the time series with a probability vector to obtain

the context vector. In contrast, in the continuous setting $V(t)$ could be a curve or realization of a continuous-time stochastic process, and S could be $[0, \tau]$ where τ is a study end time. One then weights it with a continuous density $p(t)$ (absolutely continuous wrt Lebesgue measure).

To choose p, one takes a data representation f and finds p 'similar' to f, but regularizing p. Martins et al. [2020, 2021] did this, formalizing attention mechanisms. Given a probability space $(S, \mathcal{A}, Q)$, let $\mathcal{M}_+^1(S)$ be the set of probability densities with respect to Q. Assume Q is dominated by a σ-finite measure ν (i.e. Lebesgue) and that it has density $q_0 = \frac{dQ}{d\nu}$ with respect to ν. Let $S \subseteq \mathbb{R}^D$, $\mathcal{F}$ be a function class, and $\Omega : \mathcal{M}_+^1(S) \to \mathbb{R}$ be a lower semi-continuous, proper, strictly convex functional. Given $f \in \mathcal{F}$, an *attention density* [Martins et al., 2020] $\hat{p} : \mathcal{F} \to \mathbb{R}_{\geq 0}$ solves

$$\hat{p}[f] = \arg \max_{p \in \mathcal{M}_+^1(S)} \langle p, f \rangle_{L^2(Q)} - \Omega(p). \tag{2}$$

This maximizes regularized L^2 similarity between p and a data representation f. If $\Omega(p)$ is the negative differential entropy, the attention density is Boltzmann Gibbs

$$\hat{p}[f](t) = \exp(f(t) - A(f)), \tag{3}$$

where $A(f)$ ensures $\int_S \hat{p}[f](t) dQ = 1$. If $f(t) = \theta^T \phi(t)$ for parameters and statistics $\theta \in \mathbb{R}^M, \phi(t) \in \mathbb{R}^M$ respectively, Eqn. 3 becomes an exponential family density. For f in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, it becomes a kernel exponential family density [Canu and Smola, 2006], which we propose as an alternative attention density.

One desirable class would be heavy or thin tailed exponential family-like densities. In exponential families, the support, or non-zero region of the density, is controlled by the measure Q. Letting $\Omega(p)$ be the α-Tsallis negative entropy $\Omega_\alpha(p)$ [Tsallis, 1988],

$$\Omega_\alpha(p) = \begin{cases} \frac{1}{\alpha(\alpha-1)} \left(\int_S p(t)^\alpha dQ - 1 \right), \alpha \neq 1; \\ \int_S p(t) \log p(t) dQ, \alpha = 1, \end{cases}$$

then $\hat{p}[f]$ for $f(t) = \theta^T \phi(t)$ lies in the deformed exponential family [Tsallis, 1988, Naudts, 2004]

$$\hat{p}_{\Omega_\alpha}[f](t) = \exp_{2-\alpha}(\theta^T \phi(t) - A_\alpha(f)), \tag{4}$$

where $A_\alpha(f)$ again ensures normalization and the density uses the β-exponential

$$\exp_\beta(t) = \begin{cases} [1 + (1-\beta)t]_+^{1/(1-\beta)}, \beta \neq 1; \\ \exp(t), \beta = 1. \end{cases} \tag{5}$$

For $\beta < 1$, Eqn. 5 and thus deformed exponential family densities for $1 < \alpha \leq 2$ can return 0 values. Values $\alpha > 1$ (and thus $\beta < 1$) give thinner tails than the exponential family, while $\alpha < 1$ gives fatter tails. Setting $\beta = 0$ is called *sparsemax* [Martins and Astudillo, 2016]. In this paper, we assume $1 < \alpha \leq 2$, which is the sparse case studied in Martins et al. [2020]. We again propose to replace $f(t) = \theta^T \phi(t)$ with $f \in \mathcal{H}$, which leads to the novel *kernel deformed exponential families*.

