On the Optimality of Conditional Expectation as a Bregman Predictor

Arindam Banerjee*
Department of Electrical and Computer Engineering
University of Texas at Austin
Austin, TX 78712.
abanerje@ece.utexas.edu

Xin Guo*
School of Operations Research and Industrial Engineering
Cornell University
Ithaca, NY 14583.
xinguo@orie.cornell.edu

Hui Wang
Division of Applied Mathematics
Brown University
Providence, RI 02912.
huiwang@cfm.brown.edu

Abstract

Given a probability space $(\Omega, \mathcal{F}, P)$, a $\mathcal{F}$-measurable random variable X, and a sub-σ-algebra $\mathcal{G} \subset \mathcal{F}$, it is well known that the conditional expectation $E[X|\mathcal{G}]$ is the optimal $\mathbb{L}^2$-predictor (also known as "the least mean square error" predictor) of X, among all the $\mathcal{G}$-measurable random variables [8, 11]. In this paper, we provide necessary and sufficient conditions for general loss functions under which the conditional expectation is the unique optimal predictor. We show that $E[X|\mathcal{G}]$ is the optimal predictor for all Bregman Loss Functions (BLFs), of which the $\mathbb{L}^2$-loss function is a special case. Moreover, under mild conditions, we show that the BLFs are exhaustive, i.e., if for all random variables X, the infimum of $E[F(X,Y)]$ over $\mathcal{G}$-measurable random variables Y is attained by the conditional expectation $E[X|\mathcal{G}]$, then F is a BLF.

*part of the work was done when the authors were at IBM T. J. Watson Research Center, NY.

1 Introduction

The problem of predicting the value of a random outcome based on available information arises in many contexts. To put the problem into a mathematical framework, let $(\Omega, \mathcal{F}, P)$ be a probability space and let X be a $\mathcal{F}$-measurable random variable that one wishes to predict. The available information is represented by a sub-σ-algebra of $\mathcal{F}$, say $\mathcal{G}$. Now, the question is: among all $\mathcal{G}$-measurable random variables, which one is the best predictor of X.

The notion of "best" is usually specified by a non-negative loss function F and achieved by solving a corresponding minimization problem. More precisely, the best predictor is defined as the minimizer of $E[F(X,Y)]$ over all $\mathcal{G}$-measurable random variables Y. A particularly important case is when F is the so called $\mathbb{L}^2$-loss function, also known as the mean square error, i.e., $F(x,y) \doteq \|x-y\|^2$. It is well known [8, 11] that the corresponding *unique* best predictor is given by the conditional expectation. In other words, if we write $Y \in \mathcal{G}$ for a $\mathcal{G}$-measurable random variable Y, then

$$\operatorname*{argmin}_{Y \in \mathcal{G}} E\left[\|X-Y\|^2\right] = E[X|\mathcal{G}].$$

This makes conditional expectation crucially important for prediction.

A question arises naturally: *Are there other loss functions F for which $E[X|\mathcal{G}]$ is the unique best predictor?* Some simple counter-examples lead to the general conviction that the existence of such loss functions would be rare and would have to possess very special properties. For example, if one uses the absolute error loss function ([9], Section 1.7) and take $\mathcal{G} = \{\emptyset, \Omega\}$, then any constant a satisfying $P(X \leq a) \geq 1/2 \geq P(X > a)$, i.e., the median of X and not $E[X|\mathcal{G}]$, proves to be the best predictor. Recently [1] studied the case of general convex loss functions and obtained a criterion for which a minimizing value exists when $\mathcal{G} = \mathcal{F}$.

In this paper, we provide necessary and sufficient conditions for general loss functions under which the conditional expectation is the unique optimal predictor. First, we show that the optimality property of the conditional expectation holds for *all* functions known as Bregman Loss Functions (BLFs) [3], of which the $\mathbb{L}^2$-loss function is a special case. Indeed, one can essentially create as many BLFs as convex functions, up to equivalences in linear and constant terms (see Definition 1). Secondly, we show that the class of BLFs is exhaustive under mild conditions, i.e., if the conditional expectation minimizes the expected loss function for all random variables X, then the loss function has to be a BLF.

2 Bregman Loss Functions

Definition 1 (Bregman Loss Functions) *Let $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ be a strictly convex, differentiable function. Then, the Bregman Loss Function $D_\phi : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is defined as*

$$D_\phi(x,y) = \phi(x) - \phi(y) - \langle x-y, \nabla\phi(y)\rangle.$$

Table 1: Examples of BLFs.

