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 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 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 has: 1) a value function approximating a data representation 2) an attention density chosen to be 'similar' to another data representation, encoding it into a density 3) a context c [Martins et al., 2020] taking an expectation of the value function with respect to the attention density:

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

The value function $V : S \to \mathbb{R}^D$ approximates a data representation, $T \sim p(t)$ is the random variable or vector for locations (temporal, spatial, etc) in domain S, 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. An example of this for irregularly sampled time series is Shukla and Marlin [2021]. 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 an absolutely continuous density $p(t)$ wrt Lebesgue measure. If S is a set of spatial locations and one has image data, then the value function could be a learned representation of an image.

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

$$\hat{p}[f] = \arg \max_{p \in \mathcal{M}_+^1(S)} \int_S p(t) f(t) dQ(t) - \Omega(p). \tag{2}$$

This maximizes regularized L^2 similarity between p and a data representation f. If $\Omega(p) = \int_S p(t) \log p(t) dQ(t)$ 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 $\int_S \hat{p}[f](t) dQ = 1$ (see Martins et al. [2020] for proof). 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), D be the observation dimension, and N be the number of basis functions. This involves regressing the observation matrix $\mathbf{H} \in \mathbb{R}^{D \times L}$ on a matrix $\mathbf{F} \in \mathbb{R}^{N \times L}$ of basis functions $\{\psi_n\}_{n=1}^N$ evaluated at observation locations $\{t_l\}_{l=1}^L$

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

where $\| \cdot \|_F$ is the Frobenius norm.

4 Time Warping

We now draw a connection to time warping to show an advantage of our method. One desirable summary statistic for classification is an expectation of temporal features with respect to a global density. However in many processes features may not be aligned in time, and we only observe unaligned curves. For instance, electrocardiogram (ECG) heartbeat curves have a P-wave, a QRS complex and a T-wave. These have similar patterns between heartbeats, but may have different durations and peak locations. Here we show that the expectation of a temporally aligned curve with respect to a global density is equivalent to the expectation of the unaligned curve with respect to an individualized density. However even if the global density is unimodal, the individualized density may not be. We first define the function that aligns a set of features to common reference times.

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

$$\begin{aligned} h_i(0) &= 0, h_i(\tau) = \tau \\ h_i(t_{0k}) &= t_{ik}, k = 1, \cdots K \\ h_i(t) &= t \text{ if } t \notin [0,\tau] \end{aligned}$$

Let $\{X_i\}_{i=1}^{n}, X_i : S \to \mathbb{R}$ be observed curves, each with K features occurring at individualized times $\{t_{ik}\}_{k=1}^{K} \subset [0,\tau]$ increasing in k. A set of time warping functions $\{h_i\}_{i=1}^{n}$ map reference times to individualized feature times. One can then compute aligned $X_i^*(t) = X_i(h_i(t))$. Each X_i^* has relevant features at the same times $\{t_{0k}\}_{k=1}^{K}$. 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: see Appendix B.2 for an example. Thus we require multimodal continuous attention to represent such statistics.

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.

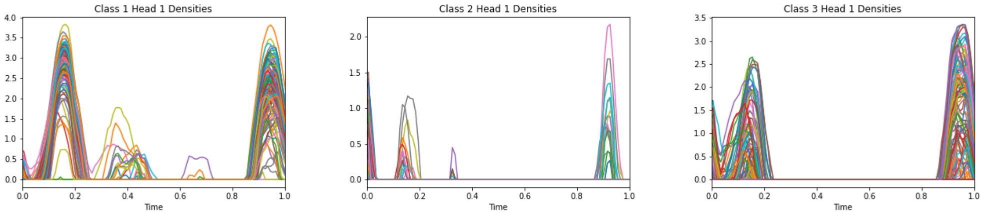

Figure 1: Attention densities for kernel deformed exponential families for the first attention head of the uWave experiment (Appendix F) and all test set participants for two classes. The densities are sparse and have support over different non-overlapping time intervals, which cannot be done with either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

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 and this is the solution to a regularized empirical risk minimization problem. 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.

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 $k(t,s) = t^3s^3 + t^4s^4$ 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 which can be multimodal. 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

Kernel deformed exponential families can approximate continuous densities satisfying a tail condition on compact domains arbitrarily well in L^p norm, Hellinger distance, and Bregman divergence.

Theorem 5.6. *Let $q_0 \in C(S)$, such that $q_0(t) > 0$ for all $t \in S$, where $S \subseteq \mathbb{R}^d$ is locally compact Hausdorff and $q_0(t)$ is the density of Q with respect to a dominating measure ν. Suppose there exists $l > 0$ such that for any $\epsilon > 0, \exists R > 0$ satisfying $|p(t) - l| \leq \epsilon$ for any t with $\|t\|_2 > R$. Define*

$$\mathcal{P}_c = \{p \in C(S) : \int_S p(t)dQ = 1, p(t) \geq 0, \forall t \in S \textit{ and } p - l \in C_0(S)\}.$$

Suppose $k(t,\cdot) \in C_0(S) \forall t \in S$ and the kernel integration condition (Eqn. 11) holds. Then kernel deformed exponential families are dense in $\mathcal{P}_c$ wrt L^r norm, Hellinger distance and Bregman divergence for the α-Tsallis negative entropy functional.

The proof (D.4) idea is that under a kernel integrability condition, deformed exponential families parametrized by $f \in \mathcal{H}$ are dense in those parametrized by $f \in C_0(S)$ [1] (we denote those parametrized by $f \in C_0(S)$ as $\mathcal{P}_0$). We can approximate $C(S)$ densities satisfying the tail condition with $\mathcal{P}_0$ densities, and thus with deformed exponential family densities. This extends Sriperumbudur et al. [2017]'s approximation to the deformed case: standard log and exponential rules cannot be applied. It requires bounding Frechet derivatives and applying the functional mean value theorem.

[1] continuous function on domain S vanishing at infinity

5.3 Using Kernels for Continuous Attention

Here we describe how to compute continuous attention mechanisms with attention densities parametrized by functions in an RKHS $\mathcal{H}$ in practice. Algorithm 1 shows the kernel deformed exponential family case: the kernel exponential family case involves a similar algorithm. Given a base density, kernel, and inducing point locations, we start by computing kernel weights $\tilde{\gamma}_i$ for $\tilde{f}(t) = Z^{\alpha-1}f(t) = \sum_{i=1}^{I} \tilde{\gamma}_i k(t, t_i)$ and estimating the matrix **B** for basis weights for the value function $V(t) = \mathbf{B}\Psi(t)$. Unlike density estimation, this form for f is simply a practical way to obtain f in an RKHS, rather than a solution to an empirical risk minimization problem. We then compute the normalizing constant $Z = \int_S \exp_{2-\alpha}(\tilde{f}(t))dQ$ via numerical integration and use it to normalize $\tilde{f}(t)$ to obtain the attention density $p(t)$. Finally we compute the context $c = \mathbb{E}_{T\sim p}[V(T)] = \mathbf{B}\mathbb{E}_p[\Psi(t)]$ by taking the expectation of $\Psi(T)$ with respect to a deformed kernel exponential family density p. Unlike Martins et al. [2020, 2021], we lack closed form expressions and use numerical integration. In the backwards pass we use automatic differentiation. Note that in some cases we have numerical underflow when computing the normalizing constant. We discuss solutions for this in Appendix E.1.

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 **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.2, 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.3.

We also show empirical convergence analysis in Appendix E.4 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

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]

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 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 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

[2] https://www.kaggle.com/datasets/shayanfazeli/heartbeat, license Open Data Commons Attribution License v1.0

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.

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 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.

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.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 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.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150, 2011.

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.

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.

Hiroki Suyari, M Tsukada, and Y Uesaka. Mathematical structure derived from tsallis entropy. In *Proc. 4th Asia-Eur. Workshop Inf. Theory (AEW)*, pages 1–4, 2004.

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]

Overview of Appendices

Appendices A to E are mostly theoretical. In Appendix A we describe how the Gaussian mixture attention from Farinhas et al. [2021] relates to the theoretical framework of Martins et al. [2020, 2021]. In Appendix B we discuss time warping, showing that sufficiently flexible continuous attention can represent time warping, that in at least in one case a multimodal density is required, and give some details describing our experiments. In Appendix C we discuss normalization for kernel exponential families, while in Appendix D we do the same along with approximation results for kernel deformed exponential families. In Appendix E we discuss rate of convergence for approximating the attention mechanism with numerical integration, and show some synthetic experiments showing rapid convergence empirically.

The remaining appendices are empirical. In Appendix F and Appendix G, we include two additional experiments. The first analyzes an accelerometer classification dataset, where we add synthetic irregular sampling. The second analyzes the IMDB dataset, a sentiment classification dataset. In Appendix H we provide additional details for the MIT-BIH experiment. Finally, in Appendix I we provide additional details for the FordA experiment. The latter two appendices in particular focus on visual comparisons of attention densities. We find that kernel deformed exponential families/sparsemax tend to learn more interesting looking attention densities than all of the other methods, and for ECG kernel sparsemax has the interpretation of selecting regions where the electrical stimuli are largest.

The licenses are: for MIT BIH open data commons attribution. For UCR time series (uWave and Ford) the exact license is difficult to find but the citation is Dau et al. [2019]. For IMDB it is again difficult to find the exact license, but the original paper is Maas et al. [2011] and the website[3] asks that it be cited.

A Gaussian Mixture Model

In this section we descibe how the Gaussian mixture attention of Farinhas et al. [2021] relates to the optimization definition of attention densities in Martins et al. [2020, 2021]. In fact their attention densities solve a related but different optimization problem. Martins et al. [2020, 2021] show that exponential family attention densities maximize a regularized linear predictor of the expected sufficient statistics of locations. In contrast, Farinhas et al. [2021] find a joint density over locations and latent states, and maximize a regularized linear predictor of the expected *joint* sufficient statistics. They then take the marginal location densities to be the attention densities.

Let $\Omega(p)$ be Shannon entropy and consider two optimization problems:

$$\arg \max_{p \in \mathcal{M}_+^1(S)} \langle \theta, \mathbb{E}_p[\phi(T)] \rangle_{l^2} - \Omega(p)$$

$$\arg \max_{p \in \mathcal{M}_+^1(S)} \langle \theta, \mathbb{E}_p[\phi(T, Z)] \rangle_{l^2} - \Omega(p)$$

The first is Eqn. 2 with $f = \theta^T \phi(t)$ and rewritten to emphasize expected sufficient statistics. If one solves the second with variables Z, we recover an Exponential family joint density

$$\hat{p}_{\Omega_\alpha}[f](t, z) = \exp(\theta^T \phi(t, z) - A(\theta)).$$

This encourages the joint density of T, Z to be similar to a *complete data* representation $\theta^T \phi(t, z)$ of both location variables T and latent variables Z, instead of encouraging the density of T to be similar to an observed data representation $\theta^T \phi(t)$. The latter optimization is equivalent to

$$\arg \max_{p \in \mathcal{M}_+^1(S)} \Omega(p)$$

$$\text{s.t.}$$

$$\mathbb{E}_{p(T,Z)}[\phi_m(T, Z)] = c_m, m = 1, \cdots, M.$$

The constraint terms c_m are determined by θ. Thus, this maximizes the joint entropy of Z and T, subject to constraints on the expected joint sufficient statistics.

[3] http://ai.stanford.edu/ amaas/data/sentiment/

To recover EM learned Gaussian mixture densities, one must select ϕ_m so that the marginal distribution of T will be a mixture of Gaussians, and relate c_m to the EM algorithm used to learn the mixture model parameters. For the first, assume that Z is a multinomial random variable with a single trial taking $|Z|$ possible values and let $\phi(t,z) = (I(z=1), \cdots, I(z=|Z|-1), I(z=1)t, I(z=1)t^2, \cdots, I(z=|Z|)t, I(z=|Z|)t^2)$. These are multinomial sufficient statistics with a single trial, followed by the sufficient statistics of $|Z|$ Gaussians multiplied by indicators for each z. Then $p(T|Z)$ will be Gaussian, $p(Z)$ will be multinomial, and $p(T)$ will be a Gaussian mixture. For contraints, Farinhas et al. [2021] have

$$\mathbb{E}_{p(T,Z)}[\phi_m(T,Z)] = \sum_{l=1}^{L} w_l \sum_{k=1}^{|Z|} p_{\text{old}}(Z=k|T=t_l)\phi_m(t_l,k), \qquad (9)$$
$$m = 1, \cdots, M$$

at each EM iteration: $p_{\text{old}}(Z|T=t_l)$ is the previous iteration's latent state density conditional on the observation location, w_l are discrete attention weights, and t_l is a discrete attention location. That EM has this constraint was shown in Wang et al. [2012]. This matches expected joint sufficient statistics to those implied by discrete attention over locations, taking into account the dependence between z and t_l given by old model parameters. An alternative is simply to let θ be the output of a neural network. While the constraints lack the intuition of Eqn. 9, it avoids the need to run EM. We focus on this case and use it for our baselines.

