An Introduction to Functional Derivatives

Béla A. Frigyik, Santosh Srivastava, Maya R. Gupta

Dept of EE, University of Washington
Seattle WA, 98195-2500

UWEE Technical Report
Number UWEETR-2008-0001
January (updated: July) 2008

Department of Electrical Engineering
University of Washington
Box 352500
Seattle, Washington 98195-2500
PHN: (206) 543-2150
FAX: (206) 543-3842
URL: http://www.ee.washington.edu

An Introduction to Functional Derivatives

Béla A. Frigyik, Santosh Srivastava, Maya R. Gupta

Dept of EE, University of Washington
Seattle WA, 98195-2500

University of Washington, Dept. of EE, UWEETR-2008-0001

January (updated: July) 2008

Abstract

This tutorial on functional derivatives focuses on Fréchet derivatives, a subtopic of functional analysis and of the calculus of variations. The reader is assumed to have experience with real analysis. Definitions and properties are discussed, and examples with functional Bregman divergence illustrate how to work with the Fréchet derivative.

1 Functional Derivatives Generalize the Vector Gradient

Consider a function f defined over vectors such that $f : \mathbb{R}^d \to \mathbb{R}$. The gradient $\nabla f = \{\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d}\}$ describes the instantaneous vector direction in which the function changes the most. The gradient $\nabla f(x_0)$ at $x_0 \in \mathbb{R}^d$ tells you that if you are starting at x_0 which direction would lead to the greatest instantaneous change in f. The inner product (dot product) $\nabla f(x_0)^T y$ for $y \in \mathbb{R}^d$ gives the directional derivative (how much f instantaneously changes) of f at x_0 in the direction defined by the vector y. One generalization of a gradient is the Jacobian, which is the matrix of derivatives for a function that map vectors to vectors ($f : \mathbb{R}^d \to \mathbb{R}^m$).

In this tutorial we consider the generalization of the gradient to functions that map functions to scalars; such functions are called functionals. For example let a functional ϕ be defined over the convex set of functions,

$$\mathcal{G} = \{g : \mathbb{R}^d \to \mathbb{R} \text{ s. t. } \int_x g(x)dx = 1, \text{ and } g(x) \geq 0 \text{ for all } x\}. \tag{1}$$

An example functional defined on this set is the entropy: $\phi : \mathcal{G} \to \mathbb{R}$ where $\phi(g) = -\int_x g(x) \ln g(x)dx$ for $g \in \mathcal{G}$.

In this tutorial we will consider functional derivatives, which are analogs of vector gradients. We will focus on the Fréchet derivative, which can be used to answer questions like, "What function g will maximize $\phi(g)$?" First we will introduce the Fréchet derivative, then discuss higher-order derivatives and some basic properties, and note optimality conditions useful for optimizing functionals. This material will require a familiarity with measure theory that can be found in any standard measure theory text or garnered from the informal measure theory tutorial by Gupta [1]. In Section 3 we illustrate the functional derivative with the definition and properties of the functional Bregman divergence [2]. Readers may find it useful to prove these properties for themselves as an exercise.

2 Fréchet Derivative

Let $\left(\mathbb{R}^d, \Omega, \nu\right)$ be a measure space, where ν is a Borel measure, d is a positive integer, and define the set of functions $\mathcal{A} = \{a \in L^p(\nu) \text{ subject to } a : \mathbb{R}^d \to \mathbb{R}\}$ where $1 \leq p \leq \infty$. The functional $\psi : L^p(\nu) \to \mathbb{R}$ is linear and continuous if

1. $\psi[\omega a_1 + a_2] = \omega\psi[a_1] + \psi[a_2]$ for any $a_1, a_2 \in L^p(\nu)$ and any real number ω
2. there is a constant C such that $|\psi[a]| \leq C\|a\|$ for all $a \in L^p(\nu)$.

Let ϕ be a real functional over the normed space $L^p(\nu)$ such that ϕ maps functions that are L^p integrable with respect to ν to the real line: $\phi : L^p(\nu) \to \mathbb{R}$. The bounded linear functional $\delta\phi[f;\cdot]$ is the Fréchet derivative of ϕ at $f \in L^p(\nu)$ if

$$\phi[f+a] - \phi[f] = \triangle\phi[f;a] = \delta\phi[f;a] + \epsilon[f,a]\,\|a\|_{L^p(\nu)} \tag{2}$$

for all $a \in L^p(\nu)$, with $\epsilon[f,a] \to 0$ as $\|a\|_{L^p(\nu)} \to 0$. Intuitively, what we are doing is perturbing the input function f by another function a, then shrinking the perturbing function a to zero in terms or its L^p norm, and considering the difference $\phi[f+a] - \phi[a]$ in this limit.

