REFERENCES

[1] S. Verdú, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[2] T. Numai and O. Kubota, "Analysis of repeated unequally spaced channels for FDM lightwave systems," *J. Lightwave Technol.*, vol. 18, no. 5, pp. 656–664, May 2000.

[3] K.-D. Chang, G.-C. Yang, and W. Kwong, "Determination of FWM products in unequally-spaced channel WDM lightwave systems," *J. Lightwave Technol.*, vol. 18, no. 12, pp. 2113–2122, Dec. 2000.

[4] R. Killey, H. Thiele, V. Mikhailov, and P. Bayvel, "Reduction of intrachannel nonlinear distortion in 40-Gbps-based WDM transmission over standard fiber," *IEEE Photon. Technol. Lett.*, vol. 12, no. 12, pp. 1624–1626, Dec. 2000.

[5] M. Hayee and A. Willner, "NRZ versus RZ in 10–40-Gb/s dispersion-managed WDM transmission systems," *IEEE Photonics Technol. Lett.*, vol. 11, no. 8, pp. 991–993, Aug. 1999.

[6] B. Konrad and K. Pertermann, "Optimum fiber dispersion in high-speed TDM systems," *IEEE Photonics Technol. Lett.*, vol. 13, no. 4, pp. 299–301, Apr. 2001.

[7] M. K. Varanasi and B. Aazhang, "Optimally near-far resistant multiuser detection in differentially coherent synchronous channels," *IEEE Trans. Inf. Theory*, vol. 37, no. 4, pp. 1006–1018, Jul. 1991.

[8] M. K. Varanasi, "Noncoherent detection in asynchronous multiuser channels," *IEEE Trans. Inf. Theory*, vol. 39, no. 1, pp. 157–176, Jan. 1993.

[9] M. K. Varanasi and A. Russ, "Noncoherent decorrelative detection for nonorthogonal multipulse modulation over the multiuser Gaussian channel," *IEEE Trans. Commun.*, vol. 46, pp. 1675–1684, 1998.

[10] M. K. Varanasi and D. Das, "Noncoherent decision-feedback multiuser detection," *IEEE Trans. Commun.*, vol. 48, no. 12, pp. 259–269, Dec. 2000.

[11] A. Russ and M. K. Varanasi, "Noncoherent multiuser detection for nonlinear modulation over the Rayleigh-fading channel," *IEEE Trans. Inf. Theory*, vol. 47, no. 1, pp. 295–306, Jan. 2001.

[12] G. P. Agrawal, *Nonlinear Fiber Optics*. San Diego, CA: Academic Press, 1995.

[13] E. Iannone, F. Matera, A. Mecozzi, and M. Settembre, *Nonlinear Optical Communication Networks*. New York: Wiley, 1998.

[14] P. A. Humblet and M. Azizoglu, "On the bit error rate of lightwave systems with optical amplifiers," *J. Lightwave Technol.*, vol. 11, pp. 1576–1582, Nov. 1991.

[15] B. Xu and M. Brandt-Pearce, "Multiuser square-law detection with applications to fiber optic communications," *IEEE Trans. Commun.*, to be published.

[16] A. Kavčić and J. M. F. Moura, "The Viterbi algorithm and Markov noise memory," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 291–301, Jan. 2000.

[17] J. Moon and J. Park, "Pattern-dependent noise prediction in signal-dependent noise," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 4, pp. 730–742, Apr. 2001.

[18] M. Schwartz, W. Bennett, and S. Stein, *Communication Systems and Techniques* New York, McGraw-Hill, 1966.

[19] B. Xu and M. Brandt-Pearce, "Comparison of FWM- and XPM-induced crosstalk using the Volterra series transfer function method," *J. Lightwave Technol.*, vol. 21, no. 1, pp. 40–53, Jan. 2003.

[20] B. Xu, "Study of fiber nonlinear effects on fiber optic communication systems," Ph.D. dissertation, Univ. Virginia, Charlottesville, 2003.

