Optimal Bregman Prediction and Jensen's Equality

Arindam Banerjee[1]
Dept of ECE
University of Texas at Austin
Austin, TX 78712
e-mail: abanerje@ece.utexas.edu

Xin Guo[1]
School of ORIE
Cornell University
Ithaca, NY 14583
e-mail: xinguo@orie.cornell.edu

Hui Wang
Division of Applied Math
Brown University
Providence, RI 02912
e-mail: huiwang@cfm.brown.edu

Abstract — **We provide necessary and sufficient conditions for general loss functions under which the conditional expectation is the unique optimal predictor of a random variable. Further, using such loss functions, we give an exact characterization of the difference between the two sides of Jensen's inequality.**

I. 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 optimal predictor of X ?

The notion of optimal 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 squared error, i.e., $F(x,y) \doteq \|x-y\|^2$. It is well known [3, 4] 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?* In this paper, we provide necessary and sufficient conditions for general loss functions under which the conditional expectation is the unique optimal predictor. Further, using such loss functions, we present Jensen's equality by an exact characterization of the difference between the two sides of the Jensen's inequality.

II. Optimal Bregman Prediction

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*

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

Some well known examples of Bregman Loss Functions (BLFs) include squared Euclidean distance, KL-divergence, Itakura-Saito distance, Mahalanobis distance, I-divergence etc., for appropriate choices of ϕ. For more discussions on BLFs, interested readers are referred to [2] and the references therein.

The first new result we present is that for all BLFs, the conditional expectation is the unique optimal predictor. More formally, among all $\mathcal{G}$-measurable random variables Y, we have $\operatorname{argmin}_{Y\in\mathcal{G}} E[d_\phi(X,Y)] = E[X|\mathcal{G}]$. Note that this renders the well known result for least squares prediction as a special case. Further, 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 $Y^* = E[X|\mathcal{G}]$, then $Y_n \to Y^*$ in probability. Finally, it can be shown that under mild assumptions BLFs are exhaustive with respect to this optimality property. More precisely, under mild conditions it can be shown that if $F : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}_+$ is a loss function such that $\operatorname{argmin}_{Y\in\mathcal{G}} \; E[F(X,Y)] = E[X|\mathcal{G}]$ for all random variables X then F has to be a BLF. Detailed theorems and proofs of these results can be found in [1].

III. Jensen's Equality

Since the conditional expectation $E[X|\mathcal{G}]$ achieves the minimum of the expected Bregman loss, we take a closer look at what this minimum value is. Since $\min_{Y\in\mathcal{G}} E[d_\phi(X,Y)] = E[d_\phi(X,E[X|\mathcal{G}])] = E[E[d_\phi(X,E[X|\mathcal{G}])|\mathcal{G}]]$, the minimum value achieved is equal to the expectation of the $\mathcal{G}$-measurable random variable $E[d_\phi(X,E[X|\mathcal{G}])|\mathcal{G}]$. In some sense, this random variable quantifies how difficult it is to predict X on $\mathcal{G}$, when the prediction accuracy is measure by a BLF. We call this random variable the *conditional Bregman information* of X denoted by $I_\phi(X|\mathcal{G})$. When $\mathcal{G}$ is the trivial σ-algebra, we call it the Bregman information of X and denoted it by $I_\phi(X)$, i.e., $I_\phi(X) = E[d_\phi(X,E[X])]$. Special cases of Bregman information include variance, mutual information etc., with appropriate choices of ϕ and X. It is not difficult to see that the (conditional) Bregman information is always non-negative. In fact, it can be shown that $E[\phi(X)] = \phi(E[X]) + I_\phi(X)$. We call this identity the *Jensen's Equality* since the Bregman information exactly quantifies the difference between two sides of the Jensen's inequality. More generally, we have conditional Jensen's equality that can be stated as $E[\phi(X)|\mathcal{G}] = \phi(E[X|\mathcal{G}]) + I_\phi(X|\mathcal{G})$.

Acknowledgments

Arindam Banerjee was under an IBM PhD fellowship when this work was done.

References

[1] A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. Submitted for journal publication, 2004.

[2] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. SIAM International Conference on Data Mining, 2004.

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

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

[1]Part of the work was done when the authors were at IBM T. J. Watson Research Center, Yorktown Heights, NY.