Bregman Power k-Means for Clustering Exponential Family Data

Adithya Vellal [1] **Saptarshi Chakraborty** [2] **Jason Xu** [1]

Abstract

Recent progress in center-based clustering algorithms combats poor local minima by implicit annealing, using a family of generalized means. These methods are variations of Lloyd's celebrated k-means algorithm, and are most appropriate for spherical clusters such as those arising from Gaussian data. In this paper, we bridge these algorithmic advances to classical work on hard clustering under Bregman divergences, which enjoy a bijection to exponential family distributions and are thus well-suited for clustering objects arising from a breadth of data generating mechanisms. The elegant properties of Bregman divergences allow us to maintain closed form updates in a simple and transparent algorithm, and moreover lead to new theoretical arguments for establishing finite sample bounds that relax the bounded support assumption made in the existing state of the art. Additionally, we consider thorough empirical analyses on simulated experiments and a case study on rainfall data, finding that the proposed method outperforms existing peer methods in a variety of non-Gaussian data settings.

1. Introduction and Background

Clustering, the task of finding naturally occurring groups within a dataset, is a cornerstone of the unsupervised learning paradigm. Among a vast literature on clustering algorithms, center-based methods remain widely popular, and k-means (MacQueen, 1967; Lloyd, 1982) remains the most prominent example 60 years after its introduction. Given n data points $\mathcal{X} = \{\boldsymbol{X}_i : i = 1, \dots, n\} \subset \mathbb{R}^p$, k-means seeks to partition the data into k groups in a way that minimizes the within-cluster variance. Representing the cluster centroids $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_k\} \subset \mathbb{R}^p$ and for some dissimilarity measure $d(\cdot, \cdot)$, k-means is formulated as the minimization of the objective function

$$f_{k\text{-means}}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \min_{1 \le j \le k} d(\boldsymbol{X}_i, \boldsymbol{\theta}_j). \tag{1}$$

Taking the squared Euclidean distance $d(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2^2$ yields the classical k-means formulation, while Bregman hard clustering (Banerjee et al., 2005b) allows d to be any Bregman divergence.

Unfortunately, k-means and its variants suffer from well-documented shortcomings such as sensitivity to initial guess (Vassilvitskii & Arthur, 2006; Bachem et al., 2017; Deshpande et al., 2020), stopping at poor local minima (Zhang et al., 1999; Xu & Lange, 2019), and fragility to outliers (Paul et al., 2021b) that continue to be addressed in recent work. In particular, a drawback we seek to address in this article is the implicit assumption behind k-means that the data can be clustered spherically, which works well in Gaussian settings but can fail to separate even simple data examples otherwise (Ng et al., 2002). To ameliorate this issue, researchers have proposed various dissimilarity measures (Banerjee et al., 2005b; De Amorim & Mirkin, 2012; Chakraborty & Das, 2017; Brécheteau et al., 2021) that admit non-elliptical contours. Among these, the choice of Bregman divergences is appealing (Telgarsky & Dasgupta, 2012; Paul et al., 2021a) as their many nice mathematical properties are amenable to analysis and effective algorithms. Their connection to exponential families makes them ideal for many common data generating mechanisms.

Like classic k-means, analogs such as Bregman hard clustering are susceptible to local optima due to non-convexity of the objective. Wrapper methods such as k-means ++ (Arthur & Vassilvitskii, 2007) and its variants alleviate the problem to an extent, though methods continue to struggle as dimension increases (Aggarwal et al., 2001). Recently, (Xu & Lange, 2019; Chakraborty et al., 2020) tackle this problem by gradually annealing the optimization landscape in the Euclidean case. Theoretical work by (Paul et al., 2021b) proposes a clustering framework that encompasses Bregman divergences, establishing desirable properties such as robustness, but does not implement or empirically analyze the Bregman case. The authors advocate generic iterative optimization, using adaptive gradient descent for the general case (Duchi et al., 2011).

[1]Department of Statistical Science, Duke University, Durham, NC, USA. [2]Department of Statistics, University of California, Berkeley, CA, USA. Correspondence to: Jason Xu <jason.q.xu@duke.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

In this paper, we propose and analyze a scalable, transparent clustering algorithm that performs annealing to target the same objective as Bregman hard clustering. That is, it inherits nice properties and interpretability while being less prone to poor local solutions. Leveraging the mean-as-minimizer property of Bregman divergences leads to a simple and elegant algorithm with closed form updates through majorization-minimization (MM). We show that it outperforms alternatives on a range of exponential family data via thorough simulation studies.

Moreover, we formulate the method so that it inherits a number of strong theoretical guarantees. Through a novel and extensive theoretical study, we bound the excess risk by appealing to the recent literature on sub-exponential concentration inequalities and the classic approach of bounding the Rademacher complexity through Dudley's entropy integral. These include generalization bounds and learning rates for a broad family of distributions, lifting a restrictive condition that assumes the data has bounded support in previous analyses (Paul et al., 2021b). Our results not only match the best known rates in literature while imposing much weaker assumptions, but also reveal an interesting dependency of the excess risk on the true cluster centers $\mathbf{\Theta}_*$ and the second moment of the underlying data distribution.

We briefly overview some relevant concepts that will be used in formulating the Bregman power k-means method.

Bregman divergences A differentiable, convex function $\phi : \mathbb{R}^p \to \mathbb{R}$ generates the *Bregman divergence* $d_\phi : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ defined by

$$d_\phi(\boldsymbol{x}, \boldsymbol{y}) = \phi(\boldsymbol{x}) - \phi(\boldsymbol{y}) - \langle \nabla\phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle. \tag{2}$$

It becomes clear that $d_\phi(\boldsymbol{x}, \boldsymbol{y}) \geq 0 \;\forall\; \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ since $\phi(\boldsymbol{x}) \geq \phi(\boldsymbol{y}) + \langle \nabla\phi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle$ is synonymous with ϕ being a convex function. From a geometric perspective, $d_\phi(\boldsymbol{x}, \boldsymbol{y})$ can be thought of as the distance between $\phi(\boldsymbol{x})$ and the first-order approximation of $\phi(\boldsymbol{x})$ centered at $\phi(\boldsymbol{y})$. In more intuitive terms, this can be described as the distance between $\phi(\boldsymbol{x})$ and the value of the tangent line to $\phi(\boldsymbol{y})$ evaluated at the point $\boldsymbol{x}$. For instance, taking $\phi(\boldsymbol{u}) = \|\boldsymbol{u}\|_2^2$ generates the Euclidean distance. Without loss of generality, one may assume $\phi(\mathbf{0}) = \nabla\phi(\mathbf{0}) = 0$.

While not necessarily symmetric like the usual Euclidean distance, Bregman divergences satisfy numerous desirable properties which make them useful for quantifying dissimilarity. They are non-negative and maintain linearity; for any convex f, g, we have $d_{\alpha f + \beta g} = \alpha d_f + \beta d_g \forall\, \alpha, \beta \in \mathbb{R}_{>0}$. Of note is that Bregman divergences obey a mean-as-minimizer property. As shown in Banerjee et al. (2005a), this can be characterized in such a way as stated in a proposition we rephrase below:

Proposition 1.1. *Let $d : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_{\geq 0}$ to be any continuous function with continuous first-order partial derivatives obeying $d(\boldsymbol{x}, \boldsymbol{x}) = 0$. Then the mean $\mathbb{E}[X]$ serves as the unique minimizer of $\mathbb{E}[d(X, \boldsymbol{y})]$ for $\boldsymbol{y} \in \mathbb{R}^p$ if and only if there exists some ϕ such that $d = d_\phi$.*

Furthermore, Bregman divergences share a one-to-one correspondence with regular exponential families, detailed in the next section, making them well-suited for learning from many common data types arising from exponential family distributions. Here the notion of Bregman information $I_\phi(X) = \min_{\boldsymbol{s} \in dom(X)} \mathbb{E}[d_\phi(X, \boldsymbol{s})]$ provides a natural measure of distortion. This is minimized at $\boldsymbol{s} = \mathbb{E}[X]$ (cf. Prop. 1.1), and $I_\phi(X)$ can thus be interpreted as a generalization of variance when spread around the mean of X is measured under d_ϕ.

Majorization-minimization The principle of MM has become increasingly popular in optimization and statistical learning (Mairal, 2015; Lange, 2016). Rather than minimizing an objective of interest f directly, an MM algorithm successively minimizes a sequence of simpler *surrogate functions* $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_n)$ that *majorize* the original objective $f(\boldsymbol{\theta})$ at the current iterate $\boldsymbol{\theta}_m$. Majorization is defined by two conditions: tangency $g(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}_m) = f(\boldsymbol{\theta}_m)$ at the current iterate, and domination $g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m) \geq f(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. The steps of the MM algorithm are defined by the rule

$$\boldsymbol{\theta}_{m+1} := \arg\min_{\boldsymbol{\theta}}\; g(\boldsymbol{\theta} \mid \boldsymbol{\theta}_m), \tag{3}$$

which immediately implies the descent property

$$f(\boldsymbol{\theta}_{m+1}) \;\leq\; g(\boldsymbol{\theta}_{m+1} \mid \boldsymbol{\theta}_m) \;\leq\; g(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}_m) \;=\; f(\boldsymbol{\theta}_m).$$

That is, a decrease in g results in a decrease in f. Note that $g(\boldsymbol{\theta}_{m+1} \mid \boldsymbol{\theta}_m) \leq g(\boldsymbol{\theta}_m \mid \boldsymbol{\theta}_m)$ does not require $\boldsymbol{\theta}_{m+1}$ to minimize g exactly, so that any descent step in g suffices. The MM principle offers a general prescription for transferring a difficult optimization task onto a sequence of simpler problems (Lange et al., 2000), and includes the well-known EM algorithm for maximum likelihood estimation under missing data as a special case (Becker et al., 1997).

Power means Power means are a class of generalized means defined $M_s(\boldsymbol{y}) = \left(\frac{1}{k}\sum_{i=1}^k y_i^s\right)^{1/s}$ for a vector $\boldsymbol{y}$. We see that $s > 1$ corresponds to the usual ℓ_s-norm of $\boldsymbol{y}$, $s = 1$ to the arithmetic mean, and $s = -1$ to the harmonic mean. Power means possess a number of nice properties: they are homogeneous, monotonic, and differentiable with

$$\frac{\partial}{\partial y_j} M_s(\boldsymbol{y}) \;=\; \left(\frac{1}{k}\sum_{i=1}^k y_i^s\right)^{\frac{1}{s}-1} \frac{1}{k} y_j^{s-1}, \tag{4}$$

and importantly they satisfy the limits

$$\lim_{s \to -\infty} M_s(\boldsymbol{y}) = \min\{y_1, \ldots, y_k\} \tag{5a}$$

$$\lim_{s\to\infty} M_s(\boldsymbol{y}) = \max\{y_1,\ldots,y_k\}. \tag{5b}$$

Further, the well-known power mean inequality holds: for any $s \leq t$, $M_s(\boldsymbol{y}) \leq M_t(\boldsymbol{y})$ (Steele, 2004).

Xu & Lange (2019) utilize these means toward clustering, proposing the power k-means objective function defined

$$f_s(\Theta) = \sum_{i=1}^{n} M_s(\|\boldsymbol{x}_i - \boldsymbol{\theta}_1\|^2, \ldots, \|\boldsymbol{x}_i - \boldsymbol{\theta}_k\|^2) \tag{6}$$

for a given power s. The algorithm then seeks to minimize f_s iteratively while sending $s \to -\infty$. Doing so approaches the original $f(\Theta)$ in (1) due to (5), coinciding with the original k-means objective and retaining its interpretation as minimizing within-cluster variance. The k-harmonic means method (Zhang et al., 1999), an early attempt to reduce the sensitivity to initialization of k-means by replacing the $\min$ appearing in (1) by the harmonic average, can be seen as the special case of (6) with $s = -1$. Power k-means clustering extends this idea to work in higher dimensions when the harmonic mean is longer a good proxy for (1), instead using a sequence of *power means* as a family of successively smoother optimization landscapes. The intermediate surfaces exhibit fewer poor local optima than (1), and each step is carried out via MM.

