Spring 1-1-2017

The Big Picture: Loss Functions at the Dataset Level

Karthik Kannan
University of Colorado at Boulder, kknicks@gmail.com

The Big Picture: Loss Functions at the Dataset Level

by

Karthik Kannan

B. Tech, VIT University, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Department of Computer Science

2017

This thesis entitled:
The Big Picture: Loss Functions at the Dataset Level
written by Karthik Kannan
has been approved for the Department of Computer Science

_____

Prof. Rafael Frongillo

_____

Prof. Stephen Becker

_____

Prof. Christian Ketelsen

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Kannan, Karthik (MS, Computer Science)

The Big Picture: Loss Functions at the Dataset Level

Thesis directed by Prof. Rafael Frongillo

Loss functions play a key role in machine learning optimization problems. Even with their widespread use throughout the field, selecting a loss function tailored to a specific problem is more art than science. Literature on the properties of loss functions that might help a practitioner make an informed choice about these loss functions is sparse.

In this thesis, we motivate research on the behavior of loss functions at the level of the dataset as a whole. We begin with a simple experiment that illustrates the differences in these loss functions. We then move on to a well-known attribute of perhaps the most ubiquitous loss function, the squared error. We will then characterize all loss functions that exhibit this property. Finally we end with extensions and possible directions of research in this field.

To Ma and Pa.

Acknowledgements

First and foremost, I would like thank Raf for being the perfect advisor and being extremely patient with me as I took my first steps in academic research. This thesis would have probably been a very different document without his support throughout my collaboration with him. I would also like to thank Chris and Stephen for being on my committee.

I would like to express my gratitude to my colleagues in the elicitation reading group, Lily and Nicole, for their feedback and conversations through our meetings. I am thankful to Apoorva, Monal, Nehal and to my roommates for always being there to encourage me and clear my head whenever I needed them.

Last, but in no way the least, I would like to thank my parents always being there. I could not have done this without them.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

Loss functions have always played a important role in machine learning. A significant task in machine learning is to model an objective function $y = f(x)$ using a prediction function $\hat{y} = \hat{f}(x)$. A loss function $L(y, \hat{y})$ tells us how close we are to this goal by measuring the discrepancy between our prediction $\hat{y}$ at every point x and the true outcome or data value y. The aim of constructing a suitable loss function is that if the value of the loss function over our predictions is "small" then, $\hat{f}(x)$ is in some sense "close" to the target $f(x)$.

Several of the algorithms used in machine learning are intrinsically associated with the loss functions that they employ: the hinge loss for support vector machines (SVMs) and squared loss for linear regression are some well known examples. While there is usually a disclaimer that the algorithms would work just as well with other losses, literature on the specific nature of the losses themselves that would help a user make this choice is sparse.

Let us take the example of the squared error or the L2 loss, which is the typical choice for linear regression. As Györfi, et al. [10] point out-

> There are two reasons for considering the L2 risk. First,... this simplifies the mathematical treatment of the whole problem ... Second, and more important, trying to minimize the L2 risk leads naturally to estimates which can be computed rapidly.

These reasons are two of many reasons that support the use of the squared error. While the reasons for choosing the squared error are sound, Steinwart and Christmann[11] acknowledge that these properties are not exclusive to the squared loss. Indeed, some properties of squared error,

such as sensitivity to outliers actually make it less suitable for algorithms such as the SVM, where the hinge loss is widely used.

Some loss functions are minimized by a specific statistic of the distribution of the outcomes over the dataset they are evaluated over. For instance, the absolute loss $|\hat{y} - y|$ is minimized at the conditional median of the distribution and the squared loss is minimized at the conditional mean of the distribution. Thus, a useful characterization of these losses is by pairing them with the statistics that minimize them. This idea is the basis of the field of property elicitation. An interesting open question in the field is given a statistic $\Gamma(x)$ of a distribution what loss functions "elicit", or are minimized at, that statistic [6].

The property elicitation framework lends itself well to the concepts of empirical risk minimization (ERM). To recall, the aim in ERM is to select an hypothesis h^* from a hypothesis class $\mathcal{H}$ that best matches a target function f. If f itself is a part of $\mathcal{H}$, ERM will yield (in the infinite dataset case) $h^* = f$ if and only if the loss function used, L, elicits $\Gamma(x)$. Here $f(x)$, would yield the statistic conditional at every point x, $\Gamma(Y|X = x)$. This is discussed in further detail in in Theorem 1. However if $f \notin \mathcal{H}$, as is often the case, the loss function that we chose plays a key role in our predictions. This is perhaps best illustrated through a simple experiment.

1.1 An experiment to motivate the study

We shall contrast two other losses, the exponential loss and the negative exponential loss, with the squared error. The equations for these losses are given in Table 2.1. Figures 1.1, 1.2, and 1.3, illustrate the behaviour of different losses when used in simple (univariate) linear regression.

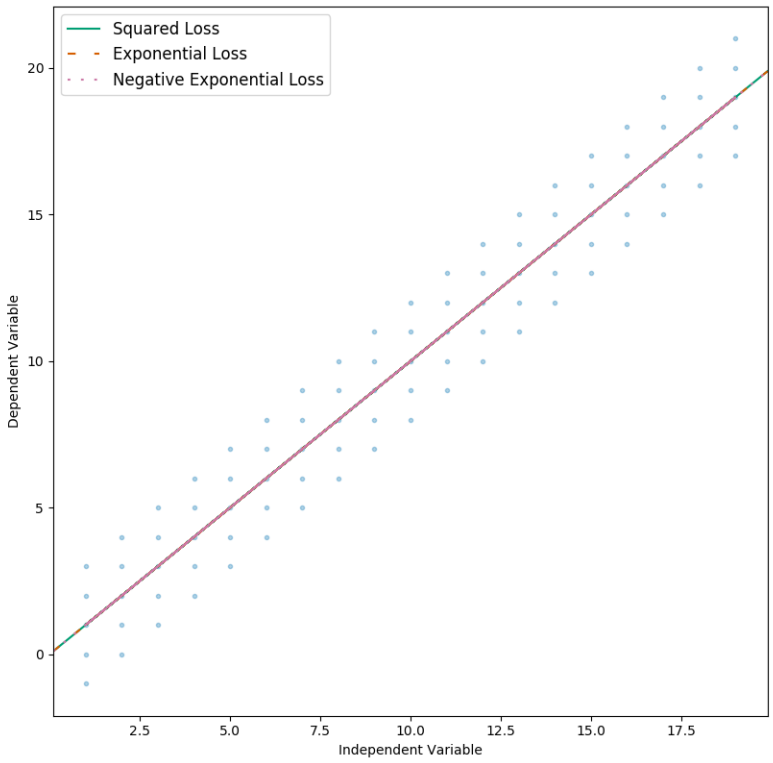

Figure 1.1: Lines yielded by different losses in Linear Regression on a synthetic dataset generated by aligning the conditional means of the independent variables in a straight line. Regression using all three losses yields identical results

