Bias-Variance Decompositions for Margin Losses

Danny Wood
University of Manchester

Tingting Mu
University of Manchester

Gavin Brown
University of Manchester

Abstract

We introduce a novel bias-variance decomposition for a range of strictly convex margin losses, including the logistic loss (minimized by the classic *LogitBoost* algorithm), as well as the squared margin loss and canonical boosting loss. Furthermore, we show that, for *all* strictly convex margin losses, the expected risk decomposes into the risk of a "central" model and a term quantifying variation in the functional margin with respect to variations in the training data. These decompositions provide a diagnostic tool for practitioners to understand model overfitting/underfitting, and have implications for additive ensemble models—for example, when our bias-variance decomposition holds, there is a corresponding "ambiguity" decomposition, which can be used to quantify model diversity.

1 INTRODUCTION

Bias-variance decompositions are broadly recognized as an important tool for understanding the generalization properties of statistical models (Geman et al., 1992; Belkin et al., 2019). The most well-known of these decompositions is for the squared loss (Geman et al., 1992), but such decompositions have been shown to hold for other losses, including KL-divergences between probability densities (Heskes, 1998; Hansen and Heskes, 2000), proper scoring rules (Buja et al., 2005) and Bregman divergences (Pfau, 2013). These decompositions allow us to see how much of our model's generalization error is due to poor model selection versus being due to randomness in the training data/learning procedure (Geman et al., 1992; Belkin et al., 2019; Adlam and Pennington, 2020).

In this work, we consider bias-variance decompositions for margin losses—i.e., surrogate loss functions for binary classification, where the task of classification is re-framed as estimation of a real value, with larger magnitudes corresponding to more confident predictions of the positive/negative class. Margin losses are used in classical algorithms and models such as AdaBoost (Freund and Schapire, 1999), LogitBoost (Friedman et al., 2000), and Support Vector Machines (SVMs), yet the issue of bias-variance decompositions for these losses has been largely overlooked. To the best of our knowledge, the only applicable previous work is Buja et al. (2005), which converts the losses into proper scoring rules. Their framework requires re-interpretation the real-valued model as a probability estimator—an undesirable intermediate step.

We present an alternative approach, not dependent on interpreting the models as probability estimators. The resulting decomposition is applicable to a broad class of losses which we characterize for the first time and refer to as *gradient-symmetric losses*. We also show that this family of losses is of independent interest, having strong connections with canonical form losses (Masnadi-Shirazi and Vasconcelos, 2010), linear odd losses (Patrini et al., 2016), and Bregman divergences. Furthermore, in the context of ensemble methods, we show that gradient-symmetric losses are exactly the sub-class of margin losses where we obtain an ambiguity decomposition (Krogh and Vedelsby, 1995) for linearly combined ensembles. A full list of our contributions is as follows.

- We derive a novel decomposition of the expected risk applicable to *any* strictly convex margin loss, separating the contributions of the expected model and variation in the functional margin.

- We introduce *gradient-symmetric* losses, a class of margin losses including the logistic, squared, Laplacian (Masnadi-Shirazi, 2011) and canonical boosting losses (Masnadi-Shirazi and Vasconcelos, 2010). For these losses, we present a novel bias-variance decomposition, which does not require interpreting models as probability estimators.
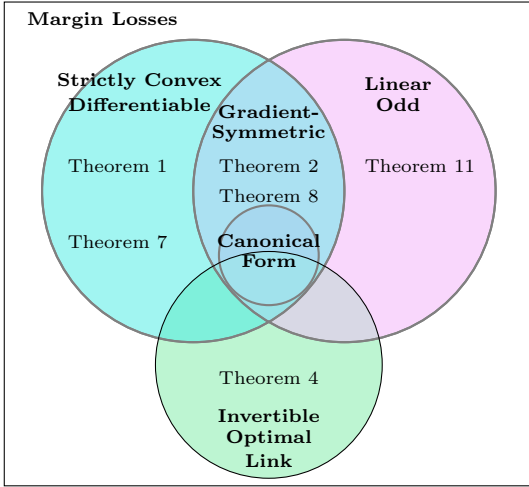
Figure 1: Visualization of How Our Contributions Relate to Known Classes of Margin Loss.

- For gradient-symmetric losses, we derive an ensemble ambiguity decomposition (Krogh and Vedelsby, 1995) for linearly combined ensembles. For non-gradient-symmetric losses, we show a similar decomposition, but requiring a *non-linear* ensemble combination rule.

- We show a close relationship between gradient-symmetric losses and Bregman divergences: gradient-symmetric losses are expressible in terms of a Bregman divergence from a representation of the label to the model.

- We examine how gradient-symmetric losses relate to canonical form losses (Masnadi-Shirazi and Vasconcelos, 2010) and linear odd losses (Patrini et al., 2016). Using the latter connection, we show that the bias term in our decomposition can be decomposed further into an expected margin term and a target-independent term—suggesting possible applications to semi-supervised settings.

The applicability of our results to various known classes of margin loss is shown in Figure 1. All the decompositions we present have a similar flavour: each separates the error into the risk of a "central" model and a stochastic component. The "central" model is the model obtained by integrating out all sources of randomness in the training data and learning algorithm, and the risk of this model is the sum of a systematic component and the error due to noise in the label distribution.
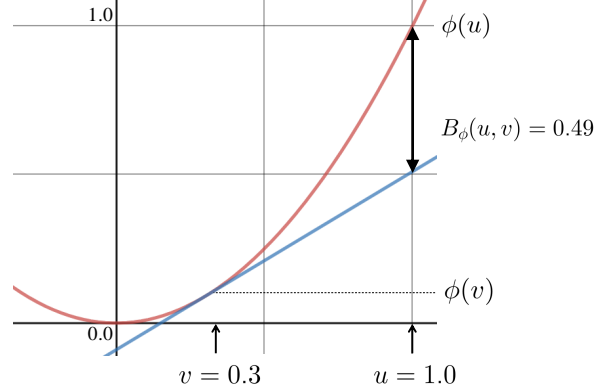

Figure 2: The Bregman divergence, illustrated for the Bregman generator $\phi(u) = u^2$.

2 PRELIMINARIES

Margin Losses We consider the following learning scenario: We have the problem of learning a function $h : \mathcal{X} \to \{-1, +1\}$, which minimizes the zero-one loss $\frac{1}{N} \sum_{j=1}^{N} [[h(\mathbf{x}_j) \neq y_j]]$ for a training set $\mathcal{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^{N}$ with feature vector $\mathbf{x}_j \in \mathcal{X}$ and target $y_j \in \{-1, +1\}$. We consider margin losses, i.e., surrogate losses of the form $\ell(yf(\mathbf{x}; \mathcal{D}))$, where $\ell : \mathbb{R} \to \mathbb{R}_+$ is an upper bound on the zero-one loss and $f(\cdot; \cdot)$ is the model, which is a function of the input $\mathbf{x} \in \mathcal{X}$ and parameterized by the training set $\mathcal{D}$. We use the model to predict a label $y \in \{-1, +1\}$ for $\mathbf{x} \in \mathcal{X}$ by mapping the model output to a class prediction with the rule $h(\mathbf{x}) = \text{sign}(f(\mathbf{x}; \mathcal{D}))$. Note that these losses have the property that the loss associated with an incorrect prediction is not affected by the class labelling scheme, i.e., $\ell((-y)(-f(\mathbf{x}; \mathcal{D}))) = \ell(yf(\mathbf{x}; \mathcal{D}))$.

The value $yf(\mathbf{x}; \mathcal{D})$ is known as the *functional margin*; positive values correspond to correct classifications, while negative values correspond to incorrect classifications. The magnitude of this margin can be interpreted as the model's confidence in its prediction, and has been associated with good generalization properties via bounds on the error (Schapire et al., 1998).

We will assume throughout this paper that the infimum of ℓ is zero. Most results can easily be modified for other finite lower bounds, but it is usually only sensible to consider losses where the minimizer(s) of ℓ are positive.

Bregman Divergences Bregman divergences (Bregman, 1967) are measures of separation between points, defined in terms of a strictly convex function ϕ, and are a frequently used tool in the study of margin losses (Zhang et al., 2004; Masnadi-Shirazi and Vasconcelos, 2009; Nock and Nielsen, 2008; Reid and Williamson, 2010). Loosely speaking, the Bregman di-

vergence $B_\phi(u, v)$ is the difference between the strictly convex function ϕ evaluated at u, and the linear approximation of ϕ around v, also evaluated at u (See Figure 2).

More formally, for some convex set S, the Bregman divergence $B_\phi : S \times \text{int}(S) \to \mathbb{R}_+$ is defined as

$$B_\phi(u, v) = \phi(u) - \phi(v) - \phi'(v)(u - v),$$

where $\text{int}(S)$ denotes the interior of the set S and ϕ is a strictly convex differentiable function. Bregman divergences are not generally symmetric, i.e., it does not necessarily hold that $B_\phi(u, v) = B_\phi(v, u)$.

Bias-Variance Decompositions In the typical supervised binary classification learning setup, a model is trained on a set of training examples $\mathcal{D} = \{(\mathbf{x}_j, y_j)\}_{j=1}^N$, where each example is drawn independently from a joint distribution over $\mathcal{X} \times \{-1, +1\}$. It is therefore natural, when considering the impact of varying the training set on model performance, to consider the training set as a random variable $D = \{(\mathbf{X}_j, Y_j)\}_{j=1}^N$, where each $(\mathbf{X}_j, Y_j)$ is independently distributed according to the aforementioned joint distribution. We consider the performance of the model $f(\cdot; D)$ over variations in D and over the true distribution of the data, which we write as the joint distribution of $(\mathbf{X}, Y)$. When we consider the random variable $f(\mathbf{X}; D)$, we will typically leave the arguments implicit, simply writing f. We also note that though $\mathbb{E}_D[\cdot]$ is the expectation over the distribution of training sets, the decomposition can just as easily apply for any source of stochasticity in the model or training data (e.g., randomization of initial weights).

For the squared loss, the setup is the same, except that $Y \in \mathbb{R}$. Here, the classic decomposition of (Geman et al., 1992) tells us that the expected risk can be decomposed as

$$\underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[(f - Y)^2\right]\right]}_{\text{expected risk}} = \underbrace{\mathbb{E}_{\mathbf{X}Y}\left[(Y - Y^*)^2\right]}_{\text{noise}} +$$
$$\underbrace{\mathbb{E}_{\mathbf{X}}\left[(Y^* - f^*)^2\right]}_{\text{bias}} + \underbrace{\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[(f^* - f)^2\right]\right]}_{\text{variance}},$$

where $Y^* = \mathbb{E}_{Y|\mathbf{X}}[Y]$ and $f^* = \mathbb{E}_D[f]$. The bias gives a measure of the distance of the "central" model f^* from the expected target, while the variance quantifies how far the typical model will be from this central model.

These quantities can be estimated in practice by training multiple models on different training sets, approximating the expectations with averages over the multiple datasets. This can either be done via training on disjoint subsets of the training set (Yang et al., 2020) or by training on bootstraps of the training data (Neal et al., 2018).

Other known bias-variance decompositions have the same structure, separating the expected risk into noise, bias and variance, with the variance term measuring the spread of models in the distribution around some central model (for example, see Pfau (2013); Heskes (1998)).