2. Bregman Power k-Means

We consider a power k-means objective function under a given Bregman divergence d_ϕ and power s:

$$f_s(\boldsymbol{\Theta}) = \sum_{i=1}^{n} M_s(d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_1), \ldots, d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_k)) \tag{7}$$

We see that this is a generalization of (6), as power k-means is recovered by taking ϕ to be the squared norm. On the other hand, Paul et al. (2021b) propose to use Adagrad to minimize a more general objective in that $M_s : \mathbb{R}^k_{\geq 0} \to \mathbb{R}_{\geq 0}$ is *any* component-wise non-decreasing function, such as a generalized mean. Though the general theoretical treatment in Paul et al. (2021b) does encompass the case where dissimiliarities $d()$ are given by Bregman divergences, the authors do not consider or implement this case explicitly. As a result, there a generic incremental optimization (such as Adagrad) is suggested, which produces a less scalable algorithm. In contrast, the geometry of Bregman divergences allows us to derive an elegant MM algorithm with closed form updates, matching the complexity of standard power k-means and Lloyd's algorithm. These properties will also lead to stronger theoretical results, detailed in the following section.

First, convexity of ϕ together with properties of power means ensures that our objective can be *majorized* by its tangent plane. That is, upon differentiating (4), one can see that the Hessian matrix of $M_s()$ is concave whenever $s \leq 1$ (Xu & Lange, 2019). This yields an upper bound which will supply a useful surrogate function:

$$\begin{aligned} f_s(\boldsymbol{\Theta}) \leq f_s(\boldsymbol{\Theta}_m) - \sum_{i=1}^{n}\sum_{j=1}^{k} w_{m,ij} \cdot d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_{m,j}) + \\ \sum_{i=1}^{n}\sum_{j=1}^{k} w_{m,ij} \cdot d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_j), \end{aligned} \tag{8}$$

where the scalars from partial differentiation abbreviated

$$w_{m,ij} = \frac{\frac{1}{k} d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_{m,j})^{s-1}}{(\frac{1}{k}\sum_{l=1}^{k}(d_\phi(\boldsymbol{x}_l, \boldsymbol{\theta}_{m,j})^s)^{1-\frac{1}{s}}} \tag{9}$$

act as weights between $\boldsymbol{x}_i$ and $\boldsymbol{\theta}_j$ at the m^{th} iteration.

Next, the mean-as-minimizer property from Prop. (1.1) suggests we may expect a closed form solution to the stationarity equations, which we derive here for completenesss. Analogous to the iteration between updating cluster label assignments and then re-defining cluster means in Lloyd's algorithm for standard k-means, we update cluster centers by minimizing the right hand side of equation (8) given weights $w_{m,ij}$ with respect to $\boldsymbol{\theta}$: for each j,

$$\begin{aligned} \nabla_{\boldsymbol{\theta}_j} \big[f_s(\boldsymbol{\Theta}_m) - \sum_{i=1}^{n}\sum_{j=1}^{k} w_{m,ij} d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_{m,j}) + \\ \sum_{i=1}^{n}\sum_{j=1}^{k} w_{m,ij} d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_j) \big] = 0 \end{aligned} \tag{10}$$

$$\sum_{i=1}^{n} w_{m,ij} \nabla^2_{\boldsymbol{\theta}_j} \phi(\boldsymbol{\theta}_j) \cdot [\boldsymbol{\theta}_j - \boldsymbol{x}_i] = 0 \tag{11}$$

$$\boldsymbol{\theta}_{m+1,j} = \frac{\sum_{i=1}^{n} w_{m,ij} \boldsymbol{x}_i}{\sum_{i=1}^{n} w_{m,ij}}. \tag{12}$$

Algorithm 1 Bregman Power k-means Pseudocode

1. Initialize $s_0 < 0$ and $\boldsymbol{\Theta}_0$, input data $\boldsymbol{x} \in \mathbb{R}^{p\times n}$, constant $\eta > 1$, iteration $m = 1$
2. **repeat**
3. $w_{m,ij} \leftarrow (\frac{1}{k}\sum_{i=1}^{k} d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_{m,j})^{s_m})^{\frac{1}{s_m}-1} d_\phi(\boldsymbol{x}_i, \boldsymbol{\theta}_{m,j})^{s_m - 1}$
4. $\boldsymbol{\theta}_{m+1,j} = (\sum_{i=1}^{n} w_{m,ij})^{-1} \sum_{i=1}^{n} w_{m,ij} \boldsymbol{x}_i$
5. $s_{m+1} \leftarrow \eta \cdot s_m$ (optional)
6. **until** convergence

Equations (9) and (12) imply a transparent, easy-to-implement method that implicitly performs annealing through a family of optimization landscapes indexed by s. The resulting iteration can be summarized concisely in Algorithm 1. By contrast, a gradient-based update for $\boldsymbol{\theta}_j$

Table 1. Examples of exponential family distributons and their corresponding Bregman divergences

Distribution	$\phi(\boldsymbol{x})$	$d_\phi(\boldsymbol{x}, \boldsymbol{y})$
Gaussian	$\\|\boldsymbol{x}\\|^2$	$\\|\boldsymbol{x}-\boldsymbol{y}\\|^2$
Multinomial	$\sum_{i=1}^m x_i \log x_i$	$\sum_{i=1}^m x_i \log \frac{x_i}{y_i}$
Gamma	$-\alpha + \alpha \log \frac{\alpha}{x}$	$\frac{\alpha}{y}(y \log \frac{y}{x}) + x - y$
Poisson	$x \log x - x$	$x \log \frac{x}{y} - (x-y)$

with step size α such as suggested in (Paul et al., 2021b) would entail

$$\boldsymbol{\theta}_{m+1,j} = \boldsymbol{\theta}_{m,j} - \alpha \sum_{i=1}^n w_{m,ij} \nabla^2_{\boldsymbol{\theta}_{m,j}} \phi(\boldsymbol{\theta}_{m,j}) \cdot [\boldsymbol{\theta}_{m,j} - \boldsymbol{x}_i],$$

both incurring higher cost at each iteration and making significantly slower progress per step. Depending on the choice of method, computing and properly tuning α must be done on a case-by-case basis and adds additional overhead.

Exponential family data A statistical motivation for our generalization comes from the connection between Bregman divergences and exponential families. Recall exponential family distributions with parameter θ and scale parameter τ take the canonical form

$$p(y|\theta,\tau) = C_1(y,\tau) \exp\left\{\frac{y\theta - \phi^*(\theta)}{C_2(\tau)}\right\}.$$

The convex conjugate of its *cumulant function* ϕ^*, which we denote ϕ, uniquely generates the Bregman divergence d_ϕ that represents the exponential family likelihood up to proportionality. With g denoting the canonical link function, the negative log-likelihood of y can be written as its Bregman divergence to the mean:

$$-\ln p(y|\theta,\tau) = d_\phi\left(y, g^{-1}(\theta)\right) + C(y,\tau).$$

As an example, the cumulant function in the Poisson likelihood is $\phi^*(x) = e^x$, whose conjugate $\phi(x) = x \ln x - x$ produces the relative entropy $d_\phi(p,q) = p\ln(p/q) - p + q$. Similarly, recall that the Bernoulli likelihood has cumulant function $\phi^*(x) = \ln(1+\exp(x))$. Its conjugate is given by $\phi(x) = x\ln x + (1-x)\ln(1-x)$, and generates $d_\phi(p,q) = p \ln \frac{p}{q} + (1-p)\ln\frac{1-p}{1-q}$. These relationships for some common distributions are summarized in Table 1 and show, for instance, that maximizing the likelihood of a generalized linear model is equivalent to minimizing a Bregman divergence between the responses and regression parameters. In the context of clustering, they allow us to understand the analog of k-means minimizing the within-cluster variance. Indeed, the Bregman hard clustering problem is equivalent to finding a partitioning of the data such that the loss in Bregman information $I_\phi(X)$ due to quantization is minimized—or equivalently, such that the within-cluster Bregman information is minimized. See Theorem 1 of Banerjee et al. (2005b) for details and a formal statement of this result. Because we target the same objective (1) as $s \to -\infty$, our formulation inherits this property immediately in the target limit. In fact, it is not difficult to show that this convergence is *uniform*:

Theorem 2.1. *For any sequence, $s_m \downarrow -\infty$ and $s_1 \leq 1$, $f_{s_m}(\cdot)$ converges uniformly on $\mathscr{C}$ to the Bregman hard clustering objective (1).*

Another desirable property of Algorithm 1 is that all iterates lie within the convex hull of the data, which suggests performance stability in addition to standard convegence and descent guarantees as a valid MM algorithm (Lange, 2016). Let $\mathscr{C}$ denote the closed convex hull of the data; the following result is inherited directly from power k-means (Xu & Lange, 2019)

Theorem 2.2. *Let $\boldsymbol{\Theta}_{n,s}$ be the (global) minimizer of $f_s(\cdot)$. Then $\boldsymbol{\Theta}_{n,s} \subset \mathscr{C}$ for all $s \leq 1$.*

These proofs are fairly straightforward and are given in full detail in the Appendix.

3. Theoretical Analysis

In addition to casting the problem in such a way that it inherits its classical guarantees, we now contribute new theoretical devices toward understanding its generalization error. The complete proofs pertaining to this section are available in the Appendix. We consider data $\{\boldsymbol{X}_i\}_{i\in[n]}$ independent and identically distributed according to some distribution P, and further assume that P has a sub-exponential ℓ_2 norm. This condition on P is strictly weaker than imposing that P has bounded support, as required in recent analyses in the literature (Paul et al., 2021b): formally,

A 1. $\{\boldsymbol{X}_i\}_{i\in[n]} \overset{\text{i.i.d.}}{\sim} P$, with

- $\sigma = \\|\\|\boldsymbol{X}\\|_2\\|_{\psi_1} \triangleq \sup_{p\in\mathbb{N}} \frac{(\mathbb{E}\\|\boldsymbol{X}\\|_2^p)^{1/p}}{p} < \infty$.
- $\sigma_\phi = \\|\phi(\boldsymbol{X})\\|_{\psi_1} \triangleq \sup_{p\in\mathbb{N}} \frac{(\mathbb{E}|\phi(\boldsymbol{X})|^p)^{1/p}}{p} < \infty$.

Note that A 1 is satisfied by many popularly used distributional models not limited to Gaussian mixtures, and always holds whenever P has bounded support. We also make the following standard assumption on regularity of the corresponding Bregman divergence.