B Time Warping

B.1 Proof of 4.2

Here we prove that continuous attention can represent time warping. We change measure to Lebesgue, use the equivalence between Lebesgue and Riemann for almost everywhere continuous functions, use integration by substitution, and then change back.

$$\begin{aligned}
\mathbb{E}_p X_i^*(U) &= \int_{[0,\tau]} p(u) X_i^*(u) dQ \\
&= \int_{[0,\tau]} p(u) q_0(u) X_i^*(u) d\nu(u) \\
&= \int_0^\tau p(u) q_0(u) X_i^*(u) du \\
&= \int_0^\tau p(g_i(t)) q_0(g_i(t)) X_i^*(g_i(t)) g_i'(t) dt \\
&= \int_0^\tau p(g_i(t)) \frac{q_0(g_i(t))}{q_0(t)} g_i'(t) X_i^*(g_i(t)) q_0(t) dt \\
&= \int_{[0,\tau]} p(g_i(t)) \frac{q_0(g_i(t))}{q_0(t)} g_i'(t) X_i^*(g_i(t)) dQ \\
&= \int_{[0,\tau]} p_i(t) X_i(t) dQ
\end{aligned}$$

Further, since h is strictly increasing, g is as well so that $g_i'(t) > 0 \forall t \in [0,\tau]$. This gives us that $p_i(t)$ is non-negative, and we have to show that it integrates to 1.

$$\begin{aligned}
\int_{[0,\tau]} p(g_i(t)) \frac{q_0(g_i(t))}{q_0(t)} g_i'(t) dQ &= \int_0^\tau p(g_i(t)) q_0(g_i(t)) g_i'(t) dt \\
&= \int_0^\tau p(u) q_0(u) du \\
&= 1
\end{aligned}$$

B.2 Multimodal Densities are Required

Consider the case where $p_{\text{global}}(t)$ is an exponential density function and $g(t)$ the inverse warping function has the following form

$$p_{\text{global}}(t) = \begin{cases} \lambda \exp(-\lambda(t)), t \geq 0 \\ 0, \text{ else} \end{cases}$$

$$g(t) = \begin{cases} C_1 \int_0^t \exp(W(u))du, t \in [0, \tau] \\ t \text{ else} \end{cases}$$

for $C_1 = \frac{\tau}{\int_0^\tau \exp(W(u))du}$ and some twice diferentiable function W. The role of C_1 is to ensure that $g(\tau) = \tau$. Then

$$p_i(t) = \lambda C_1 \exp(-\lambda C_1 \int_0^t \exp(W(u))du + W(t))$$

so that

$$\begin{aligned} p_i'(t) &= (-\lambda C_1 \exp(t) + W'(t))p_i(t) \\ p_i''(t) &= (-\lambda C_1 \exp(t) + W''(t))p_i(t) \\ &\quad + (-\lambda C_1 \exp(t) + W'(t))^2 p_i(t) \end{aligned}$$

Now note that depending on the sign and magnitude of $W''(t)$, $p_i''(t)$ may be either positive or negative and thus $p_i(t)$ may lack a unique optimum and may be multimodal.

B.3 Time Warping Experimental Details

This describes the aligned vs the observed stochastic process. The original time aligned stochastic process is given by

$$X^*(t) = \begin{cases} Z_0 \cos 9\pi t, 0 \leq t < 0.25 \\ Z_1 t^2, 0.25 \leq t < 0.5 \\ Z_2 \sin t, 0.5 \leq t < 0.75 \\ Z_3 \cos 17\pi t, 0.75 \leq t < 1 \end{cases}$$

where $Z_k \sim U(-4, 4), k = 1, 2, 3, 4$. We instead observe realizations $X_i(t) = X_i^*(g(t))$ at fixed time points with the following inverse warp function

$$g_i(t) = \begin{cases} C_i \int_0^t \exp(-s\lambda_i)ds, \lambda_i \sim U(0, 25), t \in [0, 1] \\ t \text{ else} \end{cases}$$

The global densities are $p_1 \sim U(0, 0.5)$ and $p_2 \sim U(0.5, 1)$.

C Proof Related to Proposition 5.1

C.1 Proof of Proposition 5.1

Here we prove that if the kernel evaluated twice at the same point grows sufficiently slowly with respect to base density tail decay, then we can normalize.

Proof. This proof has several parts. We first bound the RKHS function f and use the general tail bound we assumed to give a tail bound for the one dimensional marginals T_{Qd} of T_Q. Using the RKHS function bound, we then bound the integral of the unnormalized density in terms of expectations with respect to these finite dimensional marginals. We then express these expectations over finite dimensional marginals as infinite series of integrals. For each integral within the infinite series, we use the finite dimensional marginal tail bound to bound it, and then use the ratio test to show that the infinite series converges. This gives us that the original unnormalized density has a finite integral.

We first note, following Wenliang et al. [2019], that the bound on the kernel in the assumption allows us to bound f in terms of two constants and the absolute value of the point at which it is evaluated.

$$\begin{aligned}|f(t)| &= |\langle f, k(t,\cdot)\rangle_{\mathcal{H}}| \text{ reproducing property}\\ &\leq \|f\|_{\mathcal{H}}\|k(t,\cdot)\|_{\mathcal{H}} \text{ Cauchy Schwarz}\\ &= \|f\|_{\mathcal{H}}\sqrt{\langle k(t,\cdot), k(t,\cdot)\rangle_{\mathcal{H}}}\\ &= \|f\|_{\mathcal{H}}\sqrt{k(t,t)}\\ &\leq \|f\|_{\mathcal{H}}\sqrt{L_k\|t\|^{\xi} + C_k} \text{ by assumption}\\ &\leq C_0 + C_1\|t\|^{|\xi|/2} \text{ for some } C_1, C_2 > 0.\end{aligned}$$

We can write $T_Q = (T_{Q1}, \cdots, T_{QD})$. Let e_d be a standard Euclidean basis vector. Then by the assumption and setting $u = e_d$ we have

$$P(|T_{Qd}| \geq z) \leq C_q \exp(-vz^{\eta})$$

Letting Q_d be the marginal law,

$$\begin{aligned}\int_S \exp(f(t))dQ &\leq \int_S \exp(C_0 + C_1\|t\|^{\xi/2})dQ\\ &= \exp(C_0)\int_S \exp(C_1\|t\|^{\xi/2})dQ\\ &= \exp(C_0)\mathbb{E}\exp(C_1\|T_Q\|^{\xi/2})\\ &\leq \exp(C_0)\mathbb{E}\exp(C_1(\sqrt{d}\max_{d=1,\cdots,D}|T_{Qd}|)^{\xi/2})\\ &\leq \exp(C_0)\sum_{d=1}^{D}\mathbb{E}\exp(C_2|T_{Qd}|^{\xi/2})\end{aligned}$$

which will be finite if each $\mathbb{E}\exp(C_2|T_{Qd}|^{\xi/2}) < \infty$. Now letting S_d be the relevant dimension of S,

$$\begin{aligned}\mathbb{E}\exp(C_2|T_{Qd}|^{\xi/2}) &= \int_{S_d}\exp(C_2|t_d|^{\xi/2})dQ_d\\ &\leq \sum_{j=-\infty}^{-1}\int_j^{j+1}\exp(C_2|t_d|^{\xi/2})dQ_d + \sum_{j=0}^{\infty}\int_j^{j+1}\exp(C_2|t_d|^{\xi/2})dQ_d\end{aligned}$$

where the inequality follows since $S_d \subseteq \mathbb{R}$, exp is a non-negative function and probability measures are monotonic. We will show that the second sum converges. Similar techniques can be shown for the first sum. Note that for $j \geq 0$

$$\begin{aligned}Q_d([j, j+1)) &= P(T_d \geq j) - P(T_d \geq j+1)\\ &\leq P(T_d \geq j)\\ &\leq C_q \exp(-vj^{\eta}) \text{ by assumption}\end{aligned}$$

Then

$$\begin{aligned}\sum_{j=0}^{\infty}\int_j^{j+1}\exp(C_2|t_d|^{\xi/2})dQ_d &\leq \sum_{j=0}^{\infty}\exp(C_2|j|^{\xi/2})Q_d([j, j+1))\\ &\leq \sum_{j=0}^{\infty}C_Q\exp(C_2|j|^{\xi/2} - vj^{\eta})\end{aligned}$$

Let $a_j = \exp(C_2|j|^{\xi/2} - vi^{\eta})$. We will use the ratio test to show that the RHS converges. We have

$$\left|\frac{a_{j+1}}{a_j}\right| = \exp(C_2((j+1)^{\xi/2} - j^{\xi/2}) - v[(j+1)^{\eta} - j^{\eta}]). \tag{10}$$

We want this ratio to be < 1 for large j. We thus need to select η so that for sufficiently large j, we have

$$\frac{C_1}{v}((j+1)^{\xi/2} - j^{\xi/2}) < [(j+1)^{\eta} - j^{\eta}].$$

Assume that $\eta > \frac{\xi}{2}$. Then

$$\begin{aligned}\frac{(j+1)^{\eta} - j^{\eta}}{(j+1)^{\xi/2} - j^{\xi/2}} &= \frac{j^{\eta}[(1+\frac{1}{j})^{\eta} - 1]}{j^{\xi/2}[(1+\frac{1}{j})^{\xi/2} - 1]} \\ &\geq j^{\eta - \xi/2}.\end{aligned}$$

Since the RHS is unbounded for $\eta > \frac{\xi}{2}$, we have that Eqn. 10 holds for sufficiently large j. By the ratio test $\mathbb{E}_{q_d(t)} \exp(C_2|T_d|^{\xi/2}) = \sum_{j=-\infty}^{-1} \int_j^{j+1} \exp(C_2|t_d|^{\xi/2})dQ_d + \sum_{j=0}^{\infty} \int_j^{j+1} \exp(C_2|t_d|^{\xi/2})dQ_d$ is finite. Thus putting everything together we have

$$\begin{aligned}\int_S \exp(f(t))dQ &\leq \int_S \exp(C_0 + C_1\|t\|^{\xi/2})dQ \\ &< \exp(C_0) \sum_{d=1}^{D} \mathbb{E} \exp(C_2|T_{Qd}|^{\xi/2}) \\ &< \infty\end{aligned}$$

and $\tilde{p}(t)$ can be normalized.

□

C.2 Proof of Corollary 5.2

Here we prove two special cases of kernel growth rate and tail decay.

Proof. Let $\xi = 4$. Then $\eta > 2$ and

$$\begin{aligned}P(|u^T T| > t) &\leq P(|u^T T| \geq t) \text{ monotonicity} \\ &\leq C_Q \exp(-vt^{\eta}) \\ &< C_Q \exp(-vt^2).\end{aligned}$$

The second case is similar. For the uniformly bounded kernel,

$$\begin{aligned}\int_S \exp(\langle f, k(\cdot, t)\rangle_{\mathcal{H}})dQ &\leq \exp(\|f\|_{\mathcal{H}}\sqrt{C_k}) \int_S dQ \\ &= \exp(\|f\|_{\mathcal{H}}\sqrt{C_k}) \\ &< \infty.\end{aligned}$$

The first line follows from Cauchy Schwarz and $\xi = 0$ □

D Proofs Related to Kernel Deformed Exponential Family

D.1 Proof of Lemma 5.3

Here we prove a key equality for calculating normalization of deformed exponential families.

Proof. The high level idea is to express a term inside the deformed exponential family that becomes $1/Z$ once outside. Recalling the definition of β-exponential in Eqn. 5,

$$\begin{aligned}
\exp_{2-\alpha}(f(t) - \log_\alpha(Z)) &= [1 + (\alpha - 1)(f(t) - \log_\alpha Z)]_+^{\frac{1}{\alpha-1}} \\
&= [1 + (\alpha - 1)f(t) - (\alpha - 1)\frac{Z^{1-\alpha} - 1}{1 - \alpha}]_+^{\frac{1}{\alpha-1}} \\
&= [1 + (\alpha - 1)f(t) + Z^{1-\alpha} - 1]_+^{\frac{1}{\alpha-1}} \\
&= [(\alpha - 1)f(t) + Z^{1-\alpha}]_+^{\frac{1}{\alpha-1}} \\
&= [(\alpha - 1)f(t)\frac{Z^{\alpha-1}}{Z^{\alpha-1}} + Z^{1-\alpha})]_+^{\frac{1}{\alpha-1}} \\
&= \frac{1}{Z}[(\alpha - 1)f(t)Z^{\alpha-1} + 1]_+^{\frac{1}{\alpha-1}} \\
&= \frac{1}{Z}\exp_{2-\alpha}(Z^{\alpha-1}f(t))
\end{aligned}$$

□

D.2 Proof of Proposition 5.4

Here we prove that if $\tilde{f}(t)$ has compact support and the kernel grows sufficiently slowly, then its deformed exponential has a finite integral.