Note this functional derivative is linear: $\delta\phi[f;a_1+\omega a_2] = \delta\phi[f;a_1] + \omega\delta\phi[f;a_2]$.

When the second variation $\delta^2\phi$ and the third variation $\delta^3\phi$ exist, they are described by

$$\begin{aligned} \triangle\phi[f;a] &= \delta\phi[f;a] + \frac{1}{2}\delta^2\phi[f;a,a] + \epsilon[f,a]\,\|a\|^2_{L^p(\nu)} \\ &= \delta\phi[f;a] + \frac{1}{2}\delta^2\phi[f;a,a] + \frac{1}{6}\delta^3\phi[f;a,a,a] + \epsilon[f,a]\,\|a\|^3_{L^p(\nu)}\,, \end{aligned} \tag{3}$$

where $\epsilon[f,a] \to 0$ as $\|a\|_{L^p(\nu)} \to 0$. The term $\delta^2\phi[f;a,b]$ is bilinear with respect to arguments a and b, and $\delta^3\phi[f;a,b,c]$ is trilinear with respect to a, b, and c.

2.1 Fréchet Derivatives and Sequences of Functions

Consider sequences of functions $\{a_n\}, \{f_n\} \subset L^p(\nu)$, where $a_n \to a$, $f_n \to f$, and $a, f \in L^p(\nu)$. If $\phi \in \mathcal{C}^3(L^p(\nu);\mathbb{R})$ and $\delta\phi[f;a]$, $\delta^2\phi[f;a,a]$, and $\delta^3[f;a,a,a]$ are defined as above, then

$$\delta\phi[f_n;a_n] \to \delta\phi[f;a],\;\; \delta^2\phi[f_n;a_n,a_n] \to \delta^2\phi[f;a,a], \quad \text{and } \delta^3\phi[f_n;a_n,a_n,a_n] \to \delta^3\phi[f;a,a,a].$$

2.2 Strongly Positive is Analog to Positive Definite

The quadratic functional $\delta^2\phi[f;a,a]$ defined on normed linear space $L^p(\nu)$ is ***strongly positive*** if there exists a constant $k > 0$ such that $\delta^2\phi[f;a,a] \geq k\,\|a\|^2_{L^p(\nu)}$ for all $a \in \mathcal{A}$. In a finite-dimensional space, strong positivity of a quadratic form is equivalent to the quadratic form being positive definite.

From (3),

$$\begin{aligned} \phi[f+a] &= \phi[f] + \delta\phi[f;a] + \frac{1}{2}\delta^2\phi[f;a,a] + o(\|a\|^2), \\ \phi[f] &= \phi[f+a] - \delta\phi[f+a;a] + \frac{1}{2}\delta^2\phi[f+a;a,a] + o(\|a\|^2), \end{aligned}$$

where $o(\|a\|^2)$ denotes a function that goes to zero as $\|a\|$ goes to zero, even if it is divided by $\|a\|^2$. Adding the above two equations and canceling the ϕ's yields

$$0 = \delta\phi[f;a] - \delta\phi[f+a;a] + \frac{1}{2}\delta^2\phi[f;a,a] + \frac{1}{2}\delta^2\phi[f+a;a,a] + o(\|a\|^2),$$

which is equivalent to

$$\delta\phi[f+a;a] - \delta\phi[f;a] = \delta^2\phi[f;a,a] + o(\|a\|^2), \tag{4}$$

because

$$\left|\delta^2\phi[f+a;a,a] - \delta^2\phi[f;a,a]\right| \leq \|\delta^2\phi[f+a;\cdot,\cdot] - \delta^2\phi[f;\cdot,\cdot]\|\|a\|^2,$$

and we assumed $\phi \in \mathcal{C}^2$, so $\delta^2\phi[f+a;a,a] - \delta^2\phi[f;a,a]$ is of order $o(\|a\|^2)$. This shows that the variation of the first variation of ϕ is the second variation of ϕ. A procedure like the above can be used to prove that analogous statements hold for higher variations if they exist.