Computing Eqn. 1's context vector requires parametrizing $V(t)$. Martins et al. [2020] parametrize $V : S \to \mathbb{R}^D$ with $\mathbf{B} \in \mathbb{R}^{D \times N}$ as $V(t; \mathbf{B}) = \mathbf{B}\Psi(t)$ and estimate $\mathbf{B} \in \mathbb{R}^{D \times N}$ via regularized multivariate linear regression. Here $\Psi = \{\psi_n\}_{n=1}^N$ is a set of basis functions. Let L be the number of observation locations (times in a temporal setting), O be the observation dimension, and N be the number of basis functions. This involves regressing the observation matrix $\mathbf{H} \in \mathbb{R}^{O \times L}$ on a matrix $\mathbf{F} \in \mathbb{R}^{N \times L}$ of basis functions $\{\psi_n\}_{n=1}^N$ evaluated at observation locations $\{t_l\}_{l=1}^L$

$$\mathbf{B}^* = \arg \min_{\mathbf{B}} \|\mathbf{BF} - \mathbf{H}\|_F^2 + \lambda \|\mathbf{B}\|_F^2. \tag{6}$$

4 Time Warping

Martins et al. [2020, 2021] used unimodal attention densities. We can formalize one multimodal capability: aligning features in time across curves. Physical time may not properly represent dynamics: a better representation may be an individualized time scale transforming physical time. For instance, electrocardiogram (ECG) curves of heartbeats have three parts: a P-wave, a QRS complex, and a T-wave. These have similar patterns between heartbeats, but may have different durations and peak locations. Often one would like a time aligned version of such curves. Modeling this involves estimating a time warping function that lines up qualitative features, such as max or min values or first crossing times. For further reading see Ramsay and Li [1998].

Definition 4.1. (Time Warping Function) Given references times $\{t_{0j}\}_{j=1}^K$ and individualized times $\{t_{ij}\}_{j=1}^K$, both in $[0, \tau]$, a **time warping function** $h_i : S \to \mathbb{R}$ is a strictly increasing, differentiable, invertible function such that

$$h_i(0) = 0, h_i(\tau) = \tau$$
$$h_i(t_{0k}) = t_{ik}, k = 1, \cdots K.$$

and

$$h_i(t) = t \text{ if } t \notin [0, \tau]$$

Let $\{X_i\}_{i=1}^n, X_i : S \to \mathbb{R}$ be a set of observed curves, each with K features occurring at individualized times $\{t_{ij}\}_{j=1}^K \subset [0, \tau]$. A set of time warping functions $\{h_i\}_{i=1}^n$ can map individualized feature occurrence times to reference times $\{t_{0j}\}_{j=1}^K$. One can then compute aligned $X_i^*(t) = X_i(h_i(t))$. Each X_i^* has relevant features at the same times $\{t_{0j}\}_{j=1}^K$. Classically, this requires handcrafting and locating important features and estimating a warping function. We could then compute an expectation of the time warped curve with respect to a *global* fixed density $p(t)$ to obtain a summary statistic $\mathbb{E}_{T\sim p} X_i^*(T)$ of the aligned curve. The following states that multimodal continuous attention can represent such an expectation with an attention density p_i, avoiding computing $X_i^*(t)$.

Lemma 4.2. *(Continuous Attention can Represent Time Warping) Let h be a time warping function, $g = h^{-1}$ and $X_i : \mathbb{R} \to \mathbb{R}$ with support on $[0, \tau]$. Assume that Q is dominated by Lebesgue measure ν and let $q_0 = \frac{dQ}{d\nu}$. Then for any fixed density p wrt Q, if g, X_i, q_0, p are continuous almost everywhere we have*

$$\mathbb{E}_{U\sim p} X_i^*(U) = \mathbb{E}_{T\sim p_i} X_i(T) \tag{7}$$

where $p_i(t) = p(g_i(t)) \frac{q_0(g_i(t))}{q_0(t)} g_i'(t)$ and $p_i(t)$ is a valid probability density function.

See Appendix B.1 for proof. Even if $p(t)$ is unimodal, $p_i(t)$ may not be. We show an example in Appendix B.2.

5 Kernel Exponential and Deformed Exponential Families

We use kernel exponential families and a new deformed counterpart to obtain flexible attention densities solving Eqn. 2 with the same regularizers. We first review kernel exponential families. We then give a novel theoretical result describing when an unnormalized kernel exponential family density can be normalized. This says that the normalizing constant exists when the base density has fast enough tail decay relative to kernel growth. Next we introduce kernel deformed exponential families, extending kernel exponential families to have either sparse support, our focus, or fatter tails. These can attend to non-overlapping time intervals. We show similar normalization results based on kernel choice and base density. The normalizing constant exists when the unnormalized density has compact support and the kernel grows sufficiently slowly. Following this we show approximation theory. We conclude by showing how to compute attention densities in practice.