Proof.

$$\begin{aligned}
\int_S \exp_{2-\alpha}(\tilde{f}(t))dQ &= \int_S [1 + (\alpha - 1)\tilde{f}(t)]_+^{\frac{1}{\alpha-1}} dQ \\
&= \int_{\|t\| \leq C_t} [1 + (\alpha - 1)\tilde{f}(t)]_+^{\frac{1}{\alpha-1}} dQ \\
&\leq \int_{\|t\| \leq C_t} [1 + (\alpha - 1)(C_0 + C_1|C_t|^{\xi/2})]_+^{\frac{1}{\alpha-1}} dQ \\
&< \infty
\end{aligned}$$

where the 2nd to last line is due to the inequality $|f(t)| \leq C_0 + C_1\|t\|^{\xi/2}$ from the proof in Appendix C.1. □

D.3 Proof of Corollary 5.5

Here we prove that we can normalize deformed exponential families under the previous conditions.

Proof. From proposition 5.4 and the assumption,

$$\int_S \exp_{2-\alpha}(\tilde{f}(t))dQ = Z$$

for some $Z > 0$. Then

$$\begin{aligned}
\int_S \frac{1}{Z}\exp_{2-\alpha}(Z^{\alpha-1}f(t))dQ &= 1 \\
\int_S \exp_{2-\alpha}(f(t) - \log_\alpha Z)dQ &= 1
\end{aligned}$$

where the second line follows from lemma 5.3. Set $A_\alpha(f) = \log_\alpha(Z)$ and we are done. □

D.4 Approximation Theory

This section proves Theorem 5.6. We start with a Proposition, which says that under a kernel integration condition, deformed exponential families parametrized by functions in $\mathcal{H}$ can approximate similar densities parametrized by functions in $C_0(S)$ arbitrarily well.

Proposition D.1. *Define*

$$\mathcal{P}_0 = \{\pi_f(t) = \exp_{2-\alpha}(f(t) - A_\alpha(f)), t \in S : f \in C_0(S)\}$$

where $S \subseteq \mathbb{R}^d$. Suppose $k(x, \cdot) \in C_0(S), \forall x \in S$ and

$$\int \int k(x, y) d\mu(x) d\mu(y) > 0, \forall \mu \in M_b(S) \backslash \{0\}. \tag{11}$$

here $M_b(S)$ is the space of bounded measures over S. Then the set of deformed exponential families is dense in $\mathcal{P}_0$ wrt $L^r(Q)$ norm and Hellinger distance.

Proof. The kernel integration condition tells us that $\mathcal{H}$ is dense in $C_0(S)$ with respect to L^∞ norm. This was shown in Sriperumbudur et al. [2011]. For the L^r norm, we apply $\|p_f - p_g\|_{L^r} \leq 2M_{\exp}\|f - g\|_\infty$ from Lemma D.5 with $f \in C_0(S)$, $g \in \mathcal{H}$, and $f_0 = f$. L^1 convergence implies Hellinger convergence.

□

We can then use this for our main proof of Theorem 5.6. Note that our Bregman divergence result is analagous to Sriperumbudur et al. [2017]'s KL divergence result. KL divergence is Bregman divergence with the Shannon entropy functional: we show the same for Tsallis entropy, which is maximized given expected sufficient statistics by deformed exponential families [Naudts, 2004]. The Bregman divergence describes how close a density's uncertainty is to its first order approximation evaluated at another density.

Proof. For any $p \in \mathcal{P}_c$, define $p_\delta = \frac{p+\delta}{1+\delta}$. Then

$$\begin{aligned} \|p - p_\delta\|_r &= \frac{\delta}{1+\delta}\|p\|_r \\ &\to 0 \end{aligned}$$

for $1 \leq r \leq \infty$. Thus for any $\epsilon > 0, \exists \delta_\epsilon > 0$ such that for any $0 < \theta < \delta_\epsilon$, we have $\|p - p_\theta\|_r \leq \epsilon$, where $p_\theta(t) > 0$ for all $t \in S$.

Define $f = \left(\frac{1+\theta}{l+\theta}\right)^{1-\alpha} \log_{2-\alpha} p_\theta \frac{1+\theta}{l+\theta}$. Since $p \in C(S)$, so is f. Fix any $\eta > 0$ and note that

$$\begin{aligned} f(t) &\geq \eta \\ \left(\frac{1+\theta}{l+\theta}\right)^{1-\alpha} \log_{2-\alpha} p_\theta \frac{1+\theta}{l+\theta} &\geq \eta \\ \log_{2-\alpha} p_\theta \frac{1+\theta}{l+\theta} &\geq \left(\frac{1+\theta}{l+\theta}\right)^{\alpha-1} \eta \\ p_\theta \frac{1+\theta}{l+\theta} &\geq \exp_{2-\alpha}\left(\left(\frac{1+\theta}{l+\theta}\right)^{\alpha-1} \eta\right) \\ p_\theta &\geq \frac{l+\theta}{1+\theta} \exp_{2-\alpha}\left(\left(\frac{1+\theta}{l+\theta}\right)^{\alpha-1} \eta\right) \\ p - l &\geq (l+\theta)\left(\exp_{2-\alpha}\left(\left(\frac{1+\theta}{l+\theta}\right)^{\alpha-1} \eta\right) - 1\right) \end{aligned}$$

Thus

$$\begin{aligned}A &= \{t : f(t) \geq \eta\}\\ &= \left\{p - l \geq (l+\theta)\left(\exp_{2-\alpha}\left(\left(\frac{1+\theta}{l+\theta}\right)^{\alpha-1}\eta\right) - 1\right)\right\}\end{aligned}$$

Since $p - l \in C_0(S)$ the set on the second line is bounded. Thus A is bounded so that $f \in C_0(S)$. Further, by Lemma 5.3

$$p_\theta = \exp_{2-\alpha}\left(f - \log_\alpha \frac{1+\theta}{l+\theta}\right)$$

giving us $p_\theta \in \mathcal{P}_0$. By Proposition D.1 there is some p_g in the deformed kernel exponential family so that $\|p_\theta - p_g\|_{L^r(S)} \leq \epsilon$. Thus $\|p - p_g\|_r \leq 2\epsilon$ for any $1 \leq r \leq \infty$. To show convergence in Helinger distance, note

$$\begin{aligned}H^2(p, p_g) &= \frac{1}{2}\int_S (\sqrt{p} - \sqrt{p_g})^2 dQ\\ &= \frac{1}{2}\int_S (p - 2\sqrt{pp_g} + p_g) dQ\\ &\leq \frac{1}{2}\int_S (p - 2\min(p, p_g) + p_g) dQ\\ &= \frac{1}{2}\int_S |p - p_g| dQ\\ &= \frac{1}{2}\|p - p_g\|_1\end{aligned}$$

so that $L^1(S)$ convergence, which we showed, implies Hellinger convergence. Let us consider the Bregman divergence. Note the generalized triangle inequality[4] for Bregman divergence

$$B_{\Omega_\alpha}(p, p_g) = \underbrace{B_{\Omega_\alpha}(p, p_\theta)}_{I} + \underbrace{B_{\Omega_\alpha}(p_\theta, p_g)}_{II} - \underbrace{\langle p - p_\theta, \nabla\Omega_\alpha(p_\theta) - \nabla\Omega_\alpha(p_g)\rangle_2}_{III} \tag{12}$$

Term I

$$\begin{aligned}B_{\Omega_\alpha}(p, p_\theta) &= \frac{1}{\alpha(\alpha-1)}\int_S (p^\alpha - p_\theta^\alpha) dQ - \langle \nabla\Omega_\alpha(p_\theta), p - p_\theta\rangle\\ &= \frac{1}{\alpha(\alpha-1)}\int_S (p^\alpha - p_\theta^\alpha) dQ - \frac{1}{\alpha-1}\int p_\theta^{\alpha-1}(p - p_\theta) dQ\\ &\leq \frac{1}{\alpha(\alpha-1)}\int_S |p^\alpha - p_\theta^\alpha| dQ + \frac{1}{\alpha-1}\|p_\theta^{\alpha-1}\|_1 \|\|p - p_\theta\|_\infty\end{aligned}$$

Note that the Bregman divergence is non-negative and thus we only need to worry about an upper bound. The first term on the rhs clearly vanishes as $\theta \to 0$ since $p_\theta \to p$. For the second term, we already showed that $\|p - p_\theta\|_\infty \to 0$.

Term II

Fix θ. Then term II converges to 0 by Lemma D.5.

Term III

For term III,

$$\langle p - p_\theta, \nabla\Omega_\alpha(p_\theta) - \nabla\Omega_\alpha(p_g)\rangle_2 \leq \|p - p_\theta\|_\infty \|\nabla\Omega_\alpha(p_\theta) - \nabla\Omega_\alpha(p_g)\|_1$$

[4] actually an equality, see https://www2.cs.uic.edu/ zhangx/teaching/bregman.pdf for proof

Clearly the first term on the rhs converges by L^r convergence. The L^1 term for the gradient is given by

$$\begin{aligned}\|\nabla\Omega_\alpha(p_\theta) - \nabla\Omega_\alpha(p_g)\|_1 &= \frac{1}{\alpha-1}\int |p_\theta(t)^{\alpha-1} - p_g(t)^{\alpha-1}|dQ \\ &\leq \int (\|p_\theta\|_\infty + \|p_\theta - p_g\|_\infty)^{\alpha-2}\|p_\theta - p_g\|_\infty dQ \text{ Eqn. 16} \\ &= (\|p_\theta\|_\infty + \|p_\theta - p_g\|_\infty)^{\alpha-2}\|p_\theta - p_g\|_\infty\end{aligned}$$

so that the inner product terms are bounded as

$$|\langle p - p_\theta, \nabla\Omega_\alpha(p_\theta) - \nabla\Omega_\alpha(p_g)\rangle_2| \leq (\|p_\theta\|_\infty + \|p_\theta - p_g\|_\infty)^{\alpha-2}\|p_\theta - p_g\|_\infty\|p - p_\theta\|_\infty$$

□

Lemma D.2. *(Functional Taylor's Theorem) Let $F : X \to \mathbb{R}$ where X is a Banach space. Let $f, g \in X$ and let F be k times Gateaux differentiable. Then we can write*

$$F(g) = \sum_{i=0}^{k-1} F^i(f)(g-f)^i + F^k(f + c(g-f))(g-f)^k$$

for some $c \in [0, 1]$.

Proof. This is simply a consequence of expressing a functional as a function of an $\epsilon \in [0, 1]$, which restricts the functional to take input functions between two functions. We then apply Taylor's theorem to the function and apply the chain rule for Gateaux derivatives to obtain the resulting Taylor remainder theorem for functionals.

Let $G(\eta) = F(f + \eta(g - f))$. By Taylor's theorem we have

$$G(1) = G(0) + G'(0) + \cdots + G^k(c)$$

and applying the chain rule gives us

$$F(g) = \sum_{i=0}^{k-1} F^i(f)(g-f)^i + F^k(f + c(g-f))(g-f)^k$$

□

Lemma D.3. *(Functional Mean Value Theorem) Let $F : X \to \mathbb{R}$ be a Gateaux differentiable functional where $f, g \in X$ some Banach space with norm $\|\cdot\|$. Then*

$$|F(f) - F(g)| \leq \|F'(h)\|_{op}\|f - g\|$$

where $h = g + c(f - g)$ for some $c \in [0, 1]$, $F'(h)$ is the Gateaux derivative of F, and $\|\cdot\|_{op}$ is the operator norm $\|A\|_{op} = \inf\{c > 0 : \|Ax\| \leq c\|x\| \forall x \in X\}$.