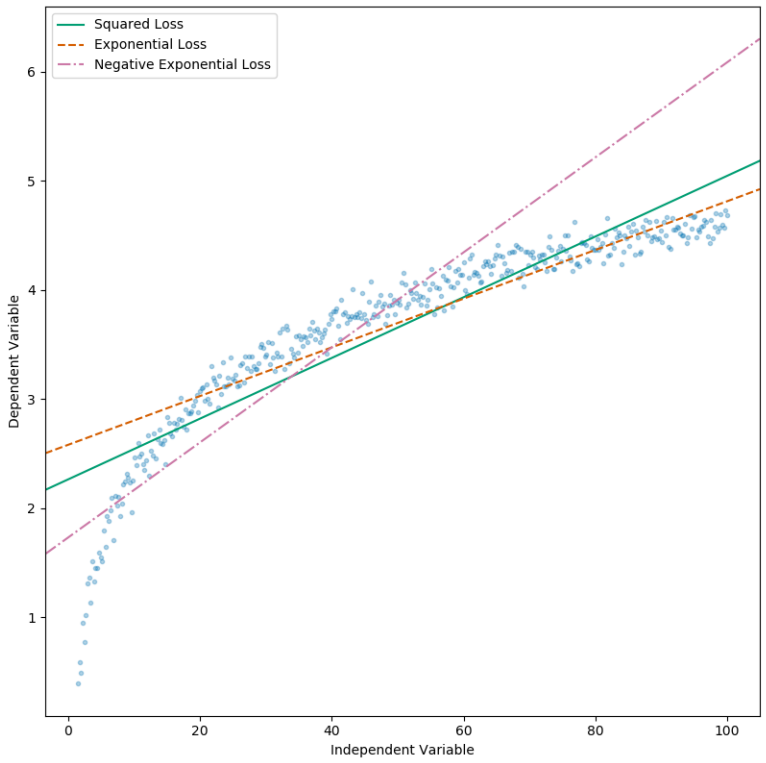

Figure 1.2: Lines yielded by different losses in Linear Regression on a synthetic dataset generated by adding noise to $f(x) = log(x)$.

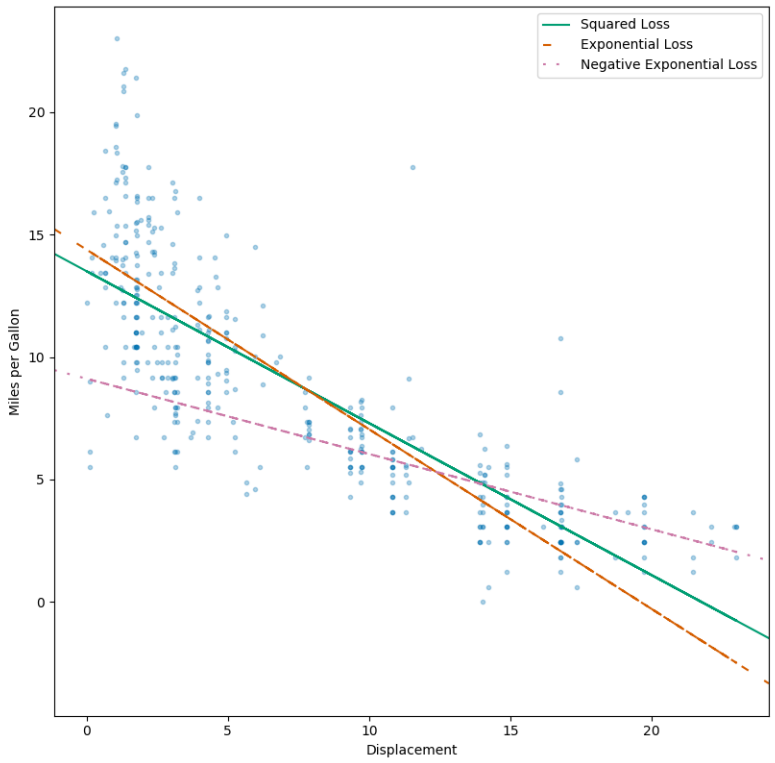

Figure 1.3: Lines yielded by different losses in Linear Regression on a subset of the Auto MPG dataset.

As we shall see in Section 2.3, all the loss functions used in the two figures above are part of a class of losses known as the Bregman Divergences, which elicit the conditional mean at every point x of the independent variable. In figure 1.1, the conditional means of the independent variable in the dataset are collinear. In this case, all three regression using all three loss functions yields the same line. The optimal hypothesis h^* is thus part of our hypothesis class $\mathcal{H}$, the set of lines. This will seldom be the case in most real world applications, though. Figures 1.2 and 1.3 illustrate the more general case where $h^* \notin \mathcal{H}$ We see that even though the loss functions are minimized at the same property of the distribution, the regression lines that we obtain from each of these losses are different. This is because every point in the data is weighted differently by each of the loss

functions. Using these weights, ERM would then prioritize a lower deviation for some values of x in the dataset more than others (see Lemma 2).

1.2 Thesis Aims and Outline

The natural question that arises from the experiment above is how losses behave in the context of the dataset as a whole. While concepts such as property elicitation study the characteristics of loss functions at the micro-level, analyzing the behaviour of loss functions conditional on the value of the independent variables in the dataset, our study aims to motivate the research of these loss functions at the dataset level. Thus given a dataset, a practitioner would then be able to decide what characteristics his final model should have and make an informed choice about the loss function based on those attributes.

The outline of the thesis is as follows,

- Chapter 2 of the thesis lays down the preliminaries that are required for the rest of the thesis.
- In the Chapter 3, we pick a well-known attribute of the squared loss error in the context of linear regression and attempt to characterize all loss functions that exhibit the same property. The results of the characterization are then discussed using synthetic and real world data.
- Chapter 4 of the thesis specifies future extensions and possible directions of research.

