Agglomerative Bregman Clustering

Matus Telgarsky MTELGARS@CS.UCSD.EDU
Sanjoy Dasgupta DASGUPTA@CS.UCSD.EDU
Department of Computer Science and Engineering, UCSD, 9500 Gilman Drive, La Jolla, CA 92093-0404

Abstract

This manuscript develops the theory of agglomerative clustering with Bregman divergences. Geometric smoothing techniques are developed to deal with degenerate clusters. To allow for cluster models based on exponential families with overcomplete representations, Bregman divergences are developed for nondifferentiable convex functions.

1. Introduction

Starting with points $\{x_i\}_{i=1}^m$ and a pairwise merge cost $\Delta(\cdot,\cdot)$, classical agglomerative clustering produces a single hierarchical tree as follows (Duda et al., 2001).

1. Start with m clusters: $C_i := \{x_i\}$ for each i.
2. While at least two clusters remain:
   (a) Choose $\{C_i, C_j\}$ with minimal $\Delta(C_i, C_j)$.
   (b) Remove $\{C_i, C_j\}$, add in $C_i \cup C_j$.

In order to build a hierarchy with low k-means cost, one can use the merge cost due to Ward (1963),

$$\Delta_w(C_i, C_j) := \frac{|C_i||C_j|}{|C_i| + |C_j|} \|\tau(C_i) - \tau(C_j)\|_2^2,$$

where $\tau(C)$ denotes the mean of cluster C.

The k-means cost, and thus the Ward merge rule, inherently prefer spherical clusters of common radius. To accommodate other cluster shapes and input domains, the squared Euclidean norm may be replaced with a relaxation sharing many of the same properties, a *Bregman divergence*.

This manuscript develops the theory of agglomerative clustering with Bregman divergences.

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

1.1. Bregman clustering

There is already a rich theory of clustering with Bregman divergences, and in particular the relationship of these divergences with exponential family distributions (Banerjee et al., 2005). The standard development has two shortcomings, the first of which is amplified in the agglomerative setting.

Degenerate divergences. Many divergences lead to merge costs which are undefined on certain inputs. This scenario is exacerbated with small clusters; for instance, with Gaussian clusters, the corresponding divergence rule is the KL divergence, which demands full rank cluster covariances. This is impossible with $\leq d$ points, but the agglomerative procedure above starts with singletons.

Minimal representations. The standard theory of exponential families and its connections to Bregman divergences depend on *minimal representations*: there is just one way to write down any particular distribution. On the other hand, the natural encoding for many problems — e.g., Ising models, and many other examples listed in the textbook of Wainwright & Jordan (2008, Section 4) — is *overcomplete*, necessitating potentially tedious conversions to invoke the theory.

1.2. Contribution

The approach of this manuscript is to carefully build a theory of Bregman divergences constructed from convex, yet nondifferentiable functions. Section 2 will present the basic definition, and verify this generalization satisfies the usual Bregman divergence properties.

Section 3 will revisit the standard Bregman hard clustering model (Banerjee et al., 2005), and show how it naturally leads to a merge cost Δ. Section 4 then constructs exponential families, demonstrating that nondifferentiable Bregman divergences, while permitting representations which are not minimal, still satisfy all the usual properties. To overcome the aforementioned small-sample cases where divergences may not be well-defined, Section 5 presents smoothing procedures which immediately follow from the preceding technical development.

To close, Section 6 presents the final algorithm, and Section 7 provides experimental validation both by measuring cluster fit, and the suitability of cluster features in supervised learning tasks.

The various appendices contain all proofs, as well as some additional technical material and examples.

1.3. Related work

A number of works present agglomerative schemes for clustering with exponential families, from the perspective of KL divergences between distributions, or the analogous goal of maximizing model likelihood, or lastly in connection to the information bottleneck method (Iwayama & Tokunaga, 1995; Fraley, 1998; Heller & Ghahramani, 2005; Garcia et al., 2010; Blundell et al., 2010; Slonim & Tishby, 1999). Furthermore, Merugu (2006) studied the same algorithm as the present work, phrased in terms of Bregman divergences. These preceding methods either do not explicitly mention divergence degeneracies, or circumvent them with Bayesian techniques, a connection discussed in Section 5.

Bregman divergences for nondifferentiable functions have been studied in a number of places. Remark 2.4 shows the relationship between the presented version, and one due to Gordon (1999). Furthermore, Kiwiel (1995) presents divergences almost identical to those here, but the manuscripts and focuses differ thereafter.

The development here of exponential families and related Bregman properties generalizes results found in a variety of sources (Brown, 1986; Azoury & Warmuth, 2001; Banerjee et al., 2005; Wainwright & Jordan, 2008); in order to cleanly document the new contribution without cluttering the text, Appendix G is devoted to bibliographic notes. Finally, parallel to the completion of this manuscript, another group has developed exponential families under similarly relaxed conditions, but from the perspective of maximum entropy and convex duality (Csiszár & Matúš, 2012).

1.4. Notation

The following concepts from convex analysis are used throughout the text; the interested reader is directed to the seminal text of Rockafellar (1970). A set is convex when the line segment between any two of its elements is again within the set. The epigraph of a function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, where $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, is the set of points bounded below by f; i.e., the set $\{(x, r) : x \in \mathbb{R}^n, r \geq f(x)\} \subseteq \mathbb{R}^n \times \bar{\mathbb{R}}$. A function is convex when its epigraph is convex, and closed when its epigraph is closed. The domain $\mathrm{dom}(f)$ of a function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$ is the subset of inputs not mapping to $+\infty$: $\mathrm{dom}(f) = \{x \in \mathbb{R}^n : f(x) < \infty\}$. A function is proper if $\mathrm{dom}(f)$ is nonempty, and f never takes on the value $-\infty$. The Bregman divergences in this manuscript will be generated from closed proper convex functions.

The conjugate of a function f is the function $f^*(\phi) := \sup_x \langle \phi, x\rangle - f(x)$; when f is closed proper convex, so is f^*, and moreover $f^{**} = f$. A subgradient g to a function f at $y \in \mathrm{dom}(f)$ provides an affine lower bound: for any $x \in \mathbb{R}^n$, $f(x) \geq f(y) + \langle g, x - y\rangle$. The set of all subgradients at a point y is denoted by $\partial f(y)$ (which is easily empty). The directional derivative $f'(y; d)$ of a function f at y in direction d is $\lim_{t\downarrow 0}(f(y + td) - f(y))/t$.

The affine hull of a set $S \subseteq \mathbb{R}^n$ is the smallest affine set containing it. If S is translated and rotated so that its affine hull is some $\mathbb{R}^d \subseteq \mathbb{R}^n$, then its interior within $\mathbb{R}^d$ is its relative interior within $\mathbb{R}^n$. Said another way, the relative interior $\mathrm{ri}(S)$ is the interior of S with respect to the $\mathbb{R}^n$ topology relativized to the affine hull of S. Although functions in this manuscript will generally be closed, their domains are often (relatively) open.

Convex functions will be defined over $\mathbb{R}^n$, but it will be useful to treat data as lying in an abstract space $\mathcal{X}$, and a statistic map $\tau : \mathcal{X} \to \mathbb{R}^n$ will embed examples in the desired Euclidean space. This map, which will also be overloaded to handle finite subsets of $\mathcal{X}$, will eventually incorporate the smoothing procedure.

The cluster cost will be denoted by ϕ, or $\phi_{f,\tau}$ to make the underlying convex function and statistic map clear; similarly, the merge cost is denoted by Δ and $\Delta_{f,\tau}$.

2. Bregman divergences

Given a convex function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, the Bregman divergence $\mathsf{B}_f(\cdot, y)$ is the gap between f and its linearization at y. Typically, f is differentiable, and so $\mathsf{B}_f(x, y) = f(x) - f(y) - \langle \nabla f(y), x - y\rangle$.

Definition 2.1. Given a convex function $f : \mathbb{R}^n \to \bar{\mathbb{R}}$, the corresponding Bregman divergence between $x, y \in \mathrm{dom}(f)$ is

$$\mathsf{B}_f(x, y) := f(x) - f(y) + f'(y; y - x). \qquad \diamond$$

Unlike gradients, directional derivatives are well-defined whenever a convex function is finite, although they can be infinite on the relative boundary of $\mathrm{dom}(f)$

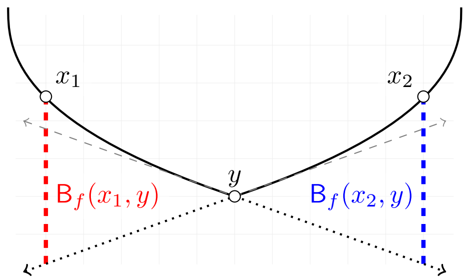


Figure 1. Bregman divergences with respect to a reference point y at which f is nondifferentiable. The thick (red or blue) dashed lines denote the divergence values themselves; they travel down from f to the negated sublinear function $x \mapsto f(y) - f'(y; y - x)$, here a pair of dotted rays. Noting Proposition 2.3 and fixing some x_i, the subgradient at y farthest from x_i is one of these dotted rays together with its dashed, gray extension. The dashed extensions, taken together, represent the sublinear function $x \mapsto f(y) + f'(y; x - y)$.

(Rockafellar, 1970, Theorems 23.1, 23.3, 23.4).

Noting that $f'(y; y - x) \geq -f'(y; x - y)$ (Rockafellar, 1970, Theorem 23.1), it may seem closer to the original expression to instead use $f(x) - f(y) - f'(y; x - y)$ (which is thus bounded above by $\mathsf{B}_f(x, y)$); please see Figure 1 for a depiction of both choices. However, $\mathsf{B}_f(\cdot, y)$ is a sum of two convex functions, and thus immediately convex; one advantage of convexity is in the computation of *Bregman projections*, which are discussed in Appendix A.3.

Example 2.2. In the case of the differentiable convex function $f_2 = \|\cdot\|_2^2$, $\mathsf{B}_{f_2}(x, y) = \|x - y\|_2^2$ follows by noting $f_2'(y; y - x) = \langle 2y, y - x\rangle$. To analyze the case of $f_1 = \|\cdot\|_1$, first consider the univariate case $h = |\cdot|$. Either by drawing a picture or checking $h'(\cdot; \cdot)$ from definition, it follows that

$$\mathsf{B}_h(x, y) := \begin{cases} 0 & \text{when } xy > 0, \\ 2|x| & \text{otherwise.} \end{cases}$$

Then noting that $f_1'(\cdot; \cdot)$ decomposes coordinate-wise, it follows that $\mathsf{B}_{f_1}(x, y) = \sum_i \mathsf{B}_h(x_i, y_i)$. Said another way, B_{f_1} is twice the l^1 distance from x to the farthest orthant containing y, which bears a resemblance to the hinge loss. ◊

B_f can also be written in terms of subgradients.