On the Optimality of Conditional Expectation as a Bregman Predictor

Arindam Banerjee, Xin Guo, and Hui Wang

***Abstract*—We consider the problem of predicting a random variable X from observations, denoted by a random variable Z. It is well known that the conditional expectation $E[X|Z]$ is the optimal $\mathbb{L}^2$ predictor (also known as "the least-mean-square error" predictor) of X, among all (Borel measurable) functions of Z. In this correspondence, we provide necessary and sufficient conditions for the general loss functions under which the conditional expectation is the unique optimal predictor. We show that $E[X|Z]$ 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 every random variable X, the infimum of $E[F(X,y)]$ over all constants y is attained by the expectation $E[X]$, then F is a BLF.**

***Index Terms*—Bregman loss functions (BLFs), conditional expectation, prediction.**

I. INTRODUCTION

THE problem of predicting a random variable 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 an $\mathcal{F}$-measurable random variable that one wishes to predict. If Z is the observation random variable, the available partial information about X that can be obtained by observing Z is represented by $\sigma(Z)$. Mathematically, $\sigma(Z)$ is the σ-algebra generated by Z and contains all Borel-measurable functions of Z. In order to ease exposition, for $Y \in \sigma(Z)$, we will simply say "Y is a function of Z" in lieu of "Y is a Borel-measurable function of Z" throughout the text. Now, the question is: among all functions of Z, which one is the best predictor of X?

The notion of *best* is usually specified by a nonnegative loss function F and achieved by solving a corresponding minimization problem. More precisely, the best predictor of X is defined as the minimizer of $E[F(X,Y)]$ over all $Y \in \sigma(Z)$. A particularly important case is when F is the so called $\mathbb{L}^2$-loss function, also known as the squared error, i.e., $F(x,y) \doteq \|x-y\|^2$. It is well known [1]–[3] that the corresponding *unique* best predictor is given by the conditional expectation. In other words

$$\arg\min_{Y \in \sigma(Z)} E[\|X-Y\|^2] = E[X|Z].$$

This makes conditional expectation crucially important for prediction.

A question arises naturally: are there other loss functions F for which $E[X|Z]$ is the unique best predictor? Some simple counterexamples lead to the general conviction that the existence of such loss

Manuscript received August 13, 2003; revised December 2, 2004. Part of the work was done when A. Banerjee and X. Guo were at IBM T. J. Watson Research Center, Yorktown Heights, NY. The material in this correspondence was presented in part at the International Symposium on Information Theory, Chicago, IL, June/July 2004.

A. Banerjee is with the Department of Electrical and Computer Engineering, University of Texas at Austin, Austin, TX 78712 USA (e-mail: abanerje@ece.utexas.edu).

X. Guo is with the School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14583 USA (e-mail: xinguo@orie.cornell.edu).

H. Wang is with the Division of Applied Mathematics, Brown University, Providence, RI 02912 USA (e-mail: huiwang@cfm.brown.edu).

Communicated by A. Kavčić, Associate Editor for Detection and Estimation.
Digital Object Identifier 10.1109/TIT.2005.850145

TABLE I
EXAMPLES OF BLFs

Domain	$\phi(x)$	$D_\phi(x,y)$	Loss
$\mathbb{R}$	x^2	$(x-y)^2$	$\mathbb{L}^2$-loss
$\mathbb{R}_{++}$	$x\log x$	$x\log(x/y)-(x-y)$	
$(0,1)$	$x\log x+(1-x)\log(1-x)$	$x\log(x/y)+(1-x)\log((1-x)/(1-y))$	
$\mathbb{R}_{++}$	$-\log x$	$x/y-\log(x/y)-1$	Itakura-Saito distance
$\mathbb{R}$	e^x	$e^x-(1+x-y)e^y$	
$\mathbb{R}^d$	$\|x\|^2$	$\|x-y\|^2$	$\mathbb{L}^2$-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

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

functions would be rare and would have to possess very special properties. For example, if one uses the absolute error loss function ([4, Sec. 1.7]), then any constant a satisfying $P(X\le a)\ge 1/2\le P(X\ge a)$, i.e., the median of X and not $E[X]$, proves to be the best constant predictor. Recently, [5] studied the case of general convex loss functions and obtained a criterion for which a best constant predictor exists.

In this correspondence, 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) [6], of which the $\mathbb{L}^2$-loss function is a special case. Indeed, one can essentially create as many BLFs as differentiable strictly convex functions, up to equivalences in linear and constant terms (see Definition 1). Second, we also show that the class of BLFs is exhaustive under mild conditions, i.e., if $\arg\min_{y\in\mathbb{R}^d}E[F(X,y)]=E[X]$ for every random variable X, then the loss function F has to be a BLF, up to an additive constant.