Bias-variance decompositions bring important insights: a large systematic component (i.e., high bias) suggests that the model is insufficiently complex (and therefore underfitting) whereas a large stochastic component (i.e., high variance) suggests over-sensitivity to the particular random sample (e.g., overfitting to training data).

3 BIAS-VARIANCE DECOMPOSITION FOR MARGIN LOSSES

3.1 Decomposition for Margin Losses

We are interested in the expected risk (or expected generalization error) of the model; for margin losses, this is defined as $\mathbb{E}_{\mathbf{X}Y}[\mathbb{E}_D[\ell(Yf(\mathbf{X}; D))]]$.

We begin with a general decomposition which has some of the properties we are looking for: it contains a term which measures the loss of a "central" model and a term which depends on how spread out the distribution of models is around this centre. However, the latter term is *not*, in general, independent of the target Y.

Theorem 1 (Margin Variance Decomposition)**.** *For a strictly convex differentiable margin loss ℓ,*

$$\underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[\ell(Yf)\right]\right]}_{\text{expected risk}} = \underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\ell(Yf^*)\right]}_{\text{risk of central model}} +$$
$$\underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[B_\ell\left(Yf, Yf^*\right)\right]\right]}_{\text{margin variance}}, \quad (1)$$

where we define the central model $f^* = \mathbb{E}_D[f]$.

Proof of this result, and all other results presented in this paper, can be found in the appendices.

The results shows that we can decompose the expected risk of the model into the loss of the expected model and a non-negative term quantifying the spread of the models around that expectation. We refer to this decomposition as the *margin variance decomposition* because for $y \in \{-1, +1\}$, the term $\mathbb{E}_D[B_\ell(yf, yf^*)]$ measures the spread of the functional margin around the expectation $\mathbb{E}_D[yf] = yf^*$.

Even with this target-dependent variance term, this decomposition gives insight into how much of the error is due the systematic part of the model (the central
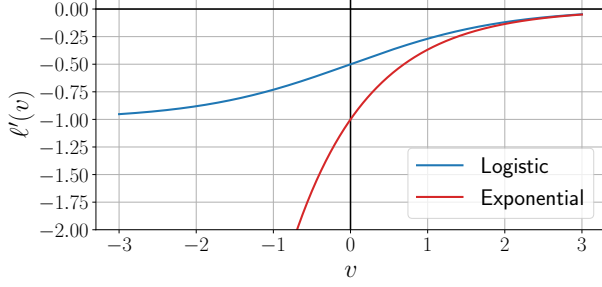
Figure 3: Comparison of Gradients of Logistic and Exponential Losses. The Logistic Gradient is Rotationally Symmetric About a Point on the y-axis, Hence we Refer to the Logistic Loss as a Gradient-Symmetric Loss.

model f^*) versus the random part (the distribution of f). However, there is a class of loss where the decomposition becomes a true bias-variance decomposition with a target-independent variance term. We consider the class of margin losses with the following property.

Definition 1 (Gradient-Symmetry). *Let* $\ell : \mathbb{R} \to \mathbb{R}_+$ *be a differentiable strictly-convex function.* ℓ *is gradient-symmetric if for all* $v \in \mathbb{R}$

$$\ell'(v) + \ell'(-v) = c \tag{2}$$

for some constant c.

Examples of gradient-symmetric losses (and notable non-gradient-symmetric ones) are given in Table 1, and illustrated in Figure 3. For practical purposes, we are only interested in losses where $c < 0$, since losses with $c > 0$ have the undesirable property of penalizing models more harshly for correct predictions than incorrect ones and losses with $c = 0$ ignore the sign of the label. With this restriction, gradient-symmetric losses are closely related to canonical form losses—each such gradient-symmetric loss is, up to a constant scaling factor, equivalent to a canonical form loss. However, the gradient-symmetry definition remains useful, as it completely captures the class of losses for which the margin-variance in (1) becomes target-independent. This is due to the following property.

Proposition 1. *For all* $u, v \in \mathbb{R}$ *and strictly convex differentiable* $\ell : \mathbb{R} \to \mathbb{R}_+$,

$$B_\ell(u, v) = B_\ell(-u, -v)$$

if and only if ℓ *is gradient-symmetric.*

This result means that for gradient-symmetric losses, the sign of Y in $B_\ell(Yf, Yf^*)$ has no effect. Applying this fact to the margin variance decomposition, we get the following bias-variance decomposition.

Theorem 2 (Bias-Variance Decomposition for Gradient-Symmetric Losses). *For a strictly convex differentiable margin loss* ℓ, *there exists a bias-variance decomposition of the form*

$$\underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[\ell(Yf)\right]\right]}_{\text{expected risk}} = \underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\ell(Yf^*)\right]}_{\text{noise + bias}} + \underbrace{\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[B_\ell(f, f^*)\right]\right]}_{\text{variance}} \tag{3}$$

if and only if ℓ *is gradient-symmetric.*

It is also possible to separate out the noise and bias (see Appendix D), but from a practical standpoint combining the two terms into a single expression is often more useful, since we do not usually know the distribution of the class labels and therefore do not have enough information to disentangle the contributions to the expected risk from the bias and the noise.

Finally, we note that it is possible to relax the condition that the loss is differentiable. If instead, we just require an injective mapping between points and their subgradients, we can extend the definitions of Bregman divergences and gradient-symmetry accordingly to retain the decompositions we have introduced.

3.2 Comparison to Buja's Decomposition

Buja et al. (2005) proposed a bias-variance decomposition for proper scoring rules, and also showed its application to margin losses via re-interpretation of the model as a probability estimator. This intermediate step requires a *link* function to perform the mapping, as well as the introduction of several other concepts. We have shown that gradient-symmetric losses do not require this intermediate step, instead having a bias-variance decomposition in terms of the raw model output, f. In order to compare the merits of the two decompositions, we now examine the properties of Buja et al. (2005)'s decomposition.

In order to present this decomposition, we must first introduce some new concepts and notation. Let $p = \mathbb{P}(Y = +1|\mathbf{X})$, i.e., p is the probability the target is $+1$ given $\mathbf{X}$, then $L(p, f)$, the expected pointwise risk of a model f when $\mathbf{X} = \mathbf{x}$, can be written

$$L(p, f) = \mathbb{E}_{Y|\mathbf{X}=\mathbf{x}}\ell(Yf) = p\ell(f) + (1 - p)\ell(-f).$$

We define the minimum risk, $\underline{L}(p)$ as the infimum of $L(p, f)$ for a fixed p, that is,

$$\underline{L}(p) = \inf_{v \in \mathbb{R}} L(p, v) = \inf_{v \in \mathbb{R}} p\ell(v) + (1 - p)\ell(-v).$$

We also define the optimal link function, $\psi(p)$ as mapping from $p \in (0, 1)$ to the $v \in \mathbb{R}$ at which the minimum is achieved, i.e.,

$$\psi(p) = \operatorname*{argmin}_{v \in \mathbb{R}} L(p, v) = \operatorname*{argmin}_{v \in \mathbb{R}} p\ell(v) + (1 - p)\ell(-v).$$

Name	Loss $\ell(v)$	Gradient $\ell'(v)$	Gradient-Symmetric?	Constant c
Squared	$(1-v)^2$	$2(v-1)$	✓	-4
Logistic	$\log(1+\exp(-v))$	$-\dfrac{1}{1+\exp(v)}$	✓	-1
Canonical Boosting Loss	$\frac{1}{2}\sqrt{v^2+4}-\frac{1}{2}v$	$\dfrac{v}{2\sqrt{v^2+4}}-\dfrac{1}{2}$	✓	-1
Laplacian Loss	$\frac{1}{2}\left(\exp(-|v|)+|v|-v\right)$	$-\frac{1}{2}+\frac{1}{2}\left(\mathrm{sgn}(v)(1-\exp(-|v|))\right)$	✓	-1
Exponential	$\exp(-v)$	$-\exp(-v)$	✗	N/A
Smooth Hinge	$\frac{1}{t}\ln(1+\exp(-t(v-1)))$	$-\dfrac{1}{1+\exp(t(v-1))}$	✗	N/A

Table 1: Examples of Margin Loss Functions and Their Derivatives

Assuming this function is invertible, we may think of the model f as implicitly defining a probability and write this probability as $q = \psi^{-1}(f)$.

Up until now, we have thought of f as being our model, with our class prediction being determined by the sign of f. If instead we think of our model as q, we can reframe the task as estimating the conditional probability of the positive class given the feature vector $\mathbf{x}$ (we could make the dependence of q on $\mathbf{x}$ and the training set $\mathcal{D}$ explicit by writing $q(\mathbf{x};\mathcal{D})$). Using the optimal link function ψ, we may now consider the loss $\ell(y\psi(q))$. This loss has the property of being a strictly proper scoring rule, i.e., $L(p,\psi(q))$ is minimized if and only if $q = p$. Here, Bregman divergences can be used to measure the separation between the desired probability estimate p and the one obtained from the model, q. This Bregman divergence measures the *excess risk*, i.e., the risk of a model minus the amount of irreducible error due to label noise.

Theorem 3. *((Zhang et al., 2004)) When ψ is invertible and differentiable, the pointwise excess risk can be written as a Bregman divergence of the form*

$$L(p,\psi(q)) - \underline{L}(p) = B_{-\underline{L}}(p,q). \qquad (4)$$

With this result, we are able to construct a bias-variance decomposition for the excess risk. This can be transformed into a bias-variance decomposition on the risk itself by noting that the minimum risk for losses with infimum zero losses is the same as the noise.

Theorem 4 (Buja's Bias-Variance Decomposition). *For loss ℓ with invertible differentiable link function ψ,*

$$\underbrace{\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[L(p,f)\right] - \underline{L}(p)\right]}_{\text{expected excess risk}} = \underbrace{\mathbb{E}_{\mathbf{X}}\left[B_{-\underline{L}}(p, q^*)\right]}_{\text{bias}}$$

$$+ \underbrace{\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[B_{-\underline{L}}(q^*,q)\right]\right]}_{\text{variance}}$$

where $q^* = [-\underline{L}']^{-1}\left(\mathbb{E}_D\left[-\underline{L}'(q)\right]\right)$.

A version of this decomposition was first noted in (Buja et al., 2005), though we formulate it here using notation and terminology from (Zhang et al., 2004) and (Reid and Williamson, 2010). Note that this decomposition can be applied for loss functions where the link function is invertible, including non-gradient-symmetric losses such as the exponential loss which minimized by AdaBoost (Friedman et al., 2000).

For the case of the logistic loss, the minimum risk is $-p\ln p - (1-p)\ln(1-p)$. This is the binary entropy, and in this case $B_{-\underline{L}}(\cdot,\cdot)$ becomes the KL-divergence. With this connection, we see that Buja's decomposition is in fact a special case of the decomposition for probability distributions discussed in (Heskes, 1998).

We will shortly examine how Theorem 4 connects to Theorem 2, but comparing it with the more general decomposition in Theorem 1 we find:

- Theorem 1 allows a more intuitive choice of "central" model $f^* = \mathbb{E}_D[f]$ which, as shown later, enables its use in a wider range of ensemble models, but its variance term is only target-independent under the gradient-symmetric condition.