Proposition 2.3. *Let a proper convex f and $y \in \operatorname{ri}(\operatorname{dom}(f))$ be given. Then for any $x \in \operatorname{dom}(f)$:*

- *$f'(y; y - x)$ and $\mathsf{B}_f(x, y)$ are finite, and*
- *$\mathsf{B}_f(x, y) := \max_{g \in \partial f(y)} f(x) - f(y) - \langle g, x - y\rangle$.*

Remark 2.4. Given $x \in \operatorname{dom}(f)$ and a dual element $g \in \mathbb{R}^n$, another nondifferentiable generalization of Bregman divergence, due to Gordon (1999), is

$$\mathsf{D}_f(x, g) := f(x) + f^*(g) - \langle g, x\rangle .$$

Now suppose there exists $y \in \operatorname{ri}(\operatorname{dom}(f))$ with $g \in \partial f(y)$; the Fenchel-Young inequality (Rockafellar, 1970, Theorem 23.5) grants $\mathsf{D}_f(x, g) = f(x) - f(y) - \langle g, x - y\rangle$. Thus, by Proposition 2.3,

$$\mathsf{B}_f(x, y) := \max\{\mathsf{D}_f(x, g) : g \in \partial f(y)\}. \qquad ◊$$

To sanity check B_f, Appendix A states and proves a number of key Bregman divergence properties, generalized to the case of nondifferentiability. The following list summarizes these properties; in general, f is closed proper convex, $y \in \operatorname{ri}(\operatorname{dom}(f))$, and $x \in \operatorname{dom}(f)$.

- $\mathsf{B}_f(\cdot, y)$ is convex, proper, nonnegative, and $\mathsf{B}_f(y, y) = 0$.
- When f is strictly convex, $\mathsf{B}_f(x, y) = 0$ iff $x = y$.
- Given $g_x \in \operatorname{ri}(\operatorname{dom}(f^*))$, $\sup_{x \in \partial f^*(g_x)} \mathsf{B}_f(x, y) = \sup_{g_y \in \partial f(y)} \mathsf{B}_{f^*}(g_y, g_x)$.
- Under some regularity conditions on f, a generalization of the Pythagorean theorem holds, with B_f replacing squared Euclidean distance.

Over and over, this section depends on relative interiors. What's so bad about the relative boundary? The directional derivatives and subgradients break down. If $y \in \operatorname{relbd}(\operatorname{dom}(f))$ and $x \in \operatorname{ri}(\operatorname{dom}(f))$, then $f'(y; y - x) = \infty = \mathsf{B}_f(x, y)$, and there exists no maximizing subgradient as in Proposition 2.3; in fact, one can not in general guarantee the existence of any subgradients at all.

In just a moment, the cluster model will be developed, where it will be very easy for the second argument argument of B_f to lie on $\operatorname{relbd}(\operatorname{dom}(f))$, rendering the divergences infinite and cluster costs meaningless. Worse, it is frequently the case $\operatorname{dom}(f)$ is relatively open, meaning the relative boundary is not in $\operatorname{dom}(f)$! The smoothing methods of Section 5 work around these issues. Their approach is simple enough: they just push relative boundary points inside the relative interior.

3. Cluster model

Let a finite collection C of points $\{x_i\}_{i=1}^m$ in some abstract space $\mathcal{X}$ — say, documents or vectors — be given. In order to cluster these with Bregman divergences, the first step is to map them into $\mathbb{R}^n$.

Definition 3.1. A statistic map τ is any function from $\mathcal{X}$ to $\mathbb{R}^n$. Given a finite set $C \subseteq \mathcal{X}$, overload τ via averages: $\tau(C) := |C|^{-1} \sum_{x\in C} \tau(x)$. $\Diamond$

For now, it suffices to think of τ as the identity map (with $\mathcal{X} = \mathbb{R}^n$), with an added convenience of computing means. Section 4, however, will rely on τ when constructing exponential families.

Definition 3.2. Given a statistic map $\tau : \mathcal{X} \to \mathbb{R}^n$ and convex function f, the cost of a single cluster C is

$$\phi_{f,\tau}(C) := \sum_{x\in C} \mathsf{B}_f(\tau(x), \tau(C)).$$

(This cost was the basis for *Bregman hard clustering* (Banerjee et al., 2005).) $\Diamond$

Example 3.3 (k-means cost)**.** Choose $\mathcal{X} = \mathbb{R}^n$, $\tau(x) = x$, and $f = \|\cdot\|_2^2$. As discussed in Example 2.2, $\mathsf{B}_f(x,y) = \|x-y\|_2^2$, and so $\phi_{f,\tau}(C) = \sum_{x\in C} \|x - \tau(C)\|_2^2$, precisely the k-means cost. $\Diamond$

As such, $\tau(C)$ plays the role of a cluster center. While this may be intuitive for the k-means cost, it requires justification for general Bregman divergences. The following definition and results bridge this gap.

Definition 3.4. A convex function f is *relatively (Gâteaux) differentiable* if, for any $y \in \mathrm{ri}(\mathrm{dom}(f))$, there exists g (necessarily any subgradient) so that, for any $x \in \mathrm{dom}(f)$, $f'(y; y-x) = \langle g, y-x\rangle$. $\Diamond$

Every differentiable function is relatively differentiable (with $g = \nabla f(y)$); fortunately, many relevant convex functions, in particular those used to construct Bregman divergences between exponential family distributions (cf. Proposition 4.5), will be relatively differentiable.

Under relative differentiability, Bregman divergences admit a bias-variance style decomposition, which immediately justifies the choice of centroid $\tau(C)$.

Lemma 3.5. *Let a proper convex relatively differentiable f, points $\{x_i\}_{i=1}^m \subset \mathbb{R}^n$, and weights $\{\alpha_i\}_{i=1}^m \subset \mathbb{R}$ be given, with $\mu := \sum_i \alpha_i x_i / (\sum_j \alpha_j) \in \mathrm{ri}(\mathrm{dom}(f))$. Then, given any point $y \in \mathrm{ri}(\mathrm{dom}(f))$,*

$$\sum_{i=1}^m \alpha_i \mathsf{B}_f(x_i, y) = \sum_{i=1}^m \alpha_i \mathsf{B}_f(x_i, \mu) + \left(\sum_{i=1}^m \alpha_i\right) \mathsf{B}_f(\mu, y).$$

Corollary 3.6. *Suppose the convex function f is relatively differentiable, let any statistic map τ and any finite cluster $C \subseteq \mathcal{X}$ be given with $\tau(C) \in \mathrm{ri}(\mathrm{dom}(f))$. Then $\phi_{f,\tau}(C) = \inf_{y\in \mathrm{ri}(\mathrm{dom}(f))} \sum_{x\in C} \mathsf{B}(\tau(x), y)$.*

Proof. Use $\mu := \tau(C) = |C|^{-1}\sum_{x\in C}\tau(x)$, Lemma 3.5, and $\mathsf{B}_f \geq 0$. $\square$

Continuing, the stage is set to finally construct the Bregman merge cost.

Definition 3.7. Given two finite subsets C_1, C_2 of $\mathcal{X}$, the cluster merge cost is simply growth in total cost:

$$\Delta_{f,\tau}(C_1, C_2) = \phi_{f,\tau}(C_1 \cup C_2) - \sum_{j\in\{1,2\}} \phi_{f,\tau}(C_j). \quad \Diamond$$

The above expression seems to imply that the computational cost of Δ scales with the number of points. But in fact, one need only look at the relevant centers.

Proposition 3.8. *Let a proper convex relatively differentiable f and two finite subsets C_1, C_2 of $\mathcal{X}$ with $\tau(C_i) \in \mathrm{ri}(\mathrm{dom}(f))$ be given. Then*

$$\Delta_{f,\tau}(C_1, C_2) = \sum_{j\in\{1,2\}} |C_j| \mathsf{B}_f(\tau(C_j), \tau(C_1 \cup C_2)).$$

Example 3.9 (Ward/k-means merge cost)**.** Continuing with the k-means cost from Example 3.3, note that for $j \in \{1,2\}$,

$$\|\tau(C_j) - \tau(C_1\cup C_2)\|_2 = \frac{|C_{3-j}| \cdot \|\tau(C_1) - \tau(C_2)\|_2}{|C_1| + |C_2|}.$$

Plugging this into the simplification of $\Delta_{f,\tau}$ provided by Proposition 3.8,

$$\begin{aligned}\Delta_{f,\tau}(C_1, C_2) &= \sum_{j\in\{1,2\}} \frac{|C_j||C_{3-j}|^2}{(|C_1|+|C_2|)^2} \|\tau(C_1) - \tau(C_2)\|_2^2 \\ &= \frac{|C_1||C_2|}{|C_1|+|C_2|} \|\tau(C_1) - \tau(C_2)\|_2^2.\end{aligned}$$

This is exactly the Ward merge cost. $\Diamond$

4. Exponential families

So far, this manuscript has developed a mathematical basis to clustering with Bregman divergences. But what does it matter, if examples of meaningful Bregman divergences are few and far between?

The primary mechanism for constructing meaningful merge costs is to model the clusters as exponential family distributions. Throughout this section, let ν be any measure over $\mathcal{X}$, and further stipulate the statistic map τ is ν-measurable.

Definition 4.1. Given a measurable statistic map τ and a vector $\theta \in \mathbb{R}^n$ of *canonical parameters*, the corresponding exponential family distribution has density

$$p_\theta(x) := \exp(\langle \tau(x), \theta\rangle - \psi(\theta)),$$

where the normalization term ψ, typically called the *cumulant* or *log partition function*, is simply

$$\psi(\theta) = \ln \int \exp(\langle \tau(x), \theta\rangle) d\nu(x). \quad \Diamond$$

Many standard distributions have this representation.

Example 4.2. Choose $\mathcal{X} = \mathbb{R}^d$ with ν being Lebesgue measure, and $n = d(d+1)$, i.e. $\mathbb{R}^n = \mathbb{R}^{d(d+1)}$. The map $\tau(x) = (x, xx^\top)$ will provide for Gaussian densities. In particular, starting from the familiar form, with mean $\mu \in \mathbb{R}^d$ and positive definite covariance $\Sigma \in \mathbb{R}^{d^2}$, the density at x, $p_\theta(x)$, is

$$\begin{aligned} &\frac{\exp(-(x-\mu)^\top \Sigma^{-1}(x-\mu)/2)}{\sqrt{(2\pi)^d |\Sigma|}} \\ &= \exp\Big(\left\langle \tau(x), (\Sigma^{-1}\mu, -\Sigma^{-1}/2) \right\rangle \\ &\quad - \frac{1}{2}\ln((2\pi)^d |\Sigma| \exp(\mu^\top \Sigma^{-1}\mu)) \Big). \end{aligned}$$

In other words, $\theta = (\Sigma^{-1}\mu, -\Sigma^{-1}/2)$. Notice that ψ (here expanded as $\frac{1}{2}\ln(\ldots)$) and θ do not make sense if Σ is merely positive semi-definite. ♢

So far so good, but where's the convex function, and does the definition of p_θ even make sense?

Proposition 4.3. *Given a measurable statistic map τ, the function ψ is well-defined, closed, convex, and never takes on the value $-\infty$.*

Remark 4.4. Notice that Proposition 4.3 did not provide that ψ is proper, only that it is never $-\infty$. Unfortunately, more can not be guaranteed: if ν is Lebesgue measure over $\mathbb{R}$ and $\tau(x) = 0$ for all x, then every parameter choice $\theta \in \mathbb{R}$ has $\psi(\theta) = \infty$. It is therefore necessary to check, for any provided τ, whether $\mathrm{dom}(\psi)$ is nonempty. ♢

Not only is ψ closed convex, it is about as well-behaved as any function discussed in this manuscript.