A 2. $\nabla\phi$ is τ_2-Lipschitz. Moreover, ϕ is τ_1-strongly convex, i.e. $\forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ and for $0 \leq \alpha \leq 1$,

$$\phi(\alpha\boldsymbol{x}+(1-\alpha)\boldsymbol{y}) \leq \alpha\phi(\boldsymbol{x})+(1-\alpha)\phi(\boldsymbol{y}) - \frac{\tau_1}{2}\alpha(1-\alpha)\|\boldsymbol{x}-\boldsymbol{y}\|_2^2.$$

Recall strong convexity of ϕ relates to the smoothness of its conjugate, i.e. the cumulant function ϕ^* of exponential families (Kakade et al., 2010; Zhou, 2018). Note that under A 2, both σ and σ_ϕ are finite when $\|\boldsymbol{X}\|_2$ is sub-Gaussian. Thus, we are also able to generalize the assumptions used in analyses of approaches such as convex clustering (Tan & Witten, 2015), as detailed in Appendix D. Now, let $\tilde{f}_{\boldsymbol{\Theta}}(\boldsymbol{x}) = M_s\left(d_\phi(\boldsymbol{x}, \boldsymbol{\theta}_1), \ldots, d_\phi(\boldsymbol{x}, \boldsymbol{\theta}_k)\right)$ and P_n be the empirical distribution based on the data $\{\boldsymbol{X}_i\}_{i\in[n]}$. That is, $P_n(A) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{\boldsymbol{X}_i \in A\}$ for any Borel set A. For simplicity, we denote $\mu g = \int g d\mu$ for any measurable function g and measure μ. Fixing these conventions, note the objective (7), upon scaling by $1/n$, can be written as $P_n\tilde{f}_{\boldsymbol{\Theta}}$. By the strong law, we know that for any $\boldsymbol{\Theta} \in \mathbb{R}^{k\times p}$, $P_n\tilde{f}_{\boldsymbol{\Theta}} \xrightarrow{a.s.} P\tilde{f}_{\boldsymbol{\Theta}}$. Thus, as n becomes large, our intuition tells us to expect that as the functions of $\boldsymbol{\Theta}$ $P_n\tilde{f}_{\boldsymbol{\Theta}}$ and $P\tilde{f}_{\boldsymbol{\Theta}}$, become close to each other, so do their respective minimizers $\widehat{\boldsymbol{\Theta}}_n$ and $\boldsymbol{\Theta}_*$. To make precise the notion of convergence of $\widehat{\boldsymbol{\Theta}}_n$ towards $\boldsymbol{\Theta}_*$, we denote the *excess risk* at any set of cluster centroids $\boldsymbol{\Theta}$ as

$$\mathfrak{R}(\boldsymbol{\Theta}) = P\tilde{f}_{\boldsymbol{\Theta}} - P\tilde{f}_{\boldsymbol{\Theta}_*}.$$

The goal of this section is to formally assert that $\mathfrak{R}(\widehat{\boldsymbol{\Theta}}_n)$ becomes very small with a high probability as one has access to more and more data.

As a first step, we prove a high probability result on $\widehat{\boldsymbol{\Theta}}_n$, showing that $\widehat{\boldsymbol{\Theta}}_n$ remains bounded with a high probability as n becomes large. To this end, we require a notion of distance between sets of cluster centroids, and following the literature (Chakraborty & Das, 2021) use the measure

$$\operatorname{dist}(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \triangleq \min_{O\in\mathscr{P}_k} \|\boldsymbol{\Theta}_1 - O\boldsymbol{\Theta}_2\|_F,$$

where $\mathscr{P}_k$ denotes the set of all $k \times k$ real permutation matrices. In particular, this accounts for the label switching problem and is agnostic to relabeling classes. Likewise, we require a standard identifiability condition (Pollard, 1981; Paul et al., 2021b):

A 3. Let $M_\epsilon = \inf\{M > 0 : \{\boldsymbol{\Theta} \in \mathbb{R}^{k\times p} : \operatorname{dist}(\boldsymbol{\Theta}, \boldsymbol{\Theta}_*) > M\} \subseteq \{\boldsymbol{\Theta} \in \mathbb{R}^{k\times p} : \mathfrak{R}(\boldsymbol{\Theta}) > \epsilon\}\}$. Then for any $\epsilon > 0$, we have $M_\epsilon < \infty$.

This states that when the distance from $\boldsymbol{\Theta}$ to $\boldsymbol{\Theta}_*$ is large, then the excess risk at $\boldsymbol{\Theta}$ is also large. The following theorem formally states a high probability bound on $\widehat{\boldsymbol{\Theta}}_n$.

Theorem 3.1. *Under assumption 1-3, $\widehat{\boldsymbol{\Theta}}_n \subset B(\xi_P + \|\boldsymbol{\Theta}_*\|_F)$, with probability at least $1 - e^{-cn}$. Here c is an absolute constant and $\xi_P = M_{P\phi+\sigma_\phi}$.*

The main idea for the proof of Theorem 3.1 is that $P\tilde{f}_{\widehat{\boldsymbol{\Theta}}_n}$ remains bounded with a high probability. Thus, $\mathfrak{R}(\widehat{\boldsymbol{\Theta}}_n)$ is also bounded with a high probability, which in turn implies that $\operatorname{dist}(\widehat{\boldsymbol{\Theta}}_n, \boldsymbol{\Theta}_*) \le \xi_P$. Before our main theorem, we recall the definitions of Rademacher complexity (Bartlett & Mendelson, 2002) and covering numbers.

Definition 3.2. (Rademacher complexity) The population Rademacher complexity of a function class $\mathcal{F}$ is defined as,

$$\mathcal{R}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{i=1}^n \epsilon_i f(\boldsymbol{X}_i),$$

where ϵ_i's are i.i.d Rademacher random variables.

Definition 3.3. (δ-cover and covering number) For a metric space (X, d), the set $X_\delta \subseteq X$ is said to be a δ-cover of X if for all $x \in X$, there is $x' \in X_\delta$, such that $d(x, x') \le \delta$. The δ-covering number of X, denoted by $N(\delta; X, d)$, is the size of the smallest δ-cover of X with respect to d.

Now consider the set $\mathcal{F} = \{\tilde{f}_{\boldsymbol{\theta}} : \boldsymbol{\Theta} \subset B(\xi_P + \|\boldsymbol{\Theta}_*\|_F)\}$, under the measure of distances between functions

$$d_{2n}(f, g) \triangleq \left(\frac{1}{n}\sum_{i=1}^n (f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i))^2\right)^{-1/2}.$$

The following theorem imposes a bound on the covering number of $\mathcal{F}$ with respect to the d_{2n} metric. The proof makes use of A 2, that $\tilde{f}_{\boldsymbol{\Theta}}$ is Lipschitz on $B(\xi_P + \|\boldsymbol{\Theta}_*\|_F)$.

Theorem 3.4. *Under assumption A 2,*

$$\mathcal{N}(\delta; \mathcal{F}, d_{2n}) \le \left(\max\left\{1, \left\lfloor \frac{(\xi_P + \|\boldsymbol{\Theta}_*\|_F)C^{1/2}}{\delta}\right\rfloor\right\}\right)^{kp};$$

$C = 2k^{2-2/s}\tau_2^2 pn^{-1}\sum_{i=1}^n(18\xi_P^2 + 18\|\boldsymbol{\Theta}_*\|_F^2 + \|\boldsymbol{X}_i\|_2^2)$.

One can now appeal to this bound on the covering number of $\mathcal{F}$ to bound the Rademacher complexity, $\mathcal{R}_n(\mathcal{F})$. More technically, we make use of Theorem 3.4 and apply Dudley's chaining arguments to produce a $\mathcal{O}(1/\sqrt{n})$ bound on the Rademacher complexity of $\mathcal{F}$.

Theorem 3.5. *Under assumptions A 1 and A 2,*

$$\mathcal{R}_n(\mathcal{F}) \le 6\tau_2 C'(\xi_P + \|\boldsymbol{\Theta}_*\|_F)\frac{k^{3/2-1/s}p}{\sqrt{n}},$$

where $C' = \sqrt{2\pi(18\xi_P^2 + 18\|\boldsymbol{\Theta}_*\|_F^2 + \mathbb{E}\|\boldsymbol{X}\|_2^2)}$.

The Rademacher complexity bound plays a key role in providing uniform concentration bounds on $\|P_n - P\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}}|P_n f - Pf|$. Since the functions in $\mathcal{F}$ are not bounded (as we do not assume $\boldsymbol{x}$ is bounded), the classical results by Bartlett & Mendelson (2002) do not directly apply. However, appealing to the sub-exponential property of $\|\boldsymbol{X}\|_2$, we apply recent concentration results derived by (Maurer & Pontil, 2021). Formally, our bound is as follows:

Theorem 3.6. *Suppose assumptions A 1-2 hold. Then for $n \ge \log(2/\delta) \ge \frac{1}{2}$, with probability at least $1 - \delta$,*

$$\|P_n - P\|_{\mathcal{F}} \le 12\tau_2 C'(\xi_P + \|\boldsymbol{\Theta}_*\|_F)\frac{k^{3/2-1/s}p}{\sqrt{n}} + 16e\sigma\tau_2 k^{1-1/s}(1 + \xi_P + \|\boldsymbol{\Theta}_*\|_F)\sqrt{\frac{2\log(2/\delta)}{n}}.$$

From Theorem 3.1, we know that with a very high probability, $\widehat{\mathbf{\Theta}}_n \subset B(\xi_p + \|\mathbf{\Theta}_*\|_F)$. Using this result, it is not difficult to then show that with a very high probability, $\mathfrak{R}(\widehat{\mathbf{\Theta}}_n) \leq 2\|P_n - P\|_{\mathcal{F}}$, which can be bounded by Theorem 3.6. Finally, the next theorem provides a bound on the excess risk.

Theorem 3.7. *Let Assumptions 1-2 hold. Then whenever $n \geq \log(2/\delta) \geq \frac{1}{2}$, with probability at least $1 - \delta - e^{-cn}$,*

$$\mathfrak{R}(\widehat{\mathbf{\Theta}}_n) \ \leq \ 24\tau_2 C'(\xi_P + \|\mathbf{\Theta}_*\|_F)\frac{k^{3/2-1/s}p}{\sqrt{n}} + 32e\sigma\tau_2 k^{1-1/s}(1 + \xi_P + \|\mathbf{\Theta}_*\|_F)\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Remark It is important to note that the bounds derived in Theorem 3.7 include the Frobenius norm of the population cluster centroid $\|\mathbf{\Theta}_*\|_F$, as well as terms such as $\mathbb{E}\|\boldsymbol{X}\|_2^2$ and $\|\|\boldsymbol{X}\|_2\|_{\psi_1}$ measuring the spread of the data. Intuitively, as spread of the data increases, it can be expected that the performance of Bregman power k-means deteriorates with the added noise. This phenomenon is reflected in the bounds on the excess risk.

Strong Consistency and $\sqrt{n}$-consistency In the classical domain of keeping k and p fixed, one can recover asymptotic results (Pollard, 1981) such as strong consistency and $\sqrt{n}$-consistency of the sample cluster centroids. We say that the sequence of the set of cluster centroids $\{\mathbf{\Theta}_n\}_{n\in\mathbb{N}}$ converges to $\mathbf{\Theta}$ if $\lim_{n\to\infty} \operatorname{dist}(\mathbf{\Theta}_n, \mathbf{\Theta}) = 0$. The following theorem asserts that indeed $\widehat{\mathbf{\Theta}}_n$ is strongly consistent for $\mathbf{\Theta}$, and moreover admits a parametric convergence rate of $\mathcal{O}(n^{-1/2})$. Before stating the result, recall that for a sequence of random variables $\{X_n\}_{n\in\mathbb{N}}$, we say that $X_n = \mathcal{O}_P(a_n)$, for a sequence of reals $\{a_n\}_{n\in\mathbb{N}}$, if X_n/a_n is *tight*, or bounded in probability.

Theorem 3.8. *If p is kept fixed, then under Assumptions 1-3, $\widehat{\mathbf{\Theta}}_n \xrightarrow{a.s.} \mathbf{\Theta}_*$. Moreover, $\mathfrak{R}(\widehat{\mathbf{\Theta}}_n) = \mathcal{O}_P(n^{-1/2})$.*

4. Empirical Performance and Results

We close our assessment of the proposed method with a thorough empirical study. Previous works on Bregman clustering largely focus on mathematical aspects, and the few that include data examples are limited to low dimensions. We extend their designs to settings as dimension increases, for a large number of clusters, and for a breadth of exponential family distributions with varying parameters, followed by application to a rainfall dataset. An open-source Python implementation of the proposed method, including reproducible code for all data generating mechanisms and experiments in this paper, is available and maintained in a repository by the first author[1].