- Although the variance term in Theorem 4 is target-independent, it requires a particular form for the "central" probability prediction q^*, which does not necessarily correspond to the prediction of f^*. In particular, q^* is defined to satisfy

$$q^* = \underset{\widehat{q}\in(0,1)}{\mathrm{argmin}}\, \mathbb{E}_D\left[B_{-\underline{L}}(\widehat{q},q)\right].$$

- These two models coincide (i.e., $f^* = \psi(q^*)$) if and only if ℓ is gradient-symmetric. In this case, the two decompositions are equivalent whenever

the conditions for both are met. This is stated more formally in Theorem 15 in Appendix E.1.

- Though Theorem 4 does not require gradient-symmetry, for some losses (such as the squared loss), the requirement for invertible ψ can necessitate restricting permitted model outputs. For example, in the case of the squared loss, $\psi(p) = 2p - 1$, so the range of ψ is $[-1, 1]$.

The strength of Buja's decomposition is that it turns the loss into a proper composite loss via the inverse link function (Reid and Williamson, 2010). These objects have since been studied in detail, both for binary classification and the multi-class case (Vernet et al., 2011). However, there are some cases, e.g., boosting, where though the models are excellent classifier, they do *not* give good probability estimates (Mease and Wyner, 2008). This makes it somewhat unnatural to apply Buja's decomposition in these cases.

3.3 Connections Between the Two Decompositions

This paper has discussed two bias-variance decompositions for gradient-symmetric losses; it is natural to ask how these two are related. We can show a deep connection between the two decompositions in Theorem 2 and Theorem 4 via the notion of *dual* Bregman divergences. Every Bregman divergence has an equivalent dual formulation defined by the convex conjugate of the generator (Banerjee et al., 2005). Let $[\phi]^*$ be the convex conjugate of the differentiable strictly convex function $\phi : S \to \mathbb{R}$, then for all $u, v \in \text{int}(S)$,

$$B_\phi(u, v) = B_{[\phi]^*}\left(\phi'(v), \phi'(u)\right).$$

Note that, for the dual formulation, we require *both* arguments be on the interior of their domain.

Given that we have two Bregman divergences of interest, B_ℓ and $B_{-\underline{L}}$, one may wonder whether they are related by this duality. Indeed they are, but the relationship is not quite as straightforward as saying that ℓ and $-\underline{L}$ are convex conjugates of each other; there is the presence of the additional scaling factor c. The actual relationship is described by the following result.

Theorem 5. *Let ℓ be a gradient-symmetric margin loss with negative minimum risk $-\underline{L}$, then*

$$[-\underline{L}]^*(v) = \ell\left(\frac{v}{c}\right).$$

Examples of this relationship can be observed in the last two columns of Table 2. With this, we can show that for gradient-symmetric losses, the two decompositions are linked by the following result.

Theorem 6. *For strictly convex invertible gradient-symmetric loss ℓ with infimum zero and with invertible $-\underline{L}'$, for all $u, v \in \mathbb{R}$,*

$$B_\ell(u, v) = B_{-\underline{L}}\left([-\underline{L}']^{-1}(cv), [-\underline{L}']^{-1}(cu)\right).$$

Hence the two decompositions are intimately linked by Bregman duality, even though the Bregman divergences are not strictly duals of each other. It is interesting to note that the connection could equally be derived via the *matching loss* approach of Helmbold et al. (1995). When conditions for both decompositions are met, i.e., for gradient-symmetric losses with a differentiable invertible link, *the variance terms in the two decompositions are exactly equivalent* (see Appendix E.2).

4 AMBIGUITY DECOMPOSITION FOR MARGIN LOSSES

It is a well-known phenomenon in machine learning that ensembles of estimators tend to outperform individual models. For ensembles of regression models, this notion is formalized by the *ambiguity decomposition* (Krogh and Vedelsby, 1995), which shows that, under the squared loss, the error of the average of distinct model outputs is guaranteed to be less than the average error of the individuals. More formally, for M model outputs, $f_1, \ldots, f_M$, combined by taking the arithmetic mean $\overline{f} = \frac{1}{M}\sum_{i=1}^M f_i$, for the label y,

$$\underbrace{(\overline{f} - y)^2}_{\text{ensemble error}} = \underbrace{\frac{1}{M}\sum_{i=1}^M (y - f_i)^2}_{\text{average error}} - \underbrace{\frac{1}{M}\sum_{i=1}^M (\overline{f} - f_i)^2}_{\text{ambiguity}}.$$

The ambiguity is a measure of ensemble diversity, independent of the target y. This decomposition can be thought of as a rearrangement of a special case of the bias-variance decomposition, taking the average over a finite set of models rather than expectation over D.

4.1 Deriving an Ambiguity Decomposition

We derive an ambiguity decomposition for margin losses following the same approach as Section 3.1, giving a more general decomposition in terms of the margin, then showing for gradient-symmetric losses, the ambiguity term becomes target-independent.

Theorem 7 (Margin Ambiguity Decomposition). *Let ℓ be a strictly convex differentiable margin loss and $\overline{f} = \frac{1}{M}\sum_{i=1}^M f_i$ be an ensemble of models, then*

$$\ell(y\overline{f}) = \underbrace{\frac{1}{M}\sum_{i=1}^M \ell(yf_i)}_{\text{average error}} - \underbrace{\frac{1}{M}\sum_{i=1}^M B_\ell\left(yf_i, y\overline{f}\right)}_{\text{margin ambiguity}}.$$

Name	Negative Minimum Risk $-\underline{L}(p)$	Conjugate Negative Minimum Risk $[-\underline{L}]^*(v)$	Loss function $\ell(v)$
Squared	$-4p(1-p)$	$\left(1 + \frac{v}{4}\right)^2$	$(1-v)^2$
Logistic	$p \ln p + (1-p) \ln(1-p)$	$\ln(1 + \exp(v))$	$\ln(1 + \exp(-v))$
Canonical Boosting Loss	$-2\sqrt{p(1-p)}$	$\frac{1}{2}\sqrt{p^2 + 4} + \frac{1}{2}p$	$\frac{1}{2}\sqrt{v^2 + 4} - \frac{1}{2}v$
Laplacian Loss	$-\frac{1}{2}(1 - \|2p - 1\|) \cdot (1 - \ln(1 - \|2p - 1\|))$	$\frac{1}{2}\left(\exp(-\|v\|) + \|v\| + v\right)$	$\frac{1}{2}\left(\exp(-\|v\|) + \|v\| - v\right)$

Table 2: Canonical Form Losses Along With Their Negative Minimum Risks and Corresponding Conjugates

Theorem 8 (Ambiguity Decomposition for Gradient-Symmetric Margin Losses). *With the same setup as in Theorem 7, we have the ambiguity decomposition*

$$\ell(y\overline{f}) = \underbrace{\frac{1}{M}\sum_{i=1}^{M}\ell(yf_i)}_{\text{average error}} - \underbrace{\frac{1}{M}\sum_{i=1}^{M}B_\ell\left(f_i, \overline{f}\right)}_{\text{ambiguity}}.$$

if and only if ℓ is gradient-symmetric.

This gives the existence of a simple ambiguity decomposition for the range of losses discussed in the previous section. Margin losses are frequently used with additive ensembles, like those constructed by LogitBoost, where ensemble members are combined by summation rather than averaging (Friedman et al., 2000). For these cases we can modify Theorem 8 by re-formulating the ensemble output as an average $f_{\text{add}} = \sum_{i=1}^{M} f_i = \frac{1}{M}\sum_{i=1}^{M} Mf_i$.

Corollary 1. *Let $f_1, \ldots, f_M$ be M model outputs and ensemble output $f_{\text{add}} = \sum_i f_i$, then*

$$\ell(yf_{\text{add}}) = \frac{1}{M}\sum_{i=1}^{M}\ell(yMf_i) - \frac{1}{M}\sum_{i=1}^{M}B_\ell\left(Mf_i,\ f_{\text{add}}\right).$$

The same trick generalizes to arbitrary weighted combination rules using $\sum_i \alpha_i f_i = \frac{1}{M}\sum_i \alpha_i M f_i$.

4.2 Ambiguity via Buja's Decomposition

One might expect that we can find an ambiguity decomposition for non-gradient-symmetric losses via proper scoring rules, in the same way that Theorem 4 gives a bias-variance decomposition for those losses. However, for linearly combined models, this is not the case. When $\overline{q}$ is the centroid combiner

$\overline{q} = [-\underline{L}']^{-1}\left(\frac{1}{M}\sum_{i=1}^{M} -\underline{L}'(q_i)\right)$, we have

$$B_{-\underline{L}}(p, \overline{q}) = \frac{1}{M}\sum_{i=1}^{M} B_{-\underline{L}}(p, q_i) - \frac{1}{M}\sum_{i=1}^{M} B_{-\underline{L}}(\overline{q}, q_i);$$

(5)

this has the form required for an ambiguity decomposition, but crucially, *only holds for the centroid combiner $\overline{q}$*. The decomposition would therefore require that $\overline{f} = \psi(\overline{q})$. Equivalently, from the definition of $\overline{q}$, we need

$$\overline{f} = \psi\left([-\underline{L}']^{-1}\left(\frac{1}{M}\sum_{i=1}^{M} -\underline{L}'(\psi^{-1}(f_i))\right)\right). \quad (6)$$

In order for this to be the case, we need a specific condition to hold, as described by the following theorem.

Theorem 9. *Let ℓ be a differentiable loss function with a continuous invertible link function, then*

$$\psi^{-1}(\overline{f}) = [-\underline{L}']^{-1}\frac{1}{M}\sum_{i=1}^{M}[-\underline{L}]'(\psi^{-1}(f_i)) \quad (7)$$

if and only if $\psi(p) = -a\underline{L}'(p)$ for some constant $a \neq 0$.

This condition turns out to be equivalent to gradient-symmetry, giving the following corollary.

Corollary 2. *Let $f_1, \ldots, f_M$ be an ensemble of models with ensemble output $\overline{f}$ a linear combination of the ensemble members. A target-independent ambiguity decomposition of form in (5) exists for linearly combined models if and only if ℓ is gradient-symmetric.*

When constructing new ensemble methods, (6) provides a natural combination rule for a given loss that guarantees that an ambiguity decomposition exists for a given ensemble. We leave exploring the merits of this rule in ensemble learning to future work.

5 CONNECTIONS TO EXISTING WORK

Gradient-symmetric losses are closely related to several other notable classes of functions. In this section, we highlight these connections and their implications:

- They are exactly the class of losses such that $\ell(yf)$ is a Bregman divergence where the first argument of the divergence is the raw model output f (or potentially the limit of such a divergence).

- They are a superset of canonical form losses (Masnadi-Shirazi and Vasconcelos, 2010).

- They are a subset of linear-odd losses (Patrini et al., 2016), for which we derive a novel decomposition of the expected risk.

5.1 Connection to Bregman Divergences

Proposition 1 explains the bias-variance decomposition for gradient-symmetric losses as a property of the Bregman divergence B_ℓ. Here, we show this is due to a deeper connection: for gradient-symmetric losses, the loss $\ell(yf)$ itself is expressible directly in terms of B_ℓ.

