SIAM J. Matrix Anal. Appl.
Vol. 29, No. 4, pp. 1120–1146

MATRIX NEARNESS PROBLEMS WITH BREGMAN DIVERGENCES*

INDERJIT S. DHILLON[†] AND JOEL A. TROPP[‡]

Abstract. This paper discusses a new class of matrix nearness problems that measure approximation error using a directed distance measure called a *Bregman divergence*. Bregman divergences offer an important generalization of the squared Frobenius norm and relative entropy, and they all share fundamental geometric properties. In addition, these divergences are intimately connected with exponential families of probability distributions. Therefore, it is natural to study matrix approximation problems with respect to Bregman divergences. This article proposes a framework for studying these problems, discusses some specific matrix nearness problems, and provides algorithms for solving them numerically. These algorithms apply to many classical and novel problems, and they admit a striking geometric interpretation.

Key words. matrix nearness problems, Bregman divergences, squared Euclidean distance, relative entropy, alternating projections

AMS subject classifications. 15A99, 65F30, 90C25

DOI. 10.1137/060649021

1. Introduction.

A recurring problem in matrix theory is to find a structured matrix that best approximates a given matrix with respect to some distance measure. For example, it may be known a priori that a certain constraint ought to hold, and yet it fails on account of measurement errors or numerical roundoff. An attractive remedy is to replace the tainted matrix by the nearest matrix that does satisfy the constraint. Matrix approximation problems typically measure the distance between matrices with a norm. The Frobenius and spectral norms are pervasive choices because they are so analytically tractable. Nevertheless, these norms are not always defensible in applications, where it may be wiser to tailor the distance measure to the context.

In this paper, we discuss a new class of matrix nearness problems that use a directed distance measure called a *Bregman divergence*. Given a differentiable, strictly convex function φ that maps matrices to the extended real numbers, we define the Bregman divergence of the matrix $\boldsymbol{X}$ from the matrix $\boldsymbol{Y}$ as

$$D_\varphi(\boldsymbol{X};\boldsymbol{Y}) \overset{\text{def}}{=} \varphi(\boldsymbol{X}) - \varphi(\boldsymbol{Y}) - \langle \nabla\varphi(\boldsymbol{Y}), \boldsymbol{X}-\boldsymbol{Y}\rangle\,,$$

where the inner product $\langle \boldsymbol{X},\boldsymbol{Y}\rangle = \operatorname{Re}\operatorname{Tr}\boldsymbol{X}\boldsymbol{Y}^*$. The two principal examples of Bregman divergences deserve immediate mention. When $\varphi(\boldsymbol{X}) = \frac{1}{2}\|\boldsymbol{X}\|_F^2$, the associated divergence is the squared Frobenius norm $\frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Y}\|_F^2$. When φ is the negative Shannon entropy, we obtain the Kullback–Leibler divergence, which is also known as relative entropy. But these two cases are just the tip of the iceberg.

Bregman divergences are well suited for nearness problems because they share many geometric properties with the squared Frobenius norm. They also exhibit an

*Received by the editors January 4, 2006; accepted for publication (in revised form) by N. J. Higham March 20, 2007; published electronically November 21, 2007.
http://www.siam.org/journals/simax/29-4/64902.html

[†]Department of Computer Sciences, University of Texas, Austin, TX 78712-1188 (inderjit@cs.utexas.edu). This author's research was supported by NSF grant CCF-0431257, NSF career award ACI-0093404, and NSF-ITR award IIS-0325116.

[‡]Applied and Computational Mathematics, California Institute of Technology, Pasadena, CA 91125-5000 (jtropp@acm.caltech.edu). This author's research was supported by an NSF graduate fellowship.

intimate relationship with exponential families of probability distributions, which recommends them for solving problems that arise in the statistical analysis of data. We will elaborate on these connections in what follows.

Let us begin with a formal statement of the Bregman nearness problem. Suppose that D_φ is a Bregman divergence, and suppose that $\{C_k\}$ is a finite collection of closed, convex sets whose intersection is nonempty. Given an input matrix $\boldsymbol{Y}$, our goal is to produce a matrix $\boldsymbol{X}$ in the intersection that diverges the least from $\boldsymbol{Y}$, i.e., to solve

$$\min_{\boldsymbol{X}} \; D_\varphi(\boldsymbol{X};\boldsymbol{Y}) \qquad \text{subject to} \qquad \boldsymbol{X} \in \bigcap\nolimits_k C_k. \tag{1.1}$$

Under mild conditions, the solution to (1.1) is unique, and it has a variational characterization analogous with the characterization of an orthogonal projection onto a convex set [10]. Minimization with respect to the second argument of the divergence enjoys rather less structure, so we refer the reader to [5] for more details. A major advantage of our problem formulation is that it admits a natural algorithm. If one possesses a method for minimizing the divergence over each of the constraint sets, then it is possible to solve (1.1) by minimizing over each constraint in turn while introducing a series of simple corrections. Several classical algorithms from the matrix literature fit into this geometric framework, but it also provides an approach to many novel problems.

We view this paper as an expository work with two central goals. First, it introduces Bregman divergences to the matrix theory literature, and it argues that they provide an important and natural class of distance measures for matrix nearness problems. Moreover, the article unifies a large class of problems into a geometrical framework, and it shows that these problems can be solved with a set of classical algorithms. Second, the paper provides specific examples of nearness problems with respect to Bregman divergences. One example is the familiar problem of producing the nearest contingency table with fixed marginals. Novel examples include computing matrix approximations using the minimum Bregman information (MBI) principle, identifying the metric graph nearest to an arbitrary graph, and determining the nearest correlation and kernel matrix with respect to matrix divergences, such as the von Neumann divergence. These applications show how Bregman divergences can be used to preserve and exploit additional structure that appears in a problem.

We must warn the reader that, in spite of the availability of some general purpose algorithms for working with Bregman divergences, they may require a substantial amount of computational effort. One basic reason is that nearness problems with respect to the Frobenius norm usually remain within the domain of linear algebra, which is a developed technology. Bregman divergences, on the other hand, transport us to the world of convex optimization, which is a rougher frontier. As outlined in section 8, there remain many unresolved research issues on the computational aspects of Bregman divergences.

Here is a brief outline of the article. Section 2 introduces Bregman divergences and Bregman projections along with their connection to exponential families of probability distributions. Matrix Bregman divergences that depend on the spectral properties of a matrix are covered in subsection 2.6. Section 3 discusses numerical methods for the basic problem of minimizing a Bregman divergence over a hyperplane. In section 4, we develop the successive projection algorithm for solving the Bregman nearness problem subject to affine constraints. Section 5 gives several examples of these problems: finding the nearest contingency table with fixed marginals, computing matrix approximations for data analysis, and determining the nearest correlation matrix with respect to the von Neumann divergence. Section 6 presents the successive projection–correction algorithm for solving the Bregman nearness problem subject to a polyhedral constraint. In section 7 we discuss two matrix nearness problems with nonaffine constraints: finding the nearest metric graph and learning a kernel matrix for data mining and machine learning applications.

2. Bregman divergences and Bregman projections.

This section develops the directed distance functions that were first studied by Bregman [8]. Our primary source is the superb article of Bauschke and Borwein [4], which studies a subclass of Bregman divergences that exhibits many desirable properties in connection with nearness problems like (1.1).

2.1. Convex analysis.

The literature on Bregman divergences involves a significant amount of convex analysis. Some standard references for this material are [35, 20]. We review some of these ideas in an effort to make this article accessible to readers who are less familiar with this field.

We will work in a finite-dimensional, real inner-product space $\mathscr{X}$. The real-linear inner product is denoted by $\langle \cdot, \cdot \rangle$ and the induced norm by $\|\cdot\|_2$. In general, the elements of $\mathscr{X}$ will be expressed with lowercase bold italic letters such as $\boldsymbol{x}$ and $\boldsymbol{y}$. We will switch to capitals, such as $\boldsymbol{X}$ and $\boldsymbol{Y}$, when it is important to view the elements of $\mathscr{X}$ as matrices.

A *convex set* is a subset C of $\mathscr{X}$ that exhibits the property

$$s\,\boldsymbol{x} + (1-s)\,\boldsymbol{y} \in C \qquad \text{for all } s \in (0,1) \text{ and } \boldsymbol{x}, \boldsymbol{y} \in C.$$

In words, the line segment connecting each pair of points in a convex set falls within the set. The *relative interior* of a convex set, abbreviated ri, is the interior of that set considered as a subset of the lowest-dimensional affine space that contains it.

In convex analysis, functions are defined on all of $\mathscr{X}$, and they take values in the extended real numbers, $\mathbb{R} \cup \{\pm\infty\}$. The *(effective) domain* of a function f is the set

$$\operatorname{dom} f \overset{\text{def}}{=} \{\boldsymbol{x} \in \mathscr{X} : f(\boldsymbol{x}) < +\infty\}.$$

A function f is *convex* if its domain is convex and it verifies the inequality

$$f(s\,\boldsymbol{x} + (1-s)\,\boldsymbol{y}) \le s\,f(\boldsymbol{x}) + (1-s)\,f(\boldsymbol{y}) \quad \text{for all } s \in (0,1) \text{ and } \boldsymbol{x}, \boldsymbol{y} \in \operatorname{dom} f.$$

If the inequality is strict, then f is *strictly convex*. In words, the chord connecting each pair of points on the graph of a (strictly) convex function lies (strictly) above the graph. A convex function is *proper* if it takes at least one finite value and never takes the value $-\infty$. A convex function f is *closed* if its lower level set $\{\boldsymbol{x} : f(\boldsymbol{x}) \le \alpha\}$ is closed for each real α. In particular, a convex function is closed whenever its domain is closed (but not conversely).

For completeness, we also introduce some technical definitions that the casual reader may prefer to glide through. A proper convex function f is called *essentially smooth* if it is everywhere differentiable on the (nonempty) interior of its domain and if $\|\nabla f(\boldsymbol{x}_t)\|$ tends to infinity for every sequence $\{\boldsymbol{x}_t\}$ from $\operatorname{ri}(\operatorname{dom} f)$ that converges to a point on the boundary of $\operatorname{dom} f$. Roughly speaking, an essentially smooth function cannot be extended to a convex function with a larger domain. The function $f(x) = -\log(x)$ with domain $(0, +\infty)$ is an example of an essentially smooth function. In what follows, we will focus on convex functions of *Legendre type.* A Legendre function is a closed, proper, convex function that is essentially smooth and also strictly convex on the relative interior of its domain.

Every convex function has a dual representation in terms of its supporting hyperplanes. This idea is formalized in the *Fenchel conjugate*, which is defined as

$$f^*(\boldsymbol{\theta}) \overset{\text{def}}{=} \sup_{\boldsymbol{x}} \left\{ \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle - f(\boldsymbol{x}) \right\}.$$

No confusion should arise from our usage of the symbol * for complex-conjugate transposition as well as Fenchel conjugation. The following facts are valuable. The conjugate of a convex function is always closed and convex. If f is a closed, convex function, then $(f^*)^* = f$. A convex function has Legendre type if and only if its conjugate has Legendre type.

Finally, we say that a convex function f is *cofinite* when

$$\lim_{\xi \to \infty} f(\xi\, \boldsymbol{x})/\xi = +\infty \qquad \text{for all nonzero } \boldsymbol{x} \text{ in } \mathscr{X}.$$

This definition means that a cofinite function grows superlinearly in every direction. For example, the function $\|\cdot\|_2^2$ is cofinite, but the function $\exp(\cdot)$ is not. It can be shown that a closed, proper, convex function f is cofinite if and only if $\operatorname{dom} f^* = \mathscr{X}$.

2.2. Divergences. Suppose that φ is a convex function of Legendre type. From every such seed function, we may construct a Bregman divergence[1]

$$D_\varphi : \operatorname{dom} \varphi \times \operatorname{ri}(\operatorname{dom} \varphi) \to [0, +\infty)$$

via the rule

$$D_\varphi(\boldsymbol{x}; \boldsymbol{y}) \overset{\text{def}}{=} \varphi(\boldsymbol{x}) - \varphi(\boldsymbol{y}) - \langle \nabla \varphi(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y} \rangle .$$

Geometrically, the divergence calculates how much the supporting hyperplane to φ at $\boldsymbol{y}$ underestimates the value of $\varphi(\boldsymbol{x})$. For an illustration, see Figure 2.1. A Bregman divergence equals zero whenever $\boldsymbol{x} = \boldsymbol{y}$, and it is positive otherwise. It is strictly convex in its first argument, and it is jointly continuous in both arguments.