Proof. Consider $G(\eta) = F(g + \eta(f - g))$. Apply the ordinary mean value theorem to obtain

$$\begin{aligned}G(1) - G(0) &= G'(c), c \in [0, 1] \\ &= F'(g + c(f - g)) \cdot (f - g)\end{aligned}$$

and thus

$$|F(f) - F(g)| \leq \|F'(h)\|_{\text{op}}\|f - g\|$$

□

Claim 1. Consider $\mathcal{P}_\infty = \{p_f = \exp_{2-\alpha}(f - A_\alpha(f)) : f \in L^\infty(S)\}$. Then for $p_f \in \mathcal{P}_\infty$, $A_\alpha(f) \leq \|f\|_\infty$.

Proof.

$$\begin{aligned}
p_f(t) &= \exp_{2-\alpha}(f(t) - A_\alpha(f)) \\
&\leq \exp_{2-\alpha}(\|f\|_\infty - A_\alpha(f)) \text{ for } 1 < \alpha \leq 2 \\
\int_S p_f(t)dQ &\leq \int_S \exp_{2-\alpha}(\|f\|_\infty - A_\alpha(f))dQ \\
1 &\leq \exp_{2-\alpha}(\|f\|_\infty - A_\alpha(f)) \\
\log_{2-\alpha} 1 &\leq \|f\|_\infty - A_\alpha(f) \\
A_\alpha(f) &\leq \|f\|_\infty
\end{aligned}$$

where for the second line recall that we assumed that throughout the paper $1 < \alpha \leq 2$. □

Lemma D.4. *Consider $\mathcal{P}_\infty = \{p_f = \exp_{2-\alpha}(f - A_\alpha(f)) : f \in L^\infty(S)\}$. Then the Frechet derivative of $A_\alpha : L^\infty \to \mathbb{R}$ exists. It is given by the map*

$$\begin{aligned}
A'(f)(g) &= \mathbb{E}_{\tilde{p}_f^{2-\alpha}}(g(T)) \\
&= \frac{\int p_f^{2-\alpha}(t)g(t)dQ}{\int p_f^{2-\alpha}(t)dQ}
\end{aligned}$$

Proof. This proof has several parts. We first derive the Gateaux differential of p_f in a direction $\psi \in L^\infty$ and as it depends on the Gateaux differential of $A_\alpha(f)$ in that direction, we can rearrange terms to recover the latter. We then show that it exists for any $f, \psi \in L^\infty$. Next we show that the second Gateaux differential of $A_\alpha(f)$ exists, and use that along with a functional Taylor expansion to prove that the first Gateaux derivative is in fact a Frechet derivative.

In Martins et al. [2020] they show how to compute the gradient of $A_\alpha(\theta)$ for the finite dimensional case: we extend this to the Gateaux differential. We start by computing the Gateaux differential of p_f.

$$\begin{aligned}
\frac{d}{d\eta}p_{f+\eta\psi}(t) &= \frac{d}{d\eta}\exp_{2-\alpha}(f(t) + \eta\psi(t) - A_\alpha(f + \eta\psi)) \\
&= \frac{d}{d\eta}[1 + (\alpha - 1)(f(t) + \eta\psi(t) - A_\alpha(f + \eta\psi))]_+^{1/(\alpha-1)} \\
&= [1 + (\alpha - 1)(f(t) + \eta\psi(t) - A_\alpha(f + \eta\psi))]_+^{(2-\alpha)/(\alpha-1)} \left(\psi(t) - \frac{d}{d\eta}A_\alpha(f + \eta\psi)\right) \\
&= p_{f+\eta\psi}^{2-\alpha}(t)\left(\psi(t) - \frac{d}{d\eta}A_\alpha(f + \eta\psi)\right)
\end{aligned}$$

evaluating at $\eta = 0$ gives us

$$dp(f;\psi)(t) = p_f^{2-\alpha}\left(\psi(t) + dA_\alpha(f;\psi)\right)$$

Note that by claim 1 we have

$$\begin{aligned}
p_{f+\eta\psi}(t) &= \exp_{2-\alpha}(f(t) + \eta\psi(t) - A_\alpha(f + \eta\psi(t))) \\
&\leq \exp_{2-\alpha}(2\|f\|_\infty + 2\eta\|\psi\|_\infty) \\
&\leq \exp_{2-\alpha}(2(\|f\|_\infty + \|\psi\|_\infty))
\end{aligned}$$

We can thus apply the dominated convergence theorem to pull a derivative with respect to η under an integral. We can then recover the Gateaux diferential of A_α via

$$\begin{aligned}
0 &= \frac{d}{d\eta}\bigg|_{\eta=0}\int p_{f+\eta\psi}(t)dQ \\
&= \int dp(f;\psi)(t)dQ \\
&= \int p_f(t)^{2-\alpha}(\psi(t) - dA_\alpha(f;\psi))dQ \\
dA_\alpha(f;\psi) &= \mathbb{E}_{\tilde{p}_f^{2-\alpha}}(\psi(T)) \\
&< \infty
\end{aligned}$$

where the last line follows as $\psi \in L^\infty$. Thus the Gateaux derivative exists in L^∞ directions. The derivative at f maps $\psi :\rightarrow \mathbb{E}_{\tilde{p}_f^{2-\alpha}}(\psi(T))$ i.e. $A'_\alpha(f)(\psi) = \mathbb{E}_{\tilde{p}_f^{2-\alpha}}(\psi(T))$. We need to show that this is a Frechet derivative. To do so, we will take take second derivatives of $p_{f+\eta\psi}(t)$ with respect to η in order to obtain second derivatives of $A_\alpha(f+\eta\psi)$ with respect to η. We will then construct a functional second order Taylor expansion. By showing that the second order terms converge sufficiently quickly, we will prove that the map $\psi :\rightarrow \mathbb{E}_{\tilde{p}_f^{2-\alpha}}(\psi(T))$ is a Frechet derivative.

$$\begin{aligned}
\frac{d^2}{d\eta^2} p_{f+\eta\psi}(t) &= \frac{d}{d\eta} p_{f+\eta\psi}(t)^{2-\alpha}\left(\psi(t) - \frac{d}{d\eta}A_\alpha(f+\eta\psi)\right) \\
&= \left(\frac{d}{d\eta} p_{f+\eta\psi}(t)^{2-\alpha}\right)\left(\psi(t) - \frac{d}{d\eta}A_\alpha(f+\eta\psi)\right) \\
&\quad - p_{f+\eta\psi}(t)^{2-\alpha}\frac{d^2}{d\eta^2}A_\alpha(f+\eta\psi) \\
&= (2-\alpha)p_{f+\eta\psi}(t)(\psi(t) - \frac{d}{d\eta}A_\alpha(f+\eta\psi))\frac{d}{d\eta}p_{f+\eta\psi}(t) \\
&\quad - p_{f+\eta\psi}(t)^{2-\alpha}\frac{d^2}{d\eta^2}A_\alpha(f+\eta\psi) \\
&= (2-\alpha)p^{3-2\alpha}_{f+\eta\psi}(\psi(t) - \frac{d}{d\eta}A_\alpha(f+\eta\psi))^2 - p_{f+\eta\psi}(t)^{2-\alpha}\frac{d^2}{d\eta^2}A_\alpha(f+\eta\psi)
\end{aligned}$$

We need to show that we can again pull the second derivative under the integral. We already showed that we can pull the derivative under once (for the first derivative) and we now need to do it again. We need to show an integrable function that dominates $p_{f+\eta\psi}(t)^{2-\alpha}(\psi(t) - \mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\eta\psi}}\psi(T))$.

$$\begin{aligned}
|p_{f+\eta\psi}(t)^{2-\alpha}(\psi(t) - \mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\eta\psi}}\psi(T))| &\leq p_{f+\eta\psi}(t)^{2-\alpha}2\|\psi\|_\infty \\
&\leq \exp_{2-\alpha}(2(\|f\|_\infty + \|\psi\|_\infty))2\|\psi\|_\infty
\end{aligned}$$

which is in $L^1(Q)$. Now applying the dominated convergence theorem

$$\begin{aligned}
0 &= \int \frac{d^2}{d\epsilon^2}p_{f+\epsilon\psi}(t)dQ \\
&= \int \left[(2-\alpha)p^{3-2\alpha}_{f+\epsilon\psi}(\psi(t) - \frac{d}{d\epsilon}A_\alpha(f+\epsilon\psi))^2 - p_{f+\epsilon\psi}(t)^{2-\alpha}\frac{d^2}{d\epsilon^2}A_\alpha(f+\epsilon\psi)\right]dQ
\end{aligned}$$

and rearranging gives

$$\begin{aligned}
\frac{d^2}{d\epsilon^2}A_\alpha(f+\epsilon\psi) &= (2-\alpha)\frac{\int p^{3-2\alpha}_{f+\epsilon\psi}(\psi(t) - \frac{d}{d\epsilon}A_\alpha(f+\epsilon\psi))^2 dQ}{\int p_{f+\epsilon\psi}(t)^{2-\alpha}dQ} \\
\frac{d^2}{d\epsilon^2}A_\alpha(f)\bigg|_{\epsilon=0} &= (2-\alpha)\frac{\int p_f^{3-2\alpha}(\psi(t) - \mathbb{E}_{\tilde{p}_f^{2-\alpha}}[\psi(T)])^2 dQ}{\int p_f(t)^{2-\alpha}dQ}
\end{aligned}$$

since $f, \psi \in L^\infty$. For the functional Taylor expansion, we have from Lemma D.2

$$A_\alpha(f+\psi) = A_\alpha(f) + A'_\alpha(f)(\psi) + \frac{1}{2}A''_\alpha(f+\epsilon\psi)(\psi)^2$$

for some $\epsilon \in [0,1]$. We thus need to show that for $\epsilon \in [0,1]$,

$$(2-\alpha)\frac{\frac{1}{\|\psi\|_\infty}\int p^{3-2\alpha}_{f+\epsilon\psi}(\psi(t) - \mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\epsilon\psi}}[\psi(T)])^2 dQ}{\int p_{f+\epsilon\psi}(t)^{2-\alpha}dQ} \overset{\psi\to 0}{\to} 0$$

It suffices to show that the numerator tends to 0 as $\psi \to 0$.

$$\begin{aligned}
&\left|\frac{1}{\|\psi\|_\infty}(\psi(t) - \mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\epsilon\psi}}[\psi(T)])^2\right| \\
&= \left|\frac{\psi(t)}{\|\psi\|_\infty}\psi(t) - \frac{\psi(t)}{\|\psi\|_\infty}2\mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\epsilon\psi}}[\psi(T)] + \frac{\mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\epsilon\psi}}[\psi(T)]}{\|\psi\|_\infty}\mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\epsilon\psi}}[\psi(T)]\right| \\
&\leq \left|\frac{\psi(t)}{\|\psi\|_\infty}\right|\left|\psi(t) - 2\mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\epsilon\psi}}[\psi(T)]\right| \\
&\quad + \left|\mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\epsilon\psi}}\frac{\psi(T)}{\|\psi\|_\infty}\right|\left|\mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\epsilon\psi}}[\psi(T)]\right| \\
&\leq \left|\psi(t) - 2\mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\epsilon\psi}}[\psi(T)]\right| + \|p_{f+\epsilon\psi}\|_{2-\alpha}^{2-\alpha}\left|\mathbb{E}_{\tilde{p}^{2-\alpha}_{f+\epsilon\psi}}[\psi(T)]\right| \\
&\to 0 \text{ as } \psi \to 0
\end{aligned}$$

and plugging this in we obtain the desired result. Thus the Frechet derivative of $A_\alpha(f)$ exists. □

Lemma D.5. *Define $\mathcal{P}_\infty = \{p_f = \exp_{2-\alpha}(f - A_\alpha(f)) : f \in L^\infty(S)\}$ where $L^\infty(S)$ is the space of almost everywhere bounded measurable functions with domain S. Fix $f_0 \in L^\infty$. Then for any fixed $\epsilon > 0$ and $p_g, p_f \in \mathcal{P}_\infty$ such that $f, g \in \overline{B}^\infty_\epsilon(f_0)$ the L^∞ closed ball around f_0, there exists constant $M_{\exp} > 0$ depending only on f_0 such that*

$$\|p_f - p_g\|_{L^r} \leq 2M_{\exp}\|f - g\|_\infty$$

Further

$$B_{\Omega_\alpha}(p_f, p_g) \leq \frac{1}{\alpha - 1}\|p_f - p_g\|_\infty[(\|p_f\|_\infty + \|p_f - p_g\|_\infty)^{\alpha-1} + \exp_{2-\alpha}(2\|g\|_\infty)]$$

Proof. This Lemma mirrors Lemma A.1 in Sriperumbudur et al. [2017], but the proof is very different as they rely on the property that $\exp(x+y) = \exp(x)\exp(y)$, which does not hold for β-exponentials. We thus had to strengthen the assumption to include that f and g lie in a closed ball, and then use the functional mean value theorem Lemma D.3 as the main technique to achieve our result.