Theorem 10 (Gradient-Symmetric Losses as Bregman Divergences). *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a strictly-convex differentiable margin loss function with infimum zero, then $\ell(yf)$ is expressible as (the limit of) a Bregman divergence with f as the first argument if and only if ℓ is gradient-symmetric. Furthermore, if $\ell(yf)$ is expressible as such a Bregman divergence, it is of the form*

$$\ell(yf) = \lim_{g \to g_y} B_\ell(f, g). \tag{8}$$

where g_y is a constant determined by y.

For losses where $\lim_{v \to \infty} \ell(v) = 0$, $g_y \in \{-\infty, \infty\}$; otherwise, the values of g_y are finite and the $B_\ell(f, g_y)$ can be evaluated directly. Note that in the former case, $\lim_{v \to \infty} \ell'(v) = 0$, and therefore $\lim_{v \to \infty} \ell'(-v) = c$, so the gradient tends to a constant in both directions.

A consequence of Theorem 10 is that where ℓ appears in our decompositions, it can be replaced with an expression of the form in the right-hand side of (8). Doing so, we find our decomposition to be closely related to the decompositions for Bregman divergences shown in (Pfau, 2013), despite arising from a very different setting.

5.2 Connection to Canonical Form Losses

Canonical form losses (Masnadi-Shirazi and Vasconcelos, 2010) (also referred to as *permissible convex*

losses (Nock and Nielsen, 2008)) are a family of losses with a tight coupling between the minimum risk and the model's interpretation as a probability estimator. In particular, a loss function ℓ is said to be in canonical form if the derivative of the negative minimum risk function coincides with the link function, i.e, $-\underline{L}'(p) = \psi(p)$ for all $p \in (0, 1)$. This condition was shown in (Masnadi-Shirazi, 2011) to be equivalent to

$$\ell'(v) + \ell'(-v) = -1. \tag{9}$$

Comparing (2) and (9), gradient-symmetric loss functions can be seen to be a more general class of losses, since the value of c is not restricted to -1. However, they are not a great deal more general: every gradient-symmetric loss with $c < 0$ can be scaled by a constant factor to become a canonical loss. More formally:

Proposition 2. *A differentiable strictly convex margin loss ℓ is gradient-symmetric with constant $c < 0$ if and only if it satisfies*

$$\underline{L}'(p) = c\psi(p) \tag{10}$$

Furthermore, the loss $\frac{1}{|c|}\ell$ is a canonical form loss.

Canonical form losses are also the class of losses for which $B_\ell(u, v) = B_{[-\underline{L}]^*}(u, v)$, so in this case Bregman divergences *are* duals of each other, though *not* because ℓ are $-\underline{L}$ convex conjugates. Instead, $-\underline{L}^*(v) = \ell(-v)$ and the equivalence is due to gradient symmetry implying that $\ell(-v) = \ell(v) - v$ and Bregman generators differing by only affine terms forming equivalence classes.

5.3 Connection to Linear Odd Losses

The Linear Odd Losses (LOLs) are an important class of loss, introduced in (Patrini et al., 2016) and characterized by the fact that they factor into a linear term and a target-independent term. This structure makes these losses of particular interest in weakly supervised and semi-supervised learning settings.

While gradient-symmetric losses are a superset of canonical form losses, they are a *subset* of the LOLs, as described in (Patrini et al., 2016). As the name suggests, a LOL is a loss where the odd part is linear, as in the following definition.

Definition 2 (Linear Odd Loss). *Any loss function $\ell : \mathbb{R} \to \mathbb{R}_+$ can be written $\ell(v) = \ell_e(v) + \ell_o(v)$, where $\ell_e(v) = \frac{1}{2}(\ell(v) + \ell(-v))$ is the even part and the odd part is $\ell_o(v) = \frac{1}{2}(\ell(v) - \ell(-v))$. ℓ is a linear odd loss if*

$$\ell_o(v) = bv$$

for some constant b, i.e., $\ell(v)$ is of the form

$$\ell(v) = \ell_e(v) + bv.$$

Since $\ell_o(v) = \frac{1}{2}(\ell(v) - \ell(-v)) = bv$, when ℓ is differentiable, taking derivatives shows this condition to be equivalent to gradient-symmetry (with $b = \frac{c}{2}$). However, LOLs are not assumed to be differentiable (nor strictly convex), making them a more general class.

Given this relationship, a natural question emerges: can the bias-variance decomposition for strictly convex gradient-symmetric losses be extended to this class? This is an especially pertinent question since some losses based on the hinge loss are linear odd, but not gradient-symmetric nor strictly convex (van Rooyen et al., 2015; Plessis et al., 2015). We begin to answer this question with the following theorem, which decomposes the expected risk into the expected margin and a target-independent component.

Theorem 11 (Decomposition for the LOLs)**.** *Let ℓ be a LOL with odd part $\ell_o(v) = bv$ and even part $\ell_e(v) = \frac{1}{2}(\ell(v) + \ell(-v))$, then the expected risk decomposes as*

$$\mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[\ell(Yf)\right]\right] = \mathbb{E}_\mathbf{X}\left[bY^*f^*\right] + \mathbb{E}_\mathbf{X}\left[\mathbb{E}_D\left[\ell_e(f)\right]\right],$$

where $Y^* = \mathbb{E}_{Y|\mathbf{X}}[Y]$ and $f^* = \mathbb{E}_D[f]$.

This reveals that for LOLs (including gradient-symmetric losses) the expected risk can be decomposed into the expected margin and a function dependent only on the distribution of f. This decomposition, an extension of the decomposition of the risk presented in (Patrini et al., 2016), is not a bias-variance decomposition in a conventional sense: since Y^*f^* can take any value in $\mathbb{R}$, there is no guarantee that the contribution of the first term is non-negative.

We can relate this decomposition to the bias-variance decomposition by noting that, for gradient-symmetric losses, the two terms can be written

$$\underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\ell(Yf^*)\right]}_{\text{bias + noise}} = \mathbb{E}_\mathbf{X}\left[bY^*f^* + \ell_e(f^*)\right] \qquad (11)$$

and

$$\underbrace{\mathbb{E}_\mathbf{X}\left[\mathbb{E}_D\left[B_\ell\left(f, f^*\right)\right]\right]}_{\text{variance}} = \mathbb{E}_\mathbf{X}\left[\mathbb{E}_D\left[\ell_e(f)\right] - \ell_e(f^*)\right]. \quad (12)$$

The first, (11), tells us that the expected risk is influenced by the target only via the expected margin. This simple form for the contribution of the target-dependent part of the risk is not dependent on the exact form of ℓ other than through the value of the constant b.

In (12), we see that for gradient symmetric losses, the variance is the the difference between the two sides of Jensen's inequality (the Jensen gap). This relationship highlights the importance of convexity; it is guaranteed that (12) is non-negative if and only ℓ_e (and by

extension ℓ) is convex. We could use (11) and (12) as a bias-variance decomposition of sorts for the LOLs. However, when ℓ is is not strictly convex, we lose the guarantee that the variance term is non-zero for non-constant f, and when ℓ is non-convex, the variance can be negative.

Finally, we note that the same exhaustive process for constructing LOLs in (Patrini et al., 2016) can be easily modified to give new gradient-symmetric losses: For any differentiable even function ℓ_e, we can construct a gradient-symmetric loss as $\ell_e(v) - \frac{c}{2}v$ for any constant c.

6 CONCLUSION

We have presented a bias-variance decomposition for a broad class of *gradient-symmetric* margin losses. Unlike previous work, our decomposition neither requires interpretation of the model as a class probability estimator nor computation of the inverse link or minimum risk functions. From this bias-variance decomposition, we have shown more general decompositions applicable to non-gradient-symmetric losses, giving decompositions applicable to the more general classes of strictly convex losses and LOLs. While we have not derived a true bias-variance decomposition for the LOLs, we have shown that it is possible to separate out the target-dependent component of the expected risk.

The framework that we have developed opens up many avenues for exploration of bias and variance in practical settings, including examination of the longstanding question of whether boosting algorithms are bias or variance reducing (Mease and Wyner, 2008), development of bias-variance decompositions for SVMs with the L2 loss (Lee and Lin, 2013), and possible application of bias-variance decompositions for alternative learning scenarios, such as positive/unlabelled classification (Plessis et al., 2015) and learning from label proportions (Quadrianto et al., 2009; Patrini et al., 2016).

References

Adlam, B. and Pennington, J. (2020). Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2020.*

Banerjee, A., Merugu, S., Dhillon, I. S., Ghosh, J.,

and Lafferty, J. (2005). Clustering with bregman divergences. *Journal of machine learning research*, 6(10).

Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854.

Bregman, L. M. (1967). 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.

Buja, A., Stuetzle, W., and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. *Working Draft*.

Bullen, P. S. (2003). *Handbook of means and their inequalities*. Springer Science & Business Media.

copper.hat (2020). Is the arg min of a strictly convex function continuous? Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/3748031 (version: 2020-07-07).

Freund, Y. and Schapire, R. E. (1999). A short introduction to boosting. In *In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, pages 1401–1406. Morgan Kaufmann.

Friedman, J., Hastie, T., Tibshirani, R., et al. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *Annals of statistics*, 28(2):337–407.

Geman, S., Bienenstock, E., and Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58.

Hansen, J. V. and Heskes, T. (2000). General bias/variance decomposition with target independent variance of error functions derived from the exponential family of distributions. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 2, pages 207–210. IEEE.

Helmbold, D., Kivinen, J., and Warmuth, M. K. K. (1995). Worst-case loss bounds for single neurons. In Touretzky, D., Mozer, M. C., and Hasselmo, M., editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press.

Heskes, T. (1998). Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10(6):1425–1433.

Krogh, A. and Vedelsby, J. (1995). Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems 7*, 7:231.

Lee, C.-P. and Lin, C.-J. (2013). A study on l2-loss (squared hinge-loss) multiclass svm. *Neural computation*, 25(5):1302–1323.

Masnadi-Shirazi, H. (2011). *The design of Bayes consistent loss functions for classification*. PhD thesis, UC San Diego.

Masnadi-Shirazi, H. and Vasconcelos, N. (2009). On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Masnadi-Shirazi, H. and Vasconcelos, N. (2010). Variable margin losses for classifier design. *Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, NIPS 2010*, pages 1–9.

Mease, D. and Wyner, A. J. (2008). Evidence contrary to the statistical view of boosting. *Journal of Machine Learning Research*, 9:131.

Neal, B., Mittal, S., Baratin, A., Tantia, V., Scicluna, M., Lacoste-Julien, S., and Mitliagkas, I. (2018). A modern take on the bias-variance tradeoff in neural networks. *CoRR*, abs/1810.08591.

Nielsen, F. (2010). Legendre transformation and information geometry.

Nock, R. and Nielsen, F. (2008). Bregman divergences and surrogates for learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2048–2059.

Patrini, G., Nielsen, F., Nock, R., and Carioni, M. (2016). Loss factorization, weakly supervised learning and label noise robustness. In *International conference on machine learning*, pages 708–717. PMLR.

Pfau, D. (2013). A Generalized Bias-Variance Decomposition for Bregman Divergences. Technical report, Columbia University.

Plessis, M. D., Niu, G., and Sugiyama, M. (2015). Convex formulation for learning from positive and unlabeled data. In Bach, F. and Blei, D., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1386–1394, Lille, France. PMLR.

Quadrianto, N., Smola, A. J., Caetano, T. S., and Le, Q. V. (2009). Estimating labels from label proportions. *Journal of Machine Learning Research*, 10(82):2349–2374.