Domain	$\phi(x)$	$D_\phi(x,y)$	**Loss**
$\mathbb{R}$	x^2	$(x-y)^2$	$\mathbb{L}^2$-loss
$\mathbb{R}_{++}$	$x\log x$	$x\log(x/y)-(x-y)$	
$(0,1)$	$x\log x+(1-x)\log(1-x)$	$x\log(x/y)+(1-x)\log((1-x)/(1-y))$	Logistic loss
$\mathbb{R}_{++}$	$-\log x$	$x/y-\log(x/y)-1$	Itakura-Saito distance
$\mathbb{R}$	e^x	$e^x-e^y-(x-y)e^y$	
$\mathbb{R}^d$	$\|x\|^2$	$\|x-y\|^2$	$\mathbb{L}^2$-loss
$\mathbb{R}^d$	x^TAx	$(x-y)^TA(x-y)$	Mahalanobis distance [1]
d-simplex	$\sum_{j=1}^d x_j\log x_j$	$\sum_{j=1}^d x_j\log(x_j/y_j)$	KL-divergence
$\mathbb{R}_+^d$	$\sum_{j=1}^d x_j\log x_j$	$\sum_{j=1}^d x_j\log(x_j/y_j)-\sum_{j=1}^d(x_j-y_j)$	Generalized I-divergence

Example 1: The well-known $\mathbb{L}^2$-loss function is perhaps the simplest and most widely used loss function. It is a special case of BLFs, with $\phi(x) \doteq \langle x,x\rangle$, so that

$$D_\phi(x,y) = \langle x,x\rangle - \langle y,y\rangle - \langle x-y, 2y\rangle = \|x-y\|^2.$$

Example 2: Another widely used BLF is the Kullback-Liebler (KL) divergence. Let $p \doteq (p_1,\ldots,p_d)$ be a discrete probability distribution so that $\sum_{j=1}^d p_j = 1$. The negative Shannon entropy, $\phi(p) \doteq \sum_{j=1}^d p_j\log p_j$, is a strictly convex function on the d-simplex. Let $q=(q_1,\ldots,q_d)$ be another probability distribution. The corresponding BLF is

$$\begin{aligned} D_\phi(p,q) &= \sum_{j=1}^d p_j\log p_j - \sum_{j=1}^d q_j\log q_j - \langle p-q, \nabla\phi(q)\rangle \\ &= \sum_{j=1}^d p_j\log p_j - \sum_{j=1}^d q_j\log q_j - \sum_{j=1}^d (p_j-q_j)(\log e+\log q_j) \\ &= \sum_{j=1}^d p_j\log\left(p_j/q_j\right), \end{aligned}$$

which is exactly the KL-divergence $KL(p\|q)$ between p and q.

Table 1 contains a list of some common convex functions and their corresponding BLFs. The following useful observation follows from the strict convexity of ϕ [6, Proposition 5.4].

Lemma 1 *For any $x,y\in\mathbb{R}^d$, $D_\phi(x,y)\geq 0$, and the equality holds if and only if $x=y$.*

Remark 1 Since a differentiable convex function is necessarily continuously differentiable [10, Theorem 25.5], the function D_ϕ is continuous. Moreover, if we write ∇_x as the gradient with respect to x, then the function

$$\nabla_x D_\phi(x,y) = \nabla\phi(x) - \nabla\phi(y)$$

[1] The matrix A is assumed to be strictly positive definite.

is also continuous. For more discussions on BLFs, interested readers are referred to [2] and the references therein.

3 The optimal Bregman predictor

In this section we will show that the conditional expectation is the unique optimal predictor for all BLFs, and that any nearly optimal predictor will converge in probability to the conditional expectation.