As a first example, consider the seed function $\varphi(\boldsymbol{x}) = \frac{1}{2} \|\boldsymbol{x}\|_2^2$, which is a Legendre function on all of $\mathscr{X}$. The associated divergence is

$$D_\varphi(\boldsymbol{x}; \boldsymbol{y}) = \tfrac{1}{2} \|\boldsymbol{x} - \boldsymbol{y}\|_2^2 .$$

We will refer to this function as the Euclidean divergence. Observe that it is symmetric in its two arguments, but it does not satisfy the triangle inequality.

Another basic example arises from the negative Shannon entropy,

$$\varphi(\boldsymbol{x}) = \sum\nolimits_n x_n \log x_n - x_n,$$

where we place the convention that $0 \log 0 = 0$. This entropy is a Legendre function on the nonnegative orthant, and it yields the divergence

$$D_\varphi(\boldsymbol{x}; \boldsymbol{y}) = \sum\nolimits_n \left[x_n \log \frac{x_n}{y_n} - x_n + y_n \right], \tag{2.1}$$

[1] It is also possible to define Bregman divergences with respect to any differentiable, strictly convex function. These divergences are not necessarily well behaved.

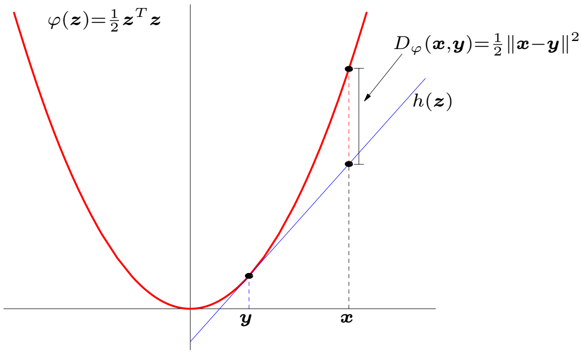
$\varphi(\boldsymbol{z})=\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}$
$D_\varphi(\boldsymbol{x},\boldsymbol{y})=\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|^2$
$h(\boldsymbol{z})$
$\boldsymbol{y}$
$\boldsymbol{x}$

FIG. 2.1. *An example of a Bregman divergence is the squared Euclidean distance. The Bregman divergence $D_\varphi(\boldsymbol{x};\boldsymbol{y})$ calculates how much the supporting hyperplane to φ at $\boldsymbol{y}$ underestimates the value of $\varphi(\boldsymbol{x})$.*

which is variously called the *relative entropy*, the *information divergence*, or the *generalized Kullback–Leibler divergence*. This divergence is not symmetric, and it does not satisfy the triangle inequality.

Bregman divergences are often referred to as *Bregman distances*, but this terminology is misleading. A Bregman divergence should not be viewed as a generalization of a metric but rather as a generalization of the preceding two examples. Like a metric, every Bregman divergence is positive except when its arguments coincide. On the other hand, divergences do not generally satisfy the triangle inequality, and they are symmetric only when the seed function φ is quadratic. In compensation, divergences exhibit other structural properties. For every three points in the interior of $\operatorname{dom}\varphi$, we have the relation

$$D_\varphi(\boldsymbol{x};\boldsymbol{z}) = D_\varphi(\boldsymbol{x};\boldsymbol{y}) + D_\varphi(\boldsymbol{y};\boldsymbol{z}) - \langle \nabla\varphi(\boldsymbol{z}) - \nabla\varphi(\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle\,.$$

When D_φ is the Euclidean divergence, one may identify this formula as the law of cosines. Later, we will also encounter a Pythagorean theorem.

We also note another expression for the divergence, which emphasizes that it is a sort of locally quadratic distance measure,

$$D_\varphi(\boldsymbol{x};\boldsymbol{y}) = (\boldsymbol{x}-\boldsymbol{y})^*\left\{\nabla^2\varphi(\boldsymbol{\xi})\right\}(\boldsymbol{x}-\boldsymbol{y}),$$

where $\boldsymbol{\xi}$ is an unknown vector that depends on $\boldsymbol{x}$ and $\boldsymbol{y}$. This formula can be obtained from the Taylor expansion of the seed function with an exact remainder term.

2.3. Exponential families. Suppose that ψ is a Legendre function. A *(full) regular exponential family* is a parameterized family of probability distributions on $\mathscr{X}$ with density function (with respect to the Lebesgue measure on $\mathscr{X}$) of the form

$$p_\psi(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = \exp\{\langle \boldsymbol{x},\boldsymbol{\theta}\rangle - \psi(\boldsymbol{\theta}) - h(\boldsymbol{x})\},$$

where the parameter $\boldsymbol{\theta}$ is drawn from the open set $\operatorname{dom}\psi$ [3]. The function ψ is called the *cumulant function* of the exponential family, and it completely determines the function h. The *expectation* of the distribution $p_\psi(\,\cdot\,|\,\boldsymbol{\theta})$ is the vector

$$\boldsymbol{\mu}(\boldsymbol{\theta}) \overset{\text{def}}{=} \int_{\mathscr{X}} \boldsymbol{x}\, p_\psi(\boldsymbol{x}\,|\,\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{x},$$

where $\mathrm{d}\boldsymbol{x}$ denotes the Lebesgue measure on $\mathscr{X}$. Many common probability distributions belong to exponential families. Prominent examples include Gaussian, Poisson, Bernoulli, and gamma distributions.

It has recently been established that there is a unique Bregman divergence that corresponds to every regular exponential family.

THEOREM 1 (Banerjee et al. [2]). *Suppose that φ and ψ are conjugate Legendre functions. Let D_φ be the Bregman divergence associated with φ, and let $p_\psi(\,\cdot \mid \boldsymbol{\theta})$ be a member of the regular exponential family with cumulant function ψ. Then*

$$p_\psi(\boldsymbol{x} \mid \boldsymbol{\theta}) = \exp\{-D_\varphi(\boldsymbol{x}; \boldsymbol{\mu}(\boldsymbol{\theta}))\}\, g_\varphi(\boldsymbol{x}),$$

where g_φ is a function uniquely determined by φ.

The spherical Gaussian distribution provides an especially interesting example of this relationship [2]. Suppose that $\boldsymbol{\mu}$ is an arbitrary vector in $\mathscr{X}$, and let σ^2 be a fixed positive number. The spherical Gaussian distributions with mean $\boldsymbol{\mu}$ and variance σ^2 form an exponential family with parameter $\boldsymbol{\theta} = \boldsymbol{\mu}/\sigma^2$ and cumulant function $\psi(\boldsymbol{\theta}) = \frac{\sigma^2}{2} \|\boldsymbol{\theta}\|_2^2$. The Fenchel conjugate of the cumulant function is $\varphi(\boldsymbol{x}) = \frac{1}{2\sigma^2} \|\boldsymbol{x}\|_2^2$, and so the Bregman divergence that appears in the bijection theorem is

$$D_\varphi(\boldsymbol{x}; \boldsymbol{\mu}) = \frac{1}{2\sigma^2} \|\boldsymbol{x} - \boldsymbol{\mu}\|_2^2 .$$

We see that the density of the distribution at a point $\boldsymbol{x}$ depends essentially on the Bregman divergence of $\boldsymbol{x}$ from the mean vector $\boldsymbol{\mu}$. This observation reinforces the intuition that the squared Euclidean norm enjoys a profound relationship with Gaussian random variables.

2.4. Bregman projections. Suppose that φ is a convex function of Legendre type, and let C be a closed, convex set that intersects $\mathrm{ri}(\mathrm{dom}\,\varphi)$. Given a point $\boldsymbol{y}$ from $\mathrm{ri}(\mathrm{dom}\,\varphi)$, we may pose the minimization problem

$$\min_{\boldsymbol{x}} \ D_\varphi(\boldsymbol{x}; \boldsymbol{y}) \qquad \text{subject to} \qquad \boldsymbol{x} \in C \cap \mathrm{ri}(\mathrm{dom}\,\varphi). \tag{2.2}$$

Since $D_\varphi(\,\cdot\,; \boldsymbol{y})$ is strictly convex, it follows from a standard argument that there exists *at most* one minimizer. It can be shown that, when φ is a Legendre function, there exists *at least* one minimizer [4, Theorem 3.12]. Therefore, the problem (2.2) has a single solution, which is called the *Bregman projection* of $\boldsymbol{y}$ onto C with respect to the divergence D_φ. Denote this solution by $P_C(\boldsymbol{y})$, and observe that we have defined a map

$$P_C : \mathrm{ri}(\mathrm{dom}\,\varphi) \to C \cap \mathrm{ri}(\mathrm{dom}\,\varphi).$$

It is evident that P_C acts as the identity on $C \cap \mathrm{ri}(\mathrm{dom}\,\varphi)$, and it can be shown that P_C is continuous.

There is also a variational characterization of the Bregman projection of a point $\boldsymbol{y}$ from $\mathrm{ri}(\mathrm{dom}\,\varphi)$ onto the set C,

$$D_\varphi(\boldsymbol{x}; \boldsymbol{y}) \geq D_\varphi(\boldsymbol{x}; P_C(\boldsymbol{y})) + D_\varphi(P_C(\boldsymbol{y}); \boldsymbol{y}) \qquad \text{for every } \boldsymbol{x} \in C \cap \mathrm{dom}\,\varphi. \tag{2.3}$$

Conversely, suppose we replace $P_C(\boldsymbol{y})$ with an arbitrary point $\boldsymbol{z}$ from $C \cap \mathrm{ri}(\mathrm{dom}\,\varphi)$ that verifies the inequality. Then $\boldsymbol{z}$ must indeed be the Bregman projection of $\boldsymbol{y}$ onto

C. When the constraint C is an affine space (i.e., a translated subspace), then the Bregman projection of $\boldsymbol{y}$ onto C has a formally stronger characterization,

$$D_\varphi(\boldsymbol{x};\boldsymbol{y}) = D_\varphi(\boldsymbol{x}; P_C(\boldsymbol{y})) + D_\varphi(P_C(\boldsymbol{y}); \boldsymbol{y}) \qquad \text{for every } \boldsymbol{x} \in C \cap \operatorname{dom}\varphi. \tag{2.4}$$

When the Bregman divergence is the Euclidean divergence, formula (2.3) reduces to the criterion for identifying the orthogonal projection onto a convex set [14, Chapter 4], while formula (2.4) is usually referred to as the Pythagorean theorem. These facts justify the assertion that Bregman projections generalize orthogonal projections.

When the constraint set C and the Bregman divergence are simple enough, it may be possible to determine the Bregman projection onto C analytically. For example, let us define the hyperplane $C = \{\boldsymbol{x} : \langle \boldsymbol{a}, \boldsymbol{x}\rangle = \alpha\}$. When $\|\boldsymbol{a}\|_2 = 1$, the projection of $\boldsymbol{y}$ onto C with respect to the Euclidean divergence is

$$P_C(\boldsymbol{y}) = \boldsymbol{y} - (\langle \boldsymbol{a}, \boldsymbol{y}\rangle - \alpha)\,\boldsymbol{a}. \tag{2.5}$$

As a second example, suppose that C contains a strictly positive vector and that $\boldsymbol{y}$ is strictly positive. Using Lagrange multipliers, we check that the projection of $\boldsymbol{y}$ onto C with respect to the relative entropy has components

$$(P_C(\boldsymbol{y}))_n = y_n \exp\{\xi\, a_n\}, \quad \text{where } \xi \text{ is chosen so that } P_C(\boldsymbol{y}) \in C. \tag{2.6}$$

In the case when all the components of $\boldsymbol{a}$ are identical (to one, without loss of generality), then $\xi = \log\alpha - \log\sum_n y_n$.

It is uncommon that a Bregman projection can be explicitly determined. In section 3, we describe numerical methods for computing the Bregman projection onto a hyperplane, which is the foundation for producing Bregman projections onto more complicated sets. For another example of a projection that can be computed analytically, turn to the end of subsection 3.3.