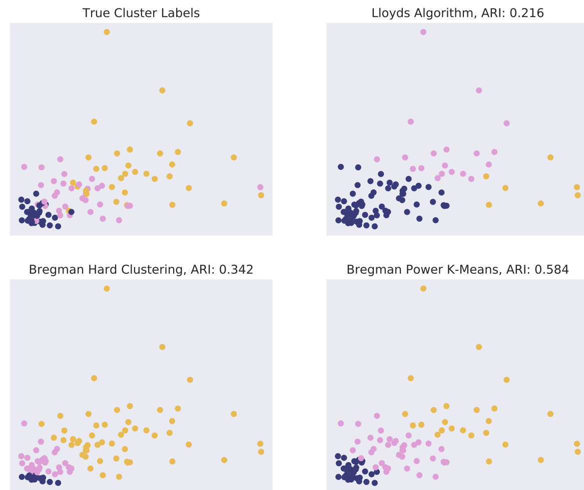

Figure 1. A visual comparison of clustering solutions.

Experiment 1: We begin by considering data simulated from various exponential families in the plane. Synthetic datasets are generated in $\mathbb{R}^2$ from Gaussian, Binomial, Poisson, and Gamma distributions from true centers at $(10, 10)$, $(20, 20)$, and $(40, 40)$. For the normal case $\sigma^2 = 16$, while the binomial parameter $n = 200$ so that clusters feature heteroskedasticity with variance implied by the mean relationship. Poisson data are sampled coordinate-wise coordinate with intensity parameter $10, 20$, and 40 respectively. Finally, each Gamma coordinate was sampled to have the same means, with shape parameters fixed at $\alpha = 15$.

As discussed in Section 2 and illustrated in Table 1, a Bregman divergence for each of these distributions may provide an ideal measure of dissimilarity for clustering data generated under its corresponding exponential family. To investigate this, we apply Lloyd's k-means algorithm, Bregman hard clustering, the original power k-means method, and our proposed Bregman power k-means algorithm on each of these four settings. Centers are randomly initialized according to a uniform distribution spanning the range of all the data points, and each peer method starts from matched initializations to ensure a fair comparison. An s_0 value of -0.2 was used for power k-means and our method.

A visual comparison of clustering solutions obtained on one simulated dataset by each peer method on Gamma data with a shape parameter $\alpha = 5$ is displayed in Figure 1. Comparing to the ground truth labels, this illustrates that Bregman power k-means is much more effective than its

[1] Publicly available at `https://github.com/avellal14/bregman_power_kmeans`

competing methods in distinguishing between points drawn from different Gamma clusters, despite the data not being perfectly separable. This is particularly apparent for Lloyd's algorithm, which fails due to seeking spherical clusters. For a closer look, we report the mean (and standard deviation) adjusted Rand index (ARI) of solutions under each algorithm computed over 250 trials in Table 2. We observe that other than performing on par in the Gaussian case, Bregman power k-means consistently achieves the best performance in the other exponential family settings.

Table 2. Mean and (standard deviation) ARI of Lloyd's algorithm, Bregman hard clustering, and their power means counterparts.

	Lloyd's	Bregman Hard	Power	Bregman Power
Gaussian	0.828 (0.012)	0.837 (0.012)	0.927 (0.003)	0.927 (0.003)
Binomial	0.730 (0.014)	0.886 (0.011)	0.915 (0.004)	0.931 (0.003)
Poisson	0.723 (0.014)	0.882 (0.010)	0.888 (0.006)	0.916 (0.004)
Gamma	0.484 (0.009)	0.868 (0.005)	0.677 (0.008)	0.879 (0.004)

Experiment 2: To better understand the behavior of each clustering method as the shape of distributions change, we now revisit the Gamma setting while varying the shape parameters $\alpha = 1, \ldots, 20$. We also increase the problem dimension to $p = 20$, with all other simulation details are unchanged from Experiment 1. Since the centers are held fixed as α is increased, higher α values correspond to less skewed Gamma distributions with lower variances. The mean ARIs (each again computed over 250 random simulations) against increasing shape parameter values are summarized in Figure 2. Due to the high skewness causing significant overlap between clusters, smaller shape parameters result in the poorest performance across all four peer methods. Each method seems to reach an inflection point somewhere in the range of $\alpha = 3 - 6$, after which further increases in α minimally change the overall shape of the distribution. As expected, Bregman power k-means achieves the best performance while Lloyd's algorithm struggles across the board. It is worth also noting that (Euclidean) power k-means eventually overtakes Bregman hard clustering, and mostly maintains better performance than Bregman hard clustering for high shape parameter values. This at first surprising result can be reconciled by the interpretation of Gamma distributions as a sum of exponential random variables (explicitly when α is an integer), so that a sum of i.i.d variables looks closer to normal as α increases. Eventually, the improvement bestowed by annealing through poor minima becomes more advantageous than information that the data are in fact Gamma as they resemble Gaussian data more and more closely. The gap in performance between our proposed method and Bregman hard clustering reiterates the merits of annealing through power means.

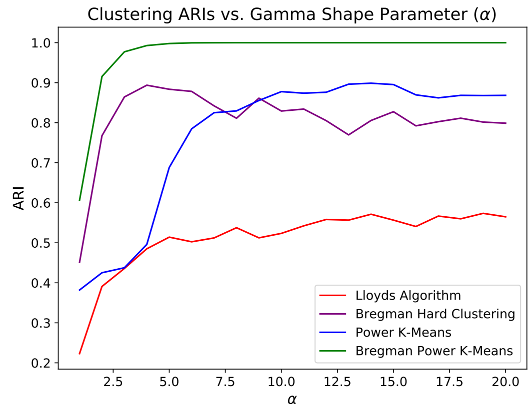

Figure 2. Performance as Gamma shape parameter varies.

Experiment 3: To better understand the effect of annealing in various feature dimensions p, we take a closer look at the Poisson setting when dimensionality ranges from $p = 2$ to $p = 50$. In this setting, true centers are given by $[40, 40]$, $[50, 50]$ and $[60, 60]$ in the planar case, while we scale the separation inversely by a $\sqrt{d}$ factor as is standard to avoid the problem becoming "too easy" in larger dimensions (Aggarwal et al., 2001). Mean ARIs and standard deviations across 250 random trials are detailed in Table 3, which also considers various initial powers s_0 for power k-means and our proposed method. We reproduce the finding of Xu & Lange (2019) under the Gaussian setting that while annealing provides advantages without the need to carefully tune s_0, more negative starting values tend to yield a more pronounced advantage as p increases. The best performing method for each dimension p is boldfaced. Similarly, we provide runtime details in Table 4. Though results vary based on implementations, we see that the proposed method is highly efficient in the data settings we consider here, and outpace competitors by an order of magnitude. Similar trends are conveyed under the other data generating mechanisms.

Experiment 4: As suggested by an anonymous reviewer, to check that empirical performance matches our theory that cluster centroids are $\sqrt{n}$-consistent (cf. Theorem 3.8), we include another experiment using the Poisson data setting ($p = 5$). We sample from the same three cluster centroids that we do in Experiment 3, and consider the Bregman divergence d_ϕ between the cluster centroids estimated by Bregman Power k-means ($\widehat{\boldsymbol{\Theta}}_n$) and the true cluster centroids ($\boldsymbol{\Theta}$) as the number of data points per cluster, n, varies from 1 to 100. For each n, we plot the lowest $d_\phi(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}_n)$ across 100 random trials to decrease how likely it is that we report the divergence at a local optimum (as the theoretical result

Table 3. Average (standard deviation) ARI across 250 trials as dimension increases, Poisson data.

	$p = 2$	$p = 5$	$p = 10$	$p = 20$	$p = 50$
Lloyd's	0.418 (0.005)	0.488 (0.006)	0.610 (0.009)	0.607 (0.012)	0.614 (0.015)
Bregman Hard	0.425 (0.005)	0.497 (0.006)	0.613 (0.009)	0.653 (0.011)	0.720 (0.014)
Power, $s_0 = -0.2$	0.451 (0.005)	**0.577 (0.005)**	0.690 (0.005)	0.561 (0.010)	0.440 (0.001)
Breg. Power, $s_0 = -.2$	**0.458 (0.005)**	**0.577 (0.005)**	0.718 (0.005)	0.734 (0.010)	0.442 (0.002)
Power, $s_0 = -1$	0.445 (0.005)	0.575 (0.005)	0.692 (0.005)	0.745 (0.005)	0.485 (0.008)
Breg. Power, $s_0 = -1$	0.452 (0.005)	0.575 (0.005)	**0.720 (0.005)**	**0.806 (0.005)**	0.610 (0.014)
Power, $s_0 = -3$	0.432 (0.005)	0.548 (0.006)	0.696 (0.005)	0.784 (0.004)	0.873 (0.004)
Breg. Power, $s_0 = -3$	0.412 (0.005)	0.555 (0.005)	0.712 (0.005)	0.804 (0.005)	0.924 (0.003)
Power, $s_0 = -9$	0.412 (0.005)	0.531 (0.006)	0.693 (0.006)	0.782 (0.006)	0.915 (0.003)
Breg. Power, $s_0 = -9$	0.321 (0.006)	0.510 (0.007)	0.696 (0.006)	0.792 (0.005)	**0.925 (0.003)**

Table 4. Average (standard deviation) runtimes (sec) across 250 trials as dimension increases, Poisson data.

	$p = 2$	$p = 5$	$p = 10$	$p = 20$	$p = 50$
Lloyd's	0.082 (0.004)	0.067 (0.002)	0.058 (0.002)	0.052 (0.001)	0.056 (0.002)
Bregman Hard	0.607 (0.012)	0.493 (0.008)	0.413 (9e-4)	0.439 (0.005)	0.506 (0.008)
Power, $s_0 = -0.2$	0.012 (0.001)	0.001 (3e-4)	0.009 (2e-4)	0.007 (4e-4)	0.004 (1e-4)
Breg. Power, $s_0 = -.2$	0.022 (0.001)	0.020 (8e-4)	0.017 (4e-4)	0.021 (7e-4)	0.009 (2e-4)
Power, $s_0 = -1$	0.006 (1e-4)	0.005 (1e-4)	0.007 (1e-4)	0.016 (4e-4)	0.006 (3e-4)
Breg. Power, $s_0 = -1$	0.011 (3e-4)	0.012 (3e-4)	0.014 (3e-4)	0.036 (7e-4)	0.019 (9e-4)
Power, $s_0 = -3$	0.005 (1e-4)	0.006 (1e-4)	0.007 (2e-4)	0.0059 (1e-4)	0.008 (2e-4)
Breg. Power, $s_0 = -3$	0.007 (1e-4)	0.01 (2e-4)	0.013 (4e-4)	0.011 (3e-4)	0.017 (4e-4)
Power, $s_0 = -9$	0.004 (9e-5)	0.006 (2e-4)	0.011 (3e-4)	0.055 (1e-4)	0.007 (2e-4)
Breg. Power, $s_0 = -9$	0.004 (2e-5)	0.01 (2e-4)	0.019 (4e-4)	0.010 (2e-4)	0.012 (3e-4)

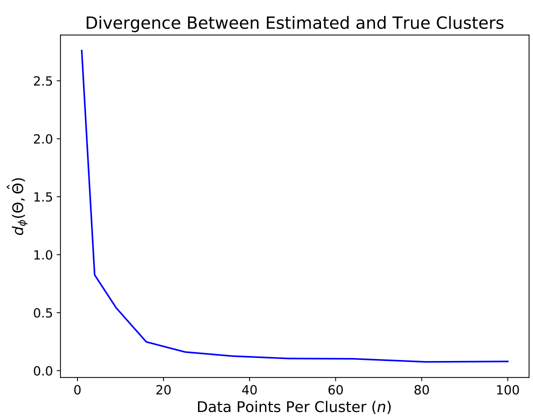

Figure 3. We see that the empirical convergence of Bregman power k-means to the true cluster centroids agrees with the $\mathcal{O}_P(n^{-1/2})$ convergence proposed in Theorem 3.8.

pertains to the true minimizers of the objective). The results are plotted in Figure 3.