Theorem 1 (Optimality Property) *Let $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ be a strictly convex, differentiable function and let D_ϕ be the corresponding BLF. Let $(\Omega, \mathcal{F}, P)$ be an arbitrary probability space and let $\mathcal{G}$ be a sub-σ-algebra of $\mathcal{F}$. Let X be any $\mathcal{F}$-measurable random variable taking values in $\mathbb{R}^d$ for which both $E[X]$ and $E[\phi(X)]$ are finite. Then, among all $\mathcal{G}$-measurable random variables, the conditional expectation is the unique minimizer (up to a.s. equivalence) of the expected Bregman loss, i.e.,*

$$\text{argmin}_{Y \in \mathcal{G}} E[D_\phi(X, Y)] = E[X|\mathcal{G}].$$

Proof: Let Y be any $\mathcal{G}$-measurable random variable, and $Y^* \doteq E[X|\mathcal{G}]$. It follows from Definition 1 that

$$\begin{aligned} & E[D_\phi(X, Y)] - E[D_\phi(X, Y^*)] \\ &= E\left[\phi(Y^*) - \phi(Y) - \langle X - Y, \nabla\phi(Y)\rangle + \langle X - Y^*, \nabla\phi(Y^*)\rangle\right]. \end{aligned}$$

Meanwhile, for any $\mathcal{G}$-measurable random variable Y, we have

$$E[\langle X - Y, \nabla\phi(Y)\rangle] = E[E[\langle X - Y, \nabla\phi(Y)\rangle|\mathcal{G}]] = E[\langle Y^* - Y, \nabla\phi(Y)\rangle].$$

In particular, $E[\langle X - Y^*, \nabla\phi(Y^*)\rangle] = 0$. Therefore,

$$\begin{aligned} E[D_\phi(X, Y)] - E[D_\phi(X, Y^*)] &= E[\phi(Y^*) - \phi(Y) - \langle Y^* - Y, \nabla\phi(Y)\rangle] \\ &= E[D_\phi(Y^*, Y)]. \end{aligned} \tag{1}$$

The theorem follows immediately from Lemma 1. ∎

Theorem 2 (Convergence in Probability) *In the setting of Theorem 1, if $\{Y_n\}$ is an infimizing sequence, i.e., Y_n is $\mathcal{G}$-measurable and*

$$E[D_\phi(X, Y_n)] \to E[D_\phi(X, Y^*)],$$

where $Y^* \doteq E[X|\mathcal{G}]$, then $Y_n \to Y^*$ in probability.

Proof: It suffices to show that for any given $\epsilon, \delta > 0$, there exists a number N such that

$$P(|Y_n - Y^*| \geq \delta) \leq \epsilon$$

$\forall n \geq N$. The integrability of X (and hence of Y^*) suggests that for a given $\epsilon > 0, \exists M$ such that

$$P(|Y^*| \geq M) \leq \epsilon/2.$$

Hence

$$\begin{aligned} P(|Y_n - Y^*| \geq \delta) &\leq P(|Y_n - Y^*| \geq \delta, |Y^*| \leq M) + P(|Y^*| \geq M) \\ &\leq P(|Y_n - Y^*| \geq \delta, |Y^*| \leq M) + \epsilon/2. \end{aligned}$$

For every $x \in \mathbb{R}^d$, if we define

$$h(x) \doteq \inf\{D_\phi(x, y) : \ y \in \mathbb{R}^d, \ |y - x| \geq \delta\},$$

then the strict convexity of ϕ implies that $h(x) > 0, \forall x \in \mathbb{R}^d$, and

$$h(x) = \inf\{D_\phi(x, y) : \ y \in \mathbb{R}^d, \ |y - x| = \delta\}.$$

Since D_ϕ is continuous (Remark 1), the infimum is always achieved. Moreover, it can be shown that

$$\alpha \doteq \inf\{h(x) : |x| \leq M\} > 0. \tag{2}$$

For now assuming (2) to be true, we have

$$P(|Y_n - Y^*| \geq \delta) \leq P(D_\phi(Y^*, Y_n) \geq \alpha) + \epsilon/2 \leq E[D_\phi(Y^*, Y_n)]/\alpha + \epsilon/2.$$

Since Y_n is an infimizing sequence, from (1) it follows that $E[D_\phi(Y^*, Y_n)] \to 0$. Hence, there exists N such that for $n \geq N$, $E[D_\phi(Y, Y_n)] \leq \epsilon\alpha/2$. Therefore, for $n \geq N$,

$$P(|Y_n - Y^*| \geq \delta) \leq \epsilon,$$

and hence we have convergence in probability.