2.3 Functional Optimality Conditions

Consider a functional J and the problem of finding the function $\hat{f}$ such that $J[\hat{f}]$ achieves a local minimum of J.

For $J[f]$ to have an extremum (minimum) at $\hat{f}$, it is necessary that

$$\delta J[f; a] = 0 \text{ and } \delta^2 J[f; a, a] \geq 0,$$

for $f = \hat{f}$ and for all admissible functions $a \in \mathcal{A}$. A sufficient condition for $\hat{f}$ to be a minimum is that the first variation $\delta J[f; a]$ must vanish for $f = \hat{f}$, and its second variation $\delta^2 J[f; a, a]$ must be strongly positive for $f = \hat{f}$.

2.4 Other Functional Derivatives

The Fréchet derivative is a common functional derivative, but other functional derivatives have been defined for various purposes. Another common one is the Gâteaux derivative, which instead of considering any perturbing function a in (2), only considers perturbing functions in a particular direction.

3 Illustrating the Fréchet Derivative: Functional Bregman Divergence

We illustrate working with the Fréchet derivative by introducing a class of distortions between any two functions called the functional Bregman divergences, giving an example for squared error, and then proving a number of properties.

First, we review the vector case. Bregman divergences were first defined for vectors [3], and are a class of distortions that includes squared error, relative entropy, and many other dissimilarities common in engineering and statistics [4]. Given any strictly convex and twice differentiable function $\tilde{\phi} : \mathbb{R}^n \to \mathbb{R}$, you can define a Bregman divergence over vectors $x, y \in \mathbb{R}^n$ that are admissible inputs to ϕ:

$$d_{\tilde{\phi}}(x, y) = \tilde{\phi}(x) - \tilde{\phi}(y) - \nabla\tilde{\phi}(y)^T(x - y). \tag{5}$$

By re-arranging the terms of (5), one sees that the Bregman divergence $d_{\tilde{\phi}}$ is the tail of the Taylor series expansion of ϕ around y:

$$\tilde{\phi}(x) = \tilde{\phi}(y) + \nabla\tilde{\phi}(y)^T(x - y) + d_{\tilde{\phi}}(x, y). \tag{6}$$

The Bregman divergences have the useful property that the mean of a set has the minimum mean Bregman divergence to all the points in the set [4].

Recently, we generalized Bregman divergence to a functional Bregman divergence [5] [2] in order to show that the mean of a set of functions minimizes the mean Bregman divergence to the set of functions. The functional Bregman divergence is a straightforward analog to the vector case. Let $\phi : L^p(\nu) \to \mathbb{R}$ be a strictly convex, twice-continuously Fréchet-differentiable functional. The Bregman divergence $d_\phi : \mathcal{A} \times \mathcal{A} \to [0, \infty)$ is defined for all $f, g \in \mathcal{A}$ as

$$d_\phi[f, g] = \phi[f] - \phi[g] - \delta\phi[g; f - g], \tag{7}$$

where $\delta\phi[g; f - g]$ is the Fréchet derivative of ϕ at g in the direction of $f - g$.

3.1 Squared Error Example

Let's consider how a particular choice of ϕ turns (7) into the total squared error between two functions. Let $\phi[g] = \int g^2 d\nu$, where $\phi : L^2(\nu) \to \mathbb{R}$, and let $g, f, a \in L^2(\nu)$. Then

$$\phi[g + a] - \phi[g] = \int (g + a)^2 d\nu - \int g^2 d\nu = 2\int gad\nu + \int a^2 d\nu.$$

Because

$$\frac{\int a^2 d\nu}{\|a\|_{L^2(\nu)}} = \frac{\|a\|^2_{L^2(\nu)}}{\|a\|_{L^2(\nu)}} = \|a\|_{L^2(\nu)} \to 0$$

as $a \to 0$ in $L^2(\nu)$, it holds that

$$\delta\phi[g; a] = 2\int gad\nu,$$

which is a continuous linear functional in a. Then, by definition of the second Fréchet derivative,

$$\begin{aligned}\delta^2\phi[g;b,a] &= \delta\phi[g+b;a]-\delta\phi[g;a]\\ &= 2\int(g+b)ad\nu-2\int gad\nu\\ &= 2\int bad\nu.\end{aligned}$$