Consider that by functional mean value inequality,

$$\begin{aligned}
\|p_f - p_g\|_{L^r} &= \|\exp_\beta(f - A_\alpha(f)) - \exp_\beta(g - A_\alpha(g))\|_{L^r} \\
&\leq \|\exp_\beta(h - A_\alpha(h))^{2-\alpha}\|_\infty(\|f - g\|_\infty + |A_\alpha(f) - A_\alpha(g)|) \qquad (13)
\end{aligned}$$

where $h = cf + (1-c)g$ for some $c \in [0,1]$. We need to bound $\exp_\beta(h - A_\alpha(h))$ and $\|A_\alpha(f) - A_\alpha(g)\|_\infty$.

We can show a bound on $\|h\|_\infty$

$$\begin{aligned}
\|h\|_\infty &= \|cf + (1-c)g - f_0 + f_0\|_\infty \\
&\leq \|c(f - f_0) + (1-c)(g - f_0) + f_0\|_\infty \\
&\leq c\|f - f_0\|_\infty + (1-c)\|g - f_0\|_\infty + \|f_0\|_\infty \\
&\leq \epsilon + \|f_0\|_\infty
\end{aligned}$$

so that h is bounded. Now we previously showed in claim 1 that $|A_\alpha(h)| \leq \|h\|_\infty \leq \epsilon + \|f_0\|_\infty$. Since $h, A_\alpha(h)$ are both bounded $\exp_\beta(h - A_\alpha(h))^{2-\alpha}$ is also.

Now note that by Lemma D.3,

$$|A_\alpha(f) - A_\alpha(g)| \leq \|A'_\alpha(h)\|_{\text{op}}\|f - g\|_\infty$$

We need to show that $\|A'_\alpha(h)\|_{\text{op}}$ is bounded for $f, g \in \overline{B}_\epsilon(f_0)$. Note that in Lemma D.4 we showed that

$$\begin{aligned}
|A'_\alpha(f)(g)| &= |\mathbb{E}_{p_f^{2-\alpha}}[g(T)]| \\
&\leq \|g\|_\infty
\end{aligned}$$

Thus $\|A'_\alpha\|_{\text{op}} = \sup\{|A'_\alpha(h)(m)| : \|m\|_\infty = 1\} \leq 1$. Let $M_{\exp}$ be the bound on $\exp_\beta(h - A_\alpha(h))$. Then putting everything together we have the desired result

$$\|p_f - p_g\|_{L^r} \leq 2M_{\exp}\|f - g\|_\infty$$

Now

$$B_{\Omega_\alpha}(p_f, p_g) = \Omega_\alpha(p_f) - \Omega_\alpha(p_g) - \langle \nabla\Omega_\alpha(p_g), p_f - p_g\rangle_2 \tag{14}$$

For the inner prodct term, first note that following Martins et al. [2020] the gradient is given by

$$\nabla\Omega_\alpha(p_g)(t) = \frac{p_g(t)^{\alpha-1}}{\alpha - 1} \tag{15}$$

Thus

$$\begin{aligned}
|\langle \nabla\Omega_\alpha(p_g), p_f - p_g\rangle_2| &\leq \|\nabla\Omega_\alpha(p_g)\|_1 \|p_f - p_g\|_\infty \\
&= \frac{1}{\alpha - 1}\int_S \exp_{2-\alpha}(g(t) - A(g))dQ\|p_f - p_g\|_\infty \\
&\leq \frac{1}{\alpha - 1}\exp_{2-\alpha}(2\|g\|_\infty)\|p_f - p_g\|_\infty
\end{aligned}$$

where the second line follows from claim 1.

Further note that by Taylor's theorem,

$$y^\alpha = x^\alpha + \alpha z^{\alpha-1}(y - x)$$

for some z between x and y. Then letting $y = p_f(t)$ and $x = p_g(t)$, we have for some $z = h(t)$ lying between $p_f(t)$ and $p_g(t)$ that

$$p_f(t)^\alpha = p_g(t)^\alpha + \alpha h(t)^{\alpha-1}(p_f(t) - p_g(t))$$

Since $f \in L^\infty$ then applying Claim 1 we have that each $p_f, p_g \in L^\infty$ and thus h is. Then

$$\begin{aligned}
|p_f(t)^\alpha - p_g(t)^\alpha| &= \alpha|h(t)|^{\alpha-1}|p_f(t) - p_g(t)| \\
&\leq \alpha\|h\|_\infty^{\alpha-1}\|p_f - p_g\|_\infty \\
&\leq \alpha\max\{\|p_f\|_\infty, \|p_g\|_\infty\}^{\alpha-1}\|p_f - p_g\|_\infty \\
&\leq \alpha(\|p_f\|_\infty + \|p_f - p_g\|_\infty)^{\alpha-1}\|p_f - p_g\|_\infty
\end{aligned} \tag{16}$$

so that

$$\begin{aligned}
|\Omega_\alpha(p_f) - \Omega_\alpha(p_g)| &= \left|\frac{1}{\alpha(\alpha - 1)}\int (p_f(t)^\alpha - p_g(t)^\alpha)dQ\right| \\
&\leq \frac{1}{\alpha - 1}(\|p_f\|_\infty + \|p_f - p_g\|_\infty)^{\alpha-1}\|p_f - p_g\|_\infty.
\end{aligned}$$

Putting it all together we obtain

$$\begin{aligned}
B_{\Omega_\alpha}(p_f, p_g) \leq{}& \frac{1}{\alpha - 1}(\|p_f\|_\infty + \|p_f - p_g\|_\infty)^{\alpha-1}\|p_f - p_g\|_\infty \\
&+ \frac{1}{\alpha - 1}\exp_{2-\alpha}(2\|g\|_\infty)\|p_f - p_g\|_\infty \\
={}& \frac{1}{\alpha - 1}\|p_f - p_g\|_\infty[(\|p_f\|_\infty + \|p_f - p_g\|_\infty)^{\alpha-1} + \exp_{2-\alpha}(2\|g\|_\infty)]
\end{aligned}$$

□

E Numerical Integration Convergence and Stability Analysis

E.1 Stable Numerical Integration

One issue is numerical underflow when computing the normalizing constant. In the kernel exponential family case, if $|f(t)|$ is very large and $f(t) < 0$ for all evaluated t, then on a computer $\exp(f(t))$

will round to 0 for all the evaluated t. Thus Z the normalizing constant will be estimated as 0 in numerical integration and the estimate of $\exp(f(t))/Z$ cannot be computed. A similar issue exists for the deformed case.

For kernel exponential family, we can use the standard technique used in discrete softmax implementations. Note

$$\begin{aligned}\exp(f(t) - A(f)) &= \exp(f(t) - C + C - A(f)) \\ &= \exp(f(t) - C)\exp(C - A(f))\end{aligned}$$

Then instead of normalizing $\exp(f(t))$ with $\exp(-A(f))$ we normalize $\exp(f(t)-C)$ with $\exp(C-A(f))$, letting $C = \sup_t f(t)$. Taking $C = \sup_t f(t)$ will prevent underflow as if $f(t) < 0 \forall t$, $f(t) - \sup f(t) \approx 0$ for some t (equality if there is a maximum, which there always is when using a finite set of t for numerical integration).

For the deformed case, recall that by Lemma 5.3, $\exp_{2-\alpha}(f(t) - \log_\alpha Z) = \frac{1}{Z}\exp_{2-\alpha}(Z^{\alpha-1}f(t))$ for $Z > 0$. Then letting $\log_\alpha C = \sup_t f(t)$ and noting that for $x, y > 0$, $\log_\alpha x - \log_\alpha y = \log_\alpha(x \oslash_\alpha y) = [x^{1-\alpha} - y^{1-\alpha} + 1]_+^{\frac{1}{1-\alpha}}$ [Suyari et al., 2004],

$$\begin{aligned}\exp_{2-\alpha}(f(t) - \log_\alpha Z) &= \exp_{2-\alpha}(f(t) - \log_\alpha C + \log_\alpha C - \log_\alpha Z) \\ &= \exp_{2-\alpha}(f(t) - \log_\alpha C - (\log_\alpha Z - \log_\alpha C)) \\ &= \exp_{2-\alpha}(f(t) - \log_\alpha C - \log_\alpha (Z \oslash_\alpha C)) \\ &= \frac{1}{Z \oslash_\alpha C}\exp_{2-\alpha}\left(\left(\frac{1}{Z \oslash_\alpha C}\right)^{\alpha-1}(f(t) - \log_\alpha C)\right)\end{aligned}$$

Now consider $\tilde{f}(t) = \left(\frac{1}{Z \oslash_\alpha C}\right)^{\alpha-1} f(t)$. Then

$$\sup_t \tilde{f}(t) = \left(\frac{1}{Z \oslash_\alpha C}\right)^{\alpha-1}\log_\alpha C$$

We can thus estimate $\tilde{f}(t)$, subtract $\sup_t \tilde{f}(t)$ (max in practice), take the deformed exponential, and normalize, similar to the kernel exponential family case.

E.2 Kernel Exponential Family Attention

Here we show conditions for which numerical integration of

$$\int_S \exp(f(t))V(t)dQ = \int_{-\infty}^{\infty} q_0(t)\exp(f(t))V(t)dt$$

using the trapezoidal rule is exponentially convergent. We start by restating a theorem. This says that the trapezoidal rule for numerical integration of holomorphic functions of sufficiently fast decay has exponential convergence. A version of this theorem comes from Trefethen and Weideman [2014], but there are slight issues with the notation and conditions. A slightly revised statement is in the course notes of Lee, which we follow here.

Theorem E.1. *Let $w : \mathbb{C} \to \mathbb{C}$ be analytic in the strip $S_b = \{z \in \mathbb{C} : |Im(z)| < a\}$ for some $a > 0$. Suppose further that $w(z) \to 0$ as $|z| \to \infty$ in the strip, and for some $M > 0$,*

$$\int_{-\infty}^{\infty} |w(x+ib)|dx \leq M \tag{17}$$

for all $b \in (-a, a)$. Then leting $I = \int_{-\infty}^{\infty} w(x)dx$ and $I_h = h\sum_{k=-\infty}^{\infty} w(kh)$,

$$|I_h - I| \leq \frac{2M}{\exp(2\pi a/h) - 1}$$

Next we give conditions on $q_0(t)$, $V(t)$, and k so that numerical integration is exponentially convergent. Specifically that q_0 and V have complex extensions to a strip and that k has exponentially decaying Fourier transform.

Corollary E.2. *Assume that q_0, V have complex extensions to a strip $S_b = \{z \in \mathbb{C} : |Im(z)| < b\}$ and V is bounded by M_V. Also assume that for any $t_i \in \mathbb{R}$ the absolute value of the Fourier transform $|\hat{k}(\xi, t_i)| = |\int_{-\infty}^{\infty} k(t, t_i)\exp(-2\pi i\xi t)dt| \leq |A(t_i)\exp(-a|\xi|)|$, where $A : \mathbb{R} \to \mathbb{C}$ and $a > 0$ is a fixed constant. Further assume that $f(t) = \sum_{i=1}^{I} \gamma_i k(t, , t_i)$, $k(t, t_i) \in C_0$, and*

$$\int_{-\infty}^{\infty} |q(x + ib)|dx \leq M \tag{18}$$

for some $M > 0$. Then if $w(z) = q_0(t)\exp(f(t))V(t)$ converges uniformly to 0 as $|z| \to \infty$, the trapezoidal rule for $w(t) = q_0(t)\exp(f(t))V(t)$ is exponentially convergent.

Proof. Note that

$$\begin{aligned}
\hat{f}(\xi) &= \int_{-\infty}^{\infty} f(t)\exp(-2\pi i\xi t)dt \\
&= \int_{-\infty}^{\infty} \sum_{i=1}^{I} \gamma_i k(t, t_i)\exp(-2\pi i\xi t)dt \\
&= \sum_{i=1}^{I} \gamma_i \int_{-\infty}^{\infty} k(t, t_i)\exp(-2\pi i\xi t)dt \\
&= \sum_{i=1}^{I} \gamma_i \hat{k}(\xi, t_i)
\end{aligned}$$

and thus

$$\begin{aligned}
|\hat{f}(\xi)| &\leq \sum_{i=1}^{I} |\gamma_i||A(t_i)|\exp(-a|\xi|) \\
&\leq \exp(-a|\xi|)I \max_{i \in I} |\gamma_i||A(t_i)|
\end{aligned}$$

Thus since the fourier transform of f has (at least) exponential decay, by Theorem 3.1 in Stein and Shakarchi [2010] the extension of $f(t)$ to $f(z)$ is holomorphic in the strip. Since compositions and products of holomorphic functions are holomorphic, $q_0(z)\exp(f(z))V(z)$ is holomorphic in the strip. Further, since for each t_i, $k(t, t_i) \in C_0$, f is and thus it is bounded and thus its complex extension is bounded, thus the complex $\exp(f)$ is bounded, say by $\exp(M_f)$. Since V is also bounded, we have

$$\begin{aligned}
\int_{-\infty}^{\infty} |w(x + ib)|dx &\leq M_V \exp(M_f)M \\
&\equiv M
\end{aligned}$$

and thus w satisfies the conditions of the previous theorem. □

Finally, we show example special cases of q_0, V, k satisfying those conditions.