Rainfall data: We now turn to a comparison of peer methods on a real dataset. Following recent Bregman clustering work of Brécheteau et al. (2021), we consider clustering rainfall data under a Gamma model with shape parameter $\alpha = 4$. Here we consider data from the Italian region of San Martino di Castrozza, collected across the years $1970-1990$ during the months of January (177 points) and June (397 points)[2]. Only days with non-zero rainfall amounts are included, and their values are often modeled by domain experts according to a Gamma distribution (Coe & Stern, 1982). For Power k-means and Bregman Power k-means, an s_0 value of -3.0 is used.

Similarly to the results in (Brécheteau et al., 2021), we observe very low separability between clusters likely caused by the large number of days with very little rainfall as evident in the ground truth plotted in Figure 4. Absolute ARIs of all methods in turn are quite low, yet Bregman power k-means offers the best performance, with a $\mathbf{48\%}$ improvement relative to the next best contender, Bregman hard clustering. Furthermore, Bregman power k-means has a runtime of 0.017 sec in comparison to the 0.975 sec runtime of Bregman hard clustering, offering a significant speedup just as we observed in Experiment 3. Both Bregman clustering methods perform an order of magnitude better than Lloyd's algorithm as well as power k-means. This is not unexpected given the highly skewed shapes of the data that bear little similarity to spherical clusters such as those arising under a Gaussian assumption. A visual comparison of the solutions can be found in Figure 4.

[2]Publicly available at `https://cran.r-project.org/web/packages/hydroTSM/vignettes/hydroTSM_Vignette-knitr.pdf`

Figure 4. A visual comparison of clustering solutions under Lloyd's algorithm for k-means, Bregman hard clustering, and the proposed method on rainfall dataset. We see the ground truth features significant overlap between the groups, yet Bregman power k-means is able to find a closer partition than its peer methods.

5. Discussion

In this article, we have demonstrated the merits of adapting the recent power k-means method to Bregman divergences both through several novel theoretical contributions and a thorough empirical study. In line with what we would expect from the theory, we find that across a number of exponential family data generating mechanisms and as cluster numbers and dimension increase, Bregman power k-means consistently outperforms its Euclidean counterparts. On the other hand, we observe a marked improvement over Bregman hard clustering without annealing, showing that the majorization-minimization scheme proposed here successfully evades poor local minima. By tailoring our algorithm to the case of Bregman divergences within the family of power means, we preserve the simplicity of Lloyd's classical method, in contrast to generic gradient-based approaches put forth for a general robust clustering context that can be applied.

A number of immediate extensions are worth exploring. The problem of feature selection alongside clustering is worthy of investigation, as discovering relevant variables may be crucial in high dimensional settings with low signal-to-noise ratio (Chakraborty et al., 2020), and may be an interpretable goal in its own right. Another useful generalization involves clustering matrix-variate data. The notion of Bregman divergence readily applies to matrices:

$$d_\phi(\boldsymbol{V}, \boldsymbol{U}) = \phi(\boldsymbol{V}) - \phi(\boldsymbol{U}) - \langle \nabla\phi(\boldsymbol{U}), \boldsymbol{V} - \boldsymbol{U} \rangle$$

where $\langle \boldsymbol{V}, \boldsymbol{U} \rangle = \mathrm{Tr}(\boldsymbol{V}\boldsymbol{U}^T)$ denotes the inner product.

In exploring penalized or robust versions where the objective may no longer admit closed form updates, it is worth leveraging the information geometry behind Bregman divergences to design effective iterative extensions of our proposed method. One can benefit from second-order rate behavior within first-order schemes—as an example, Raskutti & Mukherjee (2015) show that mirror descent performs *natural* gradient descent along the dual Riemannian manifold under a Bregman proximity term. For exponential families, the Riemannian metric of the parameter space coincides with the Fisher information, a property that has proven useful in a number of machine learning applications (Hoffman et al., 2013). Moreover, the explicit connection to exponential family likelihoods suggests it is natural to explore Bayesian approaches that may leverage the geometry of Bregman divergences (Ahn & Chewi, 2020; Duan & Dunson, 2021), especially toward model-based clustering on exponential family mixtures.

From an optimization perspective, this work falls within the unified continuous optimization perspective set forth by Teboulle (2007). There the authors suggested future extensions to various proximity measures, as well as investigation into the notion of "best" smoothing functions for given formulations. Our article contributes significant progress on the former, while our novel rate analyses which provide results dependent on $\boldsymbol{\Theta}_*$ suggest a new theoretical tool toward formally tackling the latter. Within the power means framework, this open problem is also augmented by the question of optimal annealing schedules for decreasing parameters such as s within a family of smoothing functions. We invite readers to consider these questions and possible extensions in the many exciting directions for future work.

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory*, pp. 420–434. Springer, 2001.

Ahn, K. and Chewi, S. Efficient constrained sampling via the mirror-Langevin algorithm. *arXiv preprint arXiv:2010.16212*, 2020.

Arthur, D. and Vassilvitskii, S. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics, 2007.

Bachem, O., Lucic, M., and Krause, A. Distributed and provably good seedings for k-means in constant rounds. In *International Conference on Machine Learning*, pp. 292–300. PMLR, 2017.

Banerjee, A., Guo, X., and Wang, H. On the optimality of conditional expectation as a Bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005a.

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J., and Lafferty, J. Clustering with Bregman divergences. *Journal of machine learning research*, 6(10), 2005b.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Becker, M. P., Yang, I., and Lange, K. EM algorithms without missing data. *Statistical Methods in Medical Research*, 6:38–54, 1997.

Beliakov, G., Calvo, T., and James, S. On Lipschitz properties of generated aggregation functions. *Fuzzy Sets and Systems*, 161(10):1437–1447, 2010.

Brécheteau, C., Fischer, A., and Levrard, C. Robust bregman clustering. *The Annals of Statistics*, 49(3):1679–1701, 2021.

Chakraborty, S. and Das, S. k-means clustering with a new divergence-based distance metric: Convergence and performance analysis. *Pattern Recognition Letters*, 100: 67–73, 2017.

Chakraborty, S. and Das, S. On uniform concentration bounds for bi-clustering by using the Vapnik–Chervonenkis theory. *Statistics & Probability Letters*, 175:109102, 2021.

Chakraborty, S., Paul, D., Das, S., and Xu, J. Entropy weighted power k-means clustering. In *International Conference on Artificial Intelligence and Statistics*, pp. 691–701. PMLR, 2020.

Coe, R. and Stern, R. D. Fitting models to daily rainfall data. *Journal of Applied Meteorology and Climatology*, 21(7):1024 – 1031, 1982. doi: 10.1175/1520-0450(1982)021⟨1024:FMTDRD⟩2.0.CO;2.

De Amorim, R. C. and Mirkin, B. Minkowski metric, feature weighting and anomalous cluster initializing in k-means clustering. *Pattern Recognition*, 45(3):1061–1075, 2012.

Deshpande, A., Kacham, P., and Pratap, R. Robust k-means++. In *Conference on Uncertainty in Artificial Intelligence*, pp. 799–808. PMLR, 2020.

Duan, L. L. and Dunson, D. B. Bayesian distance clustering. *Journal of Machine Learning Research*, 22(224):1–27, 2021.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12 (61):2121–2159, 2011. URL http://jmlr.org/papers/v12/duchi11a.html.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. Stochastic variational inference. *Journal of Machine Learning Research*, 14(5), 2013.

Kakade, S., Shamir, O., Sindharan, K., and Tewari, A. Learning exponential families in high-dimensions: Strong convexity and sparsity. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 381–388. JMLR Workshop and Conference Proceedings, 2010.

Lange, K. *MM optimization algorithms*, volume 147. SIAM, 2016.

Lange, K., Hunter, D. R., and Yang, I. Optimization transfer using surrogate objective functions (with discussion). *Journal of Computational and Graphical Statistics*, 9: 1–20, 2000.

Lloyd, S. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.

MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pp. 281–297. Oakland, CA, USA, 1967.

Mairal, J. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.

Maurer, A. and Pontil, M. Concentration inequalities under sub-Gaussian and sub-exponential conditions. *Advances in Neural Information Processing Systems*, 34, 2021.

Ng, A. Y., Jordan, M. I., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pp. 849–856, 2002.

Paul, D., Chakraborty, S., and Das, S. On the uniform concentration bounds and large sample properties of clustering with Bregman divergences. *Stat*, 10(1):e360, 2021a.

Paul, D., Chakraborty, S., Das, S., and Xu, J. Uniform concentration bounds toward a unified framework for robust clustering. *Advances in Neural Information Processing Systems*, 34, 2021b.

Pollard, D. Strong consistency of k-means clustering. *The Annals of Statistics*, 9(1):135–140, 1981.

Raskutti, G. and Mukherjee, S. The information geometry of mirror descent. *IEEE Transactions on Information Theory*, 61(3):1451–1457, 2015.

Rudin, W. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1964.

Steele, J. M. *The Cauchy-Schwarz master class: an introduction to the art of mathematical inequalities*. Cambridge University Press, 2004.

Tan, K. M. and Witten, D. Statistical properties of convex clustering. *Electronic journal of statistics*, 9(2):2324, 2015.

Teboulle, M. A unified continuous optimization framework for center-based clustering methods. *Journal of Machine Learning Research*, 8(1), 2007.

Telgarsky, M. and Dasgupta, S. Agglomerative bregman clustering. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, pp. 1011–1018, 2012.

Telgarsky, M. J. and Dasgupta, S. Moment-based uniform deviation bounds for k-means and friends. *Advances in Neural Information Processing Systems*, 26:2940–2948, 2013.

Van Der Vaart, A. W., van der Vaart, A. W., van der Vaart, A., and Wellner, J. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 1996.

Vassilvitskii, S. and Arthur, D. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2006.

Vershynin, R. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Xu, J. and Lange, K. Power k-means clustering. In *International Conference on Machine Learning*, pp. 6921–6931, 2019.

Zhang, B., Hsu, M., and Dayal, U. K-harmonic means—a data clustering algorithm. *Hewlett-Packard Labs Technical Report HPL-1999-124*, 55, 1999.

Zhou, X. On the Fenchel duality between strong convexity and Lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.