Kernel exponential families [Canu and Smola, 2006] extend exponential families, replacing $f(t) = \theta^T \phi(t)$ with $f \in \mathcal{H}$ a reproducing kernel Hilbert space $\mathcal{H}$ [Aronszajn, 1950]. Densities can be written

$$\begin{aligned} p(t) &= \exp(f(t) - A(f)) \\ &= \exp(\langle f, k(\cdot, t)\rangle_{\mathcal{H}}\rangle - A(f)). \end{aligned}$$

The second equality follows from the reproducing property. A challenge is to choose $\mathcal{H}, Q$ so that a normalizing constant exists, i.e., $\int_S \exp(f(t)) dQ < \infty$. These densities can approximate any continuous density over a compact domain arbitrarily well in KL divergence, Hellinger, and L^p distance [Sriperumbudur et al., 2017]. However relevant integrals including the normalizing constant require numerical integration.

To avoid infinite dimensionality one generally assumes a representation of the form $f = \sum_{i=1}^I \gamma_i k(\cdot, t_i)$, where for density estimation [Sriperumbudur et al., 2017] the t_i are the observation locations. This requires using one parameter per observation value. This model complexity may not be necessary, and often one chooses a set of *inducing points* [Titsias, 2009] $\{t_i\}_{i=1}^I$ where I is less than the number of observation locations.

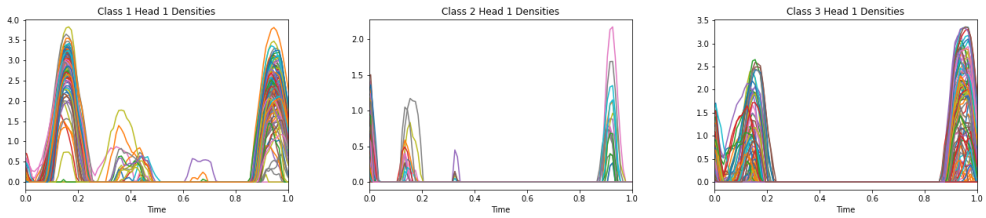

Figure 1: Attention densities for kernel deformed exponential families for the first attention head of the uWave experiment (Appendix F) and all test set participants for two classes. The densities are sparse and each have support over different non-overlapping time intervals, which cannot be done with either Gaussian mixtures or exponential families. They also attend to similar regions within each class.

For a given pair $\mathcal{H}, k$, how can we choose Q to ensure that the normalization constant exists? We first give a simple example of $\mathcal{H}, f$ and Q where it *does not*.

Example 1. Let Q be the law of a $\mathcal{N}(0,1)$ distribution and $S = \mathbb{R}$. Let $\mathcal{H} = \text{span}\{t^3, t^4\}$ with $k(t,s) = t^3 s^3 + t^4 s^4$ and $f(t) = t^3 + t^4 = k(t,1)$. Then the following integral diverges.

$$\int_S \exp(f(t))dQ = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2} + t^3 + t^4\right) dt$$

5.1 Theory for Kernel Exponential Families

We provide sufficient conditions for Q and $\mathcal{H}$ so that $A(f)$ the log-partition functional exists. We relate $\mathcal{H}$'s kernel growth rate to the tail decay of the random variable or vector T_Q with law Q.

Proposition 5.1. *Let $\tilde{p}(t) = \exp(f(t))$ where $f \in \mathcal{H}$ an RKHS with kernel k. Assume $k(t,t) \leq L_k \|t\|_2^{\xi} + C_k$ for constants $L_k, C_k, \xi > 0$. Let Q be the law of a random vector T_Q, so that $Q(A) = P(T_Q \in A)$. Assume $\forall u$ s.t. $\|u\|_2 = 1$, $z > 0$*

$$P(|u^T T_Q| \geq z) \leq C_q \exp(-vz^{\eta}) \tag{8}$$

for some constants $\eta > \frac{\xi}{2}, C_Q, v > 0$. Then

$$\int_S \tilde{p}(t)dQ < \infty.$$

See Appendix C.1 for proof. Based on $k(t,t)$'s growth, we can vary what tail decay rate for T_Q ensures we can normalize $\tilde{p}(t)$. Wenliang et al. [2019] also proved normalization conditions, but focused on exponential power density for a specific growth rate of $k(t,t)$ rather than relating tail decay to growth rate. By focusing on tail decay, our result can be applied to non-symmetric base densities. Specific kernel bound growth rate terms ξ lead to allowing different tail decay rates.