2.5. A cornucopia of divergences. In this subsection, we will present some important Bregman divergences. The *separable* divergences form the most fundamental class. A separable divergence arises from a seed function of the form

$$\varphi(\boldsymbol{x}) = \sum\nolimits_n w_n\, \varphi_n(x_n) \qquad \text{for positive weights } w_n.$$

If each φ_n is Legendre, then the weighted sum is also Legendre. In the most common situation, the weights are constant and all the φ_n are identical. In Table 2.1 we list some important Legendre functions on $\mathbb{R}$ that may be used to build separable divergences. These examples are adapted from [4] and [2]. Several of the divergences in Table 2.1 have names. We have already discussed the Euclidean divergence and the relative entropy. The bit entropy leads to a type of logistic loss, and the Burg entropy leads to the Itakura–Saito divergence.

Many of these univariate divergences are connected with well-known exponential families of probability distributions on $\mathbb{R}$. See Table 2.2 for some key examples drawn from [2].

One fundamental divergence is genuinely multidimensional. Suppose that $\boldsymbol{Q}$ is a positive-definite operator that acts on $\mathscr{X}$. We may construct a quadratic divergence on $\mathscr{X}$ from the seed function $\varphi(\boldsymbol{x}) = \frac{1}{2}\langle \boldsymbol{Q}\,\boldsymbol{x}, \boldsymbol{x}\rangle$, resulting in

$$D_\varphi(\boldsymbol{x};\boldsymbol{y}) = \tfrac{1}{2}\langle \boldsymbol{Q}\,(\boldsymbol{x}-\boldsymbol{y}), \boldsymbol{x}-\boldsymbol{y}\rangle\,.$$

TABLE 2.1
Common seed functions and the corresponding divergences.

Function name	$\varphi(x)$	$\operatorname{dom}\varphi$	$D_\varphi(x;y)$
Squared norm	$\frac{1}{2}x^2$	$(-\infty,+\infty)$	$\frac{1}{2}(x-y)^2$
Shannon entropy	$x\log x - x$	$[0,+\infty)$	$x\log\frac{x}{y} - x + y$
Bit entropy	$x\log x + (1-x)\log(1-x)$	$[0,1]$	$x\log\frac{x}{y} + (1-x)\log\frac{1-x}{1-y}$
Burg entropy	$-\log x$	$(0,+\infty)$	$\frac{x}{y} - \log\frac{x}{y} - 1$
Hellinger	$-\sqrt{1-x^2}$	$[-1,1]$	$(1-xy)(1-y^2)^{-1/2} - (1-x^2)^{1/2}$
ℓ_p quasi-norm	$-x^p \qquad (0<p<1)$	$[0,+\infty)$	$-x^p + p\,xy^{p-1} - (p-1)\,y^p$
ℓ_p norm	$\lvert x\rvert^p \qquad (1<p<\infty)$	$(-\infty,+\infty)$	$\lvert x\rvert^p - p\,x\operatorname{sgn}y\,\lvert y\rvert^{p-1} + (p-1)\,\lvert y\rvert^p$
Exponential	$\exp x$	$(-\infty,+\infty)$	$\exp x - (x-y+1)\exp y$
Inverse	$1/x$	$(0,+\infty)$	$1/x + x/y^2 - 2/y$

TABLE 2.2
Common exponential families and the corresponding divergences.

Exponential family	$\psi(\theta)$	$\operatorname{dom}\psi$	$\mu(\theta)$	$\varphi(x)$	Divergence
Gaussian (σ^2 fixed)	$\frac{1}{2}\sigma^2\theta^2$	$(-\infty,+\infty)$	$\sigma^2\theta$	$\frac{1}{2\sigma^2}x^2$	Euclidean
Poisson	$\exp\theta$	$(-\infty,+\infty)$	$\exp\theta$	$x\log x - x$	Relative entropy
Bernoulli	$\log(1+\exp\theta)$	$(-\infty,+\infty)$	$\frac{\exp\theta}{1+\exp\theta}$	$x\log x + (1-x)\log(1-x)$	Logistic loss
Gamma (α fixed)	$-\alpha\,\log(-\theta)$	$(-\infty,0)$	$-\alpha/\theta$	$-\alpha\log x + \alpha\log\alpha - \alpha$	Itakura–Saito

This divergence is connected to the exponential family of multivariate Gaussian distributions with covariance matrix $\boldsymbol{Q}^{-1}$. In the latter context, the square root of this divergence is often referred to as the *Mahalanobis distance* in statistics [29]. Other multidimensional examples arise when we compose the Euclidean norm with another function. For instance, one might consider the convex function $\varphi(\boldsymbol{x}) = -\sqrt{1 - \|\boldsymbol{x}\|_2^2}$ defined on the Euclidean unit ball. It yields the Hellinger-like divergence

$$D_\varphi(\boldsymbol{x}; \boldsymbol{y}) = \frac{1 - \langle \boldsymbol{x}, \boldsymbol{y} \rangle}{\sqrt{1 - \|\boldsymbol{y}\|_2^2}} - \sqrt{1 - \|\boldsymbol{x}\|_2^2}.$$

2.6. Matrix divergences. Hermitian matrices admit a rich variety of divergences that were first studied in [4] using the methods of Lewis [27]. Let $\mathscr{H}$ be the space of $N \times N$ Hermitian matrices equipped with the real-linear inner product $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \operatorname{Re} \operatorname{Tr} \boldsymbol{X}\boldsymbol{Y}^*$. Define the function $\boldsymbol{\lambda} : \mathscr{H} \to \mathbb{R}^N$ that maps a Hermitian matrix to the vector listing its eigenvalues in algebraically decreasing order. Let φ be a closed, proper, convex function on $\mathbb{R}^N$ that is invariant under coordinate permutation. That is, $\varphi(\boldsymbol{x}) = \varphi(\boldsymbol{P}\boldsymbol{x})$ for every permutation matrix $\boldsymbol{P}$.

By composing φ with the eigenvalue map, we induce a real-valued function on Hermitian matrices. As the following theorem elaborates, the induced map has the same convexity properties as the function φ. Therefore, the induced map can be used as a seed function to define a Bregman divergence on the space of Hermitian matrices.

THEOREM 2 (Lewis [27, 26]). *The induced map $\varphi \circ \boldsymbol{\lambda}$ has the following properties:*

1. *If φ is closed and convex, then the induced map is closed and convex.*
2. *The domain of $\varphi \circ \boldsymbol{\lambda}$ is the inverse image under $\boldsymbol{\lambda}$ of* $\operatorname{dom} \varphi$.
3. *The conjugate of the induced map satisfies the relation $(\varphi \circ \boldsymbol{\lambda})^* = \varphi^* \circ \boldsymbol{\lambda}$.*
4. *The induced map is differentiable at $\boldsymbol{X}$ if and only if φ is differentiable at $\boldsymbol{\lambda}(\boldsymbol{X})$. If $\boldsymbol{X}$ has eigenvalue decomposition $\boldsymbol{U}\{\operatorname{diag} \boldsymbol{\lambda}(\boldsymbol{X})\}\boldsymbol{U}^*$, then*

   $$\nabla(\varphi \circ \boldsymbol{\lambda})(\boldsymbol{X}) = \boldsymbol{U}\{\operatorname{diag} \nabla\varphi(\boldsymbol{\lambda}(\boldsymbol{X}))\}\boldsymbol{U}^*.$$

   In fact, this formula holds even if φ is not convex.
5. *The induced map is Legendre if and only if φ is Legendre.*

Related results hold for the singular value map, provided that φ is also absolutely invariant. That is, $\varphi(\boldsymbol{x}) = \varphi(|\boldsymbol{x}|)$ for all $\boldsymbol{x}$ in $\mathbb{R}^N$, where $|\cdot|$ is the componentwise absolute value.

Unitarily invariant matrix norms provide the most basic examples of induced maps. Indeed, item 3 of the last theorem generalizes von Neumann's famous result about dual norms of unitarily invariant prenorms [21, 438ff.].

An exquisite example of a matrix divergence arises from $\varphi(\boldsymbol{x}) = -\sum_n \log x_n$. The induced map is $(\varphi \circ \boldsymbol{\lambda})(\boldsymbol{X}) = -\log\det \boldsymbol{X}$, whose domain is the positive-definite cone. Since $\nabla(\varphi \circ \boldsymbol{\lambda})(\boldsymbol{X}) = -\boldsymbol{X}^{-1}$, the resulting divergence is

$$D_{\ell d}(\boldsymbol{X}; \boldsymbol{Y}) = \left\langle \boldsymbol{X}, \boldsymbol{Y}^{-1} \right\rangle - \log\det \boldsymbol{X}\boldsymbol{Y}^{-1} - N. \tag{2.7}$$

Intriguingly, certain projections with respect to this divergence can be computed analytically. See subsection 3.3 for details.

Another important example arises from the negative Shannon entropy $\varphi(\boldsymbol{x}) = \sum_n x_n \log x_n - x_n$. The induced map is $(\varphi \circ \boldsymbol{\lambda})(\boldsymbol{X}) = \operatorname{Tr}(\boldsymbol{X} \log \boldsymbol{X} - \boldsymbol{X})$, whose domain is the positive-semidefinite cone. This matrix function arises in quantum mechanics, where it is referred to as the *von Neumann entropy* [31]. It yields the divergence

$$D_{\mathrm{vN}}(\boldsymbol{X}; \boldsymbol{Y}) = \operatorname{Tr}[\boldsymbol{X}(\log \boldsymbol{X} - \log \boldsymbol{Y}) - \boldsymbol{X} + \boldsymbol{Y}], \tag{2.8}$$

which we will call the von Neumann divergence. In the quantum mechanics literature, this divergence is referred to as the *quantum relative entropy* [31]. This formula does not literally hold if either matrix is singular, but a limit argument shows that the divergence is finite precisely when the null space of $\boldsymbol{X}$ contains the null space of $\boldsymbol{Y}$.

When the seed function φ is separable, matrix divergences can be expressed in a way that emphasizes the distinct roles of the eigenvalues and eigenvectors. In particular, take $\varphi(\boldsymbol{x}) = \sum_n \varphi(x_n)$ and assume that $\boldsymbol{X}$ has eigenpairs $(\boldsymbol{u}_m, \mu_m)$ and that $\boldsymbol{Y}$ has eigenpairs $(\boldsymbol{v}_n, \nu_n)$. Then

$$\begin{aligned} D_{\varphi\circ\boldsymbol{\lambda}}(\boldsymbol{X}; \boldsymbol{Y}) &= \sum\nolimits_{m,n} |\langle \boldsymbol{u}_m, \boldsymbol{v}_n \rangle|^2 \left[\varphi(\mu_m) - \varphi(\nu_n) - \varphi'(\nu_n)(\mu_m - \nu_n)\right] \\ &= \sum\nolimits_{m,n} |\langle \boldsymbol{u}_m, \boldsymbol{v}_n \rangle|^2 D_\varphi(\mu_m; \nu_n). \end{aligned}$$

In words, the matrix divergence adds up the scalar divergences between pairs of eigenvalues, weighted by the squared cosine of the angle between the corresponding eigenvectors.

3. Computing Bregman projections.

It is not straightforward to compute the Bregman projection onto a general convex set. Unless additional structure is present, the best approach may be to apply standard convex optimization techniques. In this section, we discuss how to develop numerical methods for the basic problem of projecting onto a hyperplane or a halfspace. As we will see in sections 4 and 6, the projection onto an intersection of convex sets can be broken down into a sequence of projections onto the individual sets. Combining the two techniques, we can find the projection onto any affine space or polyhedral convex set.

3.1. Projection onto a hyperplane.

There is an efficient way to compute the Bregman projection onto a hyperplane. The key idea is to dualize the Bregman projection problem to obtain a nice one-dimensional problem. This approach can also be extended to produce the projection onto a halfspace because the convexity of the divergence implies that the projection lies on the boundary whenever the initial point is outside the halfspace.