Chapter 2

Preliminaries

2.1 Property Elicitation

Property elicitation is a field of study of properties that can be minimized by empirical risk minimization. We shall use the framework for property elicitation provided by Frongillo and Kash [5]. We repeat the definitions in the framework here for completeness. Let $\mathcal{Y}$ be a set of outcomes, such that $\mathcal{Y} \subset \mathbb{R}$ and $\mathcal{P}$ be a set of probability measures defined over $\mathcal{Y}$.

Definition 1. *A property is a function* $\Gamma : \mathcal{P} \to \mathbb{R}^k$ *for some* $k \in \mathbb{N}$ *which associates a desired report value with each distribution.*

In other words, a property is simply a statistic associated with a distribution. The mean, median, and mode are all properties of a distribution.

Definition 2 (Frongillo, Kash)**.** *A loss function* $L : \mathbb{R}^l \times \mathcal{Y} \to [0, \infty)$ *elicits a property* Γ *if for all* $p \in \mathcal{P}$, $\Gamma(p) = \arg\min_r \mathbb{E}[L(r, \cdot)]$ *. A property is elicitable if some loss elicits it.*

2.2 Property Elicitation and ERM

The concept of property elicitation lends itself directly to the empirical risk minimization framework. In ERM, our aim is to select an optimal hypothesis h^* that minimizes the the sum of losses over all the data points, i.e. given a dataset $\mathcal{D} \subset (\mathbb{R}^k \times \mathcal{Y})$ and a loss function $L : \mathbb{R}^k \times \mathcal{Y} \to [0, \infty)$,

$$h^* = \arg\min_{h \in \mathcal{H}} \sum_{(x,y) \in \mathcal{D}} L(h(x), y). \tag{2.1}$$

Using this definition, we can draw a direct relation between elicitation and ERM. If our optimal hypothesis h^* always passes through the conditional statistic Γ for every x, then ERM for any loss eliciting Γ will yield h^*.

Theorem 1 (Frongillo, Waggoner [7])**.** *Given a hypothesis class $\mathcal{H}$ and a dataset $\mathcal{D} \subset \mathbb{R}^m \times \mathcal{Y}$; if the function $h^* : \mathbb{R}^k \to \mathcal{Y}$ defined as $h^*(x) = \Gamma(Y|X = x)$, where $\Gamma(Y|X = x)$ is the distribution over the dependent variable Y conditional on the value of the dependent variable X being x, is in $\mathcal{H}$, then L elicits Γ only if empirical risk minimization over $\mathcal{D}$ yields h^*.*

Proof. Given that L elicits $\Gamma(Y|X = x)$, we have

$$\arg\min_{h \in \mathcal{H}} \sum_{y \in (Y|X=x)} L(h, y) = \Gamma(Y|X = x). \tag{2.2}$$

Since the loss at every x is minimized at Γ and $h^*(x) = \Gamma(Y|X = x), \forall x$,

$$\arg\min_{h \in \mathcal{H}} \sum_{x \in X} \sum_{y \in (Y|X=x)} L(h(x), y) = \sum_{x \in X} \Gamma(Y|X = x) \tag{2.3}$$

$$\implies \arg\min_{h \in \mathcal{H}} \sum_{x \in X} \sum_{y \in (Y|X=x)} L(h(x), y) = h^* \tag{2.4}$$

$$\implies \arg\min_{h \in \mathcal{H}} \sum_{(x,y) \in \mathcal{D}} L(h(x), y) \qquad = h^*. \tag{2.5}$$

□

2.3 Bregman Divergences

The well-known squared error is the most popular member of a larger family of functions known as the Bregman divergences, which can be defined as follows [3, 4].

Definition 3 (Bregman; Censor, Lent)**.** *Let $\phi : S \to \mathbb{R}$ be a strictly convex function defined on a convex set $S \subset \mathbb{R}^d$, such that ϕ is differentiable on S. For any two points $(x, y) \in S$, the Bregman Divergence associated with ϕ is defined as:*

$$\mathcal{B}_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle. \tag{2.6}$$

For the single variable case, the Bregman Divergence can be written as

$$\mathcal{B}_\phi(y, \hat{y}) = \phi(y) - \phi(\hat{y}) - \phi'(\hat{y})(y - \hat{y}). \tag{2.7}$$

2.3.1 Properties of Bregman Divergences

Some basic properties of Bregman divergences include

(1) **Non-negativity:** $\mathcal{B}_\phi(y, \hat{y}) \geq 0, \forall (y, \hat{y})$, with the equality holding if and only $\hat{y} = y$.

(2) **Convexity:** All Bregman divergences $\mathcal{B}_\phi(y, \hat{y})$ are convex in their first argument, $\hat{y}$.

(3) **Linearity:** $\mathcal{B}_{\phi_1 + \lambda\phi_2}(y, \hat{y}) = \mathcal{B}_{\phi_1}(y, \hat{y}) + \lambda\mathcal{B}_{\phi_2}(y, \hat{y})$.

A list of common convex functions and their corresponding Bregman Divergence is provided in Table 2.1.

In 2005, Banerjee et al.[1] showed that a necessary and sufficient condition for a loss function to be a Bregman divergence was that the conditional mean of the distribution was the unique minimizer for the function. In the framework of property elicitation this means all Bregman Divergences elicit the mean.

Theorem 2. *Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be a strictly convex differentiable function and $\mathcal{B}_\phi$ be the corresponding Bregman divergence. Then among all function of Z, the conditional expectation is the unique minimizer of the expected Bregman loss*

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

Theorem 3. *Let $F : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}$ be a non-negative function such that $F(x, x) = 0, \forall x \in \mathbb{R}$. If for all random variables $X, \mathbb{E}[X]$ is the unique minimizer of $\mathbb{E}[F(X, y)]$, over all outcomes $y \in \mathbb{R}$, then $F(x, y) = \mathcal{B}_\phi(x, y)$ for some strictly convex and differentiable ϕ.*

More recently Frongillo and Kash [8] generalized this result and relaxed the assumption that ϕ be differentiable.

Table 2.1: Common Convex Function and their corresponding Bregman Divergences.