Corollary 5.2. *For $\xi = 4$, T_Q can be any sub-Gaussian random vector. For $\xi = 2$ it can be any sub-exponential. For $\xi = 0$ it can have any density.*

See Appendix C.2 for proof.

5.2 Kernel Deformed Exponential Families

We now propose kernel deformed exponential families: flexible sparse non-parametric densities. These take deformed exponential families and extend them to use kernels in the deformed exponential term. This mirrors kernel exponential families. We write

$$p(t) = \exp_{2-\alpha}(f(t) - A_\alpha(f)),$$

where $f \in \mathcal{H}$ with kernel k. Fig. 1b shows that they can have support over disjoint intervals.

5.2.1 Normalization Theory

We construct a valid kernel deformed exponential family density from Q and $f \in \mathcal{H}$. We first discuss the deformed log-normalizer. In exponential family densities, the log-normalizer is the log of the normalizer. For deformed exponentials, the following holds.

Lemma 5.3. *Let $Z > 0$ be a constant. Then for $1 < \alpha \leq 2$,*

$$\frac{1}{Z} \exp_{2-\alpha}(Z^{\alpha-1} f(t)) = \exp_{2-\alpha}(f(t) - \log_\alpha Z)$$

where

$$\log_\beta t = \begin{cases} \frac{t^{1-\beta}-1}{1-\beta} \textit{ if } t > 0, \beta \neq 1; \\ \log(t) \textit{ if } t > 0, \beta = 1; \\ \textit{undefined if } t \leq 0. \end{cases}$$

See Appendix D.1 for proof. We describe a normalization sufficient condition analagous to Proposition 5.1 for the sparse deformed kernel exponential family. With Lemma 5.3, we can take an unnormalized $\exp_{2-\alpha}(\tilde{f}(t))$ and derive a valid normalized kernel deformed exponential family density. We only require that an affine function of the terms in the deformed-exponential are negative for large magnitude t.

Proposition 5.4. *For $1 < \alpha \leq 2$ assume $\tilde{p}(t) = \exp_{2-\alpha}(\tilde{f}(t))$ with $\tilde{f} \in \mathcal{H}$, $\mathcal{H}$ is a RKHS with kernel k. If $\exists C_t > 0$ s.t. for $\|t\|_2 > C_t$, $(\alpha - 1)\tilde{f}(t) + 1 \leq 0$ and $k(t,t) \leq L_k \|t\|_2^\xi + C_k$ for some $\xi > 0$, then $\int_S \exp_{2-\alpha}(\tilde{f}(t)) dQ < \infty$.*

See Appendix D.2 for proof. We now construct a valid kernel deformed exponential family density using the finite integral.

Corollary 5.5. *Under the conditions of proposition 5.4, assume $\exp_{2-\alpha}(\tilde{f}(t)) > 0$ on a set $A \subseteq S$ such that $Q(A) > 0$, then $\exists$ constants $Z > 0$, $A_\alpha(f) \in \mathbb{R}$ such that for $f(t) = \frac{1}{Z^{\alpha-1}} \tilde{f}(t)$, the following holds*

$$\int_S \exp_{2-\alpha}(f(t) - A_\alpha(f)) dQ = 1.$$

See Appendix D.3 for proof. We thus estimate $\tilde{f}(t) = Z^{\alpha-1} f(t)$ and normalize to obtain a density of the desired form.

5.2.2 Approximation Theory

For space reasons we defer the formal statements and proofs of approximation results to Appendix D.4. The intuitive idea is that under an integrability condition for the kernel, kernel deformed exponential families are dense in families of the same form $\exp_{2-\alpha}(f(t) - A_\alpha(f))$ but with a $f \in C_0(S)$ [1] instead of $f \in \mathcal{H}$. We call this class of densities $\mathcal{P}_0$. We can further approximate $C(S)$ densities satisfying a tail condition with $\mathcal{P}_0$ densities, and thus can approximate them with deformed exponential family densities. The main result is that that for a uniform base density q_0, kernel deformed exponential families can approximate continuous densities on compact domains arbitrarily well in L^p norm, Hellinger distance, and Bregman divergence.

These results extend Sriperumbudur et al. [2017]'s approximation results to the deformed setting, where standard log and exponential rules cannot be applied. Our setting requires bounding Frechet derivatives and applying the functional mean value theorem. Our Bregman divergence result is analagous to the KL divergence result in Sriperumbudur et al. [2017]. KL divergence is Bregman divergence with the Shannon entropy functional: we show the same for Tsallis entropy. The Bregman divergence describes how close a density's uncertainty is to its first order approximation evaluated at another density. Using the Tsallis entropy functional here is appropriate for deformed exponential families: they maximize it given expected sufficient statistics [Naudts, 2004].