Proposition 4.5. *Suppose* $\mathrm{dom}(\psi)$ *is nonempty. Then ψ is relatively differentiable; in fact, given any $\theta \in \mathrm{ri}(\mathrm{dom}(\psi))$, any $\hat\tau \in \partial\psi(\theta)$, and any $\xi \in \mathrm{dom}(\psi)$,*

$$\psi'(\theta; \xi - \theta) = \langle \hat\tau, \xi - \theta \rangle = \int \langle \tau(x), \xi - \theta \rangle\, p_\theta(x) d\nu(x).$$

If ψ is fully differentiable at θ, then $\nabla\psi(\theta) = \int \tau p_\theta$. Since ψ is closed, given $\hat\tau \in \partial\psi(\theta)$, it follows that $\theta \in \partial\psi^*(\hat\tau)$. There is still cause for concern that other subgradients at $\hat\tau$ lead to different densities, but as will be shown below, this does not happen.

Now that a relevant convex function ψ has been identified, the question is whether B_ψ (or B_{ψ^*}) provide a reasonable notion of distance amongst densities.

This will be answered in two ways. To start, recall the Kullback-Leibler divergence K between densities p, q:

$$\mathsf{K}(p, q) = \int p(x) \ln\left(\frac{p(x)}{q(x)}\right) d\nu(x).$$

Theorem 4.6. *Let any $\theta_1, \theta_2 \in \mathrm{ri}(\mathrm{dom}(\psi))$ and any $\hat\tau_1 \in \partial\psi(\theta_1), \hat\tau_2 \in \partial\psi(\theta_2)$ be given, where $\partial\psi^*(\hat\tau_2) \subseteq \mathrm{ri}(\mathrm{dom}(\psi))$ (for instance, if $\mathrm{dom}(\psi)$ is relatively open). Then*

$$\mathsf{K}(p_{\theta_1}, p_{\theta_2}) = \mathsf{B}_\psi(\theta_2, \theta_1) = \mathsf{B}_{\psi^*}(\hat\tau_1, \hat\tau_2).$$

Furthermore, if $\theta_1 \in \partial\psi^*(\hat\tau_2)$, then $p_{\theta_1} = p_{\theta_2}$ ν-a.e.

Motivated by Proposition 4.5 and Theorem 4.6, the choice here is to base the cluster model on B_{ψ^*}.

Given two clusters $\{C_i\}_{i=1}^2$, set $\hat\tau_i := \tau(C_i)$. When working with these clusters, it is entirely sufficient to store only these statistics and the cluster sizes, since $\tau(C_1 \cup C_2) = |C_1 \cup C_2|^{-1}(|C_1|\hat\tau_1 + |C_2|\hat\tau_2)$. Assuming for interpretability that ψ is differentiable, since ψ is closed, $\psi^{**} = \psi$, and thus $\nabla\psi(\theta_1) = \int \tau p_{\theta_1} = \hat\tau_1$; that is to say, these distributions have their (aptly named) *mean parameterizations* as their means. And as provided by Theorem 4.6, even if differentiability fails, various subgradients of the same mean all effectively represent the same distributions.

Example 4.7. Suppose $\mathcal{X}$ is a finite set, representing a vocabulary with n words, and ν is counting measure over $\mathcal{X}$. The statistic map τ converts word k into the k^{th} basis vector $\mathbf{e}_k$. Let $\hat\tau \in \mathbb{R}^n_{++}$ represent the mean parameters of a multinomial over this set; observe that

$$\begin{aligned} p_\theta(\mathbf{e}_i) &= \langle \tau(i), \hat\tau \rangle \\ &= \exp(\langle \mathbf{e}_i, \ln\hat\tau \rangle) - \ln \int \exp(\langle \tau(k), \ln\hat\tau \rangle) d\nu(k). \end{aligned}$$

That is to say, the canonical parameter vector is $\theta = \ln\hat\tau$, the coordinate-wise logarithm of the mean parameters. Proposition 4.5 can be verified directly: $(\nabla\psi(\theta))_i = e^{\theta_i} / \sum_k e^{\theta_k} = \hat\tau$. Similarly, given another multinomial with mean parameters $\hat\tau' \in \mathbb{R}^n_{++}$ and canonical parameters $\theta' = \ln\hat\tau'$,

$$\mathsf{K}(p_\theta, p_{\theta'}) = \sum_{i=1}^n \hat\tau_i \ln\left(\frac{\hat\tau_i}{\hat\tau'_i}\right).$$

The notation $\mathbb{R}^n_{++}$ means strictly positive coordinates: no word can have zero probability. Without this restriction, it is not possible to map into the canonical parameter space. This is precisely the scenario the smoothing methods of Section 5 will work around: the provided clusters are on the relative boundary of $\mathrm{dom}(\psi^*)$, which is either not part of $\mathrm{dom}(\psi^*)$ at all, or as is the case here, causes degenerate Bregman divergences (infinite valued, and lacking subgradients). ♢

Remark 4.8. The multinomials in Example 4.7 have an overcomplete representation: scaling any canonical parameter vector by a constant gives the same mean parameter. In general, if two relative interior canonical parameters $\theta_1 \neq \theta_2$ have a common subgradient $\hat{\tau} \in \partial\psi(\theta_1) \cap \partial\psi(\theta_2)$, then it follows that $\{\theta_1, \theta_2\} \subset \partial\psi^*(\hat{\tau})$ (Rockafellar, 1970, Theorem 23.5). That is to say: this scenario leads to mean parameters which have distinct subgradients, and are thus points of nondifferentiability within $\text{ri}(\text{dom}(\psi^*))$, which necessitate the generalized development of Bregman divergences in this manuscript. ◊

A further example of Gaussians appears in Appendix C.

The second motivation for $\Delta_{\psi^*,\tau}$ is an interpretation in terms of model fit.

Theorem 4.9. *Fix some measurable statistic map τ, and let two finite subsets C_1, C_2 of $\mathcal{X}$ be given with mean parameters $\{\tau(C_1), \tau(C_2)\} = \{\hat{\tau}_1, \hat{\tau}_2\} \subseteq \text{ri}(\text{dom}(\psi^*))$. Choose any canonical parameters $\theta_i \in \partial\psi^*(\hat{\tau}_i)$, and for convenience set $C_3 := C_1 \cup C_2$, with mean parameter $\hat{\tau}_3$ and any canonical parameter $\theta_3 \in \partial\psi^*(\hat{\tau}_3)$. Then*

$$\Delta_{\psi^*,\tau}(C_1, C_2) = \sum_{i\in\{1,2\}} \sum_{x\in C_i} \ln p_{\theta_i}(x) - \sum_{x\in C_3} \ln p_{\theta_3}(x).$$

5. Smoothing

The final piece of the technical puzzle is the smoothing procedure: most of the above properties — for instance, that $\mathsf{B}_f(\tau(C_1), \tau(C_2)) < \infty$ — depend on $\tau(C_2) \in \text{ri}(\text{dom}(f))$. Relative boundary points lead to degeneracies; for example, this characterizes the Gaussian degeneracy identified in the introduction.

Definition 5.1. Given a convex set S, a statistic map $\tau : \mathcal{X} \to \mathbb{R}^n$ is a *smoothing statistic map for S* if, given any nonempty finite set $C \subseteq \mathcal{X}$, $\tau(C) \in \text{ri}(S)$. ◊

It turns out to be very easy to construct smoothing statistic maps.

Theorem 5.2. *Let a nonempty convex set S be given. Let $\tau_0 : \mathcal{X} \to \mathbb{R}^n$ be a statistic map satisfying, for any finite $C \subseteq \mathcal{X}$, $\tau_0(C) \in \text{cl}(S)$. Let $z \in \text{ri}(S)$ and $\alpha \in (0,1)$ be arbitrary. Given any finite set $C \subseteq \mathcal{X}$, define the maps*

$$\tau_1(C) := (1-\alpha)\tau_0(C) + \alpha z,$$
$$\tau_2(C) := \tau_0(C) + \alpha z,$$

In general, τ_1 is a smoothing statistic map for S. If additionally S is a convex cone, then τ_2 is also a smoothing statistic map for S.

The following two examples smooth Gaussians and multinomials via Theorem 5.2. The parameters α and z are chosen from data, and moreover satisfy $\|\alpha z\| \downarrow 0$ as the total amount of available data grows; that is to say, τ will more and more closely match τ_0.

Example 5.3 (Smoothing multinomials.)**.** The mean parameters to a multinomial lie within the probability simplex, a compact convex set. As discussed in Example 4.7, only the relative interior of the simplex provides canonical parameters. According to Theorem 5.2, all that remains in fixing this problem is to determine αz.

The approach here is to interpret the provided multinomial $\tau_0(C) = \hat{\tau}$ as based on a finite sample of size m, and thus the true parameters lie within some confidence interval around $\hat{\tau}$; crucially, this confidence interval intersects the relative interior of the probability simplex. One choice is a Bernstein-based upper confidence estimate $\tau(C) = \tau_0(C) + \mathcal{O}(1/m + \sqrt{p(1-p)/m})$, where $p = 1/n$. ◊

Example 5.4 (Smoothing Gaussians.)**.** In the case of Gaussians, as discussed in Example 4.2, only positive definite covariance matrices are allowed. But this set is a convex cone, so Theorem 5.2 reduces the problem to finding a sensible element to add in.

Consider the case of singleton clusters. Adding a full-rank covariance matrix in to the observed zero covariance matrix is like replacing this singleton with a constellation of points centered at it. Equivalently, each point is replaced with a tiny Gaussian, which is reminiscent of nonparametric density estimation techniques. Therefore one option is to use bandwidth selection techniques; the experiments of Section 7 use the "normal reference rule" (Bowman & Azzalini, 1997, Section 2.4.2), trying both the approach of estimating a bandwidth for each coordinate (suffix `-nd`), and computing one bandwidth for every direction uniformly, and simply adding a rescaling of the identity matrix to the sample covariance (suffix `-n`). ◊

When there is a probabilistic interpretation of the clusters, and in particular $\tau(C)$ may be viewed as a maximum likelihood estimate, another approach is to choose some prior over the parameters, and have τ produce a MAP estimate which also lies in the relative interior. As stated, this approach will differ slightly from the one presented here: the weight on the added element will scale with the cluster size, rather than the size of the full data, and the relationship of $\tau(C_1 \cup C_2)$ to $\tau(C_1)$ and $\tau(C_2)$ becomes less clear.

6. Clustering algorithm

The algorithm appears in Algorithm 1. Letting $T_{\Delta_{f,\tau}}$ denote an upper bound on the time to calculate a

Algorithm 1 AGGLOMERATE.
Input Merge cost $\Delta_{f,\tau}$, points $\{x_i\}_{i=1}^m \subseteq \mathcal{X}$.
Output Hierarchical clustering.

Initialize forest as $\mathscr{F} := \{\{x_i\} : i \in [m]\}$.
while $|\mathscr{F}| > 1$ **do**
  Let $\{C_i, C_j\} \subseteq \mathscr{F}$ be any pair minimizing $\Delta_{f,\tau}(C_i, C_j)$, as computed by Proposition 3.8.
  Remove $\{C_i, C_j\}$ from $\mathscr{F}$, add in $C_i \cup C_j$.
end while
return the single tree within $\mathscr{F}$.

single merge cost, a brute-force implementation (over m points) takes space $\mathcal{O}(m)$ and time $\mathcal{O}(m^3 T_{\Delta_{f,\tau}})$, whereas caching merge cost computations in a min-heap requires space $\mathcal{O}(m^2)$ and time $\mathcal{O}(m^2(\lg(m) + T_{\Delta_{f,\tau}}))$. Please refer Appendix E for more notes on running times, and a depiction of sample hierarchies over synthetic data.