Domain	$\phi(x)$	$D_\phi(x, y)$	Loss
$\mathbb{R}$	x^2	$(x-y)^2$	Squared Error
$\mathbb{R}$	e^x	$e^x - e^y - e^y(x-y)$	Exponential Loss
$\mathbb{R}$	e^{-x}	$e^{-x} - e^{-y} + e^{-y}(x-y)$	Negative Exponential Loss
$\mathbb{R}_{++}$	$x\log(x)$	$x\log(x/y) - (x-y)$	
$\mathbb{R}^d$	$x^T Ax$	$(x-y)^T A(x-y)$	Mahalanobis Distance
d-simplex	$\sum_{j=1}^d x_j \log x_j$	$\sum_{j=1}^d x_j log(x_j/y_j)$	KL- Divergence

Chapter 3

On the mean of the data

3.1 Linear Regression and Squared Error

For linear regression problems, the common choice of loss function is the squared loss. The squared error is popular in part because it is solely dependent on the prediction error $y - \hat{y}$, where $\hat{y}$ is the predicted outcome and y is the true outcome. Among Bregman Divergences, the squared error is only function of this form [9].

The squared loss is perhaps the best known member of the Bregman Divergence family, obtained by plugging $\phi(x) = x^2$ in Equation 2.6. From Theorem 2, we know that the conditional mean $\mathbb{E}[Y|X = x]$, will be the unique minimizer for the squared error. In other words, the squared error elicits the mean of the distribution p, for all p defined over our outcome class $\mathcal{Y}$. This is another reason why the squared error is sometimes preferred- the mean has nicer mathematical attributes than other statistics.

Another interesting property of the squared error is that the line obtained though linear regression with squared error as the loss passes through the the point $\bar{x}, \bar{y}$. Here, $\bar{x}$ and $\bar{y}$ are obtained by the calculating the mean of the independent variable x and the dependent variable y respectively. This point will be referred to as the mean-mean point henceforth.

Definition 4. *Let $\mathcal{D} \subset \mathbb{R}^k \times \mathcal{Y}$ that contains N points of the form (x,y). The mean-mean point, denoted by $(\bar{x}, \bar{y})$ is defined as,*

$$(\bar{x}, \bar{y}) = \Big(\frac{1}{N}\sum_{x_i \in \mathcal{D}} x_i, \frac{1}{N}\sum_{y_i \in \mathcal{D}} y_i\Big). \tag{3.1}$$

Theorem 4. *Let $\mathcal{D}$ be any dataset and w^*, b^* be the optimal linear regression parameters obtained by minimizing the squared error over all the data points in $\mathcal{D}$, then we have*

$$w^* \cdot \bar{x} + b^* = \bar{y}. \tag{3.2}$$

Proof. As in ERM, the aim in linear regression is to find w^*, b^* such that the sum of losses over all point in the data is minimized.

$$w^*, b^* = \underset{w,b \in \mathbb{R}}{\arg\min} \sum_{(x,y)\in\mathcal{D}} L(w \cdot x + b, y). \tag{3.3}$$

Thus, for squared error,

$$w^*, b^* = \underset{w,b \in \mathbb{R}}{\arg\min} \sum_{(x_i,y_i)\in\mathcal{D}} (w \cdot x_i + b - y_i)^2. \tag{3.4}$$

Let $-wx_i + y_i = z_i$. We then have

$$b^* = \underset{b}{\arg\min} \sum_i (b - z_i)^2. \tag{3.5}$$

As the squared error elicits the mean [8],

$$b^* = \bar{z_i} = -w \cdot \bar{x} + \bar{y} \tag{3.6}$$

$$\implies \bar{y} = w^* \cdot \bar{x} + b^*. \tag{3.7}$$

□

3.2 Towards a General Characterization

We now formally define the meaning of a regression line passing through the mean-mean point and show that all loss functions whose regression lines pass through the mean-mean can be characterized by the inverse of their second derivative. Further, we shall show that a dataset with only three points is sufficient to completely characterize all such loss functions.

Axiom 1. *A loss function* $L : \mathbb{R}^k \times \mathcal{Y} \to [0, \infty)$ *has* ***the mean-mean attribute*** *if the regression line* $w^*x + b^*$ *obtained by minimizing* L *over the dataset* $\mathcal{D}$ *passes through the mean-mean point for all datasets* $\mathcal{D}$ *i.e.,*

$$w^* \cdot \bar{x} + b^* = \bar{y} \tag{3.8}$$

where,

$$w^*, b^* = \arg\min_{w^*, b^*} \sum_{(x_i, y_i) \in D} L(w \cdot x_i + b, y_i)$$

and

$$(\bar{x}, \bar{y}) = \Big(\frac{1}{n} \sum_{x_i \in \mathcal{D}} x_i, \frac{1}{n} \sum_{y_i \in \mathcal{D}} y_i\Big).$$

The first step in our proof is to show that the Bregman divergences are the only class of loss functions that has the mean-mean attribute.

Lemma 1. *Let* $L : \mathbb{R} \times \mathcal{Y} \to [0, \infty)$ *be a loss function and* $\mathcal{D}$ *be a set of data points such that* $D \subset \mathbb{R} \times \mathcal{Y}$*. If* L *has the mean-mean attribute, then* L *is a Bregman divergence.*

Proof. We provide a proof by construction. Suppose $\exists L$ such that L satisfies the mean-mean attribute.

We synthesize $\mathcal{D}$ to be of the form (x, y_i), i.e. with a constant independent variable and a varying outcome. In this case $(\bar{x}, \bar{y}) = (x, \bar{y})$. Also $\mathbb{E}[Y|X = x] = (x, \bar{y})$, i.e the conditional mean at x corresponds to the mean-mean point (see Figure 3.1). This means that $E[Y|X = x]$ is a unique minimizer of $L(x, y)$.

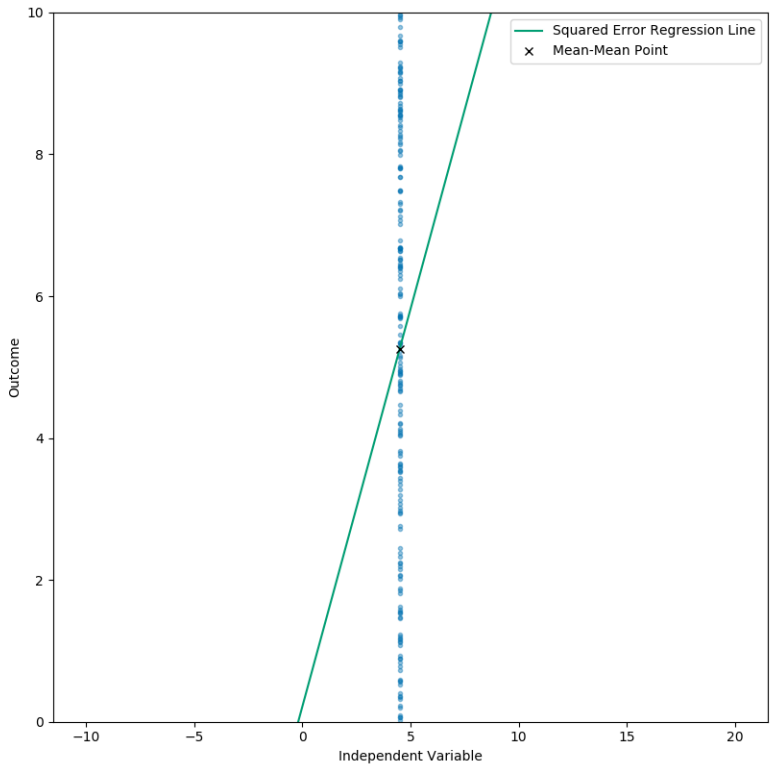

Figure 3.1: The conditional mean and mean-mean point coincide on our synthesized dataset with a constant independent variable and a variable outcome. The squared error, which is a Bregman Divergence, passes through this point.