[1] continuous function on domain S vanishing at infinity

5.3 Using Kernels for Continuous Attention

We apply kernel exponential and deformed exponential families to attention. The forward pass computes attention densities and the context vector. The backwards pass uses automatic differentiation. For kernel exponential families we compute kernel weights $\{\gamma_i\}_{i=1}^I$ for $f(t) = \sum_{i=1}^I \gamma_i k(t, t_i)$, where γ_i is computed from some representation of the data and compute $Z = \int_S \exp(f(t))dQ$ numerically. For the deformed case we compute $\tilde{\gamma}_i$ and $\tilde{f}(t) = Z^{\alpha-1} f(t) = \sum_{i=1}^I \tilde{\gamma}_i k(t, t_i)$ followed by $Z = \int_S \exp_{2-\alpha}(\tilde{f}(t))dQ$. The context $c = \mathbb{E}_{T\sim p}[V(T)] = \mathbf{B}\mathbb{E}_p[\Psi(t)]$ requires taking the expectation of $\Psi(T)$ with respect to a (deformed) kernel exponential family density p. Unlike Martins et al. [2020, 2021], difficult normalizing constants prevent us from obtaining closed form expressions. We use numerical integration and automatic differentiation for the forward/backward pass. Algorithm 1 shows how to compute a continuous attention mechanism for a kernel deformed exponential family attention density. The kernel exponential family case is similar.

Algorithm 1 Continuous Attention Mechanism via Kernel Deformed Exponential Families

Choose base density $q_0(t)$ and kernel k. Inducing point locations $\{t_i\}_{i=1}^I$
Parameters $\{\tilde{\gamma}_i\}_{i=1}^I$ the weights for $\tilde{f}(t) = (Z)^{\alpha-1} f(t) = \sum_{i=1}^I \tilde{\gamma}_i k(t, t_i)$, matrix $\mathbf{B}$ for basis weights for value function $V(t) = \mathbf{B}\Psi(t)$. I is number of inducing points.
Forward Pass
Compute $Z = \int \exp_{2-\alpha}(\tilde{f}(t))dQ(t)$ to obtain $p(t) = \frac{1}{Z}\exp_{2-\alpha}(\tilde{f}(t))$ via numerical integration
Compute $\mathbb{E}_{T\sim p}[\Psi(T)]$ via numerical integration
Compute $c = \mathbb{E}_{T\sim p}[V(T)] = \mathbf{B}\mathbb{E}_p[\Psi(T)]$
Backwards Pass use automatic differentiation

5.3.1 Numerical Integration Convergence

The trapezoidal rule's standard one-dimensional convergence rate is $O(\frac{1}{N^2})$ for an integral over a fixed interval, where N is the number of grid points. We would like better convergence guarantees. We can achieve exponential convergence for the numerical integrals of kernel exponential and deformed exponential family attention. We focus on numerical integration over the real line, leaving truncation analysis and higher dimensions to future work. We let $h > 0$ be the grid size.

For functions holomorphic in a strip with rapid decay, the trapezoidal rule has exponential convergence. For kernel exponential family attention, this gives us $O(\exp(-C/h))$ or exponential convergence for some $C > 0$ with appropriate choice of q_0, V, and k. Technical details are in E.1, and are based on extending real-valued analytic functions to complex functions analytic/holomorphic on a strip.

Kernel deformed exponential families, however, are not even differentiable, but we can construct a sequence of differentiable approximations by replacing the positive part/ReLU function in the deformed exponential with softplus for increasing values of the softplus parameter. Each differentiable approximation has exponential convergence, and by taking limits as the softplus parameter tends to infinity we can show that the numerical integral for kernel deformed exponential family attention itself has exponential convergence. Technical details are in E.2.

We also show empirical convergence analysis in Appendix E.3 and figure 2 in that appendix. Both kernel exponential and deformed exponential families see rapid convergence for 1d attention, providing excellent integral approximations with only 5-10 grid points. Further, the numerical integral using softplus is a very close approximation to that using ReLU for softplus parameters 5 and 10.