Corollary E.3. *If k is a Gaussian kernel, V is a linear combination of Gaussian RBFs, and $q_0(t) = \frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$, then the trapezoidal rule for kernel exponential family attention is exponentially convergent.*

Proof. For the Gaussian kernel, we have $k(t, t_i) = \exp(-\zeta^2(t - t_i)^2)$. Now

$$\int_{-\infty}^{\infty} \exp(-\zeta^2(t - t_i)^2)\exp(-2\pi i\xi t)dt = \frac{\exp(-i\xi t_i)}{\gamma}\sqrt{\pi}\exp(-\xi^2/(4\zeta^2))$$

and thus

$$|\int_{-\infty}^{\infty} \exp(-\zeta^2(t - t_i)^2)\exp(-2\pi i\xi t)dt| \leq \frac{\sqrt{\pi}}{\gamma}\exp(-\xi^2/(4\zeta^2))$$

which satisfies the Fourier decay condition for the kernel. A similar technique can be applied to V, and clearly both V and f bounded since they are continuous and vanish at infinity. Thus $\exp(f)$ is also bounded. For the decay of q, note

$$\begin{aligned}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}|\exp(-(x+ib)^2/2)|dx &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}|\exp((-x^2+b^2-2xbi)/2)|dx\\ &\leq \frac{1}{\sqrt{2\pi}}\exp(b^2/2)\int_{-\infty}^{\infty}\exp(-x^2/2)dx\\ &= \frac{1}{\sqrt{2\pi}}\exp(b^2/2)\end{aligned}$$

satisfying the integration condition Eqn. 17. It remains to show that $w(z)\to 0$. Note that since V, f are bounded and $q(z)\to 0$ (which was shown in proving the integration condition above) in the strip as $|z|\to\infty$, $w(z)\to 0$ as $|z|\to\infty$.

E.3 Convergence of Numerical Integration for Kernel Deformed Exponential Family Attention

E.3.1 Smooth Approximation to Kernel Deformed Exponential Family Attention

Definition E.4. Let

$$\exp_{\rho,\beta}(t) \equiv \left[\frac{1}{\rho}\log(1+\exp(\rho[1+(1-\beta)t]))\right]^{\frac{1}{1-\beta}} \tag{19}$$

for some $\rho > 0$.

Claim 2. $\exp_{\rho,\beta}(z)$ is holomorphic on the strip $\{z\in\mathbb{C} : \text{Im}(z)\in[-\pi/2,\pi/2]\}$ for the principal branch of $\log$.

Proof. Note that this proof is adapted from [https://math.stackexchange.com/users/8508/robert israel]. We show it for $\exp_{0,0}$ but the idea can easily be extended to more general ρ and β. Note that the principal branch of $\log z$ is analytic outside of $B=\{z=x+iy : -\infty < x\leq 0, y=0\}$. Thus $\exp_{0,0}$ is analytic as long as $1+\exp(1+z)$ is not in B.

Now $\exp(1+z)=\exp(x+1)\exp(iy)$, so $\exp(1+z)$ will be in $\{z=x+iy : x\in(-\infty,-1], y=0\}$ when $\exp(x+1)\geq 1, \exp(iy)=-1$, i.e. $x\geq -1, y=\pi+2\pi n, n\in\mathbb{N}$. The y condition will not be satisfied on the strip above and thus $\exp_{0,0}(z)$ is holomorphic on that strip. □

Corollary E.5. *If k is a Gaussian kernel, V is a linear combination of Gaussian RBFs, and $q_0(t)=\frac{1}{\sqrt{2\pi}}\exp(-t^2/2)$, then the trapezoidal rule for softplus approximation to kernel deformed exponential family attention is exponentially convergent.*

Proof. Apply Theorem E.1. We already showed in the kernel exponential family section that q_0 and V are holomorphic in the strip, and we showed that $\exp_{\rho,\beta}$ is in the previous claim. We also showed that V is bounded and that q_0 satisfies the integration condition. It remains to show that $\exp_{\rho,\beta}(f(t))$ is bounded (and thus its complex extension is). We show it for $\exp_{\rho,0}(f(t))$ but again the idea can be extended to general β. Note that

$$\begin{aligned}\log(1+\exp(1+f(t))) &\leq \exp(f(t))\\ &\leq \exp(M_f)\end{aligned}$$

since f is bounded. Further, fixing t, $\log(1+\exp(\rho(1+t)))/\rho$ is monotonically decreasing as a function of ρ and thus for all $\rho\geq 1$,

$$\log(1+\exp(\rho(1+t)))/\rho \leq \exp(M_f)$$

Thus w satisfies both the integration condition and the convergence condition. □

□

E.3.2 Using the Smooth Approximation to Bound the Numerical Integral for Kernel Sparsemax Attention

Let I_h be the numerical integral proportional to kernel deformed exponential family attention, $I_{h,s}$ be its softplus approximation, I be the true kernel deformed exponential family integral, and I_s be its softplus approximation. Then

$$|I_h - I| \leq |I_h - I_{h,s}| + |I_{h,s} - I_s| + |I_s - I|$$

We already bounded $|I_{h,s} - I_s|$ in the previous subsection, so we will bound the other two terms on the right hand side.

E.3.3 Bounding $I_s - I$

We first bound the difference between the softplus integral and the integral using ReLU. By Hoelder's inequality,

$$\begin{aligned}|\int_{-\infty}^{\infty} & q_0(t)V(t)[\ln(1+\exp(\rho(1+t)))/\rho - \max(0,1+t)]dt| \\ &\leq \|q_0 V\|_1 \text{esssup}_{t\in(-\infty,\infty)}[\ln(1+\exp(\rho(1+t)))/\rho - \max(0,1+t)] \\ &\leq \|q_0 V\|_1 \ln(2)/\rho\end{aligned}$$

E.3.4 Bounding $I_h - I_{h,s}$

$$\begin{aligned}|I_{h,s} - I_h| &= h\sum_{k=-\infty}^{\infty} q_0(kh)V(kh)[\ln(1+\exp(\rho(1+kh)))/\rho - \max(0,1+kh)] \\ &= h\langle q_0(h\cdot)V(h\cdot), \ln(1+\exp(\rho(1+h\cdot)))/\rho - \max(0,1+h\cdot)\rangle_{l^2} \\ &\leq h\left(\sum_{k=-\infty}^{\infty} q_0(hk)|V(hk)|\right)\ln(2)/\rho\end{aligned}$$

where the last line again uses Hoelder's inequality.

E.3.5 Putting it All Together

We now have, for any $\rho > 0$,

$$\begin{aligned}|I_h - I| &\leq |I_h - I_{h,s}| + |I_{h,s} - I_s| + |I_s - I| \\ &\leq \|q_0 V\|_1 \ln(2)/\rho + \frac{2M}{\exp(2\pi a/h) - 1} + h\left(\sum_{k=-\infty}^{\infty} q_0(hk)|V(hk)|\right)\ln(2)/\rho\end{aligned}$$

and since this holds for all $\rho > 0$ we can take $\rho \to \infty$ and we have

$$|I_h - I| \leq \frac{2M}{\exp(2\pi a/h) - 1}$$

E.4 Synthetic Experiments: Convergence

We now analyze convergence of the trapezoidal rule for multimodal continuous attention using numerical integration empirically using synthetic experiments. We define

$$f(t) = \sum_{i=1}^{I} \gamma_i k(t, t_i)$$

and

$$V(t) = \sum_{i=1}^{I} B_i k(t, t_i)$$

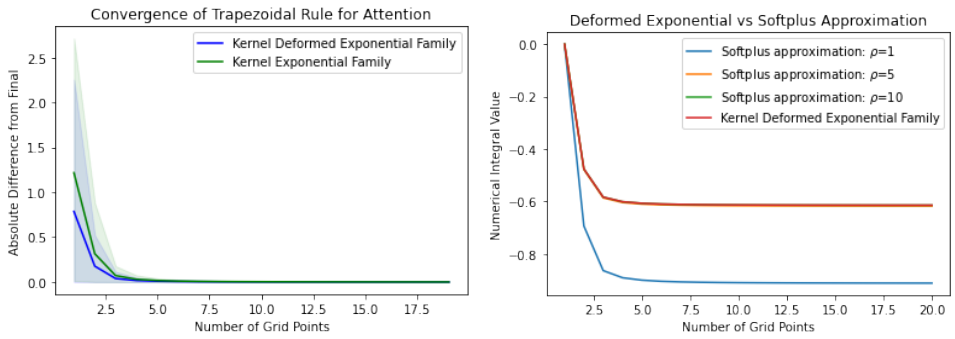

Figure 2: a) Convergence of trapezoidal rule for multimodal continuous attention. We see that both methods have very fast convergence empirically. The y-axis is the absolute difference between the numerical integral at x vs the max value of x (x ranges from 1 to 19). The kernel deformed exponential family tends to converge faster than kernel exponential family attention. b) The value of the integral using kernel deformed exponential family sparsemax vs softplus approximations with different values of ρ. We see that the attention using softplus approximation becomes indistinguishable from the deformed exponential attention for $\rho = 5, 10$.

with $\gamma_i, B_i \sim U(-1, 1)$. We set $I = 10$ and t_i to be evenly spaced in the interval $[0, 1]$. We then compute

$$\int_0^1 \exp(-t^2/2)V(t)\exp(f(t))dt$$

which is proportional to attention for kernel exponential families. We also compute

$$\int_0^1 \exp(-t^2/2)V(t)\exp_0(f(t))dt$$

proportional to the sparsemax case of kernel deformed exponential family attention. Finally, we also use the softplus approximation

$$\int_0^1 \exp(-t^2/2)V(t)\exp_{\rho,0}(f(t))dt$$

We simulate 100 times. Figure 2a compares the convergence as the number of grid points in the trapezoidal rule increases. 2b shows a single case of the integral value using deformed exponential vs the softplus approximation for various values of ρ and numbers of grid points. We see that for $\rho = 5, 10$ the integral using softplus approximation is essentially indistinguishable from that using the positive part/ReLU.

F uWave Experiment

We investigate: 1) does a large number of unimodal attention heads, as suggested in Martins et al. [2020, 2021], perform well when multimodality is needed? 2) can this method work well for irregularly sampled time series? 3) Can we learn interesting multimodal attention densities.

We analyze uWave [Liu et al., 2009]: accelerometer time series with eight gesture classes. We follow Li and Marlin [2016]'s split into 3,582 training observations and 896 test observations: sequences have length 945. We do synthetic irregular sampling and uniformly sample 10% of the observations. Because of this our results are comparable to other uWave irregular sampling papers Li and Marlin [2016], Shukla and Marlin [2019], but *not* to results using the full time series.

Table 4 shows the results. Our highest accuracy is 94.26%, the multi-head unimodal case's best is 74.69%, and the mixture's best is 81.13%. Since this dataset is small, we report ±1.96 standard deviations from 10 runs. We outperform the results of Li and Marlin [2016], who report a highest accuracy of 91.41%, and perform similarly to Shukla and Marlin [2019] (their figure suggests approximately 94% accuracy). Fig. 1 shows attention densities for one of the attention heads for the

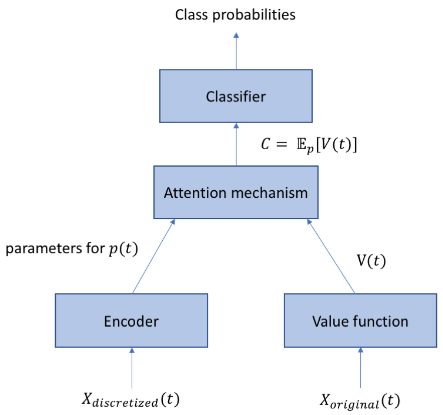

Figure 3: General architecture for classification using continuous attention mechanisms. The pipeline is trained end-to-end. The encoder takes a discretized representation of an observation (i.e. a time series itself, hidden units of an LSTM, or a layer of a CNN) and outputs parameters for an attention density. The value function takes the original (potentially irregularly sampled) time series or some representation and outputs parameters for a function $V(t)$. These are then combined in an attention mechanism by computing a context vector $c = \mathbb{E}_p[V(T)]$. For some parametrizations of p and $V(t)$ this can be computed in closed form, while for others it must be done via numerical integration. The context vector is then fed into a classifier.