Reid, M. D. and Williamson, R. C. (2010). Composite binary losses. *The Journal of Machine Learning Research*, 11:2387–2422.

S., N. (2011). Proving that $\lim_{x \to \infty} f(x) = 0$ when $\lim_{x \to \infty} f(x)$ and $\lim_{x \to \infty} f(x)$ exist. Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/42356 (version: 2011-05-31).

Schapire, R., Freund, Y., Barlett, P., and Lee, W. (1998). Boosting the margin: a new explanation for the effectiveness of voting methods. *The Annals of Statistics*, 26(5):1651 – 1686.

user61527 (2013). f is monotone and the integral is bounded. prove that lim x xf(x)=0. Mathematics Stack Exchange. URL:https://math.stackexchange.com/q/560909 (version: 2013-11-12).

van Rooyen, B., Menon, A., and Williamson, R. C. (2015). Learning with symmetric label noise: The importance of being unhinged. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Vernet, E., Reid, M. D., and Williamson, R. C. (2011). Composite multiclass losses. *Advances in Neural Information Processing Systems*, 24.

Yang, Z., Yu, Y., You, C., Steinhardt, J., and Ma, Y. (2020). Rethinking bias-variance trade-off for generalization of neural networks. In *International Conference on Machine Learning*, pages 10767–10777. PMLR.

Zhang, T. et al. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85.

Supplementary Material: Bias-Variance Decompositions for Margin Losses

A Proofs for Section 3.1

Theorem 1 (Margin Variance Decomposition). *For a strictly convex differentiable margin loss ℓ,*

$$\underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[\ell(Yf)\right]\right]}_{\text{expected risk}} = \underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\ell(Yf^*)\right]}_{\text{risk of central model}} + \\ \underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[B_\ell\left(Yf, Yf^*\right)\right]\right]}_{\text{margin variance}}, \tag{1}$$

where we define the central model $f^* = \mathbb{E}_D\left[f\right]$.

Proof. Define the random variable $V = Yf(\mathbf{X}; D) = Yf$, then by the definition of B_ℓ,

$$\begin{aligned}
\mathbb{E}_D\left[B_\ell\left(Yf, Yf^*\right)\right] &= \mathbb{E}_D\left[B_\ell\left(V, \mathbb{E}_D\left[V\right]\right)\right] \\
&= \mathbb{E}_D\left[\ell(V) - \ell(\mathbb{E}_D\left[V\right]) - \ell'(\mathbb{E}_D\left[V\right])(V - \mathbb{E}_D\left[V\right])\right] \\
&= \mathbb{E}_D\left[\ell(V)\right] - \ell(\mathbb{E}_D\left[V\right]) - \ell'(\mathbb{E}_D\left[V\right])(\mathbb{E}_D\left[V\right] - \mathbb{E}_D\left[V\right]) \\
&= \mathbb{E}_D\left[\ell(V)\right] - \ell(\mathbb{E}_D\left[V\right]).
\end{aligned}$$

Rearranging gives

$$\mathbb{E}_D\left[\ell(V)\right] = \ell(\mathbb{E}_D\left[V\right]) + \mathbb{E}_D\left[B_\ell\left(V, \mathbb{E}_D\left[V\right]\right)\right].$$

Then, expanding V from its definition,

$$\begin{aligned}
\mathbb{E}_D\left[\ell(Yf)\right] &= \ell(\mathbb{E}_D\left[Yf\right]) + \mathbb{E}_D\left[B_\ell\left(Yf, \mathbb{E}_D\left[Yf\right]\right)\right] \\
\Rightarrow \mathbb{E}_D\left[\ell(Yf)\right] &= \ell(Yf^*) + \mathbb{E}_D\left[B_\ell\left(Yf, Yf^*\right)\right].
\end{aligned}$$

Finally, taking the expectation with respect to $\mathbf{X}$ and Y completes the proof. $\square$

Proposition 1. *For all $u, v \in \mathbb{R}$ and strictly convex differentiable $\ell : \mathbb{R} \to \mathbb{R}_+$,*

$$B_\ell\left(u, v\right) = B_\ell\left(-u, -v\right)$$

if and only if ℓ is gradient-symmetric.

Proof. ($\Longrightarrow$) Assume that for all $u, v \in \mathbb{R}$, we have $B_\ell\left(u, v\right) = B_\ell\left(-u, -v\right)$, then by the expanding both sides using the definition of Bregman divergences,

$$\ell(u) - \ell(v) - \ell'(v)(u - v) = \ell(-u) - \ell(-v) - \ell'(-v)(-u + v).$$

Taking the derivative of both sides with respect to u gives

$$\begin{aligned}
\ell'(u) - \ell'(v) &= -\ell'(-u) + \ell'(-v) \\
\Longrightarrow \ell'(u) + \ell'(-u) &= \ell'(-v) + \ell'(v). \tag{13}
\end{aligned}$$

Since (13) is true for all $u, v \in \mathbb{R}$, $\ell'(u) + \ell'(-u)$ must be constant.

($\Longleftarrow$) We assume ℓ is gradient-symmetric, i.e., $\ell'(v) + \ell'(-v) = c$ for some $c \in \mathbb{R}$. We can take the anti-derivatives of both sides to get $\ell(v) - \ell(-v) = cv$ (note that the constant of integration is necessarily zero, since at $v = 0$, $\ell(v) - \ell(-v) = 0$). Considering $B_\ell(u, v)$,

$$
\begin{aligned}
B_\ell(u, v) &= \ell(u) - \ell(v) - \ell'(v)(u - v) \\
&= \ell(u) - \ell(v) - \underbrace{(c - \ell'(-v))}_{=\ell'(v)}(u - v) \\
&= \ell(u) - \ell(v) - cu + cv + \ell'(-v)(u - v) \\
&= \underbrace{\ell(-u) + cu}_{=\ell(u)} - \underbrace{(\ell(-v) + cv)}_{=\ell(v)} - cu + cv + \ell'(-v)(u - v) \\
&= \ell(-u) + cu - \ell(-v) - cv - cu + cv - \ell'(-v)(-u - (-v)) \\
&= \ell(-u) - \ell(-v) - \ell'(-v)(-u - (-v)) \\
&= B_\ell(-u, -v).
\end{aligned}
$$

$\square$

Theorem 2 (Bias-Variance Decomposition for Gradient-Symmetric Losses). *For a strictly convex differentiable margin loss ℓ, there exists a bias-variance decomposition of the form*

$$
\underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[\ell(Yf)\right]\right]}_{\text{expected risk}} = \underbrace{\mathbb{E}_{\mathbf{X}Y}\left[\ell(Yf^*)\right]}_{\text{noise + bias}} + \underbrace{\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[B_\ell(f, f^*)\right]\right]}_{\text{variance}} \tag{3}
$$

if and only if ℓ is gradient-symmetric.

Proof. ($\Longleftarrow$) If ℓ is gradient-symmetric, then by Proposition 1, $B_\ell(u, v) = B_\ell(-u, -v)$ for all $u, v \in \mathbb{R}$, and the sign of Y is irrelevant in $B_\ell(Yf, Yf^*)$, therefore the variance term in (1) can be expressed as $\mathbb{E}_D\left[B_\ell(f, f^*)\right]$ and the decomposition in (3) holds.

($\Longrightarrow$) We will show that if the decomposition holds for all choices of distributions of $\mathbf{X}$, Y and D, then for all choices of $u, v \in \mathbb{R}$, $\ell'(u) + \ell'(-u) = \ell'(v) + \ell'(v)$ and therefore ℓ is gradient-symmetric.

Let Y be the constant random variable -1 and $\mathbf{X} = \mathbf{x}$ for some constant $\mathbf{x}$. For any $u, w \in \mathbb{R}$, we can define a distribution such that $\mathbb{P}(f(X, D) = u) = \mathbb{P}(f(X, D) = w) = \frac{1}{2}$. For this distribution, we have $f^* = \frac{1}{2}(u + w)$.

Assuming the decomposition in (3) holds, then by it and (1), it must be true that

$$
\mathbb{E}_D\left[B_\ell(f, f^*)\right] = \mathbb{E}_D\left[B_\ell(-f, -f^*)\right].
$$

Expanding the expectations,

$$
\begin{aligned}
\frac{1}{2}\left(B_\ell\left(u, \frac{1}{2}(u + w)\right) + B_\ell\left(w, \frac{1}{2}(u + w)\right)\right) &= \frac{1}{2}\left(B_\ell\left(-u, -\frac{1}{2}(u + w)\right) + B_\ell\left(-w, -\frac{1}{2}(u + w)\right)\right) \\
\Longrightarrow B_\ell\left(u, \frac{1}{2}(u + w)\right) + B_\ell\left(w, \frac{1}{2}(u + w)\right) &= B_\ell\left(-u, -\frac{1}{2}(u + w)\right) + B_\ell\left(-w, -\frac{1}{2}(u + w)\right)
\end{aligned}
$$

For a fixed w, the above is true for all u (after making the appropriate changes to the distribution of D). For this to be the case, the derivatives of both sides with respect to u must be the same, i.e,

$$
\frac{\partial}{\partial u}\left[B_\ell\left(u, \frac{1}{2}(u + w)\right) + B_\ell\left(w, \frac{1}{2}(u + w)\right)\right] = \frac{\partial}{\partial u}\left[B_\ell\left(-u, -\frac{1}{2}(u + w)\right) + B_\ell\left(-w, -\frac{1}{2}(u + w)\right)\right]
$$

We now proceed to simplify both sides. Starting with the left-hand side, we find

$$\frac{\partial}{\partial u}\left[B_\ell\left(u, \frac{1}{2}(u+w)\right) + B_\ell\left(w, \frac{1}{2}(u+w)\right)\right] \tag{14}$$

$$= \frac{\partial}{\partial u}\Big[\ell(u) - \ell(\tfrac{1}{2}(u+w)) - \ell'(\tfrac{1}{2}(u+w))\Big(u - \tfrac{1}{2}(u+w)\Big)$$

$$+ \ell(w) - \ell(\tfrac{1}{2}(u+w)) - \ell'(\tfrac{1}{2}(u+w))\Big(w - \tfrac{1}{2}(u+w)\Big)\Big]$$

$$= \frac{\partial}{\partial u}\Big[\ell(u) - \ell(\tfrac{1}{2}(u+w)) + \ell(w) - \ell(\tfrac{1}{2}(u+w))$$

$$\underbrace{- \ell'(\tfrac{1}{2}(u+w))\Big(u - \tfrac{1}{2}(u+w)\Big) - \ell'(\tfrac{1}{2}(u+w))\Big(w - \tfrac{1}{2}(u+w)\Big)}_{=0}\Big]$$

$$= \frac{\partial}{\partial u}\left[\ell(u) - 2\ell(\tfrac{1}{2}(u+w))\right]$$

$$= \ell'(u) - \ell'(\tfrac{1}{2}(u+w)). \tag{15}$$

Similarly, for the right-hand side we find

$$\frac{\partial}{\partial u}\left[B_\ell\left(-u, -\frac{1}{2}(u+w)\right) + B_\ell\left(-w, -\frac{1}{2}(u+w)\right)\right] = -\ell'(-u) + \ell'(-\tfrac{1}{2}(u+w))$$

Combining the two, we get that for all $u \in \mathbb{R}$,

$$\ell'(u) - \ell'(\tfrac{1}{2}(u+w)) = -\ell'(-u) + \ell'(-\tfrac{1}{2}(u+w)).$$

Or equivalently

$$\ell'(u) + \ell'(-u) = \ell'(\tfrac{1}{2}(u+w)) + \ell'(-\tfrac{1}{2}(u+w)).$$

This argument holds for all w, so for any $v \in \mathbb{R}$, setting $w = 2v - u$ gives $v = \frac{1}{2}(w+u)$ and therefore for all u and v,

$$\ell'(u) + \ell'(-u) = \ell'(v) + \ell'(-v).$$

This means that $\ell'(f) + \ell'(-f)$ is constant and therefore ℓ is gradient-symmetric. $\square$

A.1 Proof of Buja's Decomposition

In order to prove Theorem 4, we first require the following lemma, which is a result from (Pfau, 2013).