6 Experiments

We investigate how often multimodal continuous attention outperforms unimodal, given the same architecture. We also investigate whether these methods learn rich multimodal densties. We denote kernel exponential family attention as kernel softmax and the deformed case as kernel sparsemax. Our architectures have: 1) an encoder maps a discrete time series representation to attention density parameters. 2) The value function $V(t; \mathbf{B})$ expresses an embedding of a time series as a linear transformation of basis functions. 3) The context is $c = \mathbb{E}_p[V(T)]$, which is used in 4) a classifier. Fig. 3 in the Appendices visualizes this. For kernel softmax/sparsemax the encoder outputs are the weights

Attention	Accuracy
Cts Softmax	77.72±14.20
Cts Sparsemax	77.96±9.64
Kernel Softmax (ours)	85.71±11.98
Kernel Sparsemax (ours)	88.1±1.50

Table 1: Results for 100 runs each of synthetic time warping classification experiment, $N = 64$. This involves generating $10,000$ trajectories, each of length 95 of unaligned curves.

Method	*Accuracy*	*F1*
Discrete Softmax	97.97	88.67
Cts Softmax	98.20	91.01
Cts Sparsemax	98.30	90.92
Gaussian Mixture	98.12	90.16
Kernel Softmax (ours)	**98.67**	**92.56**
Kernel Sparsemax (ours)	**98.42**	**92.07**
LSTM FCN	98.36	90.45
TST (Transformer)	98.12	90.79

Table 2: Accuracy results on MIT BIH Arrhythmia Classification dataset. The first four rows use the same architecture but different attention mechanisms. Gaussian mixture had 10 components: other choices were tried with lower performance. Rows five and six are our attention mechanisms. LSTM FCN Karim et al. [2017] is an LSTM+fully convolutional network. TST is a transformer from Zerveas et al. [2021]

for the kernel evaluations. For the Gaussian mixture case, the encoder's outputs are the mixture weights and components means and variances. Here we describe one synthetic and two real data experiments. In Appendix F we describe an additional synthetic irregularly sampled accelerometer time series experiment (uWave) and in Appendix G we describe a sentiment classification experiment. All experiments were conducted on either a single Titan X GPU or a single 1080.

6.1 Synthetic Experiment: Time Warping

We simulate time warping and do prediction for unimodal vs multimodal attention densities. Details of the original time aligned stochastic process and inverse warping function are in Appendix B.3. Given global densities for classes p_1, p_2 and an aligned X^*, the class is $\arg\max(\langle p_1, X^*\rangle, \langle p_2, X^*\rangle)$. We generate $10,000$ trajectories of length 95 observed at evenly spaced time points in the interval $[0, 1]$. We use two attention mechanisms (heads), one for each class. Letting V be the value function fit to observed data, the classifier is $\text{softmax}([\langle p_{i1}, V\rangle_{L^2}, \langle p_{i2} V\rangle_{L^2}])$ where p_{i1} and p_{i2} are the attention densities for the first and second class. Table 1 shows prediction results from 100 runs along with 1.96 standard deviation intervals. Kernel methods outperform by $7 - 10\%$.

6.2 ECG heartbeat classification

We use a kaggle version[2] of the MIT Arrhythmia Database [Goldberger et al., 2000]. Not all versions are comparable: Kachuee et al. [2018] report results under well-balanced classes (unclear how they obtain these), while Mousavi and Afghah [2019] augment the dataset with SMOTE. We do no data augmentation, so compare to two popular time series classification baselines in TSAI: a hybrid fully convolutional and LSTM networkKarim et al. [2017], and a transformer Zerveas et al. [2021].

The task is to detect abnormal heart beats from ECG. The five classes are {Normal, Supraventricular premature, Premature ventricular contraction, Fusion of ventricular and normal, Unclassifiable}. There are 87,553 training samples and 21,891 test samples. Each sample is a univariate time series of length 187: we pass this through two convolutional layers in order to obtain a multivariate representation of each time step. We then use an LSTM. The hidden layer is used to construct discrete attention using Bahdanau et al. [2015]. Following Martins et al. [2021] for unimodal continuous

[2]https://www.kaggle.com/mondejar/mitbih-database, license CC0 public domain

Method	*Accuracy*	*F1*
Discrete Softmax	51.59	34.03
Cts Softmax	77.04	77.04
Cts Sparsemax	90.14	90.15
Gaussian Mixture	51.5	34.03
Kernel Softmax (ours)	92.65	92.65
Kernel Sparsemax (ours)	**93.18**	**93.18**
LSTM FCN	90.68	90.66
TST (Transformer)	48.79	48.79