Finally, we show that $\alpha > 0$. This is proved by contradiction. Clearly $\alpha \not< 0$. Suppose $\alpha = 0$. Then there exists a sequence $\{x_n\}$ with $|x_n| \leq M$ and a sequence $\{y_n\}$ with $|y_n - x_n| = \delta$ such that

$$h(x_n) = D_\phi(x_n, y_n) \to 0.$$

Since $\{x_n\}$ and $\{y_n\}$ are both bounded, there exists a subsequence (still indexed by n) such that

$$x_n \to \bar{x}, \quad y_n \to \bar{y}.$$

Clearly $|\bar{x}| \leq M$ and $|\bar{y} - \bar{x}| = \delta$. The continuity of D_ϕ yields that $D_\phi(\bar{x}, \bar{y}) = 0$, which contradicts $h(\bar{x}) > 0$.
This completes the proof. ∎

Remark 2 Other types of convergence results may be obtained by imposing proper conditions on the function ϕ. For example, it is easy to see that $Y_n \to Y^*$ in $\mathbb{L}^2$ if the Hessian matrix of ϕ is uniformly positive definite over $\mathbb{R}^d$ (in the 1-dim case, it amounts to $\inf_{x \in \mathbb{R}} \phi''(x) > 0$).

4 The Exhaustiveness property of BLFs

In this section we establish exhaustiveness results for the class of loss functions for which the conditional expectation is the optimal predictor. More precisely, under mild regularity conditions we show that if $F : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ is a non-negative loss function such that

$$\text{argmin}_{Y \in \mathcal{G}} \; E\left[F(X,Y)\right] = E[X|\mathcal{G}], \tag{3}$$

for all random variable X, then F has to be a BLF.

We will present the results separately for the one-dimensional (Theorem 3) and the higher-dimensional (Theorem 4) case, since the latter needs slightly stronger regularity conditions; see section 5 for more discussions.

For ease of exposition, and without loss of generality, we will assume in Theorem 3 and Theorem 4 that $F(x,x) = 0, \forall x$. Indeed, if F is a loss function satisfying (3), so is $\bar{F}(x,y) \doteq F(x,y) - F(x,x)$ with $\bar{F}(x,x) \equiv 0$.

Theorem 3 ($d = 1$) *Let $F : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ be a non-negative function such that $F(x,x) = 0, \forall x \in \mathbb{R}$. Assume that F and F_x are both continuous functions. If for all random variables X, $E[X|\mathcal{G}]$ is the unique minimizer of $E[F(X,Y)]$ over all random variables $Y \in \mathcal{G}$, i.e.,*

$$\text{argmin}_{Y \in \mathcal{G}} \; E[F(X,Y)] = E[X|\mathcal{G}],$$

then $F(x,y) = D_\phi(x,y)$ for some strictly convex, differentiable function $\phi : \mathbb{R} \mapsto \mathbb{R}$.

Proof: The proof will be completed in three steps. First, we prove that $F = D_\phi$ for some convex, differentiable function ϕ, under an additional assumption that F_y is continuous; we then extend this result to the general case by a mollification argument; finally, we show that ϕ is strictly convex.

Step 1: Assume F_x and F_y are both continuous. Fix arbitrarily $a, b \in \mathbb{R}$, and $p \in [0,1]$. Consider a random variable X on some probability space $(\Omega, \mathcal{F}, P)$ such that $P(X = a) = p$ and $P(X = b) = q$ with $p + q = 1$. Let $\mathcal{G} = \{\emptyset, \Omega\}$. Fix $Y = y$, then from the assumption

$$pF(a,y) + qF(b,y) = E[F(X,Y)] \geq E[F(X,EX)] = pF(a,pa+qb) + qF(b,pa+qb)$$

for all $y \in R$. Moreover, if we consider the left-hand-side as a function of y, it equals the right-hand-side at $y = y^* \doteq E[X] = pa + qb$. Therefore, we must have

$$pF_y(a,y^*) + qF_y(b,y^*) = 0. \tag{4}$$

Substituting $p = (y^* - b)/(a - b)$ and rearranging terms yield

$$F_y(a,y^*)/(y^* - a) = F_y(b,y^*)/(y^* - b).$$

Since a, b and p are arbitrary, the above equality implies that the function

$$F_y(x,y)/(y - x)$$

is independent of x. Thus one can write, for some function H,

$$F_y(x,y) = (y-x)H(y), \tag{5}$$

where H is continuous.