A. Proof of Optimization Results from Section 2

Proof of Theorem 2.2 Let $J(\boldsymbol{x}) := d_\phi(\boldsymbol{x}, \boldsymbol{\theta})$ and let $P_\mathscr{C}(\boldsymbol{\theta})$ be the projection of $\boldsymbol{\theta}$ onto $\mathscr{C}$, i.e. $P_\mathscr{C}(\boldsymbol{\theta}) = \operatorname{argmin}_{\boldsymbol{x}\in\mathscr{C}} d_\phi(\boldsymbol{x}, \boldsymbol{\theta})$. Since $P_\mathscr{C}(\boldsymbol{\theta})$ minimizes $J(\cdot)$ over $\mathscr{C}$, there exists a subgradient $\boldsymbol{d} \in \partial J(P_\mathscr{C}(\boldsymbol{\theta}))$ such that

$$\langle \boldsymbol{d}, \boldsymbol{x} - P_\mathscr{C}(\boldsymbol{\theta})\rangle \geq 0. \tag{13}$$

We note that $J(P_\mathscr{C}(\boldsymbol{\theta})) = \{\nabla\phi(P_\mathscr{C}(\boldsymbol{\theta})) - \nabla\phi(\boldsymbol{\theta})\}$. Thus, from equation (13),

$$\langle \nabla\phi(P_\mathscr{C}(\boldsymbol{\theta})) - \nabla\phi(\boldsymbol{\theta}), \boldsymbol{x} - P_\mathscr{C}(\boldsymbol{\theta})\rangle \geq 0. \tag{14}$$

We proceed by observing that

$$d_\phi(\boldsymbol{x}, \boldsymbol{\theta}) - d_\phi(\boldsymbol{x}, P_\mathscr{C}(\boldsymbol{\theta})) - d_\phi(P_\mathscr{C}(\boldsymbol{\theta}), \boldsymbol{\theta}) \;=\; \langle \nabla\phi(P_\mathscr{C}(\boldsymbol{\theta})) - \nabla\phi(\boldsymbol{\theta}), \boldsymbol{x} - P_\mathscr{C}(\boldsymbol{\theta})\rangle \;\geq\; 0.$$

Thus, for any $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1 : \boldsymbol{\theta}_2 : \cdots : \boldsymbol{\theta}_k]^\top \in \mathbb{R}^{k\times p}$, and so

$$\begin{aligned}
& d_\phi(\boldsymbol{x}, \boldsymbol{\theta}_j) \;\geq\; d_\phi(\boldsymbol{x}, P_\mathscr{C}(\boldsymbol{\theta}_j)) + d_\phi(P_\mathscr{C}(\boldsymbol{\theta}_j), \boldsymbol{\theta}_j) \;\geq\; d_\phi(\boldsymbol{x}, P_\mathscr{C}(\boldsymbol{\theta}_j)) \quad \forall\, j = 1, \ldots, k. \\
\Longrightarrow\; & M_s\left(d_\phi(\boldsymbol{x}, P_\mathscr{C}(\boldsymbol{\theta}_1)), \ldots, d_\phi(\boldsymbol{x}, P_\mathscr{C}(\boldsymbol{\theta}_k))\right) \;\leq\; M_s\left(d_\phi(\boldsymbol{x}, \boldsymbol{\theta}_1), \ldots, d_\phi(\boldsymbol{x}, \boldsymbol{\theta}_k)\right) \\
\Longrightarrow\; & \sum_{i=1}^n M_s\left(d_\phi(\boldsymbol{X}_i, P_\mathscr{C}(\boldsymbol{\theta}_1)), \ldots, d_\phi(\boldsymbol{X}_i, P_\mathscr{C}(\boldsymbol{\theta}_k))\right) \;\leq\; \sum_{i=1}^n M_s\left(d_\phi(\boldsymbol{X}_i, \boldsymbol{\theta}_1), \ldots, d_\phi(\boldsymbol{X}_i, \boldsymbol{\theta}_k)\right) \\
\Longrightarrow\; & f_s(\boldsymbol{\Theta}') \;\leq\; f_s(\boldsymbol{\Theta}),
\end{aligned}$$

where $\boldsymbol{\Theta}' = [P_\mathscr{C}(\boldsymbol{\theta}_1) : \cdots : P_\mathscr{C}(\boldsymbol{\theta}_k)]^\top \in \mathbb{R}^{k\times p}$. Thus, $\inf_{\boldsymbol{\Theta}} f_s(\boldsymbol{\Theta}) = \inf_{\boldsymbol{\Theta}\subset C} f_s(\boldsymbol{\Theta})$., and the result follows.

Proof of Theorem 2.1 For any $\boldsymbol{\Theta} \subset \mathscr{C}$ and $s_m \downarrow -\infty$, by monotonicity of power means, $f_{s_m}(\boldsymbol{\Theta}) \downarrow f_{k-means}(\boldsymbol{\Theta})$ monotonically. The result now immediately follows from Dini's theorem (Rudin, 1964) in analysis as $\mathscr{C}$ is compact.

B. Proof of Statistical Theory Results from Section 3

Before we proceed to the proofs of main results of section 3, we first state and prove the following two lemmas. The two results hold for any $M > 0$.

Lemma B.1. *For any $\boldsymbol{\theta} \in B(M)$, $\|\nabla\phi(\boldsymbol{\theta})\|_2 \leq \tau_2 M$.*

Proof. For any $\boldsymbol{\theta} \in B(M)$,

$$\|\nabla\phi(\boldsymbol{\theta})\|_2 = \|\nabla\phi(\boldsymbol{\theta}) - \nabla\phi(\boldsymbol{0})\|_2 \leq \tau_2\|\boldsymbol{\theta}\|_2 \leq \tau_2 M.$$

□

Lemma B.2. *ϕ is $\tau_2 M$-Lipschitz on $B(M)$.*

Proof. Appealing to the mean value theorem,

$$\phi(\boldsymbol{\theta}) - \phi(\boldsymbol{\theta}') = \langle \nabla\phi(\xi), \boldsymbol{\theta} - \boldsymbol{\theta}'\rangle,$$

for some ξ in the convex combinations of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. Clearly, $\xi \in B(M)$. Thus, by Lemma B.1, $\|\nabla\phi(\xi)\|_2 \leq \tau_2 M$. Finally, appealing to Cauchy-Schwartz inequality, we get,

$$|\phi(\boldsymbol{\theta}) - \phi(\boldsymbol{\theta}')| \leq \|\nabla\phi(\xi)\|_2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \leq \tau_2 M\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$$

□

Proof of Theorem 3.1

Proof. Let $\mathbf{0}_{k\times p}$ be the $k \times p$ real matrix whose entries are all zero. By the definition of $\mathbf{\Theta}_n$,

$$P_n \tilde{f}_{\widehat{\mathbf{\Theta}}_n} \leq P_n \tilde{f}_{\mathbf{0}_{k\times p}} = \frac{1}{n}\sum_{i=1}^{n} \phi(\boldsymbol{X}_i).$$

By Bernstein's inequality, we note that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n}\phi(\boldsymbol{X}_i) - P\phi > t\right) \leq \exp\left\{-\frac{cn}{\sigma_\phi}\min\left\{\frac{t^2}{\sigma_\phi}, t\right\}\right\}. \tag{15}$$

Upon taking $t = \sigma_\phi$, we observe that the RHS of (15) reduces to e^{-cn}.

Now, let $A_n = \{\mathbf{\Theta} \in \mathbb{R}^{k\times p} : P_n \tilde{f}_{\mathbf{\Theta}} \leq P\phi + \sigma_\phi\}$. Then with probability at least $1 - e^{-cn}$, $\widehat{\mathbf{\Theta}}_n \in A_n$. Next, by definition of the set A_n, the inequality

$$P\tilde{f}_{\mathbf{\Theta}} = \mathbb{E}_{\{\boldsymbol{X}\}_{i\in[n]} \overset{\text{i.i.d.}}{\sim} P}\left[P_n \tilde{f}_{\mathbf{\Theta}}\right] \;\leq\; \mathbb{E}_{\{\boldsymbol{X}\}_{i\in[n]} \overset{\text{i.i.d.}}{\sim} P}\left[P\phi + \sigma_\phi\right] = P\phi + \sigma_\phi$$

follows if $\mathbf{\Theta} \in A_n$. Thus, for all $n \in \mathbb{N}$, we have $A_n \subseteq \{\mathbf{\Theta} \in \mathbb{R}^{k\times p} : P\tilde{f}_{\mathbf{\Theta}} \leq P\phi + \sigma_\phi\}$.

Lastly, consider the set $\mathcal{E} = \{\mathbf{\Theta} \in \mathbb{R}^{k\times p} : P\tilde{f}_{\mathbf{\Theta}} \leq P\phi + \sigma_\phi\}$. For any $\mathbf{\Theta} \in \mathcal{E}$,

$$\mathfrak{R}(\mathbf{\Theta}) = P\tilde{f}_{\mathbf{\Theta}} - P\tilde{f}_{\mathbf{\Theta}_*} \leq P\tilde{f}_{\mathbf{\Theta}} \leq P\phi + \sigma_\phi.$$

By assumption A 3, $\mathrm{dist}(\mathbf{\Theta}, \mathbf{\Theta}_*) \leq M_{P\phi+\sigma_\phi} = \xi_P$. It is easy to see that $\mathrm{dist}(\widehat{\mathbf{\Theta}}_n, \mathbf{\Theta}_*) \leq \xi_P \implies \|\hat{\boldsymbol{\theta}}_j^{(n)}\|_2 \leq \xi_P + \|\mathbf{\Theta}_*\|_F$. Thus, $\widehat{\mathbf{\Theta}}_n \subset B(\xi_P + \|\mathbf{\Theta}_*\|_F)$, with probability at least $1 - e^{-cn}$. $\square$

Before we prove Theorem 3.4, we state and prove the following lemma, which asserts the Lipschitzness of $\tilde{f}_{\mathbf{\Theta}}$.

Lemma B.3. *Suppose A 2 holds. Then, for any $\mathbf{\Theta}, \mathbf{\Theta}' \subset B(\xi_P + \|\mathbf{\Theta}_*\|_F)$ and $\boldsymbol{x} \in \mathbb{R}^p$,*

$$|\tilde{f}_{\mathbf{\Theta}}(\boldsymbol{x}) - \tilde{f}_{\mathbf{\Theta}'}(\boldsymbol{x})| \leq k^{-1/s}\tau_2(3\xi_P + 3\|\mathbf{\Theta}_*\|_F + \|\boldsymbol{x}\|_2)\sum_{j=1}^{k}\|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j\|_2.$$

Proof. For any $\boldsymbol{x} \in \mathbb{R}^p$,

$$\begin{aligned}
|\tilde{f}_{\mathbf{\Theta}}(\boldsymbol{x}) - \tilde{f}_{\mathbf{\Theta}'}(\boldsymbol{x})| =\; & k^{-1/s}\sum_{j=1}^{k}|d_\phi(\boldsymbol{x}, \boldsymbol{\theta}_j) - d_\phi(\boldsymbol{x}, \boldsymbol{\theta}'_j)| && (16)\\
\leq\; & k^{-1/s}\sum_{j=1}^{k}\Big[|\phi(\boldsymbol{\theta}'_j) - \phi(\boldsymbol{\theta}_j)| + \|\nabla\phi(\boldsymbol{\theta}'_j) - \nabla\phi(\boldsymbol{\theta}_j)\|_2(\|\boldsymbol{x}\|_2 + \xi_P + \|\mathbf{\Theta}_*\|_F)\\
& \quad + \|\nabla\phi(\boldsymbol{\theta}_j)\|_2\|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j\|\Big]\\
\leq\; & k^{-1/s}\sum_{j=1}^{k}\Big[\tau_2(\xi_P + \|\mathbf{\Theta}_*\|_F) + \tau_2(\|\boldsymbol{x}\|_2 + \xi_P + \|\mathbf{\Theta}_*\|_F) + \tau_2(\xi_P + \|\mathbf{\Theta}_*\|_F)\Big]\|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j\|_2 && (17)\\
\leq\; & k^{-1/s}\tau_2(3\xi_P + 3\|\mathbf{\Theta}_*\|_F + \|\boldsymbol{x}\|_2)\sum_{j=1}^{k}\|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j\|_2.
\end{aligned}$$

Here, equation (16) follows from (Beliakov et al., 2010) and inequality (17) follows from Lemmas B.1 and B.2. $\square$

Proof of Theorem 3.4

Proof. We begin by noting that if $|\theta_{ij} - \theta'_{ij}| \leq \rho$, then for all $i, j \in [n]$,

$$\begin{aligned} d_{2n}^2(\tilde{f}_{\mathbf{\Theta}}, \tilde{f}_{\mathbf{\Theta}'}) &= \frac{1}{n}\sum_{i=1}^{n}(\tilde{f}_{\mathbf{\Theta}}(\boldsymbol{X}_i) - \tilde{f}_{\mathbf{\Theta}'}(\boldsymbol{X}_i))^2 \\ &\leq k^{-2/s}\tau_2^2(\sum_{j=1}^{k}\|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j\|_2)^2\frac{1}{n}\sum_{i=1}^{n}(3\xi_P + 3\|\mathbf{\Theta}_*\|_F + \|\boldsymbol{X}_i\|_2)^2 \qquad (18) \\ &\leq 2k^{-2/s}\tau_2^2(\sum_{j=1}^{k}\|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j\|_2)^2\frac{1}{n}\sum_{i=1}^{n}(9(\xi_P + \|\mathbf{\Theta}_*\|_F)^2 + \|\boldsymbol{X}_i\|_2^2) \\ &\leq 2k^{2-2/s}\tau_2^2 p\rho^2\frac{1}{n}\sum_{i=1}^{n}(18\xi_P^2 + 18\|\mathbf{\Theta}_*\|_F^2 + \|\boldsymbol{X}_i\|_2^2). \qquad (19) \end{aligned}$$