Table 3: Accuracy results on FordA dataset. Kernel sparsemax outperforms baselines, while continuous and kernel softmax have similar performance. We note that several of these methods have *very* poor performance. As a sanity check we fit three classical methods: SVM with Gaussian kernel, logistic regression, and a decision tree. All have under 55% accuracy, suggesting that this dataset is problematic for some methods. Methods described in previous table.

softmax and sparsemax, we first output discrete attention weights $p = (p_1, \cdots, p_L), p \in \Delta^L$ in the probability simplex and then compute $\mu = \mathbb{E}_p[T/L]$ and $\sigma^2 = \mathbb{E}_p[(T/L)^2] - \mu^2$ where $T \sim p$. The value function uses Gaussian RBFs. The final context vector is passed through three feedforward layers. Additional details, including hyperparameters and plots of attention densities, are in Appendix H. Kernel softmax attention is not particularly interpretable. For kernel sparsemax, the attention densities tend to highlight peaks in the signal. Particularly they assign high weight to the R wave, the peak of the QRS complex, of the heartbeat.

6.3 FordA Dataset

This is a binary classification dataset for whether a sympton exists in an automotive subsystem. Each time series has 500 sensor observations, and there are 3601 training samples and 1320 test samples. Table 3 shows results. Kernel sparsemax outperforms baselines, while kernel softmax also does well. We use the same methods and architecture as the previous section, although hyperparameters are slightly different. Several methods, including continuous sparsemax and the transformer, fail completely. To sanity check we fit SVM with a Gaussian kernel, logistic regression, and a decision tree: accuracies are under 55% for these, and we conclude that this data poses difficulty for some classifiers. Appendix I provides some additional details, along with attention density plots. The kernel sparsemax plots often select peaks of a signal while learning rich sparsity patterns.

7 Discussion

In this paper we extend continuous attention mechanisms to use kernel exponential and deformed exponential family densities. The latter is a new flexible class of non-parametric densities with sparse support. We show novel existence properties for kernel exponential and deformed exponential families, prove approximation properties for the latter, and show exponential convergence of numerical integration for both attention mechanisms. We then apply these to the continuous attention framework described in Martins et al. [2020, 2021]. We show results on several datasets. Kernel attention mechanisms tend to outperform unimodal attention, often by a large margin. We also see that in many cases they exhibit multimodality. Kernel sparsemax in particular learns rich sparsity patterns while highlighting peaks of a signal. This was evident in the ECG example, where it tends to give very high weight to R waves in a signal. While this paper is more focused on general methods than a specific potentially dangerous application, potential application areas include wearable sensors and NLP, and negative impact could occur if for instance this method is applied and learns hate speech.

7.1 Limitations

A limitation of this work was the use of numerical integration, which scales poorly with location dimensionality. While we achieve exponential convergence in the 1d case, it is not clear how this will scale to higher dimensional settings. This still allows for multiple observation dimensions at a given 1d location, i.e. multivariate time series.

References

Michael Arbel and Arthur Gretton. Kernel conditional exponential family. In *International Conference on Artificial Intelligence and Statistics*, pages 1337–1346. PMLR, 2018.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

Stéphane Canu and Alex Smola. Kernel methods and the exponential family. *Neurocomputing*, 69 (7-9):714–720, 2006.

Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020.

Bo Dai, Hanjun Dai, Arthur Gretton, Le Song, Dale Schuurmans, and Niao He. Kernel exponential family estimation via doubly dual embedding. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2321–2330. PMLR, 2019.

António Farinhas, André FT Martins, and Pedro MQ Aguiar. Multimodal continuous visual attention mechanisms. *arXiv preprint arXiv:2104.03046*, 2021.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

Robert Israel (https://math.stackexchange.com/users/8508/robert israel). Determine if $f(z) = \log(e^z + 1)$ is analytic and where. Mathematics Stack Exchange. URL `https://math.stackexchange.com/q/3060260`. URL:https://math.stackexchange.com/q/3060260 (version: 2019-01-03).

Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. Ecg heartbeat classification: A deep transferable representation. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 443–444. IEEE, 2018.

Fazle Karim, Somshubra Majumdar, Houshang Darabi, and Shun Chen. Lstm fully convolutional networks for time series classification. *IEEE access*, 6:1662–1669, 2017.