Remark 1: If $\mathcal{G}$ is a sub-σ algebra of $\mathcal{F}$, and $E[X|\mathcal{G}]$ denotes the conditional expectation, all the results presented in this correspondence remain true if one replaces $E[\cdot|Z]$ with $E[\cdot|\mathcal{G}]$ and simultaneously replaces functions of Z with $\mathcal{G}$-measurable random variables.

II. BREGMAN LOSS FUNCTIONS

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

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

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 $p\doteq(p_1,\dots,p_d)$ be a discrete probability distribution so that $\sum_{j=1}^d p_j=1$. The negative Shannon entropy $\phi(p)\doteq\sum_{j=1}^d p_j\log_2 p_j$ is a strictly convex function on the d-simplex. Let $q=(q_1,\dots,q_d)$ be another probability distribution. The corresponding BLF is

$$\begin{aligned}D_\phi(p,q)&=\sum_{j=1}^d p_j\log_2 p_j-\sum_{j=1}^d q_j\log_2 q_j-\langle p-q,\nabla\phi(q)\rangle\\&=\sum_{j=1}^d p_j\log_2 p_j-\sum_{j=1}^d q_j\log_2 q_j\\&\quad-\sum_{j=1}^d(p_j-q_j)(\log_2(eq_j))\\&=\sum_{j=1}^d p_j\log_2(p_j/q_j)\end{aligned}$$

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

Table I contains a list of some common convex functions and their corresponding BLFs. The following useful observation follows from the strict convexity of ϕ [7, Proposition 5.4].

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

Remark 2: Since a differentiable convex function is necessarily continuously differentiable [8, Th. 25.5], the function D_ϕ is continuous. Moreover, if we write ∇_x as the gradient with respect to x, then the function

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

is also continuous. For more discussions on BLFs, interested readers are referred to [9] and the references therein.

III. 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 X be an arbitrary random variable taking values in $\mathbb{R}^d$ for which both $E[X]$ and $E[\phi(X)]$ are finite. Then, among all functions of Z, the conditional expectation is the unique minimizer (up to a.s. equivalence) of the expected Bregman loss, i.e.,

$$\arg\min_{Y\in\sigma(Z)}E[D_\phi(X,Y)]=E[X|Z].$$

Proof: Let Y be any function of Z, and $Y^*\doteq E[X|Z]$. It follows from Definition 1 that

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

Meanwhile, for Y being any function of Z, we have

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

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 a sequence of functions of Z such that

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

where $Y^* \doteq E[X|Z]$, 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

$$P(|Y_n - Y^*| \geq \delta) \leq \epsilon, \quad \forall\, n \geq N.$$

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

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

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

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

Since D_ϕ is continuous (Remark 2), 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

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

From the assumption on $\{Y_n\}$ and (1), it follows that $E\left[D_\phi\left(Y^*,Y_n\right)\right] \to 0$. Hence, there exists N such that for $n \geq N$, $E[D_\phi(Y,Y_n)] \leq \epsilon\alpha/2$. Therefore, for $n \geq N$

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

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 3: 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 one-dimensional (1-D) case, it amounts to $\inf_{x\in\mathbb{R}} \phi''(x) > 0$).

IV. 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 for a nonnegative loss function $F : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ if $\forall\, X, Z$,

$$\arg\,\min_{Y\in\sigma(Z)} E[F(X,Y)] = E[X|Z], \tag{3}$$

then F is a BLF.

Remark 4: Indeed, we will prove the following slightly stronger result: a nonnegative loss function F has to be a BLF if

$$\arg\,\min_{y\in\mathbb{R}^d} E[F(X,y)] = E[X] \tag{4}$$

for every random variable X. In other words, if the expectation $E[X]$ is the best *constant* predictor for every random variable X, then F is a BLF.