Lemma 1. *For any strictly convex differentiable function ϕ,*

$$\mathbb{E}_D\left[B_\phi(p,q)\right] = B_\phi(p, q^*) + \mathbb{E}_D\left[B_\phi(q^*, q)\right],$$

where $q^* = [\phi']^{-1}\left(\mathbb{E}_D\left[\phi'(q)\right]\right)$

With this result, we proceed to proof of the theorem.

Theorem 4 (Buja's Bias-Variance Decomposition). *For loss ℓ with invertible differentiable link function ψ,*

$$\underbrace{\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[L(p,f)\right] - \underline{L}(p)\right]}_{\text{expected excess risk}} = \underbrace{\mathbb{E}_{\mathbf{X}}\left[B_{-\underline{L}}\left(p, q^*\right)\right]}_{\text{bias}}$$

$$+ \underbrace{\mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[B_{-\underline{L}}\left(q^*, q\right)\right]\right]}_{\text{variance}}$$

where $q^* = [-\underline{L}']^{-1}\left(\mathbb{E}_D\left[-\underline{L}'(q)\right]\right)$.

Proof. By Theorem 3,

$$L(p, f) = \underline{L}(p) + B_{-\underline{L}}(p, q).$$

So, by application of Lemma 1, we have

$$\mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[L(p, f)\right]\right] = \mathbb{E}_{\mathbf{X}}\left[\underline{L}(p)\right] + \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[B_{-\underline{L}}(p, q)\right]\right]$$
$$= \mathbb{E}_{\mathbf{X}}\left[\underline{L}(p)\right] + \mathbb{E}_{\mathbf{X}}\left[B_{-\underline{L}}(p, q^*) + \mathbb{E}_D\left[B_{-\underline{L}}(q^*, q)\right]\right].$$

By rearrangement, we have

$$\mathbb{E}_{\mathbf{X}}\left[L(p, f) - \underline{L}(p)\right] = \mathbb{E}_{\mathbf{X}}\left[B_{-\underline{L}}(p, q^*)\right] + \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[B_{-\underline{L}}(q^*, q)\right]\right].$$

$\square$

Theorem 5. *Let ℓ be a gradient-symmetric margin loss with negative minimum risk $-\underline{L}$, then*

$$[-\underline{L}]^*(v) = \ell\left(\frac{v}{c}\right).$$

Proof. Note that for a function $\phi(p)$, we can write the convex conjugate as $[\phi]^*(v) = v[\phi']^{-1}(v) - \phi([\phi']^{-1}(v))$ (see, for instance (Nielsen, 2010)). We can therefore write the convex conjugate of the negative minimum risk as

$$(-L)^*(v) = v[-\underline{L}']^{-1}(v) + \underline{L}\left([-\underline{L}']^{-1}(v)\right).$$

Since by Proposition 2, $[-\underline{L}]'(p) = -c\psi(p)$, we have $[-\underline{L}']^{-1}(v) = \psi^{-1}\left(-\frac{v}{c}\right)$. Additionally, we remind the reader that, by definition of the minimum risk and optimal link functions, $\underline{L}(p) = p\ell(\psi(p)) + (1-p)\ell(-\psi(p))$. Putting this together,

$$[-L]^*(v) = v[-\underline{L}']^{-1}(v) + \underline{L}\left([-\underline{L}']^{-1}(v)\right)$$
$$= v\,\psi^{-1}\left(-\frac{v}{c}\right) + \underline{L}\left(\psi^{-1}\left(-\frac{v}{c}\right)\right)$$
$$= v\,\psi^{-1}\left(-\frac{v}{c}\right) + \psi^{-1}\left(-\frac{v}{c}\right)\ell\left(-\frac{v}{c}\right) + \left(1 - \psi^{-1}\left(-\frac{v}{c}\right)\right)\ell\left(\frac{v}{c}\right)$$
$$= \psi^{-1}\left(-\frac{v}{c}\right)\left(v + \ell\left(-\frac{v}{c}\right) - \ell\left(\frac{v}{c}\right)\right) + \ell\left(\frac{v}{c}\right). \tag{16}$$

We claim that the first term in the final expression is identically zero. Note that a necessary and sufficient condition for $v + \ell(-v/c) - \ell(v/c) = 0$ for all $v \in \mathbb{R}$ is that $cv + \ell(-v) - \ell(v) = 0$ for all $v \in \mathbb{R}$. The latter rearranges to $\ell(v) = \ell(-v) + cv$. For differentiable ℓ, this is equivalent to the condition for gradient-symmetric, since

$$\ell'(v) + \ell'(-v) = c$$
$$\iff \ell(v) - \ell(-v) = cv + d,$$

where d can be shown to be zero by considering that $\ell(0) - \ell(-0) = 0 = 0c + d$. Therefore, by (16),

$$[-\underline{L}]^*(v) = \ell\left(\frac{v}{c}\right).$$

$\square$

Lemma 2. *For any Bregman generator ϕ, and any $c \neq 0$, if we define ϕ_c such that for all v, $\phi_c(v) = \phi(\frac{v}{c})$, then for all $u, v \in \mathbb{R}$,*

$$B_{\phi_c}(u, v) = B_\phi\left(\frac{u}{c}, \frac{v}{c}\right).$$

Proof.

$$B_{\phi_c}(u,v) = \phi_c(u) - \phi_c(v) - \phi'_c(v)(u-v)$$
$$= \phi(u/c) - \phi(v/c) - \frac{\phi'(v/c)}{c}(u-v)$$
$$= \phi(u/c) - \phi(v/c) - \phi'(v/c)(u/c - v/c)$$
$$= B_{\phi}\left(\frac{u}{c}, \frac{v}{c}\right).$$

□

Theorem 6. *For strictly convex invertible gradient-symmetric loss ℓ with infimum zero and with invertible $-\underline{L}'$, for all $u, v \in \mathbb{R}$,*

$$B_{\ell}(u,v) = B_{-\underline{L}}\left([-\underline{L}']^{-1}(cv), [-\underline{L}']^{-1}(cu)\right).$$

Proof. We define $\ell_c(v) = \ell(\frac{v}{c})$ and note that by Theorem 5 $\ell_c(v) = \ell(\frac{v}{c}) = [-\underline{L}]^*(v)$, so

$$B_{\ell}(u,v) = B_{\ell}\left(\frac{cu}{c}, \frac{cv}{c}\right)$$
$$= B_{\ell_c}(cu, cv)$$
$$= B_{[-\underline{L}]^*}(cv, cu)$$
$$= B_{-\underline{L}}\left([-L']^{-1}(cv), [-L']^{-1}(cu)\right)$$

□

B   Proofs for Section 4.1

Theorem 7 (Margin Ambiguity Decomposition). *Let ℓ be a strictly convex differentiable margin loss and $\overline{f} = \frac{1}{M}\sum_{i=1}^{M} f_i$ be an ensemble of models, then*

$$\ell(y\overline{f}) = \underbrace{\frac{1}{M}\sum_{i=1}^{M}\ell(yf_i)}_{\text{average error}} - \underbrace{\frac{1}{M}\sum_{i=1}^{M}B_{\ell}\left(yf_i, y\overline{f}\right)}_{\text{margin ambiguity}}.$$

Proof. The ambiguity decomposition can be derived as a special case of the bias-variance decomposition. We consider an ensemble of models $f_1, \ldots, f_M$, constant $X = \mathbf{x}$ and constant $Y = y$. If we choose D such that $\mathbb{P}(f = f_i) = \frac{1}{M}$, then Theorem 1 gives

$$\mathbb{E}_D\left[\ell(Yf)\right] = \ell(Yf^*) + \mathbb{E}_D\left[B_{\ell}(f, f^*)\right]$$
$$\implies \frac{1}{M}\sum_{i=1}^{M}\ell(yf_i) = \ell(y\overline{f}) + \frac{1}{M}\sum_{i=1}^{M}B_{\ell}\left(f_i, \overline{f}\right).$$

Rearranging completes the proof.

□

Theorem 8 (Ambiguity Decomposition for Gradient-Symmetric Margin Losses). *With the same setup as in Theorem 7, we have the ambiguity decomposition*

$$\ell(y\overline{f}) = \underbrace{\frac{1}{M}\sum_{i=1}^{M}\ell(yf_i)}_{\text{average error}} - \underbrace{\frac{1}{M}\sum_{i=1}^{M}B_{\ell}\left(f_i, \overline{f}\right)}_{\text{ambiguity}}.$$

if and only if ℓ is gradient-symmetric.

Proof. ($\impliedby$) Since the ambiguity decomposition can be formulated as a special case of the bias-variance decomposition and gradient-symmetry implies that the bias-variance decomposition, gradient-symmetry also implies the ambiguity decomposition.

($\implies$) The proof of this direction is identical to the proof of the same direction in Theorem 2, except we define $\widehat{f}$ and h as the members of an ensemble with $M = 2$, rather than using them to define a probability distribution. We therefore omit the full details here. $\qquad\square$

In Section 3.3, we claim that the ambiguity decomposition for linearly combined models only exists for gradient-symmetric margin losses since (6) holds if and only if the derivative of the negative minimum risk is an scaling of the optimal link function. Here, we explain in more detail why this is the case. The key observation here will be about when the ensemble average $\frac{1}{M}\sum_{i=1}^{M} f_i$ coincides with inverse link of the central probability prediction $\overline{q} = [-L]'^{-1}\left([-L'](\psi^{-1}(f_i))\right)$. To examine this, we first need the following result about quasi-arithmetic means.[1]

Theorem 12 (Adapted from Bullen (2003), p.271, Sec. 4.1.2, Theorem 5). *Let $\mathcal{S}$ be a closed interval on the extended real line and let $\kappa : \mathcal{S} \to \mathbb{R} \cup \{-\infty, \infty\}$ and $\lambda : \mathcal{S} \to \mathbb{R} \cup \{-\infty, \infty\}$ be strictly monotonic continuous functions. Then*

$$\kappa^{-1}\left(\frac{1}{M}\sum_{i=1}^{M}\kappa(v_i)\right) = \lambda^{-1}\left(\frac{1}{M}\sum_{i=1}^{M}\lambda(v_i)\right),$$

for all $v_1, \dots, v_M \in \mathcal{S}$, if and only if $\kappa(v) = a\lambda(v) + b$, for some $a, b \in \mathbb{R}$, $a \neq 0$

We note that in Bullen (2003), the theorem is for *weighted* quasi-arithmetic means, however, the proof of Bullen (2003) still holds in both directions for the unweighted case.