We must solve the following convex program:

$$\min_{\boldsymbol{x}} D_\varphi(\boldsymbol{x}; \boldsymbol{y}) \qquad \text{subject to} \qquad \langle \boldsymbol{a}, \boldsymbol{x} \rangle = \alpha. \tag{3.1}$$

To ensure that this problem is well posed, we assume that $\operatorname{ri}(\operatorname{dom}\varphi)$ contains a feasible point. A necessary and sufficient condition on the solution $\boldsymbol{x}_\star$ of (3.1) is that the equation

$$\nabla_{\boldsymbol{x}} D_\varphi(\boldsymbol{x}; \boldsymbol{y}) = \xi \, \nabla_{\boldsymbol{x}} \left(\langle \boldsymbol{a}, \boldsymbol{x} \rangle - \alpha\right)$$

hold for a (unique) Lagrange multiplier $\xi \in \mathbb{R}$. The gradient of the divergence is $\nabla\varphi(\boldsymbol{x}) - \nabla\varphi(\boldsymbol{y})$, resulting in the equation

$$\nabla\varphi(\boldsymbol{x}_\star) = \xi \boldsymbol{a} + \nabla\varphi(\boldsymbol{y}).$$

The gradient of a Legendre function φ is a bijection from $\operatorname{dom}\varphi$ to $\operatorname{dom}\varphi^*$, and its inverse is the gradient of the conjugate [35, Thm. 26.5]. Thus we obtain an explicit expression for the Bregman projection as a function of the unknown multiplier:

$$\boldsymbol{x}_\star = \nabla\varphi^*(\xi\boldsymbol{a} + \nabla\varphi(\boldsymbol{y})). \tag{3.2}$$

Form the inner product with $\boldsymbol{a}$ and enforce the constraint to reach

$$\langle \nabla\varphi^*(\xi\boldsymbol{a} + \nabla\varphi(\boldsymbol{y})), \boldsymbol{a}\rangle - \alpha = 0. \tag{3.3}$$

Now, the left-hand side of this equation is the derivative of the strictly convex, univariate function

$$J(\xi) = \varphi^*(\xi\boldsymbol{a} + \nabla\varphi(\boldsymbol{y})) - \alpha\xi. \tag{3.4}$$

There is an implicit constraint that the argument of φ^* must lie within its domain. In view of (3.3), it becomes clear that the Lagrange multiplier is the unique minimizer of J. That is,

$$\xi_\star = \arg\min\nolimits_\xi J(\xi).$$

Once we have determined the Lagrange multiplier, we introduce it into (3.2) to obtain the Bregman projection.

The best numerical method for minimizing J depends strongly on the choice of the seed function φ. In some cases, the derivative(s) of J may be difficult to evaluate. The second derivative may even fail to exist. To that end, we offer several observations that may be valuable.

1. The domain of J contains a neighborhood of zero since $J(0) = \langle \boldsymbol{y}, \nabla\varphi(\boldsymbol{y})\rangle - \varphi(\boldsymbol{y})$.
2. Since φ^* is a Legendre function, the first derivative of J always exists. As shown in (3.3),
$$J'(\xi) = \langle \nabla\varphi^*(\xi\boldsymbol{a} + \nabla\varphi(\boldsymbol{y})), \boldsymbol{a}\rangle - \alpha.$$
3. When the Hessian of φ^* exists, we have
$$J''(\xi) = \boldsymbol{a}^* \left\{\nabla^2\varphi^*(\xi\boldsymbol{a} + \nabla\varphi(\boldsymbol{y}))\right\} \boldsymbol{a}.$$
4. When the seed function φ is separable, the Hessian $\nabla^2\varphi^*$ is diagonal.

The next two subsections provide examples that illustrate some of the issues involved in optimizing J.

3.2. Example: Relative entropy. Suppose that we wish to produce the Bregman projection of a nonnegative vector $\boldsymbol{y}$ onto the hyperplane $C = \{\boldsymbol{x} : \langle \boldsymbol{a}, \boldsymbol{x}\rangle = \alpha\}$ with respect to the relative entropy. This divergence arises from the seed function $\varphi(\boldsymbol{x}) = \sum_n x_n \log x_n - x_n$, whose conjugate is $\varphi^*(\boldsymbol{\theta}) = \sum_n \exp(\theta_n)$. To identify the Lagrange multiplier, we must minimize

$$J(\xi) = \sum\nolimits_n y_n \exp(\xi a_n) - \alpha\xi,$$

whose derivatives are

$$J'(\xi) = \sum\nolimits_n a_n y_n \exp(\xi a_n) - \alpha,$$

$$J''(\xi) = \sum\nolimits_n a_n^2 y_n \exp(\xi a_n).$$

These functions are all simple to evaluate, so it is best to use the Newton method preceded by a bracketing phase [32]. Once we have found the minimizer $\xi_\star$, the Bregman projection is

$$P_C(\boldsymbol{y}) = \boldsymbol{y} \cdot \exp(\xi_\star \boldsymbol{a}),$$

where $\cdot$ represents the Hadamard product and the exponential is performed componentwise.

3.3. Example: Log-determinant divergence. Here is a more sophisticated example that involves the log-determinant divergence. The divergence arises from the seed function $\varphi(\boldsymbol{X}) = -\log\det(\boldsymbol{X})$, whose domain is the positive-definite cone and whose gradient is $\nabla\varphi(\boldsymbol{X}) = -\boldsymbol{X}^{-1}$. The conjugate function $\varphi^*(\boldsymbol{\Theta}) = N - \log\det(-\boldsymbol{\Theta})$, whose domain is the negative-definite cone and whose gradient satisfies $\nabla\varphi^*(\boldsymbol{\Theta}) = -\boldsymbol{\Theta}^{-1}$.

Suppose we need to project the positive-definite matrix $\boldsymbol{Y}$ onto the hyperplane

$$C = \{\boldsymbol{X} : \langle \boldsymbol{A}, \boldsymbol{X} \rangle = \alpha\}, \qquad \text{where } \boldsymbol{A} = \boldsymbol{A}^*.$$

We must minimize

$$J(\xi) = N - \log\det(\boldsymbol{Y}^{-1} - \xi\boldsymbol{A}) - \alpha\xi,$$

while ensuring that $\boldsymbol{Y}^{-1} - \xi\boldsymbol{A}$ is positive definite.

Let $\boldsymbol{Y} = \boldsymbol{L}\boldsymbol{L}^*$, and abbreviate $\boldsymbol{W} = \boldsymbol{L}^*\boldsymbol{A}\boldsymbol{L}$, which is singular whenever $\boldsymbol{A}$ is rank deficient. Then the derivatives of J can be expressed as

$$J'(\xi) = \operatorname{Tr}(\boldsymbol{W}(\mathbf{I} - \xi\boldsymbol{W})^{-1}) - \alpha,$$

$$J''(\xi) = \operatorname{Tr}\big(\big(\boldsymbol{W}(\mathbf{I} - \xi\boldsymbol{W})^{-1}\big)^2\big).$$

In general, J and its derivatives are all costly. It appears that the most efficient way to calculate them for multiple values of the scalar ξ is to preprocess $\boldsymbol{W}$ to extract its eigenvalues $\{\lambda_n\}$. It follows that

$$J'(\xi) = \left(\sum\nolimits_n \frac{\lambda_n}{1 - \lambda_n\xi}\right) - \alpha,$$

$$J''(\xi) = \sum\nolimits_n \left(\frac{\lambda_n}{1 - \lambda_n\xi}\right)^2.$$

It is worth cautioning that $\operatorname{dom} J = \{\xi : \xi < 1/\max_n \lambda_n\}$ since the matrix $\mathbf{I} - \xi\boldsymbol{W}$ must remain positive definite.

Once again, we see that a guarded or damped Newton method is the best way to optimize J. Given the solution $\xi_\star$, the Bregman projection is

$$P_C(\boldsymbol{Y}) = \boldsymbol{L}(\mathbf{I} - \xi_\star\boldsymbol{W})^{-1}\boldsymbol{L}^*.$$

We can reuse the eigenvalue decomposition to accelerate this final computation.

As shown in [25], these calculations simplify massively when the constraint matrix has rank one: $\boldsymbol{A} = \boldsymbol{a}\boldsymbol{a}^*$. In this case, we can find the zero of J' analytically because $\boldsymbol{a}^*\boldsymbol{Y}\boldsymbol{a}$ is the only nonzero eigenvalue of $\boldsymbol{W}$. Then the Sherman–Morrison formula delivers an explicit expression for the projection:

$$P_C(\boldsymbol{Y}) = \boldsymbol{Y} + \frac{\boldsymbol{a}^*\boldsymbol{Y}\boldsymbol{a} - \alpha}{(\boldsymbol{a}^*\boldsymbol{Y}\boldsymbol{a})^2}(\boldsymbol{Y}\boldsymbol{a})(\boldsymbol{Y}\boldsymbol{a})^*.$$

The cost of performing the projection totals $O(N^2)$.

4. The successive projection algorithm for affine constraints. Now we describe an algorithm for solving (1.1) in the special case that the constraint sets are all *affine spaces*. In the next section, we will present some concrete problems to which this algorithm applies. The case of general convex constraint sets will be addressed afterward. We frame the following hypotheses.

Assumption A.1	
The divergence:	φ is a convex function of Legendre type $\operatorname{dom}\varphi^*$ is an open set
The constraints:	$C_1, C_2, \ldots, C_K$ are affine spaces with intersection C
Constraint qualification:	$C \cap \operatorname{ri}(\operatorname{dom}\varphi)$ is nonempty

Note that, by the results of subsection 2.3, all Bregman divergences that arise from regular exponential families satisfy Assumption A.1.

Given an input $\boldsymbol{y}_0$ from $\operatorname{ri}(\operatorname{dom}\varphi)$, we seek the Bregman projection of $\boldsymbol{y}_0$ onto the intersection C of the affine constraints. In general, it may be difficult to produce $P_C(\boldsymbol{y}_0)$. Nevertheless, if the basic sets $C_1, \ldots, C_K$ are chosen well, it may be relatively straightforward to calculate the Bregman projection onto each basic set. This heuristic suggests an algorithm: Project successively onto each basic set in the hope that the sequence of iterates will converge to the Bregman projection onto the intersection. To make this approach work in general, it is clear that we must choose every set an infinite number of times, so we add one more requirement to Assumption A.1 as follows.

Assumption A.2	
The control mapping:	$r : \mathbb{N} \to \{1, \ldots, K\}$ is a sequence that takes each output value an infinite number of times

Together, Assumptions A.1 and A.2 will be referred to as Assumption A. Here is a formal statement of the algorithm.

ALGORITHM A (successive projection). *Suppose that Assumption* A *is in force. Choose an input vector* $\boldsymbol{y}_0$ *from* $\operatorname{ri}(\operatorname{dom}\varphi)$, *and form a sequence of iterates via successive Bregman projection:*

$$\boldsymbol{y}_t = P_{C_{r(t)}}(\boldsymbol{y}_{t-1}).$$

Then the sequence of iterates $\{\boldsymbol{y}_t\}$ *converges in norm to* $P_C(\boldsymbol{y}_0)$.

We present a short proof that this algorithm is correct. We refer to the article [4] for the argument that the sequence converges, and we extend the elegant proof from [12] to show that the limit of the sequence yields the Bregman projection.