We will present the results separately for the 1-D (Theorem 3) and higher dimensional (Theorem 4) case, since the latter needs slightly stronger regularity conditions (see Section VI 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), then 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 nonnegative 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, $E[X]$ is the unique minimizer of $E[F(X,y)]$ over all constants $y \in \mathbb{R}^d$, i.e.,

$$\arg\,\min_{y\in\mathbb{R}^d} E[F(X,y)] = E[X]$$

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 such that $P(X = a) = p$ and $P(X = b) = q$ with $p + q = 1$. Then, from the assumption

$$\begin{aligned} pF(a,y) + qF(b,y) &= E[F(X,y)] \\ &\geq E[F(X,E[X])] \\ &= pF(a,pa+qb) + qF(b,pa+qb) \end{aligned}$$

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

$$pF_y(a,y^*) + qF_y(b,y^*) = 0. \tag{5}$$

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

$$F_y(x,y) = (y-x)H(y) \tag{6}$$

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 (6) 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 nonnegativity 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 nonnegative, $\mathcal{C}^\infty$ 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

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

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

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

It is standard to show that [10, Sec. 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., $E[X]$ 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

$$F_n(x,y) = \phi_n(x) - \phi_n(y) - \phi_n'(y)(x-y). \tag{7}$$

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 [8, Th. 10.1]. Setting $x = y+1$ in (7), 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 (7), 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 nonnegativity of F implies that $f(y)$ is a subgradient of ϕ [8, p. 214]. Finally, the differentiability of ϕ suggests that its subdifferential is just its derivative [8, Th. 25.1]. It follows that $\phi'(y) = f(y)$, and hence $F = D_\phi$.

Step 3: It remains to be shown 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 $P(X = \ell_1) = P(X = \ell_2) = 1/2$. 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 nonnegative 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 $E[X]$ is the unique minimizer of $E[F(X,y)]$ over all constants $y \in \mathbb{R}^d$, i.e.,

$$\arg\min_{y\in\mathbb{R}^d} E[F(X,y)] = E[X]$$

and 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 $1 \leq i,j,k \leq d$

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

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

$$\nabla\phi_i \equiv (h_{i1}, \ldots, h_{id})^T. \tag{8}$$

This follows from the given property for triplets of indexes in conjunction with the Poincaré lemma [11, Th. 8.1] applied to 1-forms, noting that every convex set is star convex. It remains to be shown that there exists a function Φ such that

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

Note that for any pair of indexes i, j, from (8) 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 such that $P(X=a)=p$ and $P(X=b)=q$ with $p+q=1$. Similar to the proof of (5), we have

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

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

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

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

$$F_{y_i}(a,a) - F_{y_i}(b,a) + \sum_{j=1}^{d} F_{y_iy_j}(a,a)(a_j - b_j) = 0$$

for every $i=1,\ldots,d$. Because F is nonnegative and $F(x,x)\equiv 0$, we have $F_{y_i}(a,a)\equiv 0$. Writing $H_{ij}(a) \doteq F_{y_iy_j}(a,a)$, and noting that a and b are arbitrary, we may rewrite the the above equation as

$$F_{y_i}(x,y) = \sum_{j=1}^{d} H_{ij}(y)(y_j - x_j), \quad \forall\, x,y \in \mathbb{R}^d. \tag{9}$$

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

$$\phi_{y_iy_j}(y) = H_{ij}(y), \quad 1 \le i,\ j \le d. \tag{10}$$

Indeed, from (9), we see that for every $k=1,\ldots,d$

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

and

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

Now, $F_{y_iy_k} = F_{y_ky_i}$ implies

$$H_{ik} \equiv H_{ki}, \quad (H_{ij})_{y_k} \equiv (H_{kj})_{y_i}. \tag{11}$$

The existence of ϕ now follows from Lemma 2.

Now, from (9), we have

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

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 nonnegativity of F.

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

V. Discussion

Throughout this correspondence, for the purpose of concise presentation, we assume that the convex function ϕ is finite on the whole Euclidean space $\mathbb{R}^d$, and the random variable X is 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 $\mathbb{R}^d$, typically $\mathrm{dom}(\phi)$, the domain of ϕ. Some examples of interest include the open half-space (for $\phi(x) = -\log x$), and the open d-simplex (for $\phi(p) = \sum_{j=1}^{d} p_j \log p_j$).

Loss functions that lead to the optimality of the conditional expectation have been studied in earlier literature. An analysis applicable to difference distortion measures, i.e., distortions of the form $F(x,y) = C(x-y)$, is presented in [12]. In particular, it is shown that if C is symmetric, i.e., $C(z) = C(-z)$, and strictly convex, and the conditional probability density of X given Z is symmetric around the conditional expectation $E[X|Z]$; then, the conditional expectation is the best predictor. Note that the third assumption regarding the symmetry of the conditional probability distribution of X given $\sigma(Z)$ is very strong and makes the optimality result very restricted. The results discussed in this correspondence apply to all random variables and hence are direct generalizations of the well-known least-squares prediction results [1], [3].