From Theorem 2 and 3, we know that a loss is a Bregman Divergence if and only if the conditional mean is a minimizer. Thus, if L has to have the mean-mean attribute, the regression line obtained by minimizing L over the $\mathcal{D}$ has to pass through $(x, \bar{y})$. Thus L has to be a Bregman Divergence. □

Lemma 2. *For any univariate linear regression problem that uses a Bregman divergence $\mathcal{B}_\phi$ as a loss function over a dataset $\mathcal{D} \subset \mathbb{R} \times \mathcal{Y}$, the optimal slope and intercept parameters, w^* and b^*, satisfy,*

$$\sum_{(x_i,y_i)\in\mathcal{D}} \phi''(w^*x_i + b^*)[y_i - (w^*x_i + b^*)] = 0$$

and,

$$\sum_{(x_i,y_i)\in\mathcal{D}} \phi''(w^*x_i + b^*)[y_i - (w^*x_i + b^*)]x_i = 0.$$

Proof. To find w^* and b^*, we need to minimize the sum of losses over all data points in $\mathcal{D}$ i.e

$$w^*, b^* = \arg\min_{w,b} \sum_{x_i,y_i\in\mathcal{D}} \mathcal{B}_\phi(y_i, wx_i + b). \tag{3.9}$$

Let

$$Z = \sum_{x_i,y_i\in\mathcal{D}} \mathcal{B}_\phi(y_i, wx_i + b). \tag{3.10}$$

This would be possible only if the partial derivatives at w^* and b^* are equal to 0. Differentiating Z with respect to w and b and applying chain rule, we obtain,

$$\frac{\partial Z}{\partial b}(w^*, b^*) = \sum_{(x_i,y_i)\in\mathcal{D}} \frac{\partial \mathcal{B}_\phi(y_i, wx_i + b)}{\partial(wx_i + b)} = 0 \tag{3.11}$$

and

$$\frac{\partial Z}{\partial w}(w^*, b^*) = \sum_{(x_i,y_i)\in\mathcal{D}} \frac{\partial \mathcal{B}_\phi(y_i, wx_i + b)}{\partial(wx_i + b)} \cdot x_i = 0. \tag{3.12}$$

Also we have,

$$\frac{\partial \mathcal{B}_\phi(\hat{y}, y)}{\partial \hat{y}} = \phi''(\hat{y})(y - \hat{y}). \tag{3.13}$$

Substituting the derivative in Equations 3.11 and 3.12 yields

$$\frac{\partial Z}{\partial b}(w^*, b^*) = \sum_{(x_i,y_i)\in\mathcal{D}} \phi''(w^*x_i + b^*)[y_i - (w^*x_i + b^*)] = 0 \tag{3.14}$$

and

$$\frac{\partial Z}{\partial w}(w^*, b^*) = \sum_{(x_i, y_i) \in \mathcal{D}} \phi''(w^* x_i + b^*)[y_i - (w^* x_i + b^*)]x_i = 0. \tag{3.15}$$

□

From Equations 3.14 and 3.15, we can see that the second derivative of ϕ acts like a weight to the points in the dataset. Each term in the sum consists of the residual at that point $y_i - (w^* x_i + b^*)$. The second derivative assigns a multiplicative penalty to each of these residuals. If we revisit Section 1.1, we can see that this intuition holds true in the plots for the the three Bregman divergences used in the experiment. For all $\hat{y}$ he squared error has $\phi''(\hat{y}) = 1$ and weighs all the points equally. For the exponential function, $\phi''(\hat{y}) = e^x$. The points having a higher value for x, the independent variable, in the dataset are thus prioritized higher, leading to a line that fits the points occurring later in the series better. On the other hand, the negative exponential function prioritizes points earlier in the series as $\phi''_{e^{-x}}(\hat{y}) = e^{-x}$.

Our next step is to derive a form that all loss functions having the mean-mean attribute must satisfy. We shall show that this can be done with a dataset of only 3 points. Let us synthesize a 3-point dataset as follows, with α, $\tilde{y}$, and c as parameters-

$$\mathcal{D} = \{(\alpha, c)(-1, \tilde{y} + c), (1 - \alpha, -\tilde{y} + c)\}$$

The parameter α controls the scaling of the points along the x-axis. The parameter $\tilde{y}$ controls the scaling of the points along the y-axis. Finally, c translates the points along the x-axis. We define the following axiom that we shall use in our subsequent proof

Axiom 2. *For any continuous loss function L, given values for α and c, there exists a value for $\tilde{y}$ for all w^*, such that linear regression using L on $\mathcal{D}$ would yield a slope of w^*.*

$$\forall L, w^*, \exists \tilde{y} : \arg\min_{w} \sum_{(x_i, y_i) \in D} L(x_i, y_i) = w^* \tag{3.16}$$

Lemma 3. *Any loss function, satisfying Axiom 2, that has the mean-mean attribute can be written as a Bregman Divergence $\mathcal{B}_\phi$, where*

$$\phi(\hat{y}) = \begin{cases} \dfrac{(A+B\hat{y})\log(A+B\hat{y})}{B^2} + C_1\hat{y} + C_2 & \text{if } B \neq 0 \\ \dfrac{\hat{y}^2}{A} + C_1\hat{y} + C_2 & \text{if } B = 0 \end{cases} \tag{3.17}$$

for some A, B, C_1 and C_2.