Yin Tat Lee. Lecture 19: Exponentially convergent trapezoidal rule. URL `https://yintat.com/teaching/cse599-winter18/19.pdf`.

Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyou Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 5243–5253. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/6775a0635c302542da2c32aa19d86be0-Paper.pdf`.

Steven Cheng-Xian Li and Benjamin M Marlin. A scalable end-to-end gaussian process adapter for irregularly sampled time series classification. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29, pages 1804–1812. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/9c01802ddb981e6bcfbec0f0516b8e35-Paper.pdf`.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020.

Jiayang Liu, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan. uwave: Accelerometer-based personalized gesture recognition and its applications. *Pervasive and Mobile Computing*, 5(6): 657–675, 2009.

Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *International Conference on Machine Learning*, pages 1614–1623. PMLR, 2016.

André Martins, António Farinhas, Marcos Treviso, Vlad Niculae, Pedro Aguiar, and Mario Figueiredo. Sparse and continuous attention mechanisms. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20989–21001. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/f0b76267fbe12b936bd65e203dc675c1-Paper.pdf`.

André FT Martins, Marcos Treviso, António Farinhas, Pedro MQ Aguiar, Mário AT Figueiredo, Mathieu Blondel, and Vlad Niculae. Sparse continuous distributions and fenchel-young losses. *arXiv preprint arXiv:2108.01988*, 2021.

Pedro Henrique Martins, Zita Marinho, and André FT Martins. ∞-former: Infinite memory transformer. *arXiv preprint arXiv:2109.00301*, 2022.

Sajad Mousavi and Fatemeh Afghah. Inter-and intra-patient ecg heartbeat classification for arrhythmia detection: a sequence to sequence deep learning approach. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1308–1312. IEEE, 2019.

Jan Naudts. Estimators, escort probabilities, and phi-exponential families in statistical physics. *arXiv preprint math-ph/0402005*, 2004.

James O Ramsay and Xiaochun Li. Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):351–363, 1998.

Satya Narayan Shukla and Benjamin Marlin. Interpolation-prediction networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2019.

Satya Narayan Shukla and Benjamin Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=4c0J6lwQ4_`.

Satya Narayan Shukla and Benjamin M Marlin. Heteroscedastic temporal variational autoencoder for irregularly sampled time series. *International Conference on Learning Representations*, 2022.

Huan Song, Deepta Rajan, Jayaraman Thiagarajan, and Andreas Spanias. Attend and diagnose: Clinical time series analysis using attention models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *The Journal of Machine Learning Research*, 18(1):1830–1888, 2017.

Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12(7), 2011.

Elias M Stein and Rami Shakarchi. *Complex analysis*, volume 2. Princeton University Press, 2010.

Dougal Sutherland, Heiko Strathmann, Michael Arbel, and Arthur Gretton. Efficient and principled score estimation with nyström kernel exponential families. In *International Conference on Artificial Intelligence and Statistics*, pages 652–660. PMLR, 2018.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In *Artificial intelligence and statistics*, pages 567–574. PMLR, 2009.

Lloyd N Trefethen and JAC Weideman. The exponentially convergent trapezoidal rule. *siam REVIEW*, 56(3):385–458, 2014.

Yao-Hung Hubert Tsai, Shaojie Bai, Makoto Yamada, Louis-Philippe Morency, and Ruslan Salakhutdinov. Transformer dissection: A unified understanding of transformer's attention via the lens of kernel. *arXiv preprint arXiv:1908.11775*, 2019.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487, 1988.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.

Shaojun Wang, Dale Schuurmans, and Yunxin Zhao. The latent maximum entropy principle. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(2):1–42, 2012.

Li Wenliang, Dougal Sutherland, Heiko Strathmann, and Arthur Gretton. Learning deep kernels for exponential family densities. In *International Conference on Machine Learning*, pages 6737–6746. PMLR, 2019.

Da Xu, Chuanwei Ruan, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. Self-attention with functional time representation learning. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 15915–15925. Curran Associates, Inc., 2019. URL `https://proceedings.neurips.cc/paper/2019/file/cf34645d98a7630e2bcca98b3e29c8f2-Paper.pdf`.

George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2114–2124, 2021.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We included the code for two experiments and are working on preparing the rest.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We did it for two of our smaller scale experiments where the datasets were small and sequence lengths were short: the synthetic time warping experiment and the uWave experiment in Appendix F.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We mention the type of resources used. These are relatively small scale experiments so we don't need to discuss total compute as environmental impacts are negligible.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [No]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]