Proof. Suppose that $\boldsymbol{a}$ is an arbitrary point in $C \cap \operatorname{dom}\varphi$. Since the seed function φ is Legendre, Bregman projections with respect to the divergence fall in the relative interior of $\operatorname{dom}\varphi$. In particular, each iterate $\boldsymbol{y}_t$ belongs to $\operatorname{ri}(\operatorname{dom}\varphi)$. Therefore, we may apply the Pythagorean theorem (2.4) to see that

$$D_\varphi(\boldsymbol{a}; \boldsymbol{y}_{t-1}) = D_\varphi(\boldsymbol{a}; \boldsymbol{y}_t) + D_\varphi(\boldsymbol{y}_t; \boldsymbol{y}_{t-1}).$$

Observe that this equation defines a recurrence, which we may solve to obtain

$$D_\varphi(\boldsymbol{a}; \boldsymbol{y}_0) = D_\varphi(\boldsymbol{a}; \boldsymbol{y}_t) + \sum\nolimits_{i=1}^{t} D(\boldsymbol{y}_i; \boldsymbol{y}_{i-1}).$$

Under Assumption A, Theorem 8.1 of [4] shows that the sequence of iterates generated by Algorithm A converges to a point $\overline{\boldsymbol{y}}$ in $C \cap \operatorname{ri}(\operatorname{dom}\varphi)$. Since the divergence is continuous in its second argument, we may take limits to reach

$$D_\varphi(\boldsymbol{a}; \boldsymbol{y}_0) = D_\varphi(\boldsymbol{a}; \overline{\boldsymbol{y}}) + \sum\nolimits_{i=1}^{\infty} D_\varphi(\boldsymbol{y}_i; \boldsymbol{y}_{i-1}).$$

We chose $\boldsymbol{a}$ arbitrarily from $C \cap \operatorname{dom}\varphi$, so we may replace $\boldsymbol{a}$ by $\overline{\boldsymbol{y}}$ to see that the infinite sum equals $D_\varphi(\overline{\boldsymbol{y}}; \boldsymbol{y}_0)$. It follows that

$$D_\varphi(\boldsymbol{a}; \boldsymbol{y}_0) = D_\varphi(\boldsymbol{a}; \overline{\boldsymbol{y}}) + D_\varphi(\overline{\boldsymbol{y}}; \boldsymbol{y}_0).$$

This equation holds for each point $\boldsymbol{a}$ in $C \cap \operatorname{dom}\varphi$, so we see that $\overline{\boldsymbol{y}}$ meets the variational characterization (2.4) of $P_C(\boldsymbol{y}_0)$. Therefore, $\overline{\boldsymbol{y}}$ is the Bregman projection of $\boldsymbol{y}_0$ onto C. □

If the sets $\{C_k\}$ are not affine, then Algorithm A will generally fail to produce the Bregman projection of $\boldsymbol{y}_0$ onto the intersection C. In section 6, we will discuss a more sophisticated iterative algorithm for solving this problem. Nevertheless, for general closed, convex constraint sets, the sequence of iterates generated by the successive projection algorithm still converges to a point in $C \cap \operatorname{ri}(\operatorname{dom}\varphi)$ [4, Theorem 8.1].

To obtain the convergence guarantee for Algorithm A, it may be necessary to work in an affine subspace of the ambient inner-product space. This point becomes important when computing the projections of nonnegative (as opposed to positive) vectors with respect to the relative entropy. It arises again when studying projections of rank-deficient matrices with respect to the von Neumann divergence. We will touch on this issue in subsections 5.1 and 5.3.

5. Examples with affine constraints.

This section presents three matrix nearness problems with affine constraints. The first requests the nearest contingency table with fixed marginals. A special case is to produce the nearest doubly stochastic matrix with respect to relative entropy. For this problem, the successive projection algorithm is identical to Kruithof's famous diagonal scaling algorithm [24, 13].

The second problem centers on a matrix nearness problem from data analysis, namely, that of finding matrix approximations based on the MBI principle, which is a generalization of Jaynes' maximum entropy principle [23].

The third problem shows how to construct the correlation matrix closest to a given positive-semidefinite matrix with respect to some matrix divergences. For reference, a correlation matrix is a positive-semidefinite matrix with a unit diagonal.

5.1. Contingency tables with fixed marginals.

A *contingency table* is an array that exhibits the joint probability mass function of a collection of discrete random variables. A nonnegative rectangular matrix may be viewed as the contingency table for two discrete random variables. We will focus on this case since higher-dimensional contingency tables essentially are no more complicated.

Suppose that p_{AB} is the joint probability mass function of two random variables A and B with sample spaces $\{1, 2, \ldots, M\}$ and $\{1, 2, \ldots, N\}$. We use $\boldsymbol{X}$ to denote the $M \times N$ contingency table whose entries are

$$x_{mn} = p_{AB}(A = m \text{ and } B = n).$$

A *marginal* of p_{AB} is a linear function of $\boldsymbol{X}$. The most important marginals of p_{AB} are the vector of row sums $\boldsymbol{X}\,\mathbf{e}$, which gives the distribution of A, and the vector of column sums $\mathbf{e}^T\boldsymbol{X}$, which gives the distribution of B. Here, $\mathbf{e}$ is a conformal vector of ones. The distribution of A conditioned on $B = n$ is given by the nth column of $\boldsymbol{X}$, and the distribution of B conditioned on $A = m$ is given by the mth row of $\boldsymbol{X}$.

However, we consider the more general case of arbitrary nonnegative matrices—we treat $\boldsymbol{X}$ as a member of the collection of $M \times N$ real matrices equipped with the inner product $\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \operatorname{Tr} \boldsymbol{X}\boldsymbol{Y}^T$. Note that, for the above probabilistic interpretation, $\boldsymbol{X}$ must be scaled so that its entries sum to 1.

A common problem is to use an initial estimate to produce a contingency table that has fixed marginals. In this setting, nearness is typically measured with relative entropy

$$D(\boldsymbol{X}; \boldsymbol{Y}) = \sum\nolimits_{m,n} \left[x_{mn} \log \frac{x_{mn}}{y_{mn}} - x_{mn} + y_{mn} \right].$$

An important special case is to find the doubly stochastic matrix nearest to a nonnegative square matrix $\boldsymbol{Y}_0$. In this case, we have two constraint sets

$$C_1 = \{\boldsymbol{X} : \boldsymbol{X}\,\mathbf{e} = \mathbf{e}\} \qquad \text{and} \qquad C_2 = \{\boldsymbol{X} : \mathbf{e}^T \boldsymbol{X} = \mathbf{e}^T\}.$$

It is clear that the intersection $C = C_1 \cap C_2$ contains the set of doubly stochastic matrices. In fact, every nonnegative matrix in C is doubly stochastic. Using (2.6), it is easy to see that Bregman projection of a matrix onto C_1 with respect to the relative entropy is accomplished by rescaling the rows so that each row sums to one. Likewise, Bregman projection of a matrix onto C_2 is accomplished by rescaling the columns. Beginning with $\boldsymbol{Y}_0$, the successive projection algorithm alternately rescales the rows and columns. This procedure, of course, is the diagonal scaling algorithm of Kruithof [24, 13], sometimes called Sinkhorn's algorithm [36]. Our approach yields a geometric interpretation of the algorithm as a method for solving a matrix nearness problem by alternating Bregman projections. It is interesting that the nonnegativity constraint is implicitly enforced by the domain of the relative entropy. This viewpoint can be traced to the work of Ireland and Kullback [22].

There is still a subtlety that requires attention. Assumption A apparently requires that C contain a matrix with strictly positive entries and that the input matrix $\boldsymbol{Y}_0$ be strictly positive. In fact, we may relax these premises. A nonnegative matrix whose zero pattern does not cover the zero pattern of $\boldsymbol{Y}_0$ has an infinite divergence from $\boldsymbol{Y}_0$. Therefore, we may as well restrict our attention to the linear space of matrices whose zero pattern covers that of $\boldsymbol{Y}_0$. Now we see that the constraint qualification in Assumption A requires that C contain a matrix with exactly the same zero pattern as $\boldsymbol{Y}_0$. If it does, the algorithm will still converge to the Bregman projection of $\boldsymbol{Y}_0$ onto the doubly stochastic matrices. Determining whether the constraint qualification holds will generally involve a separate investigation [30].

It is also worth noting that Algorithm A encompasses other iterative methods for scaling to doubly stochastic form. At each step, for example, one might rescale only the row or column whose sum is most inaccurate. Parlett and Landis have considered algorithms of this sort [33]. The problem of scaling to have other row and column sums also fits neatly into our framework, and it has the same geometric interpretation.

5.2. MBI and matrix approximation. This section discusses a novel matrix nearness problem that arises in data analysis. Given a collection of vectors $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\} \subset \operatorname{dom} \varphi$, the *Bregman information* [2] of the collection is defined to be

$$I_\varphi(X) = \sum\nolimits_{j=1}^{N} w_j \, D_\varphi(\boldsymbol{x}_j; \boldsymbol{\mu}), \tag{5.1}$$

where $w_1, w_2, \ldots, w_N$ are nonnegative weights that sum to one, and $\boldsymbol{\mu}$ is the (weighted) arithmetic mean of the collection, i.e., $\boldsymbol{\mu} = \sum_j w_j \, \boldsymbol{x}_j$. Bregman information generalizes the notion of the *variance*, $\sigma^2 = N^{-1} \sum_j \|\boldsymbol{x}_j - \boldsymbol{\mu}\|_2^2$, of a Gaussian random variable (where each $w_j = N^{-1}$). When D_φ is the relative entropy, the Bregman information that arises with an appropriate choice of weights is called *mutual information*, a fundamental quantity in information theory [11].

Bregman information exhibits an interesting connection with Jensen's inequality for a convex function φ:

$$\sum\nolimits_j w_j \, \varphi(\boldsymbol{x}_j) \geq \varphi\left(\sum\nolimits_j w_j \, \boldsymbol{x}_j\right).$$

Substituting $\boldsymbol{\mu} = \sum_j w_j \, \boldsymbol{x}_j$, we see that the difference between the two sides of the foregoing relation satisfies

$$\begin{aligned}\sum\nolimits_j w_j \, \varphi(\boldsymbol{x}_j) - \varphi(\boldsymbol{\mu}) &= \sum\nolimits_j w_j \, \varphi(\boldsymbol{x}_j) - \varphi(\boldsymbol{\mu}) - \left\langle \nabla\varphi(\boldsymbol{\mu}), \sum\nolimits_j w_j \boldsymbol{x}_j - \boldsymbol{\mu} \right\rangle \\ &= \sum\nolimits_j w_j \left[\varphi(\boldsymbol{x}_j) - \varphi(\boldsymbol{\mu}) - \langle \nabla\varphi(\boldsymbol{\mu}), \boldsymbol{x}_j - \boldsymbol{\mu}\rangle\right] \\ &= I_\varphi(X). \qquad (5.2)\end{aligned}$$

In words, the Bregman information is the disparity between the two sides of Jensen's inequality. Equation (5.2) can also be viewed as a generalization of the relationship between the variance and the arithmetic mean,

$$\sigma^2 = N^{-1} \sum\nolimits_j \|\boldsymbol{x}_j\|_2^2 - \|\boldsymbol{\mu}\|_2^2 .$$

Let us describe an application of Bregman information in data analysis. In this field, matrix approximations play a central role. Unfortunately, many common approximations destroy essential structure in the data matrix. For example, consider the k-truncated singular value decomposition (TSVD), which provides the best rank-k Frobenius-norm approximation of a matrix. In information retrieval applications, however, the matrix that describes the co-occurrence of words and documents is both sparse and nonnegative. The TSVD ruins both of these properties. In this setting, the Frobenius norm is meaningless; relative entropy is the correct divergence measure according to the unigram or multinomial language model.

We may also desire that the matrix approximation satisfy some additional constraints. For instance, it may be valuable for the approximation to preserve marginals (i.e., linear functions) of the matrix entries. Let us formalize this idea. Suppose that $\boldsymbol{Y}$ is an $M \times N$ data matrix. We seek an approximation $\widetilde{\boldsymbol{X}}$ that satisfies the constraints

$$C_k = \{\boldsymbol{X} \, : \, \langle \boldsymbol{X}, \boldsymbol{A}_k\rangle = \langle \boldsymbol{Y}, \boldsymbol{A}_k\rangle\} \qquad k = 1, \ldots, K,$$

where each $\boldsymbol{A}_k$ is a fixed constraint matrix. We will write $C = \bigcap_k C_k$. As an example, $\boldsymbol{X}$ can be required to preserve the row and/or column sums of $\boldsymbol{Y}$.

Many different matrices, including the original matrix $\boldsymbol{Y}$, may satisfy these constraints. Clearly, a good matrix approximation should involve some reduction in the number of parameters used to represent the matrix. The key question is to decide how to produce the right approximation from C. One rational approach invokes the principle of *minimum Bregman information* (MBI) [1], which states that the approximation should be the (unique) solution of the problem

$$\min_{\boldsymbol{X} \in C} \; I_\varphi(\boldsymbol{X}) \;=\; \min_{\boldsymbol{X} \in C} \; \sum\nolimits_{m,n} w_{mn} \, D_\varphi(x_{mn}, \mu), \qquad (5.3)$$

where w_{mn} are prespecified weights and $\mu = \sum_{m,n} w_{mn} x_{mn}$. If the weights w_{mn} and the matrix entries x_{mn} are both sets of nonnegative numbers that sum to one, and if the Bregman divergence is the relative entropy, then the MBI principle reduces to Jaynes' maximum entropy principle [23]. Thus, the MBI principle tries to obtain as uniform an approximation as possible subject to the specified constraints. Note that problem (5.3) can be readily solved by the successive projection algorithm.