Now define function ϕ by

$$\phi(y) \doteq \int_0^y \int_0^t H(s)dsdt.$$

Then ϕ is differentiable with $\phi(0) = \phi'(0) = 0$, $\phi''(y) = H(y)$. Since $F(x,x) = 0$, integration by parts for (5) leads to

$$F(x,y) = \int_x^y (s-x)H(s)\,ds = \phi(x) - \phi(y) - \phi'(y)(x-y).$$

It follows from the non-negativity of F that ϕ is a convex function.

Step 2: Now we show that there exists a convex function ϕ such that $F = D_\phi$ under the assumption of the theorem. Consider a sequence of mollifiers, i.e., a sequence of functions $\{g_n\}$ defined on $\mathbb{R}$, which are non-negative, $\mathcal{C}^\infty$ and with compact support such that

$$\int_{\mathbb{R}} g_n(x)\,dx = 1.$$

A classical example for such a sequence of mollifiers is as follows: let

$$g(x) \doteq \begin{cases} c\exp\left\{1/(x^2-1)\right\} & \text{if } |x| < 1, \\ 0 & \text{if } |x| \geq 1, \end{cases}$$

where the constant c is to be chosen so that $\int_{\mathbb{R}} g(x)dx = 1$, and define $g_n(x) \doteq ng(nx)$. The mollified version of F is then given by

$$F_n(x,y) \doteq \int_{\mathbb{R}} F(x-u, y-u)g_n(u)\,du = \int_{\mathbb{R}} F(x-y+u, u)g_n(y-u)\,du.$$

It is standard to show that [7, Section 7.2] F_n is continuously differentiable with respect to x and y, and that

$$\lim_{n\to\infty} F_n(x,y) = F(x,y),$$

for every $x, y \in \mathbb{R}$.

Furthermore, it is easy to see that F_n has the same property as F, i.e., $E[X|\mathcal{G}]$ is the minimizer for the loss function F_n. Therefore, by the proof in Step 1, there exists a convex, differentiable function ϕ_n such that $\phi_n(0) = \phi_n'(0) = 0$ and

$$F_n(x,y) = \phi_n(x) - \phi_n(y) - \phi_n'(y)(x-y). \tag{6}$$

In particular, $F_n(x,0) = \phi_n(x)$. Since $F_n(x,0) \to F(x,0)$ for every x, we have

$$\lim_{n\to\infty} \phi_n(x) = F(x,0) \doteq \phi(x)$$

for every x. Since ϕ_n's are convex, so is their limit ϕ. In particular, ϕ is continuous [10, Theorem 10.1]. Setting $x = y + 1$ in equation (6), we have

$$\begin{aligned} \phi_n'(y) &= F_n(y+1, y) - \phi_n(y+1) + \phi_n(y) \\ \Rightarrow \lim_{n\to\infty} \phi_n'(y) &= F(y+1, y) - \phi(y+1) + \phi(y) \doteq f(y). \end{aligned}$$

Clearly f is continuous. Letting $n \to \infty$ in both sides of equation (6), we have

$$F(x, y) = \phi(x) - \phi(y) - f(y)(x - y),$$

where ϕ is continuously differentiable, since F is continuously differentiable with respect to x. Furthermore, the non-negativity of F implies that $f(y)$ is a subgradient of ϕ [10, Page 214]. Finally, the differentiability of ϕ suggests that its subdifferential is just its derivative [10, Theorem 25.1]. It follows that $\phi'(y) = f(y)$, and hence $F = D_\phi$.

Step 3: It remains to show that ϕ is strictly convex. From step 2, we already know that ϕ is a convex function. We prove by contradiction that if ϕ is not strictly convex, the assumption of uniqueness will be violated. Suppose ϕ is not strictly convex. Then there exists an interval $I = [\ell_1, \ell_2]$ such that $\ell_1 < \ell_2$ and $\phi'(y) = \phi'(\ell_1)$ for all $y \in I$. Consider a random variable X such that $P(X = \ell_1) = P(X = \ell_2) = 1/2$, and set $\mathcal{G} = \{\emptyset, \Omega\}$. It is not difficult to check that any $y \in I$ is a minimizer. Indeed, $E[D_\phi(X, y)] \equiv 0$ for all $y \in I$. This is a contradiction, and we complete the proof. ∎

Theorem 4 ($d \geq 2$) *Let $F : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ be a non-negative function such that $F(x, x) = 0, \forall x \in \mathbb{R}^d$. Assume that $F(x, y)$ and $F_{x_i x_j}(x, y)$, $1 \leq i, j \leq d$ are all continuous. For all random variables X taking value in $\mathbb{R}^d$, if $E[X|\mathcal{G}]$ is the unique minimizer of $E[F(X, Y)]$ over all random variables $Y \in \mathcal{G}$, i.e.,*

$$\text{argmin}_{Y\in\mathcal{G}} \; E[F(X, Y)] = E[X|\mathcal{G}],$$

then $F(x, y) = D_\phi(x, y)$ for some strictly convex and differentiable function $\phi : \mathbb{R}^d \mapsto \mathbb{R}$.