Applying this result to our ambiguity decomposition, we will find that this necessarily means that the optimal link is a scaled version of $-\underline{L}'$, but first we require the following lemma.

Lemma 3. *If $\psi(p) = -a\underline{L}'(p) + b$ for some constants $a, b \in \mathbb{R}$, then $b = 0$.*

Proof. Note that from (Masnadi-Shirazi, 2011) we have that, for all $p \in (0, 1)$,

$$\psi(p) = -\psi(1 - p)$$

and since $\underline{L}(p) = \underline{L}(1 - p)$ ((Masnadi-Shirazi, 2011)), for all $p \in (0, 1)$,

$$-\underline{L}'(p) = \underline{L}'(1 - p)$$

Using these two facts, we define the linear function such that $l(p) = ap + b$, so that $\psi = \ell \circ [-\underline{L}]'$ and note that $-\underline{L}' = l^{-1} \circ \psi$. The inverse of ℓ can be written $l^{-1}(v) = \frac{1}{a}v + \frac{b}{a}$. This gives

$$
\begin{aligned}
-\underline{L}'(p) &= l^{-1} \circ \psi(p) \\
&= \frac{1}{a}\psi(p) + \frac{b}{a} \\
&= -\frac{1}{a}\psi(1 - p) + \frac{b}{a},
\end{aligned}
$$

but also

$$
\begin{aligned}
-\underline{L}'(p) &= \underline{L}'(1 - p) \\
&= -l^{-1} \circ \psi(1 - p) \\
&= -\frac{1}{a}\psi(1 - p) - \frac{b}{a}.
\end{aligned}
$$

The only way these two constraints are met is if $b = 0$. $\qquad\square$

---

[1]Quasi-arithmetic means are combiner functions of the form $\kappa^{-1}\left(\frac{1}{M}\sum_{i=1}^{M}\kappa(v_i)\right)$, where κ is a continuous monotonic function.

This lemma, combined with the previous theorem, gives the following result.

Theorem 9. *Let ℓ be a differentiable loss function with a continuous invertible link function, then*

$$\psi^{-1}(\overline{f}) = [-\underline{L}']^{-1} \frac{1}{M} \sum_{i=1}^{M} [-\underline{L}]'(\psi^{-1}(f_i)) \tag{7}$$

if and only if $\psi(p) = -a\underline{L}'(p)$ for some constant $a \neq 0$.

Proof.

$$\psi^{-1}(\overline{f}) = [-\underline{L}']^{-1} \frac{1}{M} \sum_{i=1}^{M} [-\underline{L}]'(\psi^{-1}(f_i)), \tag{17}$$

is equivalent to

$$\overline{f} = \psi\left([-\underline{L}']^{-1} \frac{1}{M} \sum_{i=1}^{M} [-\underline{L}]'(\psi^{-1}(f_i))\right). \tag{18}$$

Because ψ and $-\underline{L}'$ are both continuous and invertible,[2] the function $\psi \circ [-\underline{L}']^{-1}$ is continuous and invertible, and therefore it is a quasi-arithmetic mean and we can apply Theorem 12 to get

$$\psi\left([-\underline{L}']^{-1}(f)\right) = af + b,$$

for some constants a, b with $a \neq 0$, and by Lemma 3, $b = 0$. Defining $p = [-L']^{-1}(f)$, this gives

$$\psi(p) = a[-\underline{L}]'(p) = -a\underline{L}'(p).$$

$\square$

C   Proofs for Section 5

Before presenting proof of Theorem 10, we prove some basic facts about convex functions and limits.

Lemma 4. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a strictly convex differentiable function with $\lim_{v \to \infty} \phi(v) = c$ for some constant c. Then*

$$\lim_{v \to \infty} \phi'(v) = 0$$

Proof. Since ϕ' is monotonically increasing, it either diverges to infinity or has the upper limit $\lim_{v \to \infty} \phi'(v) = \inf(\{c \in \mathbb{R} \; : \; c > \phi'(v), \quad \forall v \in \mathbb{R}\})$. If it were the case that $\lim_{v \to \infty} \phi'(v) = \infty$ then $\lim_{v \to \infty} \phi(v) = \infty$, this contradicts our assumption, so $\lim_{v \to \infty} \phi'(v)$ must be finite. It can then be shown that if a limit exists, then it must be zero (see, for instance S. (2011)). $\square$

We also require the following lemma, the proof of which is adapted from user61527 (2013).

Lemma 5. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a differentiable convex function with $\lim_{v \to \infty} \phi(v) = 0$, then $\lim_{v \to \infty} \phi'(v)v = 0$*

Proof. We begin by noting two facts. Firstly, since ϕ is strictly convex and tending towards a finite limit, it must be strictly decreasing. Secondly, since $\lim_{v \to \infty} \phi(v) = 0$, its derivative satisfies $\lim_{N \to \infty} \int_0^N \phi'(v)dv = \lim_{N \to \infty} \phi(N) - \phi(0) = -\phi(0)$, so the derivative ϕ' has a finite integral over the positive reals.

With these facts, we prove the contrapositive of the statement in our theorem. That is, we will show that if $v\phi'(v)$ does not converge to zero (which means there must exist $\epsilon > 0$ such that for any v_0 there is always $v > v_0$ such that $|v\phi'(v)| > \epsilon$) then $\int_0^\infty \phi'(v)dv$ is not finite and therefore $\lim_{v \to \infty} \phi(v) \neq 0$.

---

[2]We know that $-\underline{L}'$ is exists and is continuous because Theorem 4 requires a differentiable optimal link, which implies the minimum risk is differentiable (Zhang et al., 2004), and because $-\underline{L}$ is strictly convex, $-\underline{L}'$ is continuous and strictly monotonic.

Fixing some $\epsilon > 0$, we define a sequence $\{v_n\}$ such that for each n in the sequence, $v_n > 2v_{n-1}$ and $|v_n\phi'(v_n)| > \epsilon$. We have

$$\left|\int_{v_{n-1}}^{v_n} \phi'(v)\,\mathrm{d}v\right| \geq \left|\int_{v_{n-1}}^{v_n} \phi'(v_n)\,\mathrm{d}v\right| = |(v_n - v_{n-1})\phi'(v_n)|$$

where we get the inequality due to the fact that ϕ' must be monotonically increasing towards 0. Now, since $v_n - v_{n-1} > \frac{1}{2}v_n$, the assumption that $|v_n\phi'(v_n)| \geq \epsilon$ gives

$$\left|\int_{v_{n-1}}^{v_n} \phi'(v)\,\mathrm{d}v\right| > \left|\frac{1}{2}v_n\phi'(v_n)\right| \geq \frac{\epsilon}{2}.$$

Since for all N,

$$\left|\int_0^\infty \phi'(v)\,\mathrm{d}v\right| > \left|\sum_{i=1}^N \int_{v_{n-1}}^{v_n} \phi'(v)\,\mathrm{d}v\right| > N\frac{\epsilon}{2},$$

the integral must be unbounded.

$\square$

Theorem 10 (Gradient-Symmetric Losses as Bregman Divergences). *Let $\ell : \mathbb{R} \to \mathbb{R}$ be a strictly-convex differentiable margin loss function with infimum zero, then $\ell(yf)$ is expressible as (the limit of) a Bregman divergence with f as the first argument if and only if ℓ is gradient-symmetric. Furthermore, if $\ell(yf)$ is expressible as such a Bregman divergence, it is of the form*

$$\ell(yf) = \lim_{g \to g_y} B_\ell(f,g). \tag{8}$$

where g_y is a constant determined by y.

Proof. The general structure of this proof is as follows. First, we show if ℓ is expressible as a Bregman divergence, then the Bregman generator is necessarily gradient-symmetric. We then show that the Bregman generator must be ℓ (up to affine terms), and therefore ℓ is gradient-symmetric. Finally, we will prove the converse: if ℓ is gradient-symmetric, then it is necessarily expressible as a Bregman divergence with generator ℓ.

Preliminaries: Note that g_y can take two values depending on the value of y. We denote the value it takes for positive y as g^+ and for negative y as g^-. We show that when (8) is satisfied, $g^- = -g^+$.

When $y = +1$, we need our Bregman divergence to satisfy

$$\ell(yf) = \ell(f) = \lim_{g \to g^+} B_\phi(f,g),$$

and similarly, when $y = -1$,

$$\ell(yf) = \ell(-f) = \lim_{g \to g^-} B_\phi(f,g).$$

Since ℓ is strictly convex with infimum zero, ℓ is zero at exactly one point (or limit) and that point is the minimum. Since Bregman divergences are zero if and only if their arguments are equal to each other, this means that when (8) holds,

$$\lim_{f \to g^+} B_\phi(f,g^+) = 0$$
$$\Rightarrow \lim_{f \to g^+} \ell(f) = 0$$
$$\Rightarrow g^+ = \operatorname*{argmin}_{f \in \mathbb{R} \cup \{-\infty,\infty\}} \ell(f), \tag{19}$$

where $\ell(\infty) = \lim_{v \to \infty} \ell(v)$, and similarly for $\ell(-\infty)$. We also have $g^- = \operatorname{argmin}_{f \in \mathbb{R} \cup \{-\infty,\infty\}} \ell(-f)$ by the same reasoning. Combining these two facts, we have that $g^- = -g^+$.

Proof that Bregman Divergence $\implies$ Gradient-Symmetry:

ϕ is Gradient-Symmetric: Using the symmetry of the labels that we have just established, we show that the generator ϕ must exhibit gradient symmetry (i.e., $\phi'(f) + \phi'(-f)$ must be constant). Observe that the margin loss has a symmetry, in that the loss is the same for the pair f and y as it is for $-f$ and $-y$, since trivially $\ell((-y)(-f)) = \ell(yf)$. This gives that $\ell(f)$ must simultaneously satisfy both

$$\ell(f) = \lim_{g \to g^+} B_\phi(f, g)$$

and

$$\ell(f) = \lim_{g \to g^-} B_\phi(-f, g).$$

Expanding both of these using the definition of Bregman divergences and rearranging slightly, we find

$$\ell(f) = \lim_{g \to g^+} \phi(f) - \phi'(g)f - \phi(g) + \phi'(g)g$$
$$\ell(f) = \lim_{g \to g^-} \phi(-f) + \phi'(g)f - \phi(g) + \phi'(g)g$$

Taking the derivatives of both expressions with respect to f, we have

$$\ell'(f) = \lim_{g \to g^+} \phi'(f) - \phi'(g)$$
$$\ell'(f) = \lim_{g \to g^-} -\phi'(-f) - \phi'(g).$$

We can equate the two right-hand sides and rearrange to get

$$\phi'(f) + \phi'(-f) = \lim_{g \to g^-} \phi'(g) - \lim_{g \to g^+} \phi'(g).$$

Since the term on the right-hand side is independent of f, the left-hand side must be constant with respect to f and therefore ϕ is gradient-symmetric.

To complete this direction of the proof, it now suffices to show that $\phi = \ell$ (up to affine terms).