If Proposition 3.8 is used to compute $\Delta_{f,\tau}$, then only the sizes and means of clusters need be stored, and computing this merge cost involves just two Bregman divergences calculations. As the new mean is a convex combination of the two old means, computing it takes time $\mathcal{O}(n)$. The Bregman cost itself can be more expensive; for instance, as discussed with Gaussians in Appendix C, it is necessary to invert a matrix, meaning $\mathcal{O}(n^3)$ steps.

7. Empirical results

Trees generated by AGGLOMERATE are evaluated in two ways. First, cluster compatibility scores are computed via dendrogram purity and initialization quality for EM upon mixtures of Gaussians. Secondly, cluster features are fed into supervised learners.

There are two kinds of data: Euclidean (points in some $\mathbb{R}^n$), and text data. There are three Euclidean data sets: UCI's `glass` (214 points in $\mathbb{R}^9$); 3s and 5s from the `mnist` digit recognition problem (1984 training digits and 1984 testing digits in $\mathbb{R}^{49}$, scaled down from the original 28x28); UCI's `spambase` data (2301 train and 2300 test points in $\mathbb{R}^{57}$). Text data is drawn from the 20 newsgroups data, which has a vocabulary of 61,188 words; a difficult dichotomy (`20n-h`), the pair `alt.atheism/talk.religion.misc` (856 train and 569 test documents); an easy dichotomy (`20n-e`), the pair `rec.sport.hockey/sci.crypt` (1192 train and 794 test documents). Finally, `20n-b` collects these four groups into one corpus.

The various trees are labeled as follows. `s-link` and `c-link` denote single and complete linkage, where l^1 distance is used for text, and l^2 distance is used for Euclidean data. `km` is the Ward/k-means merge cost. `g-n` fits full covariance Gaussians, whereas `dg-nd` fits diagonal covariance Gaussians; smoothing follows the data-dependent scheme of Example 5.4. `multi` fits multinomials, with the smoothing scheme of Example 5.3.

Table 1. Dendrogram purity on Euclidean and text data.

	c-link	s-link	km	dg-nd	g-n
glass	0.46	0.45	0.50	0.49	**0.54**
spam	0.59	0.58	0.59	**0.65**	0.60
mnist35	0.59	0.51	0.69	0.62	**0.73**
	c-link	s-link		multi	
20n-e	0.60	0.50		**0.93**	
20n-h	0.54	0.52		**0.56**	
20n-b	0.31	0.29		**0.62**	

7.1. Cluster compatibility

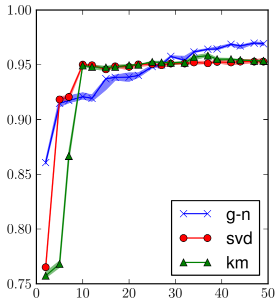
Table 1 contains cluster purity scores, a standard dendrogram quality measure, defined as follows. For any two points with the same label l, find the smallest cluster C in the tree which contains them both; the purity with respect to these two points is the fraction of C having label l. The purity of the dendrogram is the weighted sum, over all pairs of points sharing labels, of pairwise purities. The `glass`, `spam`, and `20newsgroups` data appear in Heller & Ghahramani (2005); although a direct comparison is difficult, since those experiments used subsampling and randomized purity, the Euclidean experiments perform similarly, and the text experiments fare slightly better here.

For another experiment, now assessing the viability of AGGLOMERATE as an initialization to EM applied to mixtures of Gaussians, please see Appendix F.

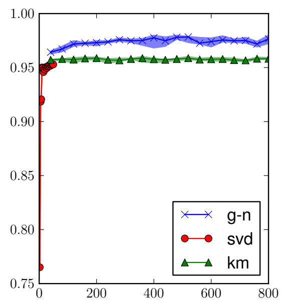
7.2. Feature generation

The final experiment is to use dendrograms, built from training data, to generate features for classification tasks. Given a budget of features k, the top k clusters $\{C_i\}_{i=1}^k$ of a specified dendrogram are chosen, and for any example x, the i^{th} feature is $\Delta(C_i, \{x\})$. Statistically, this feature measures the amount by which the model likelihood degrades when C_i is adjusted to accommodate x. The choice of tree was based on training set purity from Table 1. In all tests, the original features are discarded (i.e., only the k generated features are used).

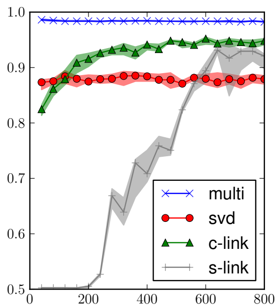
Figure 2 shows the performance of logistic regression classifiers using tree features, as well as SVD features. The stopping rule used validation set performance.



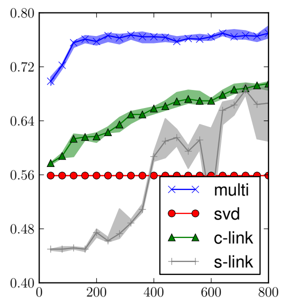
(a) `mnist35`, zoomed in.



(b) `mnist35`, zoomed out.



(c) `20n-e`.



(d) `20n-h`.

Figure 2. Comparison of dendrogram features to SVD features; y-axis denotes classification accuracy on test data, x-axis denotes #features. In the first two plots, `mnist35` was used. The SVD can only produce as many features as the dimension of the data, but the proposed tree continues to improve performance beyond this point. For the text data tasks `20n-e` and `20n-h`, tree methods strongly outperform the SVD. Please see text for details.

Acknowledgements

This work was graciously supported by the NSF under grant IIS-0713540.

References

Azoury, Katy S. and Warmuth, Manfred K. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

Banerjee, Arindam, Merugu, Srujana, Dhillon, Inderjit S., and Ghosh, Joydeep. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

Blundell, C., Teh, Y.W., and Heller, K.A. Bayesian rose trees. In *UAI*, 2010.

Borwein, Jonathan and Lewis, Adrian. *Convex Analysis and Nonlinear Optimization*. Springer Publishing Company, Incorporated, 2000.

Bowman, Adrian W. and Azzalini, Adelchi. *Applied Smoothing Techniques for Data Analysis*. Oxford University Press, USA, 1997.

Brown, Lawrence D. *Fundamentals of Statistical Exponential Families*. Insitute of Mathematical Statistics, USA, 1986.

Csiszár, Imre and Matúš, František. Generalized minimizers of convex integral functionals, Bregman distance, Pythagorean identities. 2012. `arXiv:1202.0666v1 [math.OC]`.

Duda, Richard O., Hart, Peter E., and Stork, David G. *Pattern Classification*. Wiley, 2 edition, 2001.

Folland, Gerald B. *Real analysis: modern techniques and their applicatins*. Wiley Interscience, 2 edition, 1999.

Fraley, C. Algorithms for model-based gaussian hierarchical clustering. *SIAM Journal on Scientific Computing*, 20:270–281, 1998.

Garcia, Vincent, Nielsen, Frank, and Nock, Richard. Hierarchical gaussian mixture model. In *ICASSP*, pp. 4070–4073, 2010.

Gordon, Geoff J. *Approximate Solutions to Markov Decision Processes*. PhD thesis, Carnegie Mellon University, 1999.

Heller, Katherine A. and Ghahramani, Zoubin. Bayesian hierarchical clustering. In *ICML*, pp. 297–304, 2005.

Hiriart-Urruty, Jean-Baptiste and Lemaréchal, Claude. *Fundamentals of Convex Analysis*. Springer Publishing Company, Incorporated, 2001.

Iwayama, Makoto and Tokunaga, Takenobu. Hierarchical bayesian clustering for automatic text classification. In *IJCAI*, pp. 1322–1327, 1995.

Kiwiel, Krzysztof C. Proximal minimization methods with generalized Bregman functions. *SIAM journal on control and optimization*, 35:1142–1168, 1995.

Merugu, Srujana. *Distributed Learning using Generative Models*. PhD thesis, University of Texas, Austin, 2006.

Murtagh, Fionn. A Survey of Recent Advances in Hierarchical Clustering Algorithms. *The Computer Journal*, 26(4):354–359, 1983.

Rockafellar, R. Tyrrell. *Convex Analysis*. Princeton University Press, 1970.

Slonim, Noam and Tishby, Naftali. Agglomerative information bottleneck. pp. 617–623, 1999.

Wainwright, Martin J and Jordan, Michael I. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008.

Ward, Joe H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

A. Deferred material from Section 2

A.1. Deferred Proofs

Proof of Proposition 2.3. For the first part, since $y \in \text{ri}(\text{dom}(f))$ and $x \in \text{dom}(f)$, there exists $\tau > 0$ so that $y + \tau(y - x) \in \text{dom}(f)$. Combined with the fact that the difference quotient $q(t) := (f(y+t(y-x)) - f(y))/t$ is nondecreasing (Rockafellar, 1970, Theorem 23.1), $f'(y; y-x) \leq (f(y+\tau(y-x)) - f(y))/\tau < \infty$. Since $y \in \text{ri}(\text{dom}(f))$ grants $f'(y; y-x)$ is proper (Rockafellar, 1970, Theorem 23.4), it follows that $f'(y; y-x)$ is finite. As a result of this and x, y being within $\text{dom}(f)$, $\mathsf{B}_f(x, y)$ is also finite.

For the second part, first note that $f'(y; y-x) = \sup\{\langle g, y-x \rangle : g \in \partial f(y)\}$, without appealing to $x \in \text{dom}(f)$ (Rockafellar, 1970, Theorem 23.4). But when $f'(y; y-x)$ is finite, this sup is actually a max (Borwein & Lewis, 2000, Exercise 3.1.29). □

A.2. Deferred Bregman properties

Proposition A.1 (Nonnegativity). *Suppose f is convex, proper, and $y \in \text{ri}(\text{dom}(f))$. Then $\mathsf{B}_f(\cdot, y)$ is convex, proper, nonnegative, and $\mathsf{B}_f(y, y) = 0$.*

Proof. For any $y \in \text{dom}(f)$, $\mathsf{B}_f(y, y) = f(y) - f(y) + f'(y; y-y) = 0$; in particular, there exists a point where $\mathsf{B}_f(\cdot, y)$ is finite. Next, $f'(y; \cdot)$ is convex (Rockafellar, 1970, Theorem 23.1), and now using $y \in \text{ri}(\text{dom}(f))$, $f'(y; \cdot)$ is also proper (Rockafellar, 1970, Theorem 23.4). Since f is provided as proper and convex, it follows that the sum $f(\cdot) - f(y) + f'(y; y - \cdot)$ is convex, never takes the value $-\infty$, and since it is finite at y, its domain is nonempty and thus it is proper. Finally, taking any $g \in \partial f(y)$ (which is nonempty since $y \in \text{ri}(\text{dom}(f))$ (Rockafellar, 1970, Theorem 23.4), by Proposition 2.3 and the definition of subgradient,

$$\mathsf{B}_f(x, y) \geq f(x) - f(y) - \langle g, x - y \rangle \geq 0. \qquad \square$$

Proposition A.2 (Identifiability). *Let strictly convex and proper f, $y \in \text{ri}(\text{dom}(f))$, and $x \in \text{dom}(f)$ be given. Then $\mathsf{B}_f(x, y) = 0$ iff $x = y$.*

Proof. If $x = y$, then Proposition A.1 gives the result.