Thus $\delta^2\phi[g;b,a]$ is a quadratic form, where $\delta^2\phi$ is actually independent of g and strongly positive since

$$\delta^2\phi[g;a,a]=2\int a^2d\nu=2\|a\|^2_{L^2(\nu)}$$

for all $a\in L^2(\nu)$, which implies that ϕ is strictly convex and

$$\begin{aligned}d_\phi[f,g] &= \int f^2d\nu-\int g^2d\nu-2\int g(f-g)d\nu\\ &= \int(f-g)^2d\nu\\ &= \|f-g\|^2_{L^2(\nu)}.\end{aligned}$$

3.2 Properties of Functional Bregman Divergence

Next we establish some properties of the functional Bregman divergence. We have listed these in order of easiest to prove to hardest in case the reader would like to use proving the properties as exercises.

Linearity
The functional Bregman divergence is linear with respect to ϕ.

Proof:

$$d_{(c_1\phi_1+c_2\phi_2)}[f,g]=(c_1\phi_1+c_2\phi_2)[f]-(c_1\phi_1+c_2\phi_2)[g]-\delta(c_1\phi_1+c_2\phi_2)[g;f-g]=c_1d_{\phi_1}[f,g]+c_2d_{\phi_2}[f,g]. \quad (8)$$

Convexity
The Bregman divergence $d_\phi[f,g]$ is always convex with respect to f.

Proof: Consider

$$\begin{aligned}\triangle d_\phi[f,g;a] &= d_\phi[f+a,g]-d_\phi[f,g]\\ &= \phi[f+a]-\phi[f]-\delta\phi[g;f-g+a]+\delta\phi[g;f-g].\end{aligned}$$

Using linearity in the third term,

$$\begin{aligned}&\triangle d_\phi[f,g;a]\\ &= \phi[f+a]-\phi[f]-\delta\phi[g;f-g]-\delta\phi[g;a]+\delta\phi[g;f-g],\\ &= \phi[f+a]-\phi[f]-\delta\phi[g;a],\\ &\overset{(a)}{=} \delta\phi[f;a]+\frac{1}{2}\delta^2\phi[f;a,a]+\epsilon[f,a]\,\|a\|^2_{L(\nu)}-\delta\phi[g;a]\\ &\Rightarrow \delta^2d_\phi[f,g;a,a]=\frac{1}{2}\delta^2\phi[f;a,a]>0,\end{aligned}$$

where (a) and the conclusion follows from (3).

Linear Separation
The set of functions $f\in\mathcal{A}$ that are equidistant from two functions $g_1,g_2\in\mathcal{A}$ in terms of functional Bregman divergence form a hyperplane.

Proof: Fix two non-equal functions $g_1, g_2 \in \mathcal{A}$, and consider the set of all functions in $\mathcal{A}$ that are equidistant in terms of functional Bregman divergence from g_1 and g_2:

$$\begin{aligned} & d_\phi[f, g_1] = d_\phi[f, g_2] \\ & \quad \Rightarrow \quad -\phi[g_1] - \delta\phi[g_1; f - g_1] = -\phi[g_2] - \delta\phi[g_2; f - g_2] \\ & \quad \Rightarrow \quad -\delta\phi[g_1; f - g_1] = \phi[g_1] - \phi[g_2] - \delta\phi[g_2; f - g_2]. \end{aligned}$$

Using linearity the above relationship can be equivalently expressed as

$$\begin{aligned} -\delta\phi[g_1; f] + \delta\phi[g_1; g_1] &= \phi[g_1] - \phi[g_2] - \delta\phi[g_2; f] + \\ &\quad \delta\phi[g_2; g_2], \\ \delta\phi[g_2; f] - \delta\phi[g_1; f] &= \phi[g_1] - \phi[g_2] - \delta\phi[g_1; g_1] + \\ &\quad \delta\phi[g_2; g_2]. \\ Lf &= c, \end{aligned}$$

where L is the bounded linear functional defined by $Lf = \delta\phi[g_2; f] - \delta\phi[g_1; f]$, and c is the constant corresponding to the right-hand side. In other words, f has to be in the set $\{a \in \mathcal{A} : La = c\}$, where c is a constant. This set is a hyperplane.