Bregman loss functions (BLFs) have been extensively studied in the context of convex optimization and related problems (see [13] and [14] and reference therein). Ben-Tal *et al.* [15] applied 1-D BLFs to the analysis of entropic means, where the authors studied a different optimization problem, namely $\min_{x\in\mathbb{R}} E[F(x,Y)]$ for a fixed random variable Y. An axiomatic characterization of a wide class of distortion functions, including f-divergences [16] and BLFs, was obtained in Csiszár's seminal work [14]. In another paper by Csiszár [17], BLFs were used primarily for analyzing generalized projections for nonnegative functions on convex sets. The interesting connection between reverse I divergence (which is a special case of BLFs) and arithmetic mean was briefly mentioned in [17] without further elaboration and study. Therefore, compared with Csiszár's work, we study BLFs from a different (i.e., probabilistic) perspective and with different methodologies. We provide a complete characterization of the fundamental relationship between BLFs and conditional expectation. More important, our result expands the domain of BLFs to exciting and practical application areas in information theory, such as vector quantization, as we outline next in more details.

Example 3: The key problem in `kmeans` clustering and its extensions to vector quantization [18] involves solving the following problem: given $\mathcal{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$, find a k-partitioning $\{\mathcal{X}_1,\ldots,\mathcal{X}_k\}$ of $\mathcal{X}$ such that the cost function

$$\mathcal{J}_{\mathrm{kmeans}} = \sum_{h=1}^{k} \sum_{\mathbf{x}\in\mathcal{X}_h} \|\mathbf{x} - \mu_h\|^2$$

where μ_h is the mean of $\mathcal{X}_h$, is minimized. Typically, the problem is solved by an iterative relocation scheme that alternates between assigning points to clusters $\mathcal{X}_h$ and recomputing the cluster means μ_h till convergence. From the results presented in this correspondence, one can show [9] that *the* `kmeans` *clustering algorithm*, conventionally applicable only to squared Euclidean distances, can be extended to all BLFs, and for all these BLFs, the clustering algorithm converges. In the extension, the cost function to be minimized is given by

$$\mathcal{J}_{\mathrm{BLF}} = \sum_{h=1}^{k} \sum_{\mathbf{x}\in\mathcal{X}_h} d_\phi(\mathbf{x},\mu_h).$$

An iterative relocation algorithm [9] that is a direct generalization of the `kmeans` algorithm can be employed to solve the problem. Although the cluster mean computation step remains unchanged as indicated by the results in this correspondence, the cluster assignments may be *different* depending on the exact BLF being used. Hence, the final clustering may be different for different choices of BLF. Application-dependent appropriate choices of BLFs can make significant differences in the quality of the final clusters obtained. In summary, the generalization of `kmeans` clustering to all BLFs will potentially have significant impact

on the design, analysis, and usage of vector quantization techniques in several application domains.

Another applicable area for BLFs is image processing: the paper by Li *et al.* [19] studies bounds on asymptotic performance of vector quantizers with perceptual distortion measure for which BLFs are natural candidates.

VI. Conclusion

This correspondence 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 [20]), other functions such as the Kullback–Liebler (KL) divergence or the Itakura–Saito distance (see Table I) play a dominant role. Our findings may serve as a mathematical justification for the adoption of these loss functions.

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 (11), which seems almost necessary for solving the system of (10). It will be interesting to see if the regularity condition can be relaxed.

References

[1] S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, 2nd ed. San Diego, CA: Academic, 1974.

[2] O. Knill, "Probability," Course notes from Caifornia Institute of Technology, Pasadena, CA, 1994.

[3] D. Williams, *Probability With Martingales*. Cambridge, U.K.: Cambridge Univ. Press, 1991.

[4] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York: Springer-Verlag, 1998.

[5] K. B. Athreya, "Prediction Under Convex Loss," Dept. Mathematics and Statistics, Iowa State Univ., Ames, IA, Tech. Rep. 99-2, 1999.

[6] 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. Phys.*, vol. 7, pp. 200–217, 1967.

[7] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*, ser. SIAM Classics in Applied Mathematics. Philadelphia, PA: SIAM, 1999.