Proof. Using Equations 3.14 and 3.15 and applying them to our dataset $\mathcal{D}$, we get

$$\begin{aligned} &\phi''(w^*\alpha + b^*)(-w^*\alpha) + \phi''(-w^* + b^*)(\tilde{y} + w^*) \\ &\quad + \phi''(w^*(1-\alpha) + b^*)[-\tilde{y} - w^*(1-\alpha)] = 0 \end{aligned} \tag{3.18}$$

and

$$\begin{aligned} &\phi''(w^*\alpha + b^*)(-w^*\alpha)\cdot\alpha - \phi''(-w^* + b^*)(\tilde{y} + w^*) \\ &\quad + \phi''(w^*(1-\alpha) + b^*)[-\tilde{y} - w^*(1-\alpha)]\cdot(1-\alpha)] = 0. \end{aligned} \tag{3.19}$$

The mean-mean point of $\mathcal{D}$ is $(0, c)$. Therefore for any linear regression line that passes through this point, the intercept term would be c, i.e $b^* = c$.

$$\begin{aligned} \implies &\phi''(w^*\alpha + c)(-w^*\alpha) + \phi''(-w^* + c)(\tilde{y} + w^*) \\ &\quad + \phi''(w^*(1-\alpha) + c)[-\tilde{y} - w^*(1-\alpha)] = 0 \end{aligned} \tag{3.20}$$

and

$$\begin{aligned} &\phi''(w^*\alpha + c)(-w^*\alpha)\cdot\alpha - \phi''(-w^* + c)(\tilde{y} + w^*) \\ &\quad + \phi''(w^*(1-\alpha) + c)[-\tilde{y} - w^*(1-\alpha)]\cdot(1-\alpha) = 0. \end{aligned} \tag{3.21}$$

Adding 3.20 and 3.21 and rearranging, we get

$$\frac{\phi''(w^*\alpha + c)}{\phi''((1-\alpha)w^* + c)} = \frac{(2-\alpha)[\tilde{y} + (1-\alpha)w^*]}{(\alpha+1)(-\alpha w^*)}. \tag{3.22}$$

Multiplying 3.20 by $(1-\alpha)$ and subtracting 3.21 we have

$$\frac{\phi''(w^*\alpha + c)}{\phi''(-w^* + c)} = \frac{(\tilde{y} + w^*)(2-\alpha)}{(1-2\alpha)(\alpha w^*)}. \tag{3.23}$$

We can solve 3.22 and 3.23 for $\tilde{y}$ and equate the two to get

$$\frac{\phi''(\alpha w^* + c)(\alpha+1)(-\alpha w^*)}{\phi''((1-\alpha)w^* + c)(2-\alpha)} - (1-\alpha)w^* = \frac{\phi''(\alpha w^* + c)(1-2\alpha)(\alpha w^*)}{\phi''(-w^* + c)(2-\alpha)} - w^* \tag{3.24}$$

$$\implies \frac{1-2\alpha}{\phi''(-w^* + c)} = \frac{2-\alpha}{\phi''(\alpha w^* + c)} - \frac{\alpha+1}{\phi''((1-\alpha)w^* + c)} \tag{3.25}$$

$$\implies h(-w^* + c) = \frac{2-\alpha}{1-2\alpha} h(\alpha w^* + c) - \frac{\alpha+1}{1-2\alpha} h((1-\alpha w^*) + c). \tag{3.26}$$

Here, $h(\hat{y}) = 1/\phi''(\hat{y})$. Now let $h(0)$ and $h(1)$ be two arbitrary constants.

Due to Axiom 2, we can change the value of w^* in our equations without breaking the equality in Equation 3.26. We can tweak the values of α, w^* and c such that $(\alpha w^* + c) = 0$ and $(1 - \alpha w^*) + c = 1$. Since we have two constraints and two variables, the LHS remains unrestricted and we can determine h for all $\hat{y} \in \mathbb{R}$. We have

$$\alpha w^* + c = 0 \tag{3.27}$$

$$\implies w^* = \frac{-c}{\alpha}. \tag{3.28}$$

Also,

$$(1-\alpha)w^* + c = 1. \tag{3.29}$$

From 3.27

$$\alpha = \frac{-c}{1-2c}. \tag{3.30}$$

Plugging the values in 3.26, we have

$$h(3c-1) = (2-3c)\cdot h(0) - (1-3c)\cdot h(1). \tag{3.31}$$

A change of variables $\hat{y} \Leftrightarrow 3c-1$ yields

$$h(\hat{y}) = (-\hat{y}+1)h(0) + \hat{y}h(1) \tag{3.32}$$

$$= h(0) + \hat{y}(h(1) - h(0)). \tag{3.33}$$

Setting $h(0) = A, h(1) - h(0) = B$ and plugging back $h(\hat{y}) \Leftrightarrow 1/\phi''(\hat{y})$, we obtain

$$\phi''(\hat{y}) = \frac{1}{A + B\hat{y}}. \tag{3.34}$$

Assuming $B \neq 0$ and integrating twice with respect to $\hat{y}$ on both sides we get,

$$\int\int \phi''(\hat{y})d\hat{y} = \int\int \frac{1}{A+B\hat{y}}d\hat{y} \tag{3.35}$$

$$\phi(\hat{y}) = \frac{(A+B\hat{y})\log(A+B\hat{y})}{B^2} + \frac{\hat{y}(Bc_1-1)}{B} + c_2 \tag{3.36}$$

$$= \frac{(A+B\hat{y})\log(A+B\hat{y})}{B^2} + C_1\hat{y} + C_2. \tag{3.37}$$

When $B = h(1) - h(0) = 0$ we get,

$$\int\int \phi''(\hat{y})d\hat{y} = \int\int \frac{1}{A}d\hat{y} \tag{3.38}$$

$$\implies \phi(\hat{y}) = \frac{\hat{y}^2}{A} + C_1\hat{y} + C_2. \tag{3.39}$$

□

Lemma 3, specifies the equation that all loss functions exhibiting the mean-mean attribute need to satisfy. To achieve the full characterization we need to proceed in the reverse direction, starting from the equations, to further establish constraints.

Theorem 5. *A loss function exhibits the mean-mean attribute if and only if it can be characterized as a Bregman Divergence $\mathcal{B}_\phi$, where*

$$\phi(\hat{y}) = \begin{cases} \dfrac{(A + B\hat{y})\log(A + B\hat{y})}{B^2} + C_1\hat{y} + C_2 & \text{if } B \neq 0 \\ \dfrac{\hat{y}^2}{A} + C_1\hat{y} + C_2 & \text{if } B = 0 \end{cases} \tag{3.40}$$

where $A > 0$ *and* $A + B(\hat{y}) > 0$ *for all predictions* $\hat{y}$.