The proof is divided into three analogous steps as those in Theorem 3. The only essential difference is in Step 1, which relies on the following lemma. The lemma itself is a direct consequence of the celebrated Poincaré Lemma.

Lemma 2 *Given a collection of continuous functions $\{h_{ij} : 1 \leq i, j \leq d\}$ defined on an open, convex set $U \subseteq \mathbb{R}^d$ ($d \geq 2$). If for all triples of indices $1 \leq i, j, k \leq d$,*

$$h_{ij} \equiv h_{ji}, \qquad \frac{\partial h_{ij}}{\partial x_k} \equiv \frac{\partial h_{kj}}{\partial x_i}.$$

Then there exists a function $\Phi : U \mapsto \mathbb{R}$ such that $\Phi_{x_i x_j} = h_{ij}$.

Proof: (of Lemma 2) We first show that there exists a sequence of functions $\{\phi_i : 1 \leq i \leq d\}$ defined on U such that, for every index i,

$$\nabla \phi_i \equiv (h_{i1}, \ldots, h_{id})^T. \tag{7}$$

This follows from the given property for triplets of indices in conjunction with the Poincaré Lemma [5, Theorem 8.1] applied to 1-forms, noting that every convex set is star-convex. It remains to show that there exists a function Φ such that

$$\nabla \Phi = (\phi_1, \ldots, \phi_d)^T.$$

Note that for any pair of indices i, j, from equation (7) and the given property, we have

$$\frac{\partial \phi_i}{\partial x_j} = h_{ij} = h_{ji} = \frac{\partial \phi_j}{\partial x_i}.$$

The existence of Φ now follows via the Poincaré Lemma. ∎

Proof: (of Theorem 4) **Step 1:** Assume that $F_{x_i x_j}$, $F_{x_i y_j}$ and $F_{y_i y_j}$, $1 \leq i, j \leq d$ are all continuous (i.e., F is twice continuously differentiable). Fix arbitrarily $a, b \in \mathbb{R}^d$, and $p \in [0, 1]$. Consider a random variable X on some probability space $(\Omega, \mathcal{F}, P)$ such that $P(X = a) = p$ and $P(X = b) = q$ with $p + q = 1$. Let $\mathcal{G} = \{\emptyset, \Omega\}$. Similar to the proof of equation (4), we have

$$pF_{y_i}(a, y^*) + qF_{y_i}(b, y^*) = 0, \qquad \forall\ i = 1, \ldots, d,$$

at $y^* = pa + qb$. Taking derivatives over p on both sides of the above equation and recalling $q = 1 - p$, we arrive at

$$F_{y_i}(a, y^*) - F_{y_i}(b, y^*) + \sum_{j=1}^{d} \left[pF_{y_i y_j}(a, y^*) + qF_{y_i y_j}(b, y^*)\right](a_j - b_j) = 0,$$

for every $i = 1, \ldots, d$. In particular, setting $p = 1$ leads to

$$F_{y_i}(a, a) - F_{y_i}(b, a) + \sum_{j=1}^{d} F_{y_i y_j}(a, a)(a_j - b_j) = 0, \qquad \forall\ i = 1, \ldots, d.$$

Because F is non-negative and $F(x, x) \equiv 0$, we have $F_{y_i}(a, a) \equiv 0$. Writing $H_{ij}(a) \doteq F_{y_i y_j}(a, a)$, and noting that a and b are arbitrary, we may rewrite the the above equation as

$$F_{y_i}(x, y) = \sum_{j=1}^{d} H_{ij}(y)(y_j - x_j), \qquad \forall\ x, y \in \mathbb{R}^d. \tag{8}$$