Attention	N=64	N=128
Cts Softmax	67.78±1.64	67.70± 2.49
Cts Sparsemax	74.20±2.72	74.69±3.78
Gaussian Mixture	81.13±1.76	80.99±2.79
Kernel Softmax (ours)	**93.85±0.60**	**94.26±0.75**
Kernel Sparsemax (ours)	92.32±1.09	92.15±0.79

Table 4: Accuracy results on uWave gesture classification dataset for the irregularly sampled case. Note that due to the irregular sampling, this is only comparable to Li and Marlin [2016], Shukla and Marlin [2019]. Kernel based attention substantially outperforms unimodal and mixture models. All methods use 100 attention heads. Gaussian mixture uses 100 components (and thus 300 parameters per head), and kernel methods use 256 inducing points.

first three classes. This takes one attention density for each time series in the test set of each class and plots it. Within the same class, all attention densities for the head (one for each time series) are plotted. The plot shows two things: firstly, attention densities have support over non-overlapping time intervals. This cannot be done with Gaussian mixtures, and the intervals would be the same for each density in the exponential family case. Secondly, there is high similarity of attention densities within each class, but low similarity between classes. Appendix F describes additional details.

Hyperparameter	*Value*
Batch Size	25
Value Basis Functions	64, 128
Heads	100
Inducing Points	256
Integration Grid Points	100
Optimizer	Adam
Learning Rate	1e-4
Epochs	10

Table 5: Hyperparameters for uWave Experiment

We experiment with $N = 64, 128$ basis functions, and use a learning rate of $1e-4$. We use $H = 100$ attention mechanisms, or heads. Unlike Vaswani et al. [2017], our use of multiple heads is slightly different as we use the same value function for each head, and only vary the attention densities. Additional architectural details are given below. Table 5 summarizes the hypermarameters and training details.

F.1 Value Function

The value function uses regularized linear regression on the original time series observed at random observation times (which are not dependent on the data) to obtain an approximation $V(t; \mathbf{B}) = \mathbf{B}\Psi(t) \approx X(t)$. The H in Eqn. 6 is the original time series.

F.1.1 Encoder

In the encoder, we use the value function to interpolate the irregularly sampled time series at the original points. This is then passed through a convolutional layer with 4 filters and filter size 5 followed by a max pooling layer with pool size 2. This is followed by one hidden layer with 256 units and an output v of size 256. The attention densities for each head $h = 1, \cdots, H$ are then

$$\begin{aligned}\mu_h &= w_{h,1}^T v\\ \sigma_h &= \text{softplus}(w_{h,2}^T v)\\ \gamma_h &= W^{(h)} v\end{aligned}$$

for vectors $w_{h,1}, w_{h,2}$ and matrices W^h and heads $h = 1, \cdots, H$

F.1.2 Attention Mechanism

After forming densities and normalizing, we have densities $p_1(t), \cdots, p_H(t)$, which we use to compute context scalars

$$c_h = \mathbb{E}_{p_h}[V(T)]$$

We compute these expectations using numerical integration to compute basis function expectations $\mathbb{E}_{p_h}[\psi_n(T)]$ and a parametrized value function $V(t) = B\psi(t)$ as described in section 3.

F.1.3 Classifier

The classifier takes as input the concatenated context scalars as a vector. A linear layer is then followed by a softmax activation to output class probabilities.

G IMBD Experiments

We extend Martins et al. [2020]'s code[5] for the IMDB sentiment classification dataset [Maas et al., 2011]. This starts with a document representation v computed from a convolutional neural network and uses an LSTM attention model. We use a Gaussian base density and kernel, and divide the interval $[0, 1]$ into $I = 10$ inducing points where we evaluate the kernel in $f(t) = \sum_{i=1}^{I} \gamma_i k(t, t_i)$. Table 6 shows results. On average, kernel exponential and deformed exponential family slightly outperforms the continuous softmax and sparsemax. The continuous softmax/sparsemax results are from running their code.

H MIT BIH: Additional Details

We plot attention densities for randomly selected examples under the different density classes, and find that only kernel deformed exponential families/sparsemax learn interpretable attention densities, which focus on regions where the electrical signals from the heart are strong. Figures 5 and 6 show attention densities vs original signals for continuous softmax and sparsemax, respectively. Both are only able to learn simple unimodal densities. Figure 7 shows the same for Gaussian mixture attention.

[5] Martins et al. [2020]'s repository for this dataset is https://github.com/deep-spin/quati

Attention	N=32	N=64	N=128	Mean
Cts Softmax	89.56	90.32	**91.08**	90.32
Cts Sparsemax	89.50	90.39	89.96	89.95
Kernel Softmax	**91.30**	**91.08**	90.44	**90.94**
Kernel Sparsemax	90.56	90.20	90.41	90.39

Table 6: IMDB sentiment classification dataset accuracy. Continuous softmax uses Gaussian attention, continuous sparsemax truncated parabola, and kernel softmax and sparsemax use kernel exponential and deformed exponential family with a Gaussian kernel. The latter has $\alpha = 2$ in exponential and multiplication terms. N: basis functions, $I = 10$ inducing points, bandwidth 0.01.

Hyperparameter	*Value*
Batch Size	64
LSTM Hidden Units	512
Value Basis Functions	24
Inducing Points	187
Integration Grid Points	187
Learning Rate	1e-3
Weight Decay	1e-5
Epochs	30

Table 7: Hyperparameters for MIT BIH Experiment

These do not look very multimodal, despite having one component per time point. This is likely due to lack of separation between components. However, we do see that the shapes look more flexible than in single Gaussian or truncated parabola cases. Figure 8 shows attention densities vs original signals for kernel softmax attention. While not particularly interpretable, it learns densities similar to exponential densities without them being specified, a benefit of being a non-parametric density. Figure 9 shows the same for kernel sparsemax. These show interesting highlighting of waves, which describe electrical signals passing through the heart conduction system. There is a particular focus on the R wave, the largest peak in a heartbeat, representing electrical stimulus in the main ventricular mass.

Note that given the relatively complex model structure and reasonably high capacity relative to the original time series length (512 LSTM hidden units per time step for a univariate time series of length 187), models without an interpretable attention density may still perform well by taking advantage of capacity elsewhere in the model. However, a model that selects ECG waves may be useful in convincing specialists of its value.

H.1 General Architecture

Our general architecture is shown in 4. In the first part, two convolutional layers of filter size 5 and 24 filters with padding map the original univariate time series from $\mathbb{R}^{187}$ to a multivariate representation $\mathbb{R}^{187\times 24}$. This is then passed to an LSTM. Context vectors are computed using either the original hidden states (discrete attention) or a continous-time representation (continuous attention). The context vector is then fed into a feedforward network for final classification. The entire architecture is trained end-to-end. Hyperparameters are described in Table 7.

H.2 Value Function

The value function uses regularized linear regression on the hidden states of an LSTM to obtain an approximation $V(t; \mathbf{B}) = \mathbf{B}\Psi(t) \approx h_t$. The H in Eqn. 6 is the set of all hidden states.

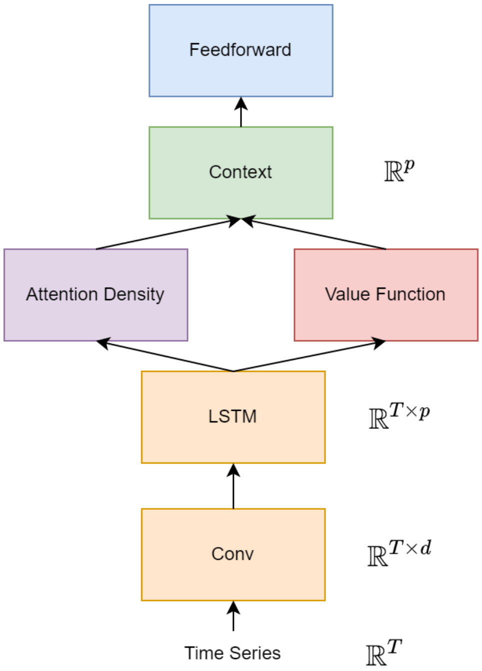

Figure 4: Architecture used for the MIT BIH and FordA experiments. The input is a univariate time series. The conv block has two conv layers, with the goal of converting a univariate time series into a multivariate one. The hidden units of the LSTM are then used to compute the attention density parameters and the value function. The feedforward network has three layers, where the first two have ReLU activation functions.

H.3 Encoder

The encoder takes the hidden layer of the LSTM as input. For the hidden units $h_t \in \mathbb{R}^p$ for a given time step, it computes

$$v_t = \tanh(W_w h_t + b_w), W_w \in \mathbb{R}^{d\times p}, b_w \in \mathbb{R}^p$$
$$\gamma_t = w_v^T v_t$$

Note that this is written slightly differently from the form in the main paper.

H.4 Attention Mechanism

The attention mechanism takes the parameters from the encoder and forms an attention density. It then computes

$$c = \mathbb{E}_p[V(T)]$$

for input to the classifier.

H.5 Classifier

The classifier has three feedforward layers, where the first two have ReLU activation functions.

I FordA: Additional Details

We again plot attention densities for randomly selected examples under each density class. We find that only the kernel methods appear truly multimodal visually (Gaussian mixture does not), and the kernel sparsemax case actually highlights specific peaks and troughs of the original signal, suggesting

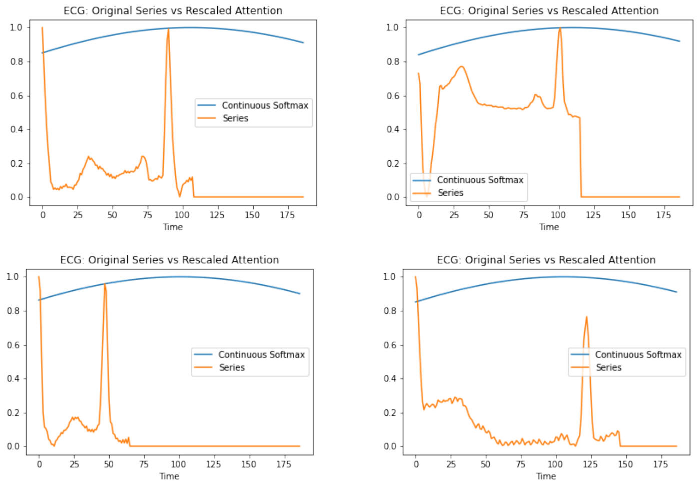

Figure 5: Original time series vs rescaled continuous softmax attention densities for four randomly selected examples from the MIT BIH dataset. Continuous softmax only learns simple unimodal densities.

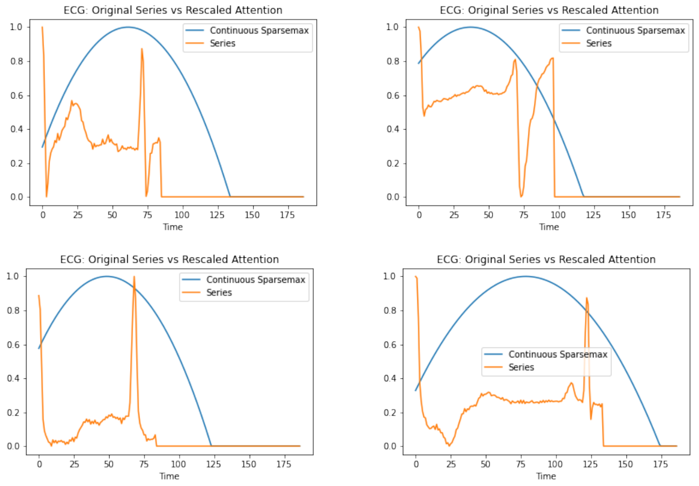

Figure 6: Original time series vs rescaled continuous sparsemax attention densities for four randomly selected examples from the MIT BIH dataset. Continuous sparsemax, similar to continuous softmax, again only learns simple unimodal densities.