Proof. We start with the equations obtained in Lemma 3. We need to prove that each of the cases in the equation satisfy the mean-mean attribute.

Case 1: $B = 0$

For this case we have

$$\phi(\hat{y}) = \frac{\hat{y}^2}{A} + C_1\hat{y} + C_2 \tag{3.41}$$

$$\implies \phi''(\hat{y}) = \frac{2}{A}. \tag{3.42}$$

In this case $\mathcal{B}_\phi$ is a variant of the squared error. Theorem 4 provides a proof for the case when $A = 1, C_1 = 0, C_2 = 0$. We shall prove the statement for the general case here. From equation 3.14 we have the a condition for the optimal linear regression line. This equation is sufficient to show that the squared error has the mean-mean attribute. We have

$$\sum_{(x_i,y_i)\in\mathcal{D}} \phi''(w^*x_i + b^*)[y_i - (w^*x_i + b^*)] = 0 \tag{3.43}$$

$$\implies \sum_{(x_i,y_i)\in\mathcal{D}} \frac{2}{A}[y_i - (w^*x_i + b^*)] = 0 \tag{3.44}$$

$$\implies \sum_{(x_i,y_i)\in\mathcal{D}} [y_i - (w^*x_i + b^*)] = 0 \tag{3.45}$$

$$\implies \sum_{y_i\in\mathcal{D}} y_i = w^* \sum_{x_i\in\mathcal{D}} + b^* \sum_{i=0}^{N} \cdot 1 \tag{3.46}$$

$$\implies \sum_{y_i\in\mathcal{D}} y_i = w^* \sum_{x_i\in\mathcal{D}} + Nb^*. \tag{3.47}$$

Dividing by N on both sides, where N is the number of points in $\mathcal{D}$

$$\bar{y} = w^*\bar{x} + b^*. \tag{3.48}$$

In this case, the resulting regression line passes through the mean-mean point for all values of $A > 0$.

Case 2 : $B \neq 0$

We have

$$\phi(\hat{y}) = \frac{(A + B\hat{y}) \log(A + B\hat{y})}{B^2} + C_1\hat{y} + C_2 \tag{3.49}$$

$$\implies \phi''(\hat{y}) = \frac{1}{A + B\hat{y}}. \tag{3.50}$$

Again, from equation 3.14 we have

$$\sum_{(x_i,y_i)\in\mathcal{D}} \phi''(w^*x_i + b^*)[y_i - (w^*x_i + b^*)] = 0 \tag{3.51}$$

$$\implies \sum_{(x_i,y_i)\in\mathcal{D}} \frac{1}{A + B(w^*x_i + b^*)}[y_i - (w^*x_i + b^*)] = 0 \tag{3.52}$$

$$\implies \sum_{(x_i,y_i)\in\mathcal{D}} \frac{y_i}{A + B(w^*x_i + b^*)} = \sum_{(x_i,y_i)\in\mathcal{D}} \frac{w^*x_i + b^*}{A + B(w^*x_i + b^*)}. \tag{3.53}$$

Multiplying by B on both sides; adding and subtracting A to the numerator of the RHS

$$\sum_{(x_i,y_i)\in\mathcal{D}} \frac{By_i}{A + B(w^*x_i + b^*)} = \sum_{(x_i,y_i)\in\mathcal{D}} \frac{B(w^*x_i + b^*) + A - A}{A + B(w^*x_i + b^*)} \tag{3.54}$$

$$\implies \sum_{(x_i,y_i)\in\mathcal{D}} \frac{By_i}{A + B(w^*x_i + b^*)} = \sum_{i=0}^{N} 1 - \sum_{(y_i)\in\mathcal{D}} \frac{A}{A + B(w^*x_i + b^*)} \tag{3.55}$$

$$\implies \sum_{(x_i,y_i)\in\mathcal{D}} \frac{By_i + A}{A + B(w^*x_i + b^*)} = N. \tag{3.56}$$

Now from equation 3.15 we have

$$\sum_{(x_i,y_i)\in\mathcal{D}} \phi''(w^*x_i + b^*)[y_i - (w^*x_i + b^*)]x_i = 0 \tag{3.57}$$

$$\implies \sum_{(x_i,y_i)\in\mathcal{D}} \frac{[y_i - (w^*x_i + b^*)]x_i}{A + B(w^*x_i + b^*)} = 0 \tag{3.58}$$

$$\implies \sum_{(x_i,y_i)\in\mathcal{D}} \frac{[By_i - (B(w^*x_i + b^*) + A - A)]x_i}{A + B(w^*x_i + b^*)} = 0 \tag{3.59}$$

$$\implies \sum_{(x_i,y_i)\in\mathcal{D}} \frac{(By_i + A)x_i}{A + B(w^*x_i + b^*)} - \sum_{(x_i)\in\mathcal{D}} x_i = 0. \tag{3.60}$$

Multiplying by w^*/N on both sides, where N is the number of points in $\mathcal{D}$

$$\frac{w^*}{N} \sum_{(x_i,y_i)\in\mathcal{D}} \frac{(By_i + A)x_i}{A + B(w^*x_i + b^*)} = w^*\bar{x}. \tag{3.61}$$

Multiplying by B on both sides; adding and subtracting b^* and A to the numerator of the LHS yields

$$\frac{1}{N} \sum_{(x_i,y_i)\in\mathcal{D}} \frac{(By_i + A)[B(w^*x_i + b^* - b^*) + A - A]}{A + B(w^*x_i + b^*)} = Bw^*\bar{x} \tag{3.62}$$

$$\implies \frac{1}{N} \sum_{(x_i,y_i)\in\mathcal{D}} (By_i + A) - \frac{1}{N} \sum_{(x_i,y_i)\in\mathcal{D}} \frac{(Bb^* + A)(By_i + A)}{A + B(w^*x_i + b^*)} = Bw^*\bar{x}. \tag{3.63}$$