Next, we consider an important and natural source of constraints. *Clustering* is the problem of partitioning a set of objects into clusters, where each cluster contains "similar" objects. Data matrices often capture the relationships between two sets of objects, such as word–document matrices in information retrieval and gene-expression matrices in bioinformatics. In such applications, it is often desirable to solve the coclustering problem, i.e., to simultaneously cluster the rows and columns of a data matrix. Formally, a co-clustering (ρ, γ) is a partition of the rows into I row clusters $\rho_1, \ldots, \rho_I$ and the columns into J column clusters $\gamma_1, \ldots, \gamma_J$, i.e.,

$$\bigcup\nolimits_{i=1}^{I} \rho_i = \{1, 2, \ldots, M\}, \qquad \text{where} \qquad \rho_i \cap \rho_\ell = \emptyset \quad \text{for } i \neq \ell,$$

$$\bigcup\nolimits_{j=1}^{J} \gamma_j = \{1, 2, \ldots, N\}, \qquad \text{where} \qquad \gamma_j \cap \gamma_\ell = \emptyset \quad \text{for } j \neq \ell.$$

Given a coclustering, the rows belonging to row cluster ρ_1 can be arranged first, followed by rows belonging to row cluster ρ_2, etc. Similarly the columns can be re-ordered. This re-ordering has the effect of dividing the matrix into $I \cdot J$ subblocks, each of which is called a *cocluster*.

The coclustering problem is to search for the "best" possible row and column clusters. One way to measure the quality of a coclustering is to associate it with its MBI matrix approximation. A natural constraint set $C_{(\rho,\gamma)}$ for the coclustering problem contains matrices that preserve marginals of all the $I \cdot J$ coclusters (local information) in addition to row and column marginals (global information). With this constraint set, a formal objective for the coclustering problem is to find (ρ, γ), which corresponds to the best possible MBI approximation:

$$\min_{\rho,\gamma} D_\varphi(\boldsymbol{Y}; \boldsymbol{X}_{(\rho,\gamma)}), \qquad \text{where} \qquad \boldsymbol{X}_{(\rho,\gamma)} = \arg \min_{\boldsymbol{X} \in C_{(\rho,\gamma)}} I_\varphi(\boldsymbol{X}). \tag{5.4}$$

This formulation yields an optimal coclustering as well as its associated MBI matrix approximation. The quality of such matrix approximations is a topic for further study. Note that problem (5.4) requires a combinatorial search, and it is known to be NP-complete. The most familiar clustering formulation, namely, the k-means problem, is the special case of (5.4) obtained from the Euclidean divergence, the choice $J = N$, and the condition of preserving cocluster sums.

As an example, consider the nonnegative matrix

$$\boldsymbol{Y} = \begin{bmatrix} 5 & 5 & 5 & 0 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 5 & 5 & 5 \\ 0 & 0 & 0 & 5 & 5 & 5 \\ 4 & 4 & 0 & 4 & 4 & 4 \\ 4 & 4 & 4 & 0 & 4 & 4 \end{bmatrix}.$$

On using coclustering (three row clusters and two column clusters), preserving row sums, column sums, and cocluster sums, the MBI principle (with relative entropy as the Bregman divergence) yields the matrix approximation

$$\boldsymbol{X}_1 = \begin{bmatrix} 5.4 & 5.4 & 4.2 & 0 & 0 & 0 \\ 5.4 & 5.4 & 4.2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4.2 & 5.4 & 5.4 \\ 0 & 0 & 0 & 4.2 & 5.4 & 5.4 \\ 3.6 & 3.6 & 2.8 & 2.8 & 3.6 & 3.6 \\ 3.6 & 3.6 & 2.8 & 2.8 & 3.6 & 3.6 \end{bmatrix}.$$

Note that this approximation has rank two and preserves nonnegativity as well as most of the nonzero structure of $\boldsymbol{Y}$. It can be verified that all the cocluster sums, row sums, and column sums of $\boldsymbol{X}_1$ match those of $\boldsymbol{Y}$. In contrast, the rank-two SVD approximation

$$\boldsymbol{X}_2 = \begin{bmatrix} 5.09 & 5.09 & 4.66 & -0.69 & 0.29 & 0.29 \\ 5.09 & 5.09 & 4.66 & -0.69 & 0.29 & 0.29 \\ 0.29 & 0.29 & -0.69 & 4.66 & 5.09 & 5.09 \\ 0.29 & 0.29 & -0.69 & 4.66 & 5.09 & 5.09 \\ 3.04 & 3.04 & 1.98 & 3.51 & 4.41 & 4.41 \\ 4.41 & 4.41 & 3.51 & 1.98 & 3.04 & 3.04 \end{bmatrix}$$

preserves neither the nonnegativity, the nonzero structure, nor the marginals of $\boldsymbol{Y}$.

5.3. The nearest correlation matrix.

A *correlation matrix* is a (real) positive-semidefinite matrix with a unit diagonal. Correlation matrices arise in statistics and applications such as finance, where they display the normalized second-order statistics (i.e., pairwise correlation coefficients) of a collection of random variables. In the deterministic setting, a correlation matrix may be viewed as the Gram matrix of a collection of unit vectors.

Higham has recently studied the nearest correlation matrix problem measuring distances using a type of weighted Frobenius norm [19]. Higham solves the problem by means of the Dykstra–Han algorithm given in section 6, alternating between the positive-semidefinite cone and the set of matrices with unit diagonal. We have observed that the nearest correlation matrix problem can be posed with Bregman divergences and, in particular, with matrix divergences.

Let us consider the problem of producing the correlation matrix closest to a given positive-semidefinite matrix with respect to the von Neumann divergence

$$D_{\mathrm{vN}}(\boldsymbol{X};\boldsymbol{Y}) = \mathrm{Tr}\,[\boldsymbol{X}(\log\boldsymbol{X} - \log\boldsymbol{Y}) - \boldsymbol{X} + \boldsymbol{Y}].$$

In case $\boldsymbol{Y}$ is singular, we must restrict our attention to the linear space of matrices whose null space contains the null space of $\boldsymbol{Y}$. After taking this step, one must interpret the formulae with care. These remarks signal our reason for employing the von Neumann divergence to measure the disparity between correlation matrices. A matrix $\boldsymbol{X}$ has an infinite divergence from $\boldsymbol{Y}$ unless the null space of $\boldsymbol{X}$ contains the null space of $\boldsymbol{Y}$. In particular, the rank of the Bregman projection of $\boldsymbol{Y}$ onto the correlation matrices cannot exceed the rank of $\boldsymbol{Y}$. See also the examples at the end of this subsection.

The correlation matrices can be viewed as the intersection of the set of unit-diagonal matrices with the positive-semidefinite cone. This cone is also the domain of the von Neumann divergence, so we do not need to explicitly enforce the positive-semidefinite constraint. In fact, we need only project onto the set C of matrices whose diagonal entries all equal one. It is natural to view C as the intersection of the affine constraint sets

$$C_k = \{\boldsymbol{X} : x_{kk} = 1\}.$$

There is no explicit formula for the projection of a matrix $\boldsymbol{Y}$ onto the set C_k, but the discussion in section 3 shows that we can solve the problem by minimizing the function (3.4), which, in this example, reads

$$J(\xi) = \operatorname{Tr}\exp\{\log \boldsymbol{Y} + \xi\, \mathbf{e}_k \mathbf{e}_k^T\} - \xi, \tag{5.5}$$

where $\mathbf{e}_k$ is the kth canonical basis vector. Given the minimizer $\xi_\star$, the projection of $\boldsymbol{Y}$ onto C_k is

$$P_{C_k}(\boldsymbol{Y}) = \exp\{\log \boldsymbol{Y} + \xi_\star\, \mathbf{e}_k \mathbf{e}_k^T\}. \tag{5.6}$$

Beware that one cannot read these formulae literally when $\boldsymbol{Y}$ is rank deficient! In any case, the numerical calculations are not trivial to perform. In order to apply the Newton method, the second derivative of J is needed, which is more involved due to the noncommutativity of matrix multiplication.

Unfortunately, treating these issues in detail is beyond the scope of this paper.

There is an interesting special case that can be treated without optimization: the von Neumann projection of a matrix with constant diagonal onto the correlation matrices can always be obtained by rescaling. In particular, the projection preserves the zero pattern of the matrix and the eigenvalue distribution. To verify this point, suppose the diagonal entries of $\boldsymbol{Y}$ equal α, and set $\boldsymbol{X} = \alpha^{-1}\boldsymbol{Y}$. According to the Karush–Kuhn–Tucker conditions, $\boldsymbol{X}$ is the Bregman projection of $\boldsymbol{Y}$ onto the set C provided that $\nabla_{\boldsymbol{X}} D_{\mathrm{vN}}(\boldsymbol{X}; \boldsymbol{Y})$ is diagonal. The latter gradient equals $\log \boldsymbol{X} - \log \boldsymbol{Y} + \mathbf{I}$, and a short calculation completes the argument. In contrast, the Frobenius norm projection of a matrix with constant diagonal does not preserve its nonzero structure or eigenvalue distribution. As an example, let $\boldsymbol{Y}$ be the 4×4 symmetric tridiagonal Toeplitz matrix with 2's on the diagonal and -1's on the off-diagonal. The nearest correlation matrix to it, in the Frobenius norm, equals (to the figures shown)

$$\begin{bmatrix} 1.0000 & -0.8084 & 0.1916 & 0.1068 \\ -0.8084 & 1.0000 & -0.6562 & 0.1916 \\ 0.1916 & -0.6562 & 1.0000 & -0.8084 \\ 0.1068 & 0.1916 & -0.8084 & 1.0000 \end{bmatrix}.$$

As a second example, draw a random orthogonal matrix $\boldsymbol{Q}$ and form the rank-deficient matrix $\boldsymbol{Y} = \boldsymbol{Q}\,\operatorname{diag}(1, 10^{-3}, 10^{-6}, 0)\,\boldsymbol{Q}^T$. For instance,

$$\boldsymbol{Y} = \begin{bmatrix} .18335 & -.15180 & .08258 & -.34620 \\ -.15180 & .12606 & -.06887 & .28655 \\ .08258 & -.06887 & .03786 & -.15582 \\ -.34620 & .28655 & -.15582 & .65373 \end{bmatrix}.$$

The correlation matrix nearest to $\boldsymbol{Y}$ in the Frobenius norm is obtained by simply shifting the diagonal:

$$\boldsymbol{X}_1 = \begin{bmatrix} 1.0000 & -.15180 & .08258 & -.34620 \\ -.15180 & 1.0000 & -.06887 & .28655 \\ .08258 & -.06887 & 1.0000 & -.15582 \\ -.34620 & .28655 & -.15582 & 1.0000 \end{bmatrix}.$$

Meanwhile, the nearest correlation matrix with respect to the von Neumann divergence has the same range space as $\boldsymbol{Y}$ and thus is also of rank 3:

$$\boldsymbol{X}_2 = \begin{bmatrix} 1.0000 & -.77271 & .59020 & -.99995 \\ -.77271 & 1.0000 & -.96847 & .77080 \\ .59020 & -.96847 & 1.0000 & -.58778 \\ -.99995 & .77080 & -.58778 & 1.0000 \end{bmatrix}.$$

Note that due to space limitations, all of the above matrices are shown only to five digits of accuracy. The eigenvalues of $\boldsymbol{X}_1$ are 0.611, 0.851, 0.951, and 1.588, while the nonzero eigenvalues of $\boldsymbol{X}_2$ are $0.457 \times 10^{-6}, 0.650 \times 10^{-2}$, and 3.350. Thus we see that the Frobenius norm solution does not preserve the small eigenvalues, while the von Neumann divergence solution preserves the rank and also tries to preserve the eigenvalue distribution.