Generalized Pythagorean Inequality For any $f, g, h \in \mathcal{A}$,

$$d_\phi[f, h] = d_\phi[f, g] + d_\phi[g, h] + \delta\phi[g; f - g] - \delta\phi[h; f - g].$$

Proof:

$$\begin{aligned} & d_\phi[f, g] + d_\phi[g, h] \\ & \quad = \quad \phi[f] - \phi[h] - \delta\phi[g; f - g] - \delta\phi[h; g - h] \\ & \quad = \quad \phi[f] - \phi[h] - \delta\phi[h; f - h] + \delta\phi[h; f - h] \\ & \qquad\quad -\delta\phi[g; f - g] - \delta\phi[h; g - h] \\ & \quad = \quad d_\phi[f, h] + \delta\phi[h; f - g] - \delta\phi[g; f - g], \end{aligned}$$

where the last line follows from the definition of the functional Bregman divergence and the linearity of the fourth and last terms.

Equivalence Classes
Partition the set of strictly convex, differentiable functions $\{\phi\}$ on $\mathcal{A}$ into classes with respect to functional Bregman divergence, so that ϕ_1 and ϕ_2 belong to the same class if $d_{\phi_1}[f, g] = d_{\phi_2}[f, g]$ for all $f, g \in \mathcal{A}$. For brevity we will denote $d_{\phi_1}[f, g]$ simply by d_{ϕ_1}. Let $\phi_1 \sim \phi_2$ denote that ϕ_1 and ϕ_2 belong to the same class, then $\sim$ is an equivalence relation because it satisfies the properties of *reflexivity* (because $d_{\phi_1} = d_{\phi_1}$), *symmetry* (because if $d_{\phi_1} = d_{\phi_2}$, then $d_{\phi_2} = d_{\phi_1}$), and *transitivity* (because if $d_{\phi_1} = d_{\phi_2}$ and $d_{\phi_2} = d_{\phi_3}$, then $d_{\phi_1} = d_{\phi_3}$).

Further, if $\phi_1 \sim \phi_2$, then they differ only by an affine transformation.

Proof: It only remains to be shown that if $\phi_1 \sim \phi_2$, then they differ only by an affine transformation. Note that by assumption, $\phi_1[f] - \phi_1[g] - \delta\phi_1[g; f - g] = \phi_2[f] - \phi_2[g] - \delta\phi_2[g; f - g]$, and fix g so $\phi_1[g]$ and $\phi_2[g]$ are constants. By the linearity property, $\delta\phi[g; f - g] = \delta\phi[g; f] - \delta\phi[g; g]$, and because g is fixed, this equals $\delta\phi[g; f] + c_0$ where c_0 is a scalar constant. Then $\phi_2[f] = \phi_1[f] + (\delta\phi_2[g; f] - \delta\phi_1[g; f]) + c_1$, where c_1 is a constant. Thus,

$$\phi_2[f] = \phi_1[f] + Af + c_1,$$

where $A = \delta\phi_2[g; \cdot] - \delta\phi_1[g; \cdot]$, and thus $A : \mathcal{A} \to \mathbb{R}$ is a linear operator that does not depend on f.

Dual Divergence
Given a pair (g, ϕ) where $g \in L^p(\nu)$ and ϕ is a strictly convex twice-continuously Fréchet-differentiable functional,

then the function-functional pair (G, ψ) is the Legendre transform of (g, ϕ) [6], if

$$\phi[g] = -\psi[G] + \int g(x)G(x)d\nu(x), \tag{9}$$

$$\delta\phi[g; a] = \int G(x)a(x)d\nu(x), \tag{10}$$

where ψ is a strictly convex twice-continuously Fréchet-differentiable functional, and $G \in L^q(\nu)$, where $\frac{1}{p} + \frac{1}{q} = 1$.