Suppose $x \neq y$, and set $z := (x - y)/2$. By strict convexity,

$$f(y + z) < \frac{1}{2}(f(x) + f(y)).$$

Choosing any subgradient $g \in \partial f(y)$ and making use of Proposition 2.3 and definition of subgradient,

$$\begin{aligned} \mathsf{B}_f(x, y) &\geq f(x) - f(y) - \langle g, x - y \rangle \\ &\geq f(x) - f(y) - 2\langle g, z \rangle \\ &\geq f(x) - f(y) - 2(f(y+z) - f(y)) > 0. \qquad \square \end{aligned}$$

Proposition A.3 (Duality). *Let closed proper convex f, $y \in \text{ri}(\text{dom}(f))$ and $g_x \in \text{ri}(\text{dom}(f^*))$ be given. Then $\sup_{x \in \partial f^*(g_x)} \mathsf{B}_f(x, y) = \sup_{g \in \partial f(y)} \mathsf{B}_{f^*}(g_y, g_x)$.*

Proof. Since f is closed, $f = f^{**}$, and by the Fenchel-young inequality (Rockafellar, 1970, Theorem 23.5.d),

$$\begin{aligned} &\sup_{x \in \partial f^*(g_x)} \mathsf{B}_f(x, y) \\ &= \sup_{x \in \partial f^*(g_x)} \sup_{g_y \in \partial f(y)} f(x) - f(y) - \langle g_y, x - y \rangle \\ &= \sup_{g_y \in \partial f(y)} \sup_{x \in \partial f^*(g_x)} \langle g_x, x \rangle - f^*(g_x) + f^*(g_y) \\ &\quad - \langle g_y, y \rangle - \langle g_y, x - y \rangle \\ &= \sup_{g_y \in \partial f(y)} \sup_{x \in \partial f^*(g_x)} f^*(g_y) - f^*(g_x) - \langle x, g_y - g_x \rangle \\ &= \sup_{g_y \in \partial f(y)} \mathsf{B}_{f^*}(g_y, g_x). \qquad \square \end{aligned}$$

A.3. Proof of Pythagorean property

The Pythagorean theorem is concerned with Bregman projections onto a set, which are simply the points of that set which minimize the Bregman divergence from some reference point. In general, these projections are not guaranteed to exist, however the following lemma will provide some sufficient conditions.

For a counterexample to the general existence, consider $\mathbb{R}^2$ with $f(x) = x_2^2$, the second coordinate squared, meaning $\mathsf{B}_f(x, y) = (x_2 - y_2)^2$. The target set is

$$S := \{(x_1, x_2) \in \mathbb{R}^2 : x_1 \in \mathbb{R}, x_2 \leq -e^{x_1}\},$$

and the reference point is $y := (0, 1)$. Note that

$$\inf_{x \in S} \mathsf{B}_f(x, y) = 1;$$

on the other hand, every $x \in S$ has $\mathsf{B}_f(x, y) > 1$, meaning the infimum is not attained, and projections do not exist.

Lemma A.4. *Let closed proper convex relative differentiable f, closed convex set S with $\text{ri}(S) \cap \text{ri}(\text{dom}(f)) \neq \emptyset$, any $y \in \text{ri}(\text{dom}(f))$, and any $g_y \in \partial f(y)$ be given. Define the set*

$$\Lambda_f^S(y) := \{x \in S : \mathsf{B}_f(x, y) = \inf_{z \in S} \mathsf{B}_f(z, y)\}.$$

If any of the following statements hold, then $\Lambda_f^s(y)$ is nonempty.

- *S is compact.*
- *f is* 1-coercive*, meaning*

$$\lim_{\|x\|\to\infty} \frac{f(x)}{\|x\|} = \infty.$$

- *S is affine, and $g_y \in \mathrm{ri}(\mathrm{dom}(f^*))$ (it suffices for $\mathrm{dom}(f^*)$ to be affine, thus relatively open).*
- *Most generally,*

$$\begin{gathered}\mathrm{ri}(\mathrm{dom}(\phi \mapsto \sup_{x\in S} \langle -\phi, x\rangle)) \\ \cap \\ \mathrm{ri}(\mathrm{dom}(\phi \mapsto f^*(\phi + g_y) + f(y) - \langle g_y, y\rangle)) \\ \neq \emptyset.\end{gathered}$$

Now suppose $\Lambda_f^s(y)$ is nonempty, and let $\bar{x} \in \Lambda_f^S(y)$ be arbitrary. Then every $g_{\bar{x}} \in \partial f(\bar{x})$ satisfies

$$\sup_{z\in S} \langle g_y - g_{\bar{x}}, z - \bar{x}\rangle = 0.$$

Note that the above existence conditions are simply constraint qualification for the dual problem present in the proof; as such, applications falling outside the above sufficient conditions can simply choose a more specialized constraint qualification, but otherwise repeat the same proof.

To see the value of the above cases, consider again $f(x) = x_2^2$ as above, but now S is any affine set. In general, S is not compact, and this f is not 1-coercive. However, $\mathrm{dom}(f^*) = \{0\} \times \mathbb{R}$, which is affine, and so the third condition above grants the existence of Bregman projections.

Proof. Let ι_S be the indicator function for S, which is closed proper convex since S is closed convex and nonempty. Let $g_y \in \partial f(y)$ be arbitrary, and define $z(x; g_y) := f(x) - f(y) - \langle g_y, x - y\rangle$; since f is closed proper convex, $z(x)$, which simply adds an affine function, is also closed proper convex (Rockafellar, 1970, Theorem 9.3). As such, Fenchel's duality theorem (Rockafellar, 1970, Theorem 31.2.1) gives

$$\inf_{x\in\mathbb{R}^n} z(x; g_y) + \iota_S(x) \tag{A.5}$$

$$= \sup_{\phi\in\mathbb{R}^n} -z^*(\phi; g_y) - \iota_S^*(-\phi). \tag{A.6}$$

Next, $\iota_S^*(-\phi) = \sup_{x\in S} \langle -\phi, x\rangle$. On the other hand

$$z^*(\phi; g_y) = f^*(\phi + g_y) + f(y) - \langle g_y, y\rangle$$

(Rockafellar, 1970, Theorem 12.3). If it can be shown that $\mathrm{ri}(\mathrm{dom}(z^*(\cdot; g_y))) \cap \mathrm{ri}(\mathrm{dom}(\iota_S^*(-\cdot))) \neq \emptyset$, then the primal problem (A.5) has minimizers (Rockafellar, 1970, Theorem 31.2.1), which will lead to the eventual statement about Λ_f^S. To this end, first notice that last of the four conditions in the theorem is exactly this intersection condition, and thus it automatically grants minimizers. On the other hand, if S is compact, then ι_S^* is finite everywhere, which combined with the fact that $z^*(\cdot; g_y)$ is proper (cf. (Rockafellar, 1970, Theorem 12.2)) grants that the desired intersection is nonempty, and minimizers to (A.5) exist. Next, suppose f is 1-coercive, from which it follows that $z(\cdot; g_y)$ is also 1-coercive. But this entails that $z^*(\cdot; g_y)$ is finite everywhere (Hiriart-Urruty & Lemaréchal, 2001, Proposition E.1.3.8, since $z(\cdot; g_y) \geq 0$), which means the desired intersection is again nonempty (since $\mathrm{dom}(\iota_S^*(-\cdot))$ is nonempty, and thus has nonempty relative interior), and so minimizers to (A.5) exist. Lastly, suppose $g_y \in \mathrm{ri}(\mathrm{dom}(f^*)$ and S is affine. Write $S = L + a$, where L is a linear subspace, and $a \in S$; then $\iota_S^*(-\phi) = \iota_{L^\perp}(\phi) - \langle \phi, a\rangle$ (Rockafellar, 1970, Theorem 12.3 and the discussion preceding it), and in particular $\mathrm{dom}(\iota_S^*(-\cdot)) = L^\perp$. Next, $g_y \in \mathrm{ri}(\mathrm{dom}(f^*))$ and the form of $z^*(\cdot; g_y)$ provides $0 \in \mathrm{ri}(\mathrm{dom}(z^*(\cdot; g_y)))$, so 0 lies within their intersection, meaning in particular it is nonempty, and minimizers thus exist.

Now suppose the existence of a (an arbitrary) primal minimizer $\bar{x}$, and since $\mathrm{ri}(\mathrm{dom}(f)) \cap \mathrm{ri}(\mathrm{dom}(\iota_S)) \neq \emptyset$ by assumption, the dual problem (A.6) possesses maximizers (Rockafellar, 1970, Theorem 31.2.1). It must hold (Rockafellar, 1970, Theorems 23.8 and 31.3) that every maximizer can be written as

$$\bar{\phi} \in \partial z(\bar{x}; g_y) = \{\phi - g_y : \phi \in \partial f(\bar{x})\}.$$

Correspondingly, let $g_{\bar{x}} \in \partial f(\bar{x})$ be arbitrary, so now $\bar{\phi} = g_{\bar{x}} - g_y$ maximizes the dual problem; then, by (Rockafellar, 1970, Theorem 23.5) and the above duality relation,

$$\begin{aligned}
&z(\bar{x}; g_y) \\
&= -z^*(\bar{\phi}; g_y) - \sup_{x\in S} \left\langle -\bar{\phi}, x\right\rangle \\
&= -\sup_x \left(\langle g_{\bar{x}} - g_y, x\rangle - f(x) + f(y) + \langle g_y, x - y\rangle\right) \\
&\quad - \sup_{x\in S} \langle g_y - g_{\bar{x}}, x\rangle \\
&= -\sup_x(\langle g_{\bar{x}}, x\rangle - f(x)) - f(y) + \langle g_y, y\rangle \\
&\quad - \sup_{x\in S} \left(\langle g_y - g_{\bar{x}}, x\rangle\right) \\
&= f(\bar{x}) - \langle \bar{x}, g_{\bar{x}}\rangle - f(y) - \langle g_y, \bar{x}\rangle + \langle g_y, \bar{x}\rangle + \langle g_y, y\rangle \\
&\quad - \sup_{x\in S} \left(\langle g_y - g_{\bar{x}}, x\rangle\right) \\
&= z(\bar{x}; g_y) - \sup_{x\in S} \langle g_y - g_{\bar{x}}, x - \bar{x}\rangle .
\end{aligned}$$

Since $z(\bar{x}; g_y)$ is finite, it can be cancelled from both side, yielding

$$\sup_{x \in S} \langle g_y - g_{\bar{x}}, x - \bar{x} \rangle = 0.$$

To finish, relative differentiability grants that $z(\cdot; g_y) = \mathsf{B}_f(\cdot, y)$ over $\text{dom}(f)$, and the domain of $\inf_{x \in S} \mathsf{B}_f(x, y)$ can be safely restricted to $\text{dom}(f) \cap S$. ☐

Lemma A.7. *Let a closed proper convex relative differentiable f, $y \in \text{ri}(\text{dom}(f))$, closed convex set S with $\text{ri}(S) \cap \text{ri}(\text{dom}(f)) \neq \emptyset$, and any $\bar{x} \in \Lambda_f^S(y) \neq \emptyset$ be given. Let $g_y \in \partial f(y)$ and $g_{\bar{x}} \in \partial f(\bar{x})$ be arbitrary. Finally, set*

$$H := \{z \in \mathbb{R}^n : \langle g_y - g_{\bar{x}}, z \rangle = \langle g_y - g_{\bar{x}}, \bar{x} \rangle\}.$$

If $g_y = g_{\bar{x}}$, then $H = \mathbb{R}^n$ and $\mathsf{B}_f(\bar{x}, y) = 0 = \min_{z \in H} \mathsf{B}_f(z, y)$. Otherwise, when $g_y \neq g_{\bar{x}}$, then H is a supporting hyperplane to S, and $\mathsf{B}_f(\bar{x}, y) = \min_{z \in H} \mathsf{B}_f(z, y)$.