[8] R. T. Rockafellar, *Convex Analysis*, ser. Princeton Landmarks in Mathematics. Princeton, NJ: Princeton Univ. Press, 1970.

[9] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," in *Proc. SIAM Int. Conf. Data Mining*, 2004, pp. 234–245.

[10] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 3rd ed. New York: Springer-Verlag, 2001.

[11] C. H. Edwards, *Advanced Calculus of Several Variables*. Mineola, NY: Dover, 1995.

[12] H. L. V. Trees, *Detection, Estimation and Modulation Theory (Part I)*. New York: Wiley, 1968.

[13] Y. Censor and S. Zenios, *Parallel Optimization: Theory, Algorithms, and Applications*. London, U.K.: Oxford Univ. Press, 1998.

[14] I. Csiszár, "Why least sqaures and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Annals Stat.*, vol. 19, no. 4, pp. 2032–2066, 1991.

[15] A. Ben-Tal, A. Charnes, and M. Teboulle, "Entropic means," *J. Math. Anal. Appl.*, vol. 139, pp. 537–551, 1989.

[16] I. Csisár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar*, vol. 2, pp. 299–318, 1967.

[17] I. Csiszár, "Generalized projections for nonnegative functions," *Acta Mathematica Hungarica*, vol. 68, no. 1–2, pp. 161–185, 1995.

[18] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwell, MA: Kluwer, 1991.

[19] J. Li, N. Chaddha, and R. M. Gray, "Asymptotic performance of vector quantizers with a perceptual distortion measure," *IEEE Trans. Inf. Theory*, vol. 45, pp. 1082–1091, 1999.

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

Information Bounds and Quickest Change Detection in Decentralized Decision Systems

Yajun Mei

***Abstract*—The quickest change detection problem is studied in decentralized decision systems, where a set of sensors receive independent observations and send summary messages to the fusion center, which makes a final decision. In the system where the sensors do not have access to their past observations, the previously conjectured asymptotic optimality of a procedure with a monotone likelihood ratio quantizer (MLRQ) is proved. In the case of additive Gaussian sensor noise, if the signal-to-noise ratios (SNR) at some sensors are sufficiently high, this procedure can perform as well as the optimal centralized procedure that has access to all the sensor observations. Even if all SNRs are low, its detection delay will be at most $\pi/2 - 1 \approx 57\%$ larger than that of the optimal centralized procedure. Next, in the system where the sensors have full access to their past observations, the first asymptotically optimal procedure in the literature is developed. Surprisingly, the procedure has the same asymptotic performance as the optimal centralized procedure, although it may perform poorly in some practical situations because of slow asymptotic convergence. Finally, it is shown that neither past message information nor the feedback from the fusion center improves the asymptotic performance in the simplest model.**

***Index Terms*—Asymptotic optimality, CUSUM, multisensor, quantization, sensor networks, sequential detection.**

I. Introduction

The problem of quickest change detection has a variety of applications, including industrial quality control, reliability, fault detection, and signal detection. The classical or centralized version of this problem, where all observations are available at a single central location, is a well-developed area (see, e.g., [1], [7], and [17]). Recently, this problem has been applied in decentralized or distributed decision systems, which have many important applications, including multisensor data fusion, mobile and wireless communication, surveillance systems, and distributed detection.

Fig. 1 illustrates the general setting of decentralized decision systems. In such a system, at time n, each of a set of L sensors S_l receives an observation $X_{l,n}$ and then sends a sensor message $U_{l,n}$ to a central processor, called the *fusion center*, which makes a final decision when observations are stopped. In order to reduce cost and increase reliability, it is required that the sensor messages belong to a finite alphabet

Manuscript received November 21, 2002; revised November 10, 2004. This work was supported in part by the National Institutes of Health under Grant R01 AI055343. The material in this correspondence was presented in part at the IEEE International Symposium on Information Theory, Chicago, IL, June/July 2004.

The author was with the Department of Mathematics, California Institute of Technology, Pasadena, CA USA. He is now with the Department of Biostatistics, Fred Hutchinson Cancer Research Center, Seattle, WA 98109 USA (e-mail: ymei@fhcrc.org).

Communicated by A. Kavčić, Associate Editor for Detection and Estimation.
Digital Object Identifier 10.1109/TIT.2005.850159