Substituting from Equation, 3.56, we get-

$$\frac{1}{N} \sum_{(x_i,y_i)\in\mathcal{D}} (By_i + A) = B(w^*\bar{x} + b^*) + A \tag{3.64}$$

$$\implies \frac{1}{N} \sum_{(x_i,y_i)\in\mathcal{D}} (y_i) = w^*\bar{x} + b^* \tag{3.65}$$

$$\implies \bar{y} = w^*\bar{x} + b^*. \tag{3.66}$$

As ϕ is a convex function, an additional constraint is imposed by the requirement that the second derivative always be positive. We have for all $x_i \in D$

$$\phi''(w^*x_i + b*) = \frac{1}{A + B(w^*x_i + b^*)} > 0 \tag{3.67}$$

$$\implies A + B(w^*x_i + b^*) > 0. \tag{3.68}$$

The fact that this constraint can always be satisfied is not obvious and merits discussion. For any loss passing satisfying the mean-mean attribute, the outcome space $\mathcal{Y}$ will always need to be a subset of the prediction space $\hat{\mathcal{Y}}$, which is a set of all possible values the hypothesis h^* can predict. This is because the mean-mean point always needs to be in $\hat{\mathcal{Y}}$ for a loss to have the mean-mean attribute. If $\mathcal{Y}$ was not a subset of $\hat{\mathcal{Y}}$, it would be possible to construct a dataset such that none of the outcomes lie in the the outcome space $\hat{\mathcal{Y}}$, meaning that mean-mean point would not be in $\hat{\mathcal{Y}}$. Thus, we have $\mathcal{Y} \subseteq \hat{\mathcal{Y}}$. For the loss given by equation 3.40, we have $\mathcal{Y} = \mathbb{R}^+$. Thus, $\hat{\mathcal{Y}}$ will always contain $\mathbb{R}^+$, and the constraint given by equation 3.68 can always be satisfied. □

3.3 Discussion

3.3.1 Looking at the Second Derivative of $x \log x$

When $B \neq 0$ in Equation 3.40, we get a Bregman Divergence $\mathcal{B}_\phi$ where ϕ is a variant of the function $x \log x$. Along with the squared error, this function is the only other function that has the mean-mean attribute. From Lemma 2, we know that the value of $\phi''(x)$, serves as a penalty to the residual at each point. For $\phi(x) = x \log x$, we have $\phi''(x) = 1/x$.

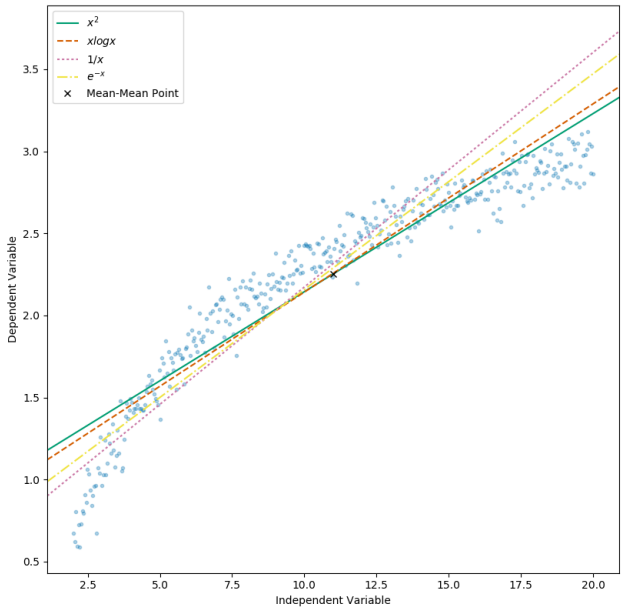

Figure 3.2: Linear Regression using Bregman Divergences with different values of ϕ. Notice that $x \log x$ and x^2 intersect at the mean-mean point.

Figure 3.2 uses our synthetic dataset and performs linear regression using various Bregman divergences of it. We can see that the effect that weighting of $\phi''(x)$ has on the slope of the line, favoring a lower residual for the points occuring earlier in the series.

3.3.2 Inspecting the domain of $h(x)$

Since ϕ is a convex function its second derivative $h(x)$ needs to be non-negative. This places certain restrictions on our choice of function based on the domain of $h(x)$.

We have from Equation 3.33

$$h(\hat{y}) = h(0) + \hat{y}(h(1) - h(0)). \tag{3.69}$$

For $h(\hat{y})$ for all $\hat{y} \in \mathbb{R}$ to be strictly positive, we would need the second term in the RHS to be $0 \implies h(1) = h(0)$ and $h(\hat{y})$ is a constant. Thus we have

Corollary 1. *The only loss function that has the mean-mean attribute in the domain of* $\mathbb{R}$ *is the squared error.*

If we restrict $\hat{y}$ to $\mathbb{R}^+$ we remove the need for $h(\hat{y})$ to be a constant and thus $\phi(\hat{y})$ can be characterized by equation 3.40.

Chapter 4

Future Work

4.1 The Mean-Mean attribute in the multidimensional case

Our work in Chapter 3 focused on the mean-mean point. We restricted ourselves to the single-dimensional case and characterized all loss functions that exhibit the mean-mean attribute. The single dimension is easier to visualize and makes for equations with lesser degrees of freedom. An obvious extension would then be to extend our results and generalize them to the multi-dimensional case.

4.2 Understanding the importance of the $\phi''(x)$

At this point, we have an intuition of the role the second derivative of ϕ plays in deciding the slope and intercept of the regression line. A rigorous mathematical treatment is required to quantify the exact impact of $\phi''(x)$ and any other factors in the optimality conditions defined in Equations 3.14 and 3.15 on the regression line.

Bibliography

[1] Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with bregman divergences. Journal of machine learning research, 6(Oct):1705–1749, 2005.

[2] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. Journal of the American Statistical Association, 101(473):138–156, 2006.

[3] Lev 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 mathematical physics, 7(3):200–217, 1967.

[4] Yair Censor and Arnold Lent. An iterative row-action method for interval convex programming. Journal of Optimization theory and Applications, 34(3):321–353, 1981.

[5] Rafael Frongillo and Ian Kash. On elicitation complexity. In Advances in Neural Information Processing Systems, pages 3240–3248, 2015.

[6] Rafael Frongillo, Ian Kash, and Stephen Becker. Open problem: Property elicitation and elicitation complexity. In 29th Annual Conference on Learning Theory, pages 1655–1658, 2016.

[7] Rafael Frongillo and Bo Wagonner. Elicitation and machine learning, Jul 2016.

[8] Rafael M Frongillo and Ian A Kash. Vector-valued property elicitation. In COLT, pages 710–727, 2015.

[9] Tilmann Gneiting. Making and evaluating point forecasts. Journal of the American Statistical Association, 106(494):746–762, 2011.

[10] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. A distribution-free theory of nonparametric regression. Springer Science & Business Media, 2006.

[11] Ingo Steinwart and Andreas Christmann. Support vector machines. Springer Science & Business Media, 2008.