The recent literature contains a substantial amount of work on numerical methods for calculating nearest correlation matrices with respect to the Frobenius norm. Higham describes an alternating projection method, as well as an approach via semidefinite programming [19]. Malick [28] and Boyd and Xiao [7] study efficient algorithms for solving the dual of a more general projection problem, while Qi and Sun [34] develop a generalized Newton method for the nearest correlation matrix problem.

In contrast, the problem of finding nearest correlation matrices with respect to a Bregman divergence is virtually unstudied. The main motivation for studying this problem is that it leads to correlation matrices that have a very different character, which may be more appropriate in applications. For example, as shown above, the method for solving the von Neumann nearness problem may yield low-rank correlation matrices. This type of solution has immense practical value because it explains the data using a small number of *factors* [17]. In contrast, the Frobenius norm solution may increase the rank of the matrix. Unfortunately, to apply our technique, the initial matrix must lie in the domain of the von Neumann divergence, i.e., the positive-semidefinite cone. One remedy is to preprocess the matrix by performing a Frobenius projection onto the positive-semidefinite cone.

The broad scope of the present article limits the amount of detail we can provide, so we have been only able to sketch one algorithm for solving the nearest correlation matrix problem. It would be valuable to devise more powerful algorithms by invoking ideas from the papers cited above.

6. The successive projection–correction algorithm for convex constraints. This section describes an algorithm for solving the Bregman nearness problem (1.1) in the case where the constraints are closed, convex sets. In the succeeding section, we will present some nearness problems to which this algorithm applies. We frame the following hypotheses:

Assumption B	
The divergence:	φ is a convex function of Legendre type φ is cofinite, i.e., $\operatorname{dom}\varphi^* = \mathscr{X}$
The constraints:	$C_1, \dots, C_K$ are closed, convex sets with intersection C
Constraint qualification:	$\operatorname{ri}(C_1) \cap \cdots \cap \operatorname{ri}(C_K) \cap \operatorname{ri}(\operatorname{dom}\varphi)$ is nonempty
The control mapping:	$r : \mathbb{N} \to \{1, 2, \dots, K\}$ is a sequence that takes each output value at least once during each T consecutive input values

Given an input $\boldsymbol{y}_0$ from $\operatorname{ri}(\operatorname{dom}\varphi)$, we seek the Bregman projection of $\boldsymbol{y}_0$ onto C with respect to the divergence D_φ. As before, the algorithm projects successively onto each constraint set. Since the sets are no longer affine, it is also necessary to introduce a correction term to guide the algorithm toward the Bregman projection. This algorithm generalizes a method for the Euclidean divergence that was developed independently by Dykstra [16] and Han [18].

Algorithm B (successive projection–correction). *Suppose that Assumption* B *is in force, and let* $\boldsymbol{y}_0 \in \operatorname{ri}(\operatorname{dom}\varphi)$. *The algorithm performs the following steps:*

1. *Initialize the correction variables:* $\boldsymbol{q}^k = \boldsymbol{0}$ *for each* $k = 1, \dots, K$.
2. *Construct the next iterate via the rule*

$$\boldsymbol{y}_{t+1} \leftarrow P_{C_{r(t)}}\left(\nabla\varphi^*\left(\nabla\varphi(\boldsymbol{y}_t) + \boldsymbol{q}^{r(t)}\right)\right).$$

3. *Update the correction term*

$$\boldsymbol{q}^{r(t)} \leftarrow \boldsymbol{q}^{r(t)} + \nabla\varphi(\boldsymbol{y}_t) - \nabla\varphi(\boldsymbol{y}_{t+1}).$$

4. *Return to step* 2.

Then the sequence of iterates converges in norm to the Bregman projection of $\boldsymbol{y}_0$ *onto* C *with respect to* D_φ,

$$P_C(\boldsymbol{y}_0) = \lim_{t\to\infty} \boldsymbol{y}_t.$$

The proof that this algorithm succeeds is quite burdensome, and none of the arguments in the literature are especially intuitive. The correctness of the algorithm that we have presented here follows from Tseng's general framework [37]. His paper contains only the development for the Euclidean divergence; see [6, 9] for comments on the extension. The literature contains several other proofs with somewhat different hypotheses [6, 9]. We feel that the above version offers the best tradeoff between applicability and accessibility.

The following connections may help the reader understand the algorithm somewhat better. It is possible to identify this procedure as a generalization of Bregman's algorithm for minimizing strictly convex functions [10, 9]. Bregman's algorithm is a primal-dual method that maximizes with respect to one dual variable (the $\boldsymbol{q}^k$) at a time, while maintaining the Karush–Kuhn–Tucker conditions on the primal problem. It is also possible to view the algorithm as a coordinate ascent algorithm for an optimization problem that is dual to the projection problem [37]. It is for this reason that the update in step 2 closely resembles the dual function J obtained in section 3.

6.1. Comparing the algorithms. Let us take a moment to weigh the successive projection–correction algorithm (Algorithm B) against the successive projection algorithm (Algorithm A). It is most important to note that Algorithm A applies only to the case where the constraints are affine, while Algorithm B succeeds for general closed, convex constraints. It can be shown that the corrections in Algorithm B are unnecessary when the constraints are affine, so it reduces to Algorithm A [9].

Although it may appear that Algorithm A has a weaker constraint qualification, the difference here is purely formal. We remark that the constraint qualification in Algorithm B can be weakened when some of the constraint sets are polyhedral, i.e., can be written as a finite intersection of halfspaces. In that case, we may remove the relative interior from the polyhedral constraint sets in the constraint qualification.

The methods also place different hypotheses on the divergence; Algorithm B asks more from the seed function φ than Algorithm A. The former requires that $\operatorname{dom}\varphi^* = \mathscr{X}$ while the latter needs only $\operatorname{dom}\varphi^*$ to be open. For example, the Burg entropy $\varphi(x) = -\log(x)$ is admissible for Algorithm A but not for Algorithm B.

Finally, the control mapping for Algorithm B is more restrictive than the control mapping for Algorithm A. The former requires that the projections be performed in almost cyclic order, while the latter requires only that each constraint set should appear an infinite number of times.

7. Examples with convex constraints. This section discusses two matrix nearness problems that involve nonaffine constraints. First, we discuss the *metric nearness problem*, which elicits the closest metric graph to a given weighted graph. We have already studied this problem with respect to norms in [15]. Here, we expand our treatment to Bregman divergences.

Second, we study an important problem in data analysis, namely learning a so-called "kernel" or similarity matrix that satisfies constraints that arise from knowledge of the underlying application domain.

7.1. The metric nearness problem. We recently encountered a striking new matrix nearness problem [15] while studying an application in computational biology. In this article, we extend the problem to Bregman divergences and show that it can be solved using the successive projection–correction algorithm (Algorithm B).

Suppose that $\boldsymbol{X}$ is the adjacency matrix of an undirected, weighted graph on N vertices. That is, x_{mn} registers the weight of the edge between vertices m and n. Since the graph is undirected, $\boldsymbol{X}$ is a symmetric matrix. We will also assume that $\boldsymbol{X}$ is *hollow* (i.e., has a zero diagonal). If one interprets the weights as distances, it is natural to ask whether the graph can be embedded in a metric space. Indeed, the embedding is possible if and only if the triangle inequalities hold, i.e.,

$$x_{mn} \le x_{m\ell} + x_{\ell n} \qquad \text{for each triple of distinct vertices } (\ell, m, n). \tag{7.1}$$

Note that the condition (7.1) implies that the weights are nonnegative, provided that $\boldsymbol{X}$ is symmetric. We will refer to any hollow, symmetric matrix that satisfies (7.1) as a *metric adjacency matrix*.

The *metric nearness problem* is to find the metric adjacency matrix closest to a given adjacency matrix. We view this nearness problem as an agnostic method for learning a metric from noisy distance measurements. It is entirely distinct from multidimensional scaling, which requests an ensemble of points in a *specified* metric space (usually Euclidean) that realizes a given set of distances. In our first report on this problem [15], we used weighted matrix norms to measure the distance between adjacency matrices. In this article, we will use Bregman divergences. Note that the divergence is unrelated to the metric encoded in the entries of the adjacency matrix; the divergence is used to determine how much one adjacency matrix (i.e., graph) differs from another.

By this point, it should be clear how we propose to solve the metric nearness problem. We will work in the space of hollow, symmetric matrices. It is evident that the metric adjacency matrices from a closed, convex cone C. Clearly, C is the intersection of $\binom{N}{3}$ halfspaces:

$$C_{\ell mn} = \{\boldsymbol{X} : x_{mn} - x_{m\ell} - x_{\ell n} \leq 0\},$$

where ℓ, m, and n index distinct vertices. Therefore, we may apply Algorithm B.

To be concrete, we will consider Bregman projections with respect to the relative entropy. For reference, the seed function is

$$\varphi(\boldsymbol{X}) = \sum\nolimits_{mn} \left[x_{mn} \log x_{mn} - x_{mn}\right],$$

which has Fenchel conjugate

$$\varphi^*(\boldsymbol{Y}) = \sum\nolimits_{mn} \exp y_{mn}.$$

The divergence is

$$D_\varphi(\boldsymbol{X}; \boldsymbol{Y}) = \sum\nolimits_{mn} \left[x_{mn} \log \frac{x_{mn}}{y_{mn}} - x_{mn} + y_{mn}\right].$$

This divergence has an interesting advantage over the Frobenius norm. If the original adjacency matrix does not contain zero distances, then the projection on the metric adjacency matrices will not contain any zero distances. This fact ensures that the final matrix defines a genuine metric, rather than a pseudometric.

Algorithm B requires that we compute the Bregman projection of a matrix that has the form $\boldsymbol{X} = \nabla\varphi^*(\nabla\varphi(\boldsymbol{Y}_t) + \boldsymbol{Q}^{\ell mn})$, where $\boldsymbol{Q}^{\ell mn}$ is a dual variable. It is easy to check that this expression reduces to

$$\boldsymbol{X} = \boldsymbol{Y}_t \cdot \exp\cdot(\boldsymbol{Q}^{\ell mn}),$$

where $\cdot$ is the Hadamard (i.e., componentwise) product and $\exp\cdot$ is the Hadamard exponential. We will see that the dual variable $\boldsymbol{Q}^{\ell mn}$ has at most six nonzero entries. Therefore, the matrix $\boldsymbol{X}$ differs from $\boldsymbol{Y}_t$ in at most six places.

It is straightforward to calculate the Bregman projection $\boldsymbol{Y}_{t+1}$ of the matrix $\boldsymbol{X}$ onto the constraint $C_{\ell mn}$. If $\boldsymbol{X}$ already falls in the constraint set, then the projection $\boldsymbol{Y}_{t+1} = \boldsymbol{X}$. Otherwise, set $\delta = \sqrt{(x_{m\ell} + x_{\ell n})/x_{mn}}$. The entries of the projection $\boldsymbol{Y}_{t+1}$ are identical to those of $\boldsymbol{X}$ except for the following six:

$$\begin{aligned} y_{mn} &= \delta\, x_{mn} & \qquad y_{nm} &= y_{mn} \\ y_{m\ell} &= x_{m\ell}/\delta & \qquad y_{\ell m} &= y_{m\ell} \\ y_{\ell n} &= x_{\ell n}/\delta & \qquad y_{n\ell} &= y_{\ell n}. \end{aligned}$$

In words, the projection determines how much the triangle inequality is violated, and it distributes the deficit multiplicatively among the three edges.

Finally, the algorithm updates the dual variable $\boldsymbol{Q}^{\ell mn}$ associated with the constraint using the formula

$$\boldsymbol{Q}^{\ell mn} \leftarrow \boldsymbol{Q}^{\ell mn} + \log\cdot(\boldsymbol{Y}_t) - \log\cdot(\boldsymbol{Y}_{t+1})$$

where $\log\cdot$ is the Hadamard logarithm. This update affects only six entries of $\boldsymbol{Q}^{\ell mn}$. In practice, we would store only the upper triangle of the adjacency matrices, so the update touches only three entries.