Above, equation (18) follows from Lemma B.3. We take $\rho = \delta^{-1}\big(2k^{2-2/s}\tau_2^2 pn^{-1}\sum_{i=1}^{n}(18\xi_P^2 + 18\|\mathbf{\Theta}_*\|_F^2 + \|\boldsymbol{X}_i\|_2^2)\big)^{-1/2}$, which simplifies the RHS of (19) to equal δ^2. Next, let

$$\mathfrak{N}_\rho = \begin{cases} \{-(\xi_P + \|\mathbf{\Theta}_*\|_F) + \frac{i}{2\rho} : 1 \leq i \leq \lfloor\frac{\xi_P + \|\mathbf{\Theta}_*\|_F}{\rho}\rfloor\} & \text{if } \rho < \xi_P + \|\mathbf{\Theta}_*\|_F \\ \{\mathbf{0}\} & \text{Otherwise.} \end{cases}$$

We consider a $(k\rho\sqrt{p})$-net of $B(\xi_P + \|\mathbf{\Theta}_*\|_F)$, $\mathfrak{M}_\rho = \{\mathbf{\Theta} \in \mathbb{R}^{k\times p} : \theta_{ij} \in \mathfrak{N}_\rho\}$. From (19), $\mathfrak{F}_\rho = \{\tilde{f}_{\mathbf{\Theta}} : \mathbf{\Theta} \in \mathfrak{M}_\rho\}$ constitutes a δ-cover of $\mathcal{F}$. Thus,

$$\mathcal{N}(\delta; \mathcal{F}, d_{2n}) \leq |\mathfrak{F}_\rho| \leq |\mathfrak{M}_\rho| \leq \left(\max\left\{1, \left\lfloor\frac{\xi_P + \|\mathbf{\Theta}_*\|_F}{\rho}\right\rfloor\right\}\right)^{kp}.$$

Plugging in the value of ρ gives us the desired result. ☐

Proof of Theorem 3.5

Proof. It is easy to show using lemma B.3 that

$$\begin{aligned} \mathrm{diam}(\mathcal{F}) &= \sup_{f,g\in\mathcal{F}} d_{2n}(f,g) \\ &= \sup_{\mathbf{\Theta}\subset(\xi_P + \|\mathbf{\Theta}_*\|_F)} \sqrt{2k^{-2/s}\tau_2^2\Big[\sum_{j=1}^{k}\|\boldsymbol{\theta}_j - \boldsymbol{\theta}'_j\|_2\Big]^2\frac{1}{n}\sum_{i=1}^{n}9(\xi_P + \|\mathbf{\Theta}_*\|_F)^2 + \|\boldsymbol{X}_i\|_2^2} \\ &\leq 2C^{1/2}(\xi_P + \|\mathbf{\Theta}_*\|_F). \end{aligned}$$

Now applying Dudley's chaining (Van Der Vaart et al., 1996),

$$\begin{aligned} \widehat{\mathcal{R}}_n(\mathcal{F}) &\leq \frac{12}{\sqrt{n}}\int_0^{\mathrm{diam}(\mathcal{F})}\sqrt{\log\mathcal{N}(\delta;\mathcal{F},d_{2n})}d\delta \\ &\leq \frac{12\sqrt{kp}}{\sqrt{n}}\int_0^{2C^{1/2}(\xi_P + \|\mathbf{\Theta}_*\|_F)}\sqrt{\log\left(\max\left\{1, \left\lfloor\frac{M_\epsilon C^{1/2}}{\delta}\right\rfloor\right\}\right)}d\delta \\ &= \frac{12\sqrt{kp}}{\sqrt{n}}C^{1/2}(\xi_P + \|\mathbf{\Theta}_*\|_F)\Gamma(3/2) \\ &= 6\sqrt{\pi C}M_\epsilon\sqrt{\frac{kp}{n}} \\ &= 6\tau_2\sqrt{2\pi\sum_{i=1}^{n}\left(18\xi_P^2 + 18\|\mathbf{\Theta}_*\|_F^2 + n^{-1}\|\boldsymbol{X}_i\|_2^2\right)}\,(\xi_P + \|\mathbf{\Theta}_*\|_F)\frac{k^{3/2-1/s}p}{\sqrt{n}} \end{aligned}$$

Finally,

$$
\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= \mathbb{E}\widehat{\mathcal{R}}_n(\mathcal{F}) \\
&\leq 6\tau_2 \mathbb{E}\sqrt{2\pi \sum_{i=1}^{n} \big(18\xi_P^2 + 18\|\boldsymbol{\Theta}_*\|_F^2 + n^{-1}\|\boldsymbol{X}_i\|_2^2\big)(\xi_P + \|\boldsymbol{\Theta}_*\|_F)} \frac{k^{3/2-1/s}p}{\sqrt{n}} \\
&\leq 6\tau_2 \sqrt{2\pi \sum_{i=1}^{n} \big(18\xi_P^2 + 18\|\boldsymbol{\Theta}_*\|_F^2 + \mathbb{E}\|\boldsymbol{X}\|_2^2\big)(\xi_P + \|\boldsymbol{\Theta}_*\|_F)} \frac{k^{3/2-1/s}p}{\sqrt{n}}.
\end{aligned}
$$

□

Proof of Theorem 3.6

Proof. Let $g((\boldsymbol{X}_i)_{i\in[n]}) = \sup_{\boldsymbol{\Theta}\subset B(\xi_P+\|\boldsymbol{\Theta}_*\|_F)} P_n\tilde{f}_{\boldsymbol{\Theta}} - P\tilde{f}_{\boldsymbol{\Theta}}$. It follows that

$$
g((\boldsymbol{X}_i)_{i\in[n]}) = \big[g((\boldsymbol{X}_i)_{i\in[n]}) - \mathbb{E}g((\boldsymbol{X}_i)_{i\in[n]})\big] + \mathbb{E}g((\boldsymbol{X}_i)_{i\in[n]}).
$$

Following Maurer & Pontil (2021), we bound the first term through concentration inequalities and the second term through symmetrization. It is easy to see that

$$
\mathbb{E}g((\boldsymbol{X}_i)_{i\in[n]}) \leq 2\mathcal{R}_n(\mathcal{F}) \tag{20}
$$

Next, consider

$$
\mathcal{B} = \{h : \mathcal{F} \to \mathbb{R} : \sup_{f\in\mathcal{F}} |h(f)| < \infty\}.
$$

Clearly, $\mathcal{B}$ is a normed vector space with $\|h\|_{\mathcal{B}} = \sup_{f\in\mathcal{F}} |h(f)|$. For any $\boldsymbol{X}_i$, we define a corresponding $Y_i \in \mathcal{B}$ by

$$
Y_i(f) = \frac{1}{n}\left(f(\boldsymbol{X}_i) - \mathbb{E}f(\boldsymbol{X}_i)\right) \quad f \in \mathcal{F}.
$$

Therefore, $\mathbb{E}[Y_i] \equiv 0$, and $g\left((\boldsymbol{X}_i)_{i\in[n]}\right) = \|\sum_{i=1}^{n} Y_i\|_{\mathcal{B}}$. We now note that

$$
\begin{aligned}
&\|\|Y_i\|_{\mathcal{B}}\|_{\psi_1} \\
&= \frac{1}{n}\|\sup_{f\in\mathcal{F}} \left(f(\boldsymbol{X}_i) - \mathbb{E}f(\boldsymbol{X}_i)\right)\|_{\psi_1} \\
&= \frac{1}{n}\|\sup_{f\in\mathcal{F}} \mathbb{E}\left(f(\boldsymbol{X}_i) - \mathbb{E}f(\boldsymbol{X}_i')|(\boldsymbol{X}_i)_{i\in[n]}\right)\|_{\psi_1} \\
&= \frac{1}{n}\|\sup_{\boldsymbol{\Theta}\subset B(\xi_P+\|\boldsymbol{\Theta}_*\|_F)} \mathbb{E}\left(\tilde{f}_{\boldsymbol{\Theta}}(\boldsymbol{X}_i) - \tilde{f}_{\boldsymbol{\Theta}}(\boldsymbol{X}_i')|(\boldsymbol{X}_i)_{i\in[n]}\right)\|_{\psi_1} \\
&\leq \frac{k^{-1/s}}{n}\|\sup_{\boldsymbol{\Theta}\subset B(\xi_P+\|\boldsymbol{\Theta}_*\|_F)} \sum_{j=1}^{k} \mathbb{E}\left(d_\phi(\boldsymbol{X}_i, \boldsymbol{\theta}_j) - d_\phi(\boldsymbol{X}_i', \boldsymbol{\theta}_j)|(\boldsymbol{X}_i)_{i\in[n]}\right)\|_{\psi_1} \\
&= \frac{k^{-1/s}}{n}\|\sup_{\boldsymbol{\Theta}\subset B(\xi_P+\|\boldsymbol{\Theta}_*\|_F)} \sum_{j=1}^{k} \mathbb{E}\big((\phi(\boldsymbol{X}_i) - \phi(\boldsymbol{X}_i')) - \langle\nabla\phi(\boldsymbol{\theta}_j), \boldsymbol{X}_i - \boldsymbol{X}_i'\rangle|(\boldsymbol{X}_i)_{i\in[n]}\big)\|_{\psi_1} \\
&\leq \frac{k^{1-1/s}}{n}\|\mathbb{E}\left(|(\phi(\boldsymbol{X}_i) - \phi(\boldsymbol{X}_i')|\big|(\boldsymbol{X}_i)_{i\in[n]}\right)\|_{\psi_1} + \frac{k^{-1/s}}{n}\|\sup_{\boldsymbol{\Theta}\subset B(\xi_P+\|\boldsymbol{\Theta}_*\|_F)} \sum_{j=1}^{k} |\langle\nabla\phi(\boldsymbol{\theta}_j), \boldsymbol{X}_i - \boldsymbol{X}_i'\rangle|\big|(\boldsymbol{X}_i)_{i\in[n]}\big)\|_{\psi_1} \\
&\leq \frac{k^{1-1/s}}{n}\|(\phi(\boldsymbol{X}_i) - \phi(\boldsymbol{X}_i')\|_{\psi_1} + \frac{k^{-1/s}}{n}\|\sup_{\boldsymbol{\Theta}\subset B(\xi_P+\|\boldsymbol{\Theta}_*\|_F)} \mathbb{E}\sum_{j=1}^{k} \|\nabla\phi(\boldsymbol{\theta}_j)\|_2\|\boldsymbol{X}_i - \boldsymbol{X}_i'\|_2\big|(\boldsymbol{X}_i)_{i\in[n]}\big)\|_{\psi_1} \qquad (21) \\
&\leq \frac{2k^{1-1/s}}{n}\|\phi(\boldsymbol{X})\|_{\psi_1} + \frac{k^{1-1/s}\tau_2(\xi_P + \|\boldsymbol{\Theta}_*\|_F)}{n}\|\sup_{\boldsymbol{\Theta}\subset B(\xi_P+\|\boldsymbol{\Theta}_*\|_F)} \mathbb{E}\|\boldsymbol{X}_i - \boldsymbol{X}_i'\|_2\big|(\boldsymbol{X}_i)_{i\in[n]}\big)\|_{\psi_1}
\end{aligned}
$$

$$\leq \frac{2k^{1-1/s}}{n}\|\phi(\boldsymbol{X})\|_{\psi_1} + \frac{2k^{1-1/s}\tau_2(\xi_P + \|\boldsymbol{\Theta}_*\|_F)}{n}\|\|\boldsymbol{X}\|_2\|_{\psi_1}$$
$$= \frac{2k^{1-1/s}}{n}(\|\phi(\boldsymbol{X})\|_{\psi_1} + \tau_2(\xi_P + \|\boldsymbol{\Theta}_*\|_F)\|\|\boldsymbol{X}\|_2\|_{\psi_1}).$$