Proof. First consider the case $g_y = g_{\bar{x}}$, which means $H = \mathbb{R}^n$, and thus $\inf_{z \in H} \mathsf{B}_f(z, y) = \mathsf{B}_f(y, y) = 0$. Furthermore, by nonnegativity of B_f and relative differentiability of f,

$$\begin{aligned} 0 &\leq \mathsf{B}_f(\bar{x}, y) + \mathsf{B}_f(y, \bar{x}) \\ &= f(\bar{x}) - f(y) - \langle g_y, \bar{x} - y \rangle \\ &\quad + f(y) - f(\bar{x}) - \langle g_{\bar{x}}, y - \bar{x} \rangle \\ &= \langle g_y - g_{\bar{x}}, y - \bar{x} \rangle \\ &= 0, \end{aligned}$$

which means in particular that $\mathsf{B}_f(\bar{x}, y) = 0 = \inf_{z \in H} \mathsf{B}_f(z, y)$.

On the other hand, if $g_y \neq g_{\bar{x}}$, then H is a hyperplane. By Lemma A.4 and the definition of H, for any $z \in H$ and $q \in S$,

$$\langle g_y - g_{\bar{x}}, q \rangle \leq \langle g_y - g_{\bar{x}}, \bar{x} \rangle = \langle g_y - g_{\bar{x}}, z \rangle .$$

Thus, since all of S falls on one side of H, and moreover since H actually contacts S ($\bar{x} \in S \cap H$), then H is a supporting hyperplane to S.

For the minimality statement (when H is a hyperplane), note $\mathsf{B}_f(\bar{x}, y) \geq \inf_{z \in H} \mathsf{B}_f(z, y)$ since again $\bar{x} \in H$. For the other direction, taking an arbitrary $z \in H$ and noting $0 = \langle g_y - g_{\bar{x}}, z - \bar{x} \rangle$ by definition of H, and the definition of subgradient,

$$\begin{aligned} &\mathsf{B}_f(\bar{x}, y) \\ &= f(\bar{x}) - f(y) - \langle g_y, \bar{x} - y \rangle \\ &= f(\bar{x}) - f(y) - \langle g_y, \bar{x} - y \rangle - \langle g_y - g_{\bar{x}}, z - \bar{x} \rangle \\ &= f(\bar{x}) + \langle g_{\bar{x}}, z - \bar{x} \rangle - f(y) - \langle g_y, (\bar{x} - y) + (z - \bar{x}) \rangle \\ &\leq f(z) - f(y) - \langle g_y, z - y \rangle \\ &\leq \sup\{f(z) - f(y) - \langle g, z - y \rangle : g \in \partial f(y)\} \\ &= \mathsf{B}_f(z, y). \end{aligned}$$

Since z was arbitrary, the result follows. ☐

Theorem A.8 (Pythagorean Theorem)**.** *Let f, S, y, and $\Lambda_f^S(y)$ be as discussed in Lemma A.4 and Lemma A.7. Then, for any $\bar{x} \in \Lambda_f^S(y)$ and any $q \in S$,*

$$\mathsf{B}_f(q, y) \geq \mathsf{B}_f(q, \bar{x}) + \mathsf{B}_f(\bar{x}, y).$$

If S is affine, then

$$\mathsf{B}_f(q, y) = \mathsf{B}_f(q, \bar{x}) + \mathsf{B}_f(\bar{x}, y).$$

Proof. To start, using relative differentiability to rewrite Bregman divergences in terms of subgradients $g_y \in \partial f(y)$ and $g_{\bar{x}} \in \partial f(\bar{x})$,

$$\begin{aligned} &\mathsf{B}_f(q, y) - \mathsf{B}_f(q, \bar{x}) - \mathsf{B}_f(\bar{x}, y) \\ &= \langle g_y, y - q \rangle + \langle g_{\bar{x}}, q - \bar{x} \rangle + \langle g_y, \bar{x} - y \rangle \\ &= \langle g_{\bar{x}} - g_y, q - \bar{x} \rangle . \end{aligned} \tag{A.9}$$

Without any assumptions on S, it follows by Lemma A.4 that

$$\text{(A.9)} \geq \inf_{z \in S} \langle g_{\bar{x}} - g_y, z - \bar{x} \rangle \geq 0,$$

which provides the first part of the theorem.

Now suppose S is affine. If $S = \mathbb{R}^n$, then $\bar{x} = y$, and so $\mathsf{B}_f(q, y) - \mathsf{B}_f(q, \bar{x}) - \mathsf{B}_f(\bar{x}, y) = 0$ directly. When $S \subset \mathbb{R}^n$, if $g_{\bar{x}} = g_y$, then eq. (A.9) is exactly zero, and the second result follows again. Finally, if $S \subset \mathbb{R}^n$ and $g_{\bar{x}} \neq g_y$, the set H as defined in Lemma A.7 will be a supporting hyperplane to S, and thus contain S (since H has dimension $n-1$, and S is an affine set with dimension at most $n-1$). But then $q \in S$ means $q \in H$, so by definition of H,

$$\text{(A.9)} = \langle g_{\bar{x}} - g_y, q - \bar{x} \rangle = 0.$$ ☐

B. Deferred material from Section 3

Remark B.1. It is stated in the definition of relative differentiability that the desired element of $\mathbb{R}^n$ is necessarily any subgradient; this remark provides an argument.

To be precise, f is relative differentiable if for any $y \in \text{ri}(\text{dom}(f))$, there is a q so that for any $x \in \text{dom}(f)$, $f'(y; y-x) = \langle q, y-x \rangle$. Note that f always has subgradients at relative interior points (Rockafellar, 1970, Theorem 23.4); it will be shown both that any subgradient can act as q, and furthermore that q must be a subgradient.

To start, suppose there is a y and a subgradient g so that the statement fails, meaning there is an $x' \in \text{dom}(f)$ so that $f'(y; y-x') \neq \langle g, y-x' \rangle$. Meanwhile, suppose contradictorily that f is relatively differentiable, granting the existence of a desired q.

Since $y \in \text{ri}(\text{dom}(f))$ and $x' \in \text{dom}(f)$, there is a tiny $\tau > 0$ so that

$$z := y + \tau(y - x') \in \text{dom}(f).$$

Since $f'(y; r) = \sup\{\langle g, r \rangle : g \in \partial(y)\}$ (Rockafellar, 1970, Theorem 23.4),

$$\begin{aligned} \langle q, y-x' \rangle &= f'(y; y-x') > \langle g, y-x' \rangle, \\ \langle q, y-z \rangle &= f'(y; y-z) \geq \langle g, y-z \rangle. \end{aligned}$$

Unwrapping the definition of z, this means

$$\begin{aligned} \langle q, y-z \rangle &\geq \langle g, y-z \rangle \\ &= -\tau \langle g, y-x' \rangle \\ &> -\tau \langle q, y-x' \rangle \\ &= \langle q, y-z \rangle, \end{aligned}$$

a contradiction.

For the other direction, let $y \in \text{ri}(\text{dom}(f))$ be arbitrary, along with an element q granting relative differentiability, and any $\bar{x} \in \mathbb{R}^n$. Now, for any $g \in \partial f(y)$,

$$f(\bar{x}) \geq f(y) + \langle g, \bar{x} - y \rangle.$$

Taking the supremum over $\partial f(y)$ on both sides and noting $f'(y; \bar{x} - y) = \sup\{\langle g, \bar{x} - y \rangle : g \in \partial f(y)\}$ (Rockafellar, 1970, Theorem 23.4),

$$\begin{aligned} f(\bar{x}) &\geq \sup\{f(y) + \langle g, \bar{x} - y \rangle : g \in \partial f(y)\} \\ &= f(y) + f'(y; \bar{x} - y) \\ &= f(y) + \langle q, \bar{x} - y \rangle. \end{aligned}$$

Since $\bar{x}$ was arbitrary, it follows that q is a subgradient at y. ◊

Proof of Lemma 3.5. Choosing any $g_\mu \in \partial f(\mu)$ and $g_y \in \partial f(y)$ and recalling the definition of relative differentiability, the result follows by noting both sides are equal to

$$\sum_i \alpha_i \left(f(x_i) \underbrace{- f(\mu) + f(\mu)}_{=0} - f(y) - \langle g_y, x_i - y \rangle \right) + \underbrace{\left\langle g_\mu, \left(\sum_i \alpha_i x_i \right) - \mu \sum_j \alpha_j \right\rangle}_{=0}.$$ □

Proof of Proposition 3.8. First note by the decomposition property that $\tau(C_1 \cup C_2)$ is a convex combination of $\tau(C_1)$ and $\tau(C_2)$; since convex combinations of relatively interior points stay within the relative interior (Rockafellar, 1970, Theorem 6.2), then $\tau(C_1 \cup C_2) \in \text{ri}(\text{dom}(f))$. Thus Lemma 3.5 can be applied for each $j \in \{1, 2\}$, yielding

$$\begin{aligned} &|C_j| \mathsf{B}_f(\tau(C_j), \tau(C_1 \cup C_2)) \\ &= \sum_{x_i \in C_j} \mathsf{B}_f(\tau(x_i), \tau(C_1 \cup C_2)) - \phi_{f,\tau}(C_j). \end{aligned}$$

Summing over both $j \in \{1, 2\}$, the result follows. □

C. Deferred material from Section 4

C.1. Deferred technical material

When dealing with the cumulant ψ, it is convenient to first prove most properties for the *moment generating function*

$$\rho(\theta) := \int \exp(\langle \tau(x), \theta \rangle) d\nu(x) = \int \exp(\langle \tau, \theta \rangle).$$

Proposition C.1. *The moment generating function ρ is well-defined, positive, convex, and closed.*

Proof. Since τ is a measurable and $\exp(\langle \cdot, \theta \rangle)$ is continuous, it follows that $x \mapsto \exp(\langle \tau(x), \theta \rangle)$ is measurable. Furthermore, since $\exp > 0$, it follows that the integral $\int \exp(\langle \tau, \theta \rangle)$ (and thus ρ) is always well-defined and positive (although it may be $+\infty$).

Convexity of ρ follows the convexity of exp and linearity of $\int$ and $\langle \tau(x), \cdot \rangle$. Closure is shown by establishing lower semi-continuity via Fatou's lemma and the continuity of $\exp(\langle \tau(x), \cdot \rangle)$; indeed, given any convergent sequence $\theta_i \to \theta$,

$$\begin{aligned} \liminf_i \int \exp(\langle \tau, \theta_i \rangle) &\geq \int \liminf_i \exp(\langle \tau, \theta_i \rangle) \\ &= \int \exp(\langle \tau, \theta \rangle). \end{aligned}$$ □

Most of these properties of ρ immediately carry over to ψ.

Proof of Proposition 4.3. Since ρ is well-defined and positive, $\psi = \ln(\rho(\theta))$ is also well-defined, and never takes the value $-\infty$. Similarly, lower semi-continuity (closedness) follows from the continuity of ln and the lower semi-continuity of ρ.