Since F_{y_i} is continuously differentiable for every i, it follows easily that H_{ij} is also continuously differentiable for all $1 \leq i, j \leq d$. We now claim that there exists a function $\phi : y \in \mathbb{R}^d \mapsto H(y) \in \mathbb{R}$ such that

$$\phi_{y_i y_j}(y) = H_{ij}(y), \qquad 1 \leq i, j \leq d. \tag{9}$$

Indeed, from equation (8), we see that for every $k = 1, \ldots, d$,

$$F_{y_i y_k}(x, y) = \sum_{j=1}^{d} (H_{ij})_{y_k}(y)(y_j - x_j) + H_{ik}(y),$$

and

$$F_{y_k y_i}(x, y) = \sum_{j=1}^{d} (H_{kj})_{y_i}(y)(y_j - x_j) + H_{ki}(y),$$

Now, $F_{y_i y_k} = F_{y_k y_i}$ implies

$$H_{ik} \equiv H_{ki}, \qquad (H_{ij})_{y_k} \equiv (H_{kj})_{y_i}. \tag{10}$$

The existence of ϕ now follows from Lemma 2.

Now, from equation (8) we have

$$F_{y_i}(x, y) = \sum_{j=1}^{d} \phi_{y_i y_j}(y)(y_j - x_j) = \frac{\partial}{\partial y_i} \left[-\phi(y) - \langle \nabla \phi(y), x - y \rangle \right],$$

which, combined with the condition $F(x, x) \equiv 0$, readily yields

$$F(x, y) = \phi(x) - \phi(y) - \langle \nabla \phi(y), x - y \rangle = D_\phi(x, y).$$

The convexity of ϕ is implied by the non-negativity of F.

Step 2 and **Step 3:** Now repeating the same steps as those in the proof of Theorem 3, Theorem 4 is immediate. ∎

5 Conclusion

Our paper provides necessary and sufficient conditions for loss functions under which the conditional expectation is the unique optimal predictor. Beyond its mathematical interest, the expansion from the $\mathbb{L}^2$-loss function to the general class of BLFs has its own distinctive value. In areas such as image and speech codings where the $\mathbb{L}^2$-loss function is no longer an appropriate or even meaningful measure of error (as was pointed out in [4]), other functions such as the KL-divergence, the Itakura-Saito distance (see Table 1) etc., play a dominant role. Our findings may serve as a mathematical justification for the adoption of these loss functions.

It is worth pointing out that throughout the paper, for the purpose of concise presentation, we assume that the convex function ϕ is finite on the whole Euclidean space $\mathbb{R}^d$. The random variable X is also allowed to take values in the whole $\mathbb{R}^d$. However, the same methodology with very minor modifications will lead to similar results when $\mathbb{R}^d$ is replaced by an open convex subset of $\mathbb{R}^d$. Some examples of interest include the open half-space (for $\phi(x) = -\log x$), and the open d-simplex (for $\phi(p) = \sum_{j=1}^{d} p_j \log p_j$).

Finally, as was alluded earlier, the stronger regularity condition for the high-dimensional case (Theorem 4) is used in a crucial way to verify the compatibility condition (10), which seems almost necessary for solving the system of equations (9). It will be interesting to see if the regularity condition can be relaxed.

References

[1] K. B. Athreya. Prediction under convex loss. Technical Report 99-2, Department of Mathematics and Statistics, Iowa State University, 1999.

[2] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. Technical Report TR-03-19, Department of Computer Science, University of Texas at Austin, 2003.

[3] 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 Computational Mathematics and Physics*, 7:200–217, 1967.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 1991.

[5] C. H. Edwards. *Advanced Calculus of Several Variables.* Academic Press, 1973.

[6] I. Ekeland and R. Témam. *Convex Analysis and Variational Problems.* SIAM, Classics in Applied Mathematics, 1999.

[7] D. Gilbarg and N. Trudinger. *Elliptic Partial Differential Equations of Second Order.* Springer-Verlag, New York, 3nd edition, 2001.

[8] S. Karlin and H. M. Taylor. *A First Course in Stochastic Processes.* Academic Press, 2nd edition, 1974.

[9] E. L. Lehmann and G. Casella. *Theory of Point Estimation.* Springer, 2nd edition, 1998.

[10] R. T. Rockafeller. *Convex Analysis.* Princeton Landmarks in Mathematics. Princeton University Press, 1970.

[11] D. Williams. *Probability with Martingales.* Cambridge University Press, 1991.