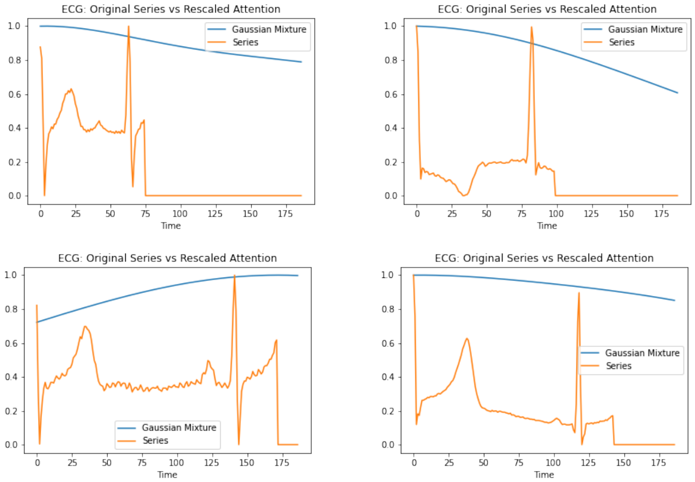

Figure 7: Original time series vs rescaled Gaussian mixture attention densities for four randomly selected examples from the MIT BIH dataset. Number of mixture components equal to number of time points. This does not appear very multimodal, likely because the mixture components are not well separated. However, the attention densities seem to have more flexible shapes than for continuous softmax/sparsemax.

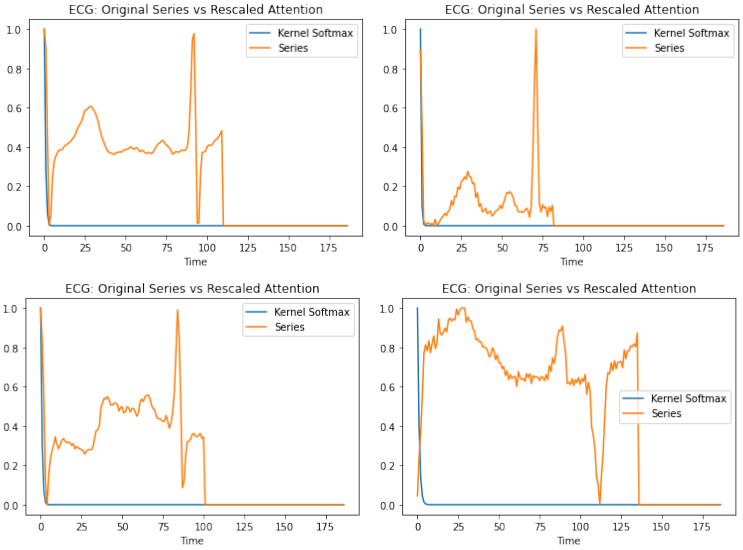

Figure 8: Original time series vs rescaled kernel softmax attention densities for four randomly selected examples from the MIT BIH dataset. While this is not particularly interpretable, it still learns a density that looks similar to an exponential density without explicitly specifying this shape. Further, the empirical performance beats the other attention mechanisms other than kernel sparsemax.

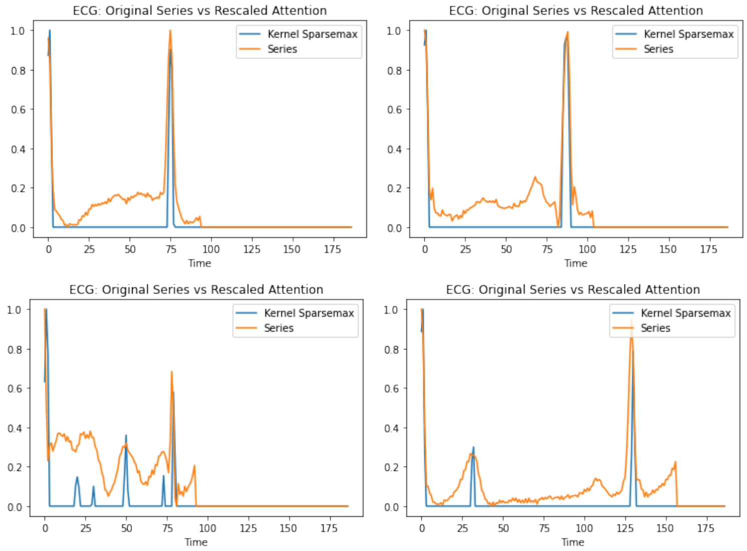

Figure 9: Original time series vs rescaled kernel sparsemax attention densities for four randomly selected examples from the MIT BIH dataset. All cases are multimodal. Further, this is much more interpretable than any of the other methods, and tends to select local peaks, or waves. These waves represent the electrical stimulus as they pass through different parts of the heart conduction system. There is a particular focus on the R wave, the largest peak, which represents the electrical stimulus in the main ventricular mass.

Hyperparameter	*Value*
Batch Size	64
LSTM Hidden Units	128
Value Basis Functions	256
Inducing Points	500
Integration Grid Points	500
Learning Rate	1e-3
Weight Decay	1e-5
Epochs	30

Table 8: Hyperparameters for FordA Experiment

the ability to attend to higher frequencies. Figure 13 shows rescaled kernel softmax attention densities along with the original series. These are generally smooth and show rich multimodality. Figure 14 shows the same for kernel sparsemax. These are very interesting, and learn rich sparsity patterns and often highlight the peaks and troughs of the series, while taking 0 values over many regions where the series becomes negative.

This dataset uses the same architecture as the MIT BIH dataset. However, several hyperaparameters are different, as mentioned in Table 8. In particular, we use more value basis functions (256) as the sequence is longer, and a smaller number of LSTM hidden units (128). On this dataset, claiming interpretability is more difficult because it is engine noise while the class meanings are unknown: only the binary labels are present. However it is still interesting that kernel sparsemax is able to select individual waves and exhibit rich sparsity patterns.

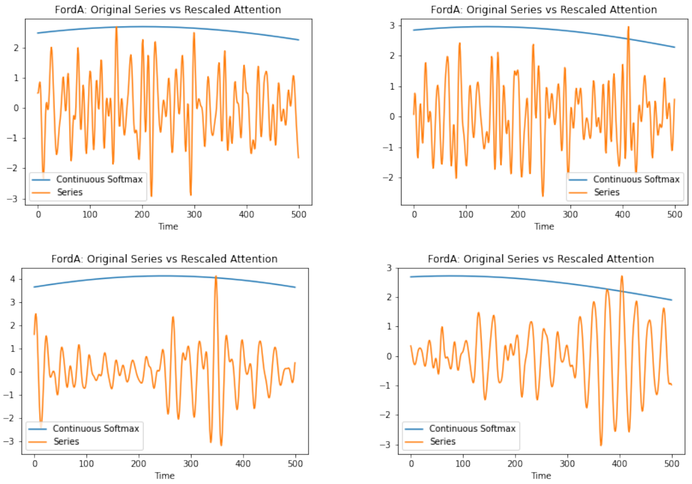

Figure 10: Original time series vs rescaled attention densities for four randomly selected examples from the FordA dataset using kernel softmax. The densities are rescaled so that they have the same max as the signal. The densities are simple and do not attend to fine portions of the signal.

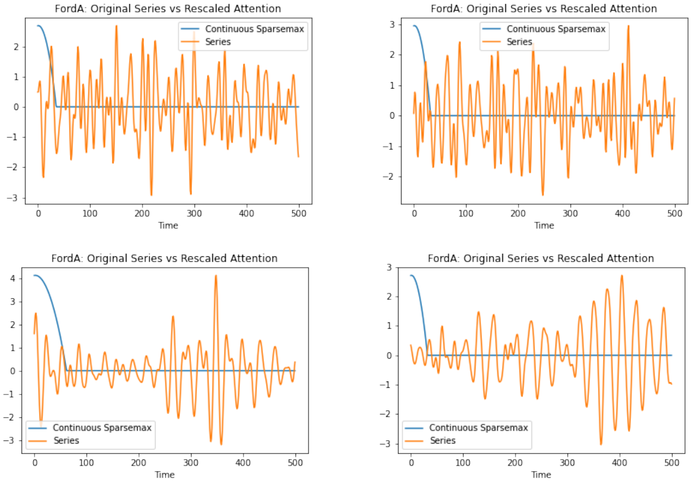

Figure 11: Original time series vs rescaled attention densities for four randomly selected examples from the FordA dataset using continuous sparsemax. The densities are rescaled so that they have the same max as the signal. The densities are again simple and do not attend to fine portions of the signal, although have more focus than in the continuous softmax case.

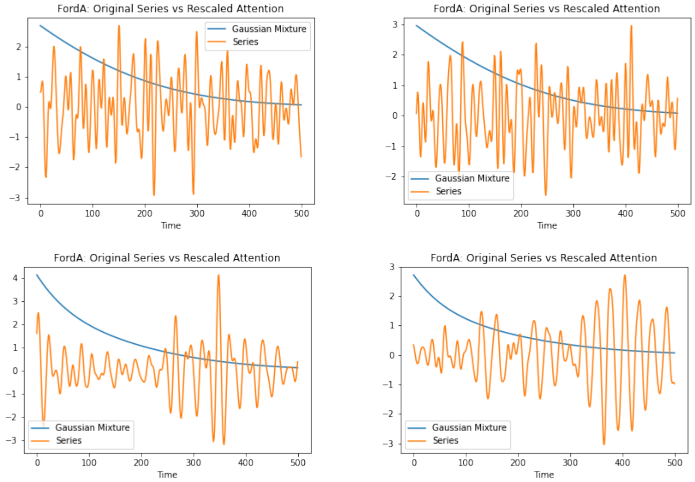

Figure 12: Original time series vs rescaled attention densities for four randomly selected examples from the FordA dataset using Gaussian mixture. The densities are rescaled so that they have the same max as the signal. The densities do not appear multimodal, again likely due to lack of separation between components. However, the shape looks different from the simple shapes of the Gaussian and truncated parabola from continuous softmax and sparsemax.

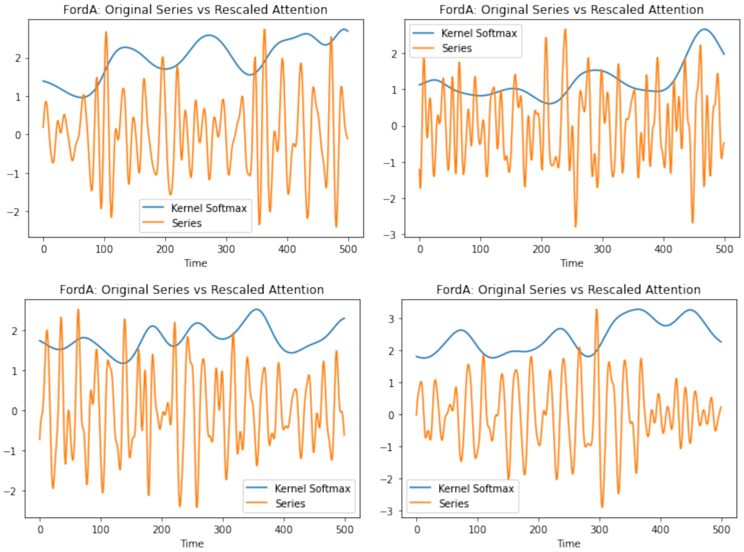

Figure 13: Original time series vs rescaled attention densities for four randomly selected examples from the FordA dataset using kernel softmax. The densities are rescaled so that they have the same max as the signal. The densities exhibit complex multimodal shapes, although they do not appear to select obvious features of the signal.

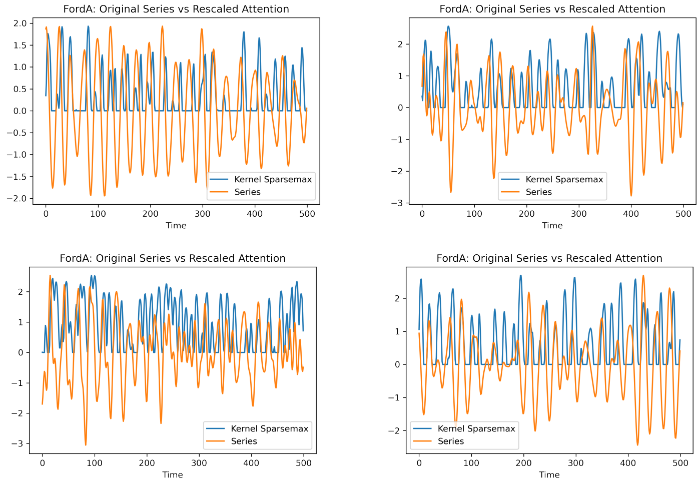

Figure 14: Original time series vs rescaled attention densities for four randomly selected examples from the FordA dataset using kernel sparsemax. The densities are rescaled so that they have the same max as the signal. The densities exhibit very similar patterns to the signal itself, often selecting the peaks or troughs. They also have complex sparsity patterns.