Finally, let any $\theta_0, \theta_1 \in \mathbb{R}^n$ and any $\alpha \in [0,1]$ be given, and define $\theta_\alpha = (1-\alpha)\theta_0 + \alpha\theta_1$. By Hölder's inequality with conjugate exponents $1/\alpha, 1/(1-\alpha)$,

$$\begin{aligned}&\int \exp(\langle \tau, \theta_\alpha \rangle)\\ &= \int \exp(\langle \tau, \theta_0 \rangle)^\alpha \exp(\langle \tau, \theta_1 \rangle)^{(1-\alpha)}\\ &\leq \left(\int \exp(\langle \tau, \theta_0 \rangle)^{\frac{\alpha}{\alpha}}\right)^\alpha \left(\int \exp(\langle \tau, \theta_1 \rangle)^{\frac{1-\alpha}{1-\alpha}}\right)^{1-\alpha}.\end{aligned}$$

Applying ln to both sides gives convexity. ☐

The combination of convexity and lower semi-continuity provide $\rho = \rho^{**}$ and $\psi = \psi^{**}$.

Next, the useful fact that means according to the density $\exp(\langle \tau, \theta \rangle)$ can be bounded in terms of nearby values of ρ.

Lemma C.2. *Let $\theta \in \mathrm{dom}(\rho)$ and any $d \in \mathbb{R}^n$ and $\epsilon > 0$ be given so that $\theta \pm \epsilon d \in \mathrm{dom}(\rho)$. Then*

$$\int |\langle \tau, d \rangle| \exp(\langle \tau, \theta \rangle) < \frac{1}{\epsilon}(\rho(\theta - \epsilon d) + \rho(\theta + \epsilon d)) < \infty.$$

Proof. For any $y \in \mathbb{R}$, $|y| = \epsilon|y|/\epsilon < e^{\epsilon|y|}/\epsilon$. Thus, for any $x \in \mathcal{X}$, by positivity of exp,

$$\begin{aligned}&|\langle \tau(x), d \rangle| e^{\langle \tau(x), y \rangle}\\ &< \frac{1}{\epsilon} e^{\langle \tau(x), y \rangle + \epsilon |\langle \tau(x), d \rangle|}\\ &\leq \frac{1}{\epsilon}\left(e^{\langle \tau(x), y \rangle + \epsilon \langle \tau(x), d \rangle} + e^{\langle \tau(x), y \rangle - \epsilon \langle \tau(x), d \rangle}\right).\end{aligned}$$

Correspondingly,

$$\int |\langle \tau, d \rangle| e^{\langle \tau, \theta \rangle} < \frac{1}{\epsilon}(\rho(\theta + \epsilon d) + \rho(\theta - \epsilon d)) < \infty,$$

the final inequality using $\theta \pm \epsilon d \in \mathrm{dom}(\rho)$. ☐

These bounds allow directional derivatives to be controlled.

Lemma C.3. *Let $\theta \in \mathrm{dom}(\rho)$ and any $d \in \mathbb{R}^n$ be given so that $\theta \pm d \in \mathrm{dom}(\rho)$. Then*

$$\rho'(\theta; d) = \int \langle \tau, d \rangle \exp(\langle \tau, \theta \rangle),$$

where the integral is well-defined and finite.

Proof. Let θ and d be given as specified above. Define $f(x; h) = \exp(\langle \tau(x), \theta + hd \rangle)$, where $h \in [-1, +1]$. Since $\mathrm{dom}(\rho)$ is convex, $\int f(x; h) < \infty$ for all $h \in [-1, +1]$. Furthermore, $f(x; \cdot)$ is convex for all x, and $f'(x; h) = \frac{\partial}{\partial h} f(x; h) = \langle \tau(x), d \rangle f(x; h)$.

Now define $g(x) := 4(f(x; -1) + f(x; +1))$; note $\int |g| < \infty$. For any $(x, h) \in \mathcal{X} \times [0, 1/2]$, by Lemma C.2 (applied to $(\theta + hd) + \pm d/2$) and convexity of $f(x; \cdot)$,

$$\begin{aligned}|f'(x; h)| &= |\langle \tau(x), d \rangle| f(x; h)\\ &< 2(f(x; h + 1/2) + f(x; h - 1/2))\\ &\leq 4g(x).\end{aligned}$$

Now let any sequence $h_n \downarrow 0$ be given (without loss of generality, $h_n \in (0, 1/2]$), and define

$$q_n(x) := \frac{f(x; h_n) - f(x; 0)}{h_n}.$$

By construction, $f'(x; 0) = \lim_n q_n(x)$. By the mean value theorem and construction of g,

$$|q_n(x)| \leq \sup_{h \in [0, 1/2]} |f'(x; h)| \leq g(x).$$

So, by the dominated convergence theorem (with dominating function g),

$$\begin{aligned}\lim_{n \to \infty} \frac{\rho(\theta + h_n d) - \rho(\theta)}{h_n} &= \lim_{n \to \infty} \int q_n\\ &= \int \lim_{n \to \infty} q_n\\ &= \int \langle \tau, d \rangle \exp(\langle \tau, \theta \rangle).\end{aligned}$$

Since $h_n \downarrow 0$ was arbitrary, the result follows. ☐

From directional derivatives, subgradients follow.

Lemma C.4. *If $\theta \in \mathrm{ri}(\mathrm{dom}(\rho))$, then $\partial\rho(\theta) \neq \emptyset$, and moreover every $g \in \partial\rho(\theta)$ satisfies, for every $d \in \mathbb{R}^n$ with $\theta + d \in \mathrm{dom}(\rho)$,*

$$\langle g, d \rangle = \int \langle \tau, d \rangle \exp(\langle \tau, \theta \rangle).$$

If $\theta \in \mathrm{int}(\mathrm{dom}(\rho))$, then $\nabla\rho(\theta)$ exists, and satisfies

$$\nabla\rho(\theta) = \int \tau \exp(\langle \tau, \theta \rangle).$$

Proof. Let $\theta \in \mathrm{ri}(\mathrm{dom}(\rho))$ be given. Thus $\partial\rho(\theta)$ is nonempty (Rockafellar, 1970, Theorem 23.4), and every $g \in \partial\rho(\theta)$ satisfies $\langle g, \cdot \rangle \leq \rho'(\theta; \cdot)$. Now let $d \in \mathbb{R}^n$ be given with $\theta + d \in \mathrm{dom}(\rho)$; by definition of relative interior, there must exist $\epsilon > 0$ so that $\theta - \epsilon d \in \mathrm{dom}(\rho)$. Combining these pieces with Lemma C.3,

$$\langle g, \epsilon d\rangle \le \rho'(\theta; \epsilon d) = \int \langle \tau, \epsilon d\rangle \exp(\langle \tau, \theta\rangle),$$
$$\langle g, -\epsilon d\rangle \le \rho'(\theta; -\epsilon d) = \int \langle \tau, -\epsilon d\rangle \exp(\langle \tau, \theta\rangle);$$

i.e., $\langle g, d\rangle = \int \langle \tau, d\rangle \exp(\langle \tau, \theta\rangle)$.

If $\theta \in \mathrm{int}(\mathrm{dom}(\rho))$, then repeating the argument above with directions $(\mathbf{e}_i)_{i=1}^n$ gives the desired unique subgradient (hence a gradient). □

The analogous proofs for ψ follow directly.

Proof of Proposition 4.5. For the directional derivative, since $\psi(\theta) = \ln(\theta)$, it suffices to combine $\rho'(\theta; \xi - \theta)$ from Lemma C.3 with the fact that directional derivatives have a chain rule when the outer function is differentiable. For the subgradient and gradient properties, it suffices to repeat the arguments of Lemma C.4 with ψ replacing ρ. □

In order to connect K and B_ψ (i.e., to prove Theorem 4.6), the following standard result is useful.

Lemma C.5. *Given $p, q \in L^1(\nu)$ with $p, q \ge 0$ and $\int p = \int q > 0$, then $\mathsf{K}(p, q) = 0$ iff $p = q$ ν-a.e.*

Proof. If $p = q$, the result is immediate. If $\mathsf{K}(p, q) = 0$, then $p(x)\ln(p(x)/q(x)) = 0$ for ν-almost-every x. Set $S_p := \{x : p(x) > 0\}$, where $\nu(S_p) > 0$ since $\|p\|_1 > 0$. The two preceding statements provide for $S_p' \subseteq S_p$ with $\nu(S_p') = \nu(S_p)$ and $p = q$ on S_p. Combining these statements,

$$\int q = \int p = \int_{S_p'} p = \int_{S_p'} q.$$

As such, $\nu(\{x : q(x) > 0, x \notin S_p'\}) = 0$, and thus $p = q$ ν-a.e. □

Proof of Theorem 4.6. Let any $\theta_1, \theta_2 \in \mathrm{ri}(\mathrm{dom}(\psi))$ be given, and let $\hat\tau_1 \in \partial\psi(\theta_1)$ be arbitrary. Then

$$\begin{aligned}
&\mathsf{K}(p_{\theta_1}, p_{\theta_2})\\
&= \int p_{\theta_1}(x) \ln\left(\frac{p_{\theta_1}(x)}{p_{\theta_2}(x)}\right) d\nu(x)\\
&= \int p_{\theta_1}(x)\Big(\langle \tau(x), \theta_1\rangle - \psi(\theta_1)\\
&\qquad - \langle \tau(x), \theta_2\rangle + \psi(\theta_2)\Big) d\nu(x)\\
&= \psi(\theta_2) - \psi(\theta_1) - \int \langle \tau(x), \theta_2 - \theta_1\rangle\, p_{\theta_1}(x) d\nu(x)\\
&= \psi(\theta_2) - \psi(\theta_1) - \langle \hat\tau_1, \theta_2 - \theta_1\rangle \qquad \text{(C.6)}\\
&= \mathsf{B}_\psi(\theta_2, \theta_1).
\end{aligned}$$

where the penultimate equality used Proposition 4.5, and the final equality used the fact that ψ is relatively differentiable (again via Proposition 4.5), meaning the sup in the subgradient definition of B_f (cf. Proposition 2.3) is attained with every subgradient.

Now let $\hat\tau_2 \in \partial\psi(\theta_2)$ be arbitrary. By two applications of the Fenchel-Young inequality to ψ with elements $(\theta_1, \hat\tau_1)$ and $(\theta_2, \hat\tau_2)$,

$$\begin{aligned}
\text{(C.6)} &= \langle \theta_2, \hat\tau_2\rangle - \psi^*(\hat\tau_2) - \langle \theta_1, \hat\tau_1\rangle + \psi^*(\hat\tau_1)\\
&\quad - \langle \hat\tau_1, \theta_2 - \theta_1\rangle\\
&= \psi^*(\hat\tau_1) - \psi^*(\hat\tau_2) - \langle \hat\tau_1 - \hat\tau_2, \theta_2\rangle\,, \qquad \text{(C.7)}
\end{aligned}$$

where the final statement used $\theta_2 \in \partial\psi^*(\hat\tau_2)$ (since $\hat\tau_2 \in \partial\psi(\theta_2)$ and ψ^* is closed (Rockafellar, 1970, Theorem 23.5).