Given Legendre transformation pairs $f, g \in L^p(\nu)$ and $F, G \in L^q(\nu)$,

$$d_\phi[f, g] = d_\psi[G, F].$$

Proof: The proof begins by substituting (9) and (10) into (7):

$$\begin{aligned} d_\phi[f, g] &= \phi[f] + \psi[G] - \int g(x)G(x)d\nu(x) - \int G(x)(f - g)(x)d\nu(x) \\ &= \phi[f] + \psi[G] - \int G(x)f(x)d\nu(x). \end{aligned} \tag{11}$$

Applying the Legendre transformation to (G, ψ) implies that

$$\psi[G] = -\phi[g] + \int g(x)G(x)d\nu(x) \tag{12}$$

$$\delta\psi[G; a] = \int g(x)a(x)d\nu(x). \tag{13}$$

Using (12) and (13), $d_\psi[G, F]$ can be reduced to (11).

Non-negativity
The functional Bregman divergence is non-negative.

Proof: To show this, define $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$ by $\tilde{\phi}(t) = \phi\left[tf + (1 - t)g\right]$, $f, g \in \mathcal{A}$. From the definition of the Fréchet derivative,

$$\frac{d}{dt}\tilde{\phi} = \delta\phi[tf + (1 - t)g; f - g]. \tag{14}$$

The function $\tilde{\phi}$ is convex because ϕ is convex by definition. Then from the mean value theorem there is some $0 \leq t_0 \leq 1$ such that

$$\tilde{\phi}(1) - \tilde{\phi}(0) = \frac{d}{dt}\tilde{\phi}(t_0) \geq \frac{d}{dt}\tilde{\phi}(0). \tag{15}$$

Because $\tilde{\phi}(1) = \phi[f]$, $\tilde{\phi}(0) = \phi[g]$, and (14), subtracting the right-hand side of (15) implies that

$$\phi[f] - \phi[g] - \delta\phi[g, f - g] \geq 0. \tag{16}$$

If $f = g$, then (16) holds in equality. To finish, we prove the converse. Suppose (16) holds in equality; then

$$\tilde{\phi}(1) - \tilde{\phi}(0) = \frac{d}{dt}\tilde{\phi}(0). \tag{17}$$

The equation of the straight line connecting $\tilde{\phi}(0)$ to $\tilde{\phi}(1)$ is $\ell(t) = \tilde{\phi}(0) + (\tilde{\phi}(1) - \tilde{\phi}(0))t$, and the tangent line to the curve $\tilde{\phi}$ at $\tilde{\phi}(0)$ is $y(t) = \tilde{\phi}(0) + t\frac{d}{dt}\tilde{\phi}(0)$. Because $\tilde{\phi}(\tau) = \tilde{\phi}(0) + \int_0^\tau \frac{d}{dt}\tilde{\phi}(t)dt$ and $\frac{d}{dt}\tilde{\phi}(t) \geq \frac{d}{dt}\tilde{\phi}(0)$ as a direct consequence of convexity, it must be that $\tilde{\phi}(t) \geq y(t)$. Convexity also implies that $\ell(t) \geq \tilde{\phi}(t)$. However, the assumption that (16) holds in equality implies (17), which means that $y(t) = \ell(t)$, and thus $\tilde{\phi}(t) = \ell(t)$, which is not strictly convex. Because ϕ is by definition strictly convex, it must be true that $\phi[tf + (1 - t)g] < t\phi[f] + (1 - t)\phi[g]$ unless $f = g$. Thus, under the assumption of equality of (16), it must be true that $f = g$.

4 Further Reading

For further reading, try the text by Gelfand and Fomin [6], and the wikipedia pages on functional derivatives, Fréchet derivatives, and Gâteaux derivatives. Readers may also find our paper [2] helpful, which further illustrates the use of functional derivatives in the context of the functional Bregman divergence, conveniently using the same notation as this introduction.

References

[1] M. R. Gupta, "A measure theory tutorial: Measure theory for dummies," *Univ. of Washington Technical Report 2006-0008*, Available at idl.ee.washington.edu/publications.php.

[2] B. A. Frigyik, S. Srivastava, and M. R. Gupta, "Functional Bregman divergence and Bayesian estimation of distributions," *To appear: IEEE Trans. on Information Theory*, available at idl.ee.washington.edu/publications.php.

[3] L. Bregman, "The relaxation method of finding the common points of convex sets and its application to the solution of problems in convex programming," *USSR Computational Mathematics and Mathematical Physics*, vol. 7, pp. 200–217, 1967.

[4] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[5] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian quadratic discriminant analysis," *Journal of Machine Learning Research*, vol. 8, pp. 1287–1314, 2007.

[6] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*. USA: Dover, 2000.