Consider the following adjacency matrix, which fails to be a metric graph,

$$\boldsymbol{Y} = \begin{bmatrix} 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 10000 & 1 & 1 & 1 \\ 1 & 10000 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}.$$

The nearest metric adjacency matrix in relative entropy is found to be

$$\boldsymbol{X}_1 = \begin{bmatrix} 0 & 5.49 & 5.49 & 1.00 & 1.00 & 1.00 \\ 5.49 & 0 & 10.99 & 5.49 & 5.49 & 5.49 \\ 5.49 & 10.99 & 0 & 5.49 & 5.49 & 5.49 \\ 1.00 & 5.49 & 5.49 & 0 & 1.00 & 1.00 \\ 1.00 & 5.49 & 5.49 & 1.00 & 0 & 1.00 \\ 1.00 & 5.49 & 5.49 & 1.00 & 1.00 & 0 \end{bmatrix}.$$

Note that the effect of the outlier edge has dissipated, and the resulting metric graph does not have very large edge weights. On the other hand, the outlier edge leads to a significant change in edge weights when the Euclidean divergence is used:

$$\boldsymbol{X}_2 = \begin{bmatrix} 0 & 1667.33 & 1667.33 & 1.00 & 1.00 & 1.00 \\ 1667.33 & 0 & 3334.67 & 1667.33 & 1667.33 & 1667.33 \\ 1667.33 & 3334.67 & 0 & 1667.33 & 1667.33 & 1667.33 \\ 1.00 & 1667.33 & 1667.33 & 0 & 1.00 & 1.00 \\ 1.00 & 1667.33 & 1667.33 & 1.00 & 0 & 1.00 \\ 1.00 & 1667.33 & 1667.33 & 1.00 & 1.00 & 0 \end{bmatrix}.$$

7.2. Learning a kernel matrix.

In data mining and machine learning applications, linear separators or hyperplanes are often used to cluster or classify data. However, linear separators are inadequate when the data is not linearly separable. To overcome this problem, the data can first be mapped (nonlinearly) to a higher-dimensional feature space, after which linear separators can be used in the transformed feature space.

Suppose the data belong to the set Ω, and $f : \Omega \to \mathscr{X}$ maps the data to an inner-product space $\mathscr{X}$, called the *feature space*. Given data objects $\{u_1, u_2, \ldots, u_N\} \subset \Omega$, the Gram matrix $\boldsymbol{X}$ is the $N \times N$ matrix of inner products in the feature space, $x_{mn} = \langle f(u_m), f(u_n)\rangle = g(u_m, u_n)$. This Gram matrix is also called the *kernel matrix*, and it captures the similarity between the objects u_m and u_n. When the data space Ω is an inner-product space, common kernels include the polynomial kernel $g(\boldsymbol{u}_m, \boldsymbol{u}_n) = \langle \boldsymbol{u}_m, \boldsymbol{u}_n\rangle^d$ and the Gaussian kernel $g(\boldsymbol{u}_m, \boldsymbol{u}_n) = \exp\left\{-\frac{1}{2}\left\|\boldsymbol{u}_m - \boldsymbol{u}_n\right\|_2^2 / \sigma^2\right\}$. These kernels are both positive definite. Conversely, any positive-definite matrix can be thought of as a kernel matrix [38]. In general, the set Ω can be arbitrary. For example, Ω might contain nucleotide sequences of varying lengths or phylogenetic trees or arbitrary graphs.

In many such situations, the choice of the kernel matrix is unclear. There is often an approximate kernel matrix $\boldsymbol{Y}_0$ that we wish to modify based on our information about the underlying data objects. This information may take various forms:

- known values for kernel entries ($x_{mn} = \alpha$),
- known distances between objects in the feature space ($x_{mm}+x_{nn}-2x_{mn} = \beta$), or
- known bounds on kernel entries ($x_{mn} \le x_{rs}$) or distances ($x_{mm} + x_{nn} - 2x_{mn} \le \gamma$).

Such constraints are typically obtained from the application domain, such as information about whether a pair of genes or proteins is functionally more similar than another pair.

Suppose that we are given an approximate kernel matrix $\boldsymbol{Y}_0$. Our problem is to find the nearest positive-definite matrix to $\boldsymbol{Y}_0$ that satisfies linear equality and inequality constraints. The von Neumann divergence can be used as the nearness measure:

$$D_{\mathrm{vN}}(\boldsymbol{X};\boldsymbol{Y}) = \operatorname{Tr}[\boldsymbol{X}(\log \boldsymbol{X} - \log \boldsymbol{Y}) - \boldsymbol{X} + \boldsymbol{Y}].$$

Using the von Neumann divergence appears to be advantageous when the initial kernel matrix $\boldsymbol{Y}_0$ is of low rank and it is desired that its null space be preserved [25]. Recall that, in the low-rank case, the von Neumann divergence $D_{\mathrm{vN}}(\boldsymbol{X};\boldsymbol{Y}_0)$ is finite only when the null space of $\boldsymbol{X}$ contains the null space of $\boldsymbol{Y}_0$. Hence, both the null space constraint and positive semidefiniteness are automatically enforced by the successive projection–correction algorithm.

8. Open problems and conclusions.

The Bregman nearness problem is relatively unstudied, so it opens a rich vein of new questions. Here are some specific challenges that deserve attention.

1. The matrix divergences described in subsection 2.6 offer an intriguing way to compute distances between Hermitian matrices. It would be valuable to characterize different types of projections onto important sets of matrices, such as the positive-semidefinite cone, the nonnegative cone, or the set of diagonal matrices. This could lead to more efficient numerical methods for key problems.
2. The algorithms described in this paper apply only to projections onto polyhedral convex sets. Some important constraint sets—such as the positive-semidefinite cone—are not so simple. In this work, we avoided trouble by incorporating the positive-semidefinite constraint into the divergence, but this approach is not always warranted. For more general problems, a different approach is necessary.
3. A more serious problem with the successive projection approach is that it offers only linear convergence. For applications, it may be critical to develop algorithms with superlinear convergence.
4. The matrix functions that arise from the study of matrix divergences lead to another challenge. We are not aware of a sophisticated approach to calculating a function such as $\exp(\log \boldsymbol{Y} + \boldsymbol{A})$ other than to work with the corresponding eigendecompositions. Expressions of this form frequently arise in Bregman nearness problems, and we would like to have more robust,

efficient techniques for their computation. Moreover, the numerical stability of various techniques needs to be studied.

5. In applications, it is most important to determine what divergence is appropriate. This choice is likely to depend on domain expertise, coupled with a nuanced understanding of the properties of different divergences.
6. One can also imagine the problem of *learning* a divergence from data. This method would be the ultimate way to match the distance measure with the application. The connection between divergences and exponential families even provides a theoretical justification for this approach.

In conclusion, we have offered evidence that Bregman divergences provide a powerful way to measure the distance between matrices. They can react to structure in the matrix in a way that the Frobenius norm does not. This property makes them extremely valuable for applications, although it may take some effort to determine what divergence is appropriate. Moreover, the numerical methods for computing Bregman projections are still in their infancy. These challenges must be faced before divergences can occupy their potential role in data analysis.

Acknowledgments. We would like to thank Nick Higham and two anonymous referees for a thorough reading and helpful suggestions.

REFERENCES

[1] A. Banerjee, I. S. Dhillon, J. Ghosh, S. Merugu, and D. S. Modha, *A generalized maximum entropy approach to Bregman co-clustering and matrix approximation*, J. Mach. Learn. Res., 8 (2007), pp. 1919–1986. Available online at http://jmlr.csail.mit.edu/papers/volume8/banerjee07a/banerjee07a.pdf.

[2] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, *Clustering with Bregman divergences*, J. Mach. Learn. Res., 6 (2005), pp. 1705–1749.

[3] O. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*, John Wiley, New York, 1978.

[4] H. H. Bauschke and J. M. Borwein, *Legendre functions and the method of random Bregman projections*, J. Convex Anal., 4 (1997), pp. 27–67.

[5] H. H. Bauschke and P. L. Combettes, *Iterating Bregman retractions*, SIAM J. Optim., 13 (2003), pp. 1159–1173.

[6] H. H. Bauschke and A. S. Lewis, *Dykstra's algorithm with Bregman projections: A convergence proof*, Optimization, 48 (2000), pp. 409–427.

[7] S. Boyd and L. Xiao, *Least-squares covariance matrix adjustment*, SIAM J. Matrix Anal. Appl., 27 (2005), pp. 532–546.

[8] L. M. Bregman, *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 200–217.

[9] L. M. Bregman, Y. Censor, and S. Reich, *Dykstra's algorithm as the nonlinear extension of Bregman's optimization method*, J. Convex Anal., 6 (1999), pp. 319–333.

[10] Y. Censor and S. A. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*, Numer. Math. Sci. Comput., Oxford University Press, Oxford, UK, 1997.

[11] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley, New York, 1991.

[12] I. Csiszár, *I-divergence geometry of probability distributions and minimization problems*, Ann. Probab., 3 (1975), pp. 146–158.

[13] W. E. Deming and F. F. Stephan, *On a least squares adjustment of a sampled frequency table when the expected marginal totals are known*, Ann. Math. Statist., 11 (1943), pp. 427–444.

[14] F. Deutsch, *Best Approximation in Inner Product Spaces*, Springer-Verlag, New York, 2001.

[15] I. S. Dhillon, S. Sra, and J. A. Tropp, *Triangle fixing algorithms for the metric nearness problem*, in Proceedings of the Eighteenth Annual Conference on Neural Information Processing Systems (NIPS), MIT Press, Cambridge, MA, 2005, pp. 361–368.

[16] R. L. Dykstra, *An algorithm for restricted least squares regression*, J. Amer. Statist. Assoc., 78 (1983), pp. 837–842.

[17] I. Grubišić and R. Pietersz, *Efficient rank reduction of correlation matrices*, Linear Algebra Appl., 422 (2007), pp. 629–653.

[18] S.-P. Han, *A successive projection method*, Math. Programming, 40 (1988), pp. 1–14.

[19] N. J. Higham, *Computing the nearest correlation matrix—a problem from finance*, IMA J. Numer. Anal., 22 (2002), pp. 329–343.

[20] J.-B. Hiriart-Urruty and C. Lemaréchal, *Fundamentals of Convex Analysis*, Springer, Berlin, 2001.

[21] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[22] C. T. Ireland and S. Kullback, *Contingency tables with given marginals*, Biometrika, 55 (1968), pp. 179–188.

[23] E. T. Jaynes, *Information theory and statistical mechanics*, Phys. Rev., 106 (1957), pp. 620–630.

[24] R. Kruithof, *Telefoonverkeersrekening*, De Ingenieur, 52 (1937), pp. E15–E25.

[25] B. Kulis, M. Sustik, and I. S. Dhillon, *Learning low-rank kernel matrices*, in Proceedings of the Twenty-Third International Conference on Machine Learning (ICML), Morgan Kaufmann, San Francisco, 2006, pp. 505–512.

[26] A. S. Lewis, *The convex analysis of unitarily invariant matrix functions*, J. Convex Anal., 2 (1995), pp. 173–183.

[27] A. S. Lewis, *Convex analysis on the Hermitian matrices*, SIAM J. Optim., 6 (1996), pp. 164–177.

[28] J. Malick, *A dual approach to semidefinite least-squares problems*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 272–284.

[29] K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis*, Academic Press, London, 1979.

[30] M. V. Menon, *Reduction of a matrix with positive elements to a doubly stochastic matrix*, Proc. Amer. Math. Soc., (1967), pp. 244–247.

[31] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*, Cambridge University Press, Cambridge, UK, 2000.

[32] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer, New York, 2006.

[33] B. N. Parlett and T. L. Landis, *Methods for scaling to doubly stochastic form*, Linear Algebra Appl., 48 (1982), pp. 53–79.

[34] H. Qi and D. Sun, *A quadratically convergent Newton method for computing the nearest correlation matrix*, SIAM J. Matrix Anal. Appl., 28 (2006), pp. 360–385.

[35] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[36] R. Sinkhorn, *A relationship between arbitrary positive matrices and doubly stochastic matrices*, Ann. Math. Statist, 35 (1964), pp. 876–879.

[37] P. Tseng, *Dual coordinate ascent methods for non-strictly convex minimization*, Math. Programming, 59 (1993), pp. 231–247.

[38] V. N. Vapnik, *Statistical Learning Theory*, John Wiley, New York, 1998.