Now consider any element $\theta_2' \in \partial\psi^*(\hat\tau_2)$ (and associate it with subgradient $\hat\tau_2 \in \partial\psi(\theta_2')$). By the assumption on $\partial\psi^*(\hat\tau_2)$, $\theta_2' \in \mathrm{ri}(\mathrm{dom}(\psi))$, which means the chain of equalities leading to Equation (C.7) can be instantiated with with $(\theta_2', \hat\tau_2)$ in place of $(\theta_1, \hat\tau_1)$, providing

$$\mathsf{K}(p_{\theta_2'}, p_{\theta_2}) = \psi^*(\hat\tau_2) - \psi^*(\hat\tau_2) - \langle \hat\tau_2 - \hat\tau_2, \theta_2\rangle = 0,$$

which by Lemma C.5 means $p_{\theta_2'} = p_{\theta_2}$ ν-a.e. In particular, this provides that $\theta_1 \in \partial\psi^*(\hat\tau_2)$ implies $p_{\theta_1} = p_{\theta_2}$ ν-a.e.

Again taking $\theta_2' \in \partial\psi^*(\hat\tau_2)$ to be arbitrary, $p_{\theta_2'} = p_{\theta_2}$ ν-a.e., Lemma C.5, and the chain of equalities leading to (C.7) yield

$$\begin{aligned}
&\psi^*(\hat\tau_1) - \psi^*(\hat\tau_2) - \langle \hat\tau_1 - \hat\tau_2, \theta_2\rangle\\
&= \mathsf{K}(p_{\theta_1}, p_{\theta_2})\\
&= \mathsf{K}(p_{\theta_1}, p_{\theta_2'})\\
&= \psi^*(\hat\tau_1) - \psi^*(\hat\tau_2) - \langle \hat\tau_1 - \hat\tau_2, \theta_2'\rangle\,.
\end{aligned}$$

As such, it follows that the choice of subgradient within $\partial\psi^*(\hat\tau_2)$ is irrelevant, meaning $\mathsf{K}(p_{\theta_1}, p_{\theta_2}) = \mathsf{B}_{\psi^*}(\hat\tau_1, \hat\tau_2)$. □

Proof of Theorem 4.9. First note that $\hat{\tau}_3 \in \text{ri}(\text{dom}(\psi^*))$, since it is the convex combination of two relative interior elements (Rockafellar, 1970, Theorem 6.1), and further that relative interior elements always have subgradients (Rockafellar, 1970, Theorem 23.4).

Decomposing $\tau(C_3)$ and using bilinearity of inner products,

$$\begin{aligned} 0 &= \sum_{i\in\{1,2\}} |C_i|(\tau(C_i) - \tau(C_1 \cup C_2)) \\ &= \sum_{i\in\{1,2\}} |C_i| \langle \theta_3, \tau(C_i) - \tau(C_1 \cup C_2)\rangle . \end{aligned}$$

Next, for any $i \in \{1,2,3\}$, the Fenchel-Young inequality (Rockafellar, 1970, Theorem 23.5) and decomposition $\hat{\tau}_i = |C|^{-1}\sum_{x\in C_i}\tau(x)$ grant

$$\begin{aligned} \sum_{x\in C_i} \ln p_{\theta_i}(x) &= \sum_{x\in C_i} (\langle \tau(x), \theta_i\rangle - \psi(\theta_i)) \\ &= |C_i|(\langle \hat{\tau}_i, \theta_i\rangle - \psi(\theta_i)) \\ &= |C_i|\psi^*(\hat{\tau}_i). \end{aligned}$$

Therefore, using the form of $\Delta_{\psi^*,\tau}$ from Proposition 3.8, noting that relative differentiability of ψ^* (cf. Proposition 4.5) provides that θ_3 is a maximizing subgradient of ψ^* at $\hat{\tau}_3$, and combining these pieces,

$$\begin{aligned} &\Delta_{\psi^*,\tau}(C_1, C_2) \\ &= \sum_{i\in\{1,2\}} |C_i| \mathsf{B}_{\psi^*}(\hat{\tau}_i, \hat{\tau}_3) \\ &= \sum_{i\in\{1,2\}} |C_i|(\psi^*(\hat{\tau}_i) - \psi^*(\hat{\tau}_3) - \langle \theta_3, \hat{\tau}_i - \hat{\tau}_3\rangle) \\ &= \sum_{i\in\{1,2\}}\sum_{x\in C_i} \ln p_{\theta_i}(x) - \sum_{x\in C_1\cup C_2} \ln p_{\theta_3}(x) - 0. \quad \square \end{aligned}$$

C.2. Deferred examples

Example C.8 (Gaussians)**.** Using the exponential family form of the Gaussian density from Example 4.2, the KL divergence between two Gaussians p, q with mean parameters $(\hat{\tau}_p, \hat{\tau}_q) = ((\mu_p, \Sigma_p + \mu_p\mu_p^\top), (\mu_q, \Sigma_q + \mu_q\mu_q^\top))$ is

$$\begin{aligned} \mathsf{K}(p,q) = \frac{1}{2}\Big(& \ln|\Sigma_q| - \ln|\Sigma_p| - n + \big\langle \Sigma_p, \Sigma_q^{-1}\big\rangle \\ &+ (\mu_q - \mu_p)^\top \Sigma_q^{-1}(\mu_q - \mu_p)\Big). \end{aligned}$$

(By Theorem 4.6, this also gives the appropriate Bregman divergence: $\mathsf{B}_{\psi^*}(\hat{\tau}_p, \hat{\tau}_q) = \mathsf{K}(p,q)$.) As mentioned numerous times, these quantities are not defined for points on the relative boundary of ψ^* (i.e., covariances which are positive semi-definite but not positive definite). ♢

D. Deferred material from Section 5

Proof of Theorem 5.2. Letting $\beta \in (0,1]$ be arbitrary, since τ_0 is a statistic map,

$$\begin{aligned} \beta\tau_0(C) + \alpha z &= \beta|C|^{-1}\sum_{x\in C}\tau_0(x) + \alpha z \\ &= |C|^{-1}\sum_{x\in C}(\beta\tau_0(x) + \alpha z), \end{aligned}$$

whence it follows that τ_1 and τ_2 are statistic maps.

Next, $\tau_1(C) \in \text{ri}(S)$ immediately since a nontrivial convex combination of an element from the relative interior of a set with an element of its closure is always in the relative interior (Rockafellar, 1970, Theorem 6.1).

Now suppose S is a convex cone. Then for any element $z \in \text{ri}(S)$ and $x \in \text{cl}(S)$, the definition of cone provides that $z + x$ and $z + 2x$ are both within $\text{cl}(S)$. But if $z + x \in \text{relbd}(S)$, then $z + 2x \notin \text{cl}(S)$, a contradiction. $\square$

E. Deferred material from Section 6

As stated in Section 6, letting $T_{\Delta_{f,\tau}}$ denote an upper bound on the time to calculate a single merge cost, caching merge cost computations in a min-heap requires space $\mathcal{O}(m^2)$ and time $\mathcal{O}(m^2(\lg(m) + T_{\Delta_{f,\tau}}))$. The initial heap construction takes time $\mathcal{O}(m^2(\lg(m) + T_{\Delta_{f,\tau}}))$. Updating the heap can simply compute and insert $\mathcal{O}(m)$ new merge costs, which takes time $\mathcal{O}(m(\lg(m) + T_{\Delta_{f,\tau}}))$. When extracting the min element, it is possible that invalid pairs are encountered (i.e., pairs where a cluster has already been merged). But across all iterations, at most $\mathcal{O}(m^2)$ invalid pairs are encountered. The total time follows by combining all these terms.

In general, any deterministic scheme may require $\Omega(m^2)$ comparisons (otherwise it can be forced to ignore a crucial pair), but some costs admit an efficient $\mathcal{O}(m^2)$ time and $\mathcal{O}(m)$ space algorithm (Murtagh, 1983). As $\mathcal{O}(m^2)$ space is prohibitive, a heuristic alternative, which was very effective when producing Section 7, is to store a fixed number (much smaller than m) of the smallest merges merges for each cluster.

Lastly, please see Figure 3 for a pair of sample hierarchies over synthetic data.

F. Deferred material from Section 7

Table 2 shows the final log-likelihood of a mixture of ten Gaussians, trained by EM, but initialized in various ways. Amongst the baselines tried, the best was `em-ra`, which chooses random responsibilities (a random E-step), then runs an M-step, and proceeds as

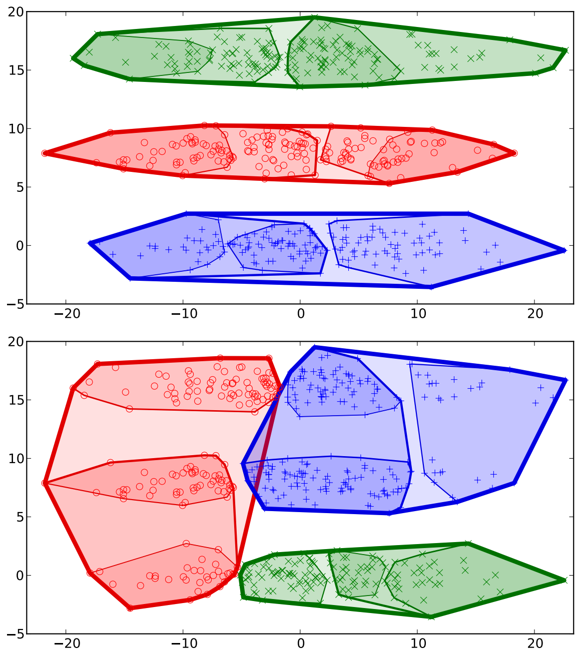


Figure 3. AGGLOMERATE applied to 500 points in 3 long, thin Gaussians; the top instance fit full covariance Gaussians, whereas the bottom instance fit spherical Gaussians (i.e., Ward/k-means cost).

Table 2. Final Gaussian mixture log-likelihood with EM, given various initializations.

	em-ra	km	dg-nd	g-n
glass	-1617.42	-1619.55	**-1614.91**	-1628.61
spam	-156079	-157733	-156311	**-155905**
mnist35	-296491	-296025	**-291731**	-295636

usual. The score reported is the best over ten random trials, whereas the trees, being deterministic, represent a single run.

G. Bibliographic notes

This appendix aims to collect various theorems and results which are generalized by the presented text.

- The Pythagorean theorem for Bregman divergences, presented here in Appendix A.3, appears in many places. The authors have not encountered the approach followed here, which first proves a strong duality result, and obtains the remaining parts as corollaries.

- Corollary 3.6, which provides that centroids minimize Bregman discrepancy, was also proved by Banerjee et al. (2005); the extension to nondifferentiability has little impact on the proof, however the approach here relies on the bias-variance decomposition in Lemma 3.5.

- The basic properties of ψ, captured in Proposition 4.3, follow proofs very similar to those provided by Brown (1986); however, the authors have come across the same proofs in a number of different sources — they appear fairly standard at this point.

- Properties of directional derivatives and subgradients of ψ, which appear in Proposition 4.5, have been derived for the differentiable case in numerous places. The approach here is again very close to the work of Brown (1986), however the application of the dominated convergence theorem follows a general blueprint for passing derivatives through integrals (Folland, 1999, proof of Theorem 2.27).

- The fundamental connections between Bregman divergences and KL divergences for exponential families, Theorem 4.6, was stated for differentiable functions by Azoury & Warmuth (2001). The version here, in addition to handling nondifferentiability, also settles certain identifiability concerns.

- In general, many of the results regarding exponential families appear in some form in the text of Brown (1986) and the text of Wainwright & Jordan (2008); the treatment here removes the need for differentiability, and replaces interiors with relative interiors, which furthermore removes the dependence upon minimality.