Inequality 21 follows from Lemma 6 of (Maurer & Pontil, 2021). Thus, by Proposition $7(ii)$ of (Maurer & Pontil, 2021),

$$\begin{aligned}
g((\boldsymbol{X}_i)_{i\in[n]}) - \mathbb{E}g((\boldsymbol{X}_i)_{i\in[n]}) \;&\leq\; 8e\|g((\boldsymbol{X}_i)_{i\in[n]})\|_{\psi_1}\sqrt{\frac{2\log(1/\delta)}{n}} \\
&\leq\; 8e\|\|\sum_{i=1}^{n} Y_i\|_{\mathcal{B}}\|_{\psi_1}\sqrt{\frac{2\log(1/\delta)}{n}} \\
&\leq\; 8e\sum_{i=1}^{n}\|\|Y_i\|_{\mathcal{B}}\|_{\psi_1}\sqrt{\frac{2\log(1/\delta)}{n}} \\
&\leq\; 16ek^{1-1/s}(\|\phi(\boldsymbol{X})\|_{\psi_1} + \tau_2(\xi_P + \|\boldsymbol{\Theta}_*\|_F)\|\|\boldsymbol{X}\|_2\|_{\psi_1})\sqrt{\frac{2\log(1/\delta)}{n}}.
\end{aligned} \tag{22}$$

Recalling the definition

$$g((\boldsymbol{X}_i)_{i\in[n]}) = \sup_{\boldsymbol{\Theta}\subset B(\xi_P+\|\boldsymbol{\Theta}_*\|_F)} P_n\tilde{f}_{\boldsymbol{\Theta}} - P\tilde{f}_{\boldsymbol{\Theta}},$$

with probability at least $1-\delta$,

$$g((\boldsymbol{X}_i)_{i\in[n]}) \leq 2\mathcal{R}_n(\mathcal{F}) + 16ek^{1-1/s}(\|\phi(\boldsymbol{X})\|_{\psi_1} + \tau_2(\xi_P + \|\boldsymbol{\Theta}_*\|_F)\|\|\boldsymbol{X}\|_2\|_{\psi_1})\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Similarly, we can show that with probability at least $1-\delta$,

$$\sup_{\boldsymbol{\Theta}\subset B(\xi_P+\|\boldsymbol{\Theta}_*\|_F)} P\tilde{f}_{\boldsymbol{\Theta}} - P_n\tilde{f}_{\boldsymbol{\Theta}} \leq 2\mathcal{R}_n(\mathcal{F}) + 16ek^{1-1/s}(\|\phi(\boldsymbol{X})\|_{\psi_1} + \tau_2(\xi_P + \|\boldsymbol{\Theta}_*\|_F)\|\|\boldsymbol{X}\|_2\|_{\psi_1})\sqrt{\frac{2\log(1/\delta)}{n}}.$$

Combining the above and replacing δ by $\delta/2$, we get that with probability at least $1-\delta$,

$$\sup_{\boldsymbol{\Theta}\subset B(\xi_P+\|\boldsymbol{\Theta}_*\|_F)} |P_n\tilde{f}_{\boldsymbol{\Theta}} - P\tilde{f}_{\boldsymbol{\Theta}}| \leq \mathcal{R}_n(\mathcal{F}) + 16ek^{1-1/s}(\|\phi(\boldsymbol{X})\|_{\psi_1} + \tau_2(\xi_P + \|\boldsymbol{\Theta}_*\|_F)\|\|\boldsymbol{X}\|_2\|_{\psi_1})\sqrt{\frac{2\log(2/\delta)}{n}}.$$

Now by bounding $\mathcal{R}_n(F)$ using Theorem 3.5, we obtain our desired result. □

C. Proof of Theorem 3.7

Proof. We note the following:

$$\begin{aligned}
\mathfrak{R}(\widehat{\boldsymbol{\Theta}}_n) &= P\tilde{f}_{\widehat{\boldsymbol{\Theta}}_n} - P\tilde{f}_{\boldsymbol{\Theta}_*} \\
&\leq P\tilde{f}_{\widehat{\boldsymbol{\Theta}}_n} - P_n\tilde{f}_{\widehat{\boldsymbol{\Theta}}_n} + P_n\tilde{f}_{\widehat{\boldsymbol{\Theta}}_n} - P\tilde{f}_{\boldsymbol{\Theta}_*} \\
&\leq P\tilde{f}_{\widehat{\boldsymbol{\Theta}}_n} - P_n\tilde{f}_{\widehat{\boldsymbol{\Theta}}_n} + P_n\tilde{f}_{\boldsymbol{\Theta}_*} - P\tilde{f}_{\boldsymbol{\Theta}_*} \\
&\leq 2\sup_{\widehat{\boldsymbol{\Theta}}_n\subset B(\xi_P+\|\boldsymbol{\Theta}_*\|_F)} |P_n\tilde{f}_{\boldsymbol{\Theta}} - Pf_{\boldsymbol{\Theta}}| && (23) \\
&\leq 24\tau_2\sqrt{2\pi\sum_{i=1}^{n}\left(18\xi_P^2 + 18\|\boldsymbol{\Theta}_*\|_F^2 + \mathbb{E}\|\boldsymbol{X}\|_2^2\right)(\xi_P + \|\boldsymbol{\Theta}_*\|_F)}\frac{k^{3/2-1/s}p}{\sqrt{n}} \\
&\quad + 32ek^{1-1/s}\left[\sigma_\phi + \sigma\tau_2(\xi_P + \|\boldsymbol{\Theta}_*\|_F)\|\right]\sqrt{\frac{2\log(2/\delta)}{n}}. && (24)
\end{aligned}$$

Here, (23) holds with probability at least $1-e^{-cn}$ by Theorem 3.1 and (24) holds with probability at least $1-\delta$ by a simple application of Theorem 3.6. □

Proof of Theorem 3.8

Proof. For simplicity of notations, let $C_1 = \max\left\{24\tau_2 C'(\xi_P + \|\mathbf{\Theta}_*\|_F)k^{3/2-1/s}p, 32e\sigma\tau_2 k^{1-1/s}(1+\xi_P+\|\mathbf{\Theta}_*\|_F)\right\}$. From Theorem 3.7, we know that with probability at least $1-\delta-e^{-cn}$,

$$\mathfrak{R}(\widehat{\mathbf{\Theta}}_n) \le \frac{C_1}{\sqrt{n}} + C_1\sqrt{\frac{\log(2/\delta)}{n}}. \tag{25}$$

Now, fix $\epsilon > 0$: if $n \ge \max\left\{\frac{4C_1^2}{\epsilon^2}, \frac{\epsilon^4}{4C_1^4}\right\}$ and $\delta = 2\exp\left(-\frac{\sqrt{n}\epsilon^2}{2C^2}\right)$, the RHS of (25) becomes no bigger than ϵ. Thus,

$$\mathbb{P}\left(|P\tilde{f}_{\widehat{\mathbf{\Theta}}_n} - P\tilde{f}_{\mathbf{\Theta}^*}| > \epsilon\right) \le 2\exp\left(-\frac{\sqrt{n}\epsilon^2}{2C^2}\right), \quad \forall n \ge 4C^2/\epsilon^2.$$

Since the series $\sum_{n=1}^{\infty}\exp\left(-\frac{\sqrt{n}\epsilon^2}{2C^2}\right)$ is convergent from the above equation, so is $\mathbb{P}\left(|P\tilde{f}_{\widehat{\mathbf{\Theta}}_n} - P\tilde{f}_{\mathbf{\Theta}^*}| > \epsilon\right)$. Hence, $P\tilde{f}_{\widehat{\mathbf{\Theta}}_n} \xrightarrow{a.s.} P\tilde{f}_{\mathbf{\Theta}^*}$. Thus, for any $\epsilon > 0$, it follows that $P\tilde{f}_{\widehat{\mathbf{\Theta}}_n} \le p\tilde{f}_{\mathbf{\Theta}^*} + \epsilon$ almost surely with respect to $[P]$ for n sufficiently large. From assumption A 3, for any fixed $\eta > 0$ and n large, we also know $\text{dist}(\widehat{\mathbf{\Theta}}_n, \mathbf{\Theta}^*) \le \eta$ almost surely with respect to $[P]$. Together, $\text{dist}(\widehat{\mathbf{\Theta}}_n, \mathbf{\Theta}^*) \xrightarrow{a.s.} 0$, which proves the result. □

D. Assumption of sub-Gaussianity of $\|\boldsymbol{X}\|_2$

In this section, we show that a concrete sufficient condition for A 1 to be satisfied is that $\|\boldsymbol{X}\|_2$ is sub-Gaussian. We state and prove this result in the following theorem.

Theorem D.1. *If $\|\boldsymbol{X}\|_2$ is sub-Gaussian then under assumption A 2,*

(a) $\|\|\boldsymbol{X}\|_2\|_{\psi_1} < \infty$.

(b) $\|\phi(\boldsymbol{X})\|_{\psi_1} < \infty$.

Proof. *(a)* We know that $\|\|\boldsymbol{X}\|_2\|_{\psi_2} \triangleq \sup_{p\in\mathbb{N}} \frac{(\mathbb{E}\|\boldsymbol{X}\|_2^p)^{1/p}}{\sqrt{p}} < \infty$ from Proposition 2.5.2 of (Vershynin, 2018). Thus,

$$\|\|\boldsymbol{X}\|_2\|_{\psi_1} = \sup_{p\in\mathbb{N}} \frac{(\mathbb{E}\|\boldsymbol{X}\|_2^p)^{1/p}}{p} \le \sup_{p\in\mathbb{N}} \frac{(\mathbb{E}\|\boldsymbol{X}\|_2^p)^{1/p}}{\sqrt{p}} < \infty.$$

(b) We note that

$$\begin{aligned}
\|\phi(\boldsymbol{X})\|_{\psi_1} &= \sup_{p\in\mathbb{N}} \frac{(\mathbb{E}|\phi(\boldsymbol{X})|^p)^{1/p}}{p} \\
&= \sup_{p\in\mathbb{N}} \frac{(\mathbb{E}|d_\phi(\boldsymbol{X},\mathbf{0})|^p)^{1/p}}{p} \\
&\le \sup_{p\in\mathbb{N}} \frac{(\mathbb{E}\tau_2\|\boldsymbol{X}\|_2^{2p})^{1/p}}{p} \\
&= \tau_2 \sup_{p\in\mathbb{N}} \frac{(\mathbb{E}\|\boldsymbol{X}\|_2^{2p})^{1/p}}{p} \\
&= \tau_2 \sup_{p\in\mathbb{N}} \left(\frac{(\mathbb{E}\|\boldsymbol{X}\|_2^{2p})^{1/2p}}{\sqrt{p}}\right)^2 \\
&= 2\tau_2 \left(\sup_{p\in\mathbb{N}} \frac{(\mathbb{E}\|\boldsymbol{X}\|_2^{2p})^{1/2p}}{\sqrt{2p}}\right)^2 \\
&\le 2\tau_2 \|\|\boldsymbol{X}\|_2\|_{\psi_2}^2 < \infty.
\end{aligned} \tag{26}$$

Inequality (26) follows from Lemma B.1 appearing in the Supplement of Telgarsky & Dasgupta (2013). □