Necessity of $\phi(f) = \ell(f)$: We assume that the loss can be expressed as a Bregman divergence and show that in this case, we must have $\phi(f) = \ell(f)$. If $\ell(yf)$ can be expressed as a Bregman divergence with generator ϕ, then, taking $y = +1$, it must satisfy

$$\ell(yf) = \ell(f) = \lim_{g \to g^+} \phi(f) - \phi'(g)f - \phi(g) + \phi'(g)g$$

Note that, by definition, $\lim_{g \to g^+} \phi(g) = 0$ and by Lemma 4, $\lim_{g \to g^+} \phi'(g) = 0$. Additionally, $\lim_{g \to g^+} \phi'(g)g = 0$ (trivially when g^+ is finite and by Lemma 5 when g^+ is infinite). Therefore, taking the limit of these terms, we find

$$\ell(f) = \phi(f) \tag{20}$$

This completes one direction of the proof: If $\ell(yf)$ is expressible as a Bregman divergence, then ℓ is its Bregman generator, and ℓ is gradient-symmetric.

Proof that Gradient Symmetry $\implies$ Bregman divergence: We now show that if ℓ is gradient-symmetric, it can be written as the limit of a Bregman divergence. We do this by showing that

$$\lim_{g \to g^+} B_\ell(f, g) = \lim_{g \to g^+} [\ell(f) - \ell(g) - \ell'(g)(f - g)]$$
$$= \lim_{g \to g^+} [\ell(f) - \ell'(g)f - \ell(g) + \ell'(g)g]$$
$$= \ell(f) - \lim_{g \to g^+} [\underbrace{\ell'(g)f}_{\to 0} - \underbrace{\ell(g)}_{\to 0} - \underbrace{\ell'(g)g}_{\to 0}]$$
$$= \ell(f) = \ell(yf)|_{y=1}$$

In the third line, we claim that three terms each converge to zero. The first of these is due to Lemma 4, the second due to the condition we place on ℓ that its infimum is zero and the third due to Lemma 5.

Now we do the same process for the negative limit $\lim_{g \to g^-} B_\ell(f, g)$. Since $\lim_{g \to g^+} \ell'(g) = 0$ by Lemma 4, the condition on the derivative gives $\lim_{g \to g^-} \ell'(g) = c$. Using this fact, along with the fact that gradient-symmetry implies $\ell(f) = \ell(-f) + cf$, we have

$$
\begin{aligned}
\lim_{g \to g^-} B_\ell(f, g) &= \lim_{g \to g^-} [\ell(f) - \ell'(g)f - \ell(g) + \ell'(g)g] \\
&= \underbrace{\ell(-f) + cf}_{=\ell(f)} - cf - \lim_{g \to g^-} \ell(g) + \lim_{g \to g^-} \ell'(g)g \\
&= \ell(-f) - \lim_{g \to g^-} \ell(g) + \lim_{g \to g^-} \ell'(g)g.
\end{aligned}
$$

Using the same identity on $\ell(g)$ and the identity $\ell'(g) = c - \ell'(-g)$,

$$
\begin{aligned}
&= \ell(-f) - \lim_{g \to g^-} \ell(-g) - \lim_{g \to g^-} cg + \lim_{g \to g^-} (c - \ell'(-g))g \\
&= \ell(-f) - \lim_{g \to g^+} \ell(g) - \lim_{g \to g^-} cg + \lim_{g \to g^-} cg - \lim_{g \to g^+} \ell'(g))g \\
&= \ell(-f) - \lim_{g \to g^+} \ell'(g)g \\
&= \ell(yf)|_{y=-1},
\end{aligned}
$$

where the last equality is due to Lemma 5. Combining the $y = +1$ case and the $y = -1$ case, we have that for gradient-symmetric losses, $\ell(yf) = \lim_{g \to g_y} B_\ell(f, g)$, completing the proof.

$\square$

Proposition 2. *A differentiable strictly convex margin loss ℓ is gradient-symmetric with constant $c < 0$ if and only if it satisfies*

$$\underline{L}'(p) = c\psi(p) \tag{10}$$

Furthermore, the loss $\frac{1}{|c|}\ell$ is a canonical form loss.

Proof. If a loss is gradient-symmetric, it satisfies $\ell'(f) + \ell'(-f) = c$ for some c, and if $c < 0$ the loss $\widehat{\ell}(f) = \ell(f)/|c|$ satisfies $\widehat{\ell}'(f) + \widehat{\ell}'(-f) = -1$. This new loss has the same optimal link and is a canonical form loss. Since the minimum risk of this new loss is $1/|c|$ times the original minimum risk,

$$\frac{1}{|c|}\underline{L}(p) = \widehat{\underline{L}}(p),$$

where $\widehat{\underline{L}}$ is the minimum risk of $\widehat{\ell}$. Since $\widehat{\ell}$ is in canonical form, by definition we have

$$-\widehat{\underline{L}}'(p) = \psi(p) \implies -\frac{1}{|c|}\underline{L}'(p) = \psi(p),$$

which immediately gives $\underline{L}'(p) = c\psi(p)$.

Reasoning in the other direction, if we start with a loss satisfying $\underline{L}'(p) = c\psi(p)$, then it would share its optimal link with the loss $\widehat{\ell}$ and its minimum risk would be $\frac{1}{|c|}$ the minimum risk of $\widehat{\ell}$. Since $\widehat{\ell}$ is in canonical form (by virtue the negative derivative of its minimum risk being equal to its optimal link), it satisfies $\widehat{\ell}'(f) + \widehat{\ell}'(-f) = -1$. Therefore, $\ell(f) = |c|\widehat{\ell}(f)$ satisfies $\ell'(f) + \ell'(-f) = c$. $\square$

Theorem 11 (Decomposition for the LOLs). *Let ℓ be a LOL with odd part $\ell_o(v) = bv$ and even part $\ell_e(v) = \frac{1}{2}(\ell(v) + \ell(-v))$, then the expected risk decomposes as*

$$\mathbb{E}_{\mathbf{X}Y}[\mathbb{E}_D[\ell(Yf)]] = \mathbb{E}_{\mathbf{X}}[bY^*f^*] + \mathbb{E}_{\mathbf{X}}[\mathbb{E}_D[\ell_e(f)]],$$

where $Y^* = \mathbb{E}_{Y|\mathbf{X}}[Y]$ and $f^* = \mathbb{E}_D[f]$.

Proof. If we look at the expected risk

$$\mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[\ell(Yf)\right]\right] = \mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[\underbrace{\frac{1}{2}(\ell(Yf) + \ell(-Yf))}_{\text{even part}} + \underbrace{\frac{1}{2}(\ell(Yf) - \ell(-Yf))}_{\text{odd part}}\right]\right]$$

$$= \mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[\ell_e(Yf) + \ell_o(Yf)\right]\right]$$

$$= \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[\ell_e(f)\right]\right] + \mathbb{E}_{\mathbf{X}Y}\left[\mathbb{E}_D\left[bYf\right]\right]$$

$$= \mathbb{E}_{\mathbf{X}}\left[\mathbb{E}_D\left[\ell_e(f)\right]\right] + bY^*f^*,$$

$\square$

D   Separating Bias and Noise

In the bias-variance decomposition in Theorem 2, the first term of the decomposition, $\mathbb{E}_{\mathbf{X}Y}\left[\ell(Yf^*)\right]$, is the risk of the expected model, which is comprised of the noise, i.e., the systematic error that occurs due to uncertainty in the target; and the bias, i.e., the difference in loss between the predictor f^* and the ideal predictor for the variable Y. It is not usually possible to experimentally measure the noise in a dataset, so for practical application of the decomposition, it makes sense to keep these quantities combined in a single term. However, from a theoretical standpoint, it is useful to further decompose this term. In this appendix, we discuss how this may be achieved.

Theorem 3 shows that the excess risk can be written as a Bregman divergence, which is one of the key insights in constructing a bias-variance decomposition for these losses. However, rearranging the relationship in (4), we see that the pointwise risk for a fixed model f can be written

$$\mathbb{E}_{Y|\mathbf{X}}\left[\ell(Yf)\right] = L(p, f) = B_{-\underline{L}}\left(p, \psi^{-1}(f)\right) + \underline{L}(p).$$

This decomposition of the pointwise risk has some striking similarities to the decompositions we have already seen: the expected error on the left-hand side is decomposed into two terms: the first term measures the separation between a point $\psi^{-1}(f)$ and the central point of a distribution $p = \mathbb{E}_Y\left[Y_{0/1}\right]$, where $Y_{0/1} = 1$ when $Y = 1$ and $Y_{0/1} = 0$ otherwise. More importantly, the observation allows us to break down the pointwise risk term and minimum risk terms in the following manner.

Theorem 13 (Pointwise and Minimum Risks as Bregman Divergences)**.** *Let ℓ be a loss function with a differentiable minimum risk such that $\underline{L}(0) > -\infty$ and invertible link function ψ. We have*

$$L(p, f) = \mathbb{E}_Y\left[B_{-\underline{L}}\left(Y_{0/1}, q\right)\right] + \underline{L}(0), \tag{21}$$

where $q = \psi^{-1}(f)$. Furthermore, for $p \in (0, 1)$,

$$\underline{L}(p) = \mathbb{E}_Y\left[B_{-\underline{L}}\left(Y_{0/1}, p\right)\right] + \underline{L}(0). \tag{22}$$

For fair losses (i.e., losses where $\underline{L}(0) = \underline{L}(1) = 0$), the second term on the right-hand side vanishes.

Proof. We begin by showing the result for the minimum risk. To avoid confusion with signs, we use the notation $\phi = -\underline{L}$ and consider the generator ϕ. Using the definition of a Bregman divergence and the fact that $\mathbb{E}_{Y_{0/1}}\left[Y_{0/1}\right] = p$, we have

$$\mathbb{E}_Y\left[B_\phi(Y_{0/1}, p)\right] = \mathbb{E}_Y\left[\phi(Y_{0/1}) - \phi(p) - \phi'(p)(Y_{0/1} - p)\right]$$

$$= \mathbb{E}_Y\left[\phi(Y_{0/1})\right] - \phi(p) - \phi'(p)(p - p)$$

$$= \mathbb{E}_Y\left[\phi(Y_{0/1})\right] - \phi(p)$$

$$= p\phi(1) + (1 - p)\phi(0) - \phi(p)$$

$$= -p\underline{L}(1) - (1 - p)\underline{L}(0) + \underline{L}(p). \tag{23}$$

For margin losses, $\underline{L}(1) = \underline{L}(0)$, so $p\underline{L}(1) + (1 - p)\underline{L}(0) = \underline{L}(0)$, and (23) yields

$$\mathbb{E}_Y\left[B_{-\underline{L}}\left(Y_{0/1}, p\right)\right] = \underline{L}(p) - \underline{L}(0).$$

Rearranging gives the result

$$\underline{L}(p) = \mathbb{E}_Y \left[B_{-\underline{L}} \left(Y_{0/1}, p \right) \right] + \underline{L}(0). \tag{24}$$

Using this, we can prove the more general result for $L(p, f)$. Starting from the Bregman divergence for the excess risk (4) and applying (22), we have

$$L(p, f) = B_{-\underline{L}} (p, q) + \underline{L}(p)$$
$$L(p, f) = B_{-\underline{L}} (p, q) + \mathbb{E}_Y \left[B_{-\underline{L}} \left(Y_{0/1}, p \right) \right] + \underline{L}(0).$$

Using the definition of Bregman divergences and again defining $\phi = -\underline{L}$, and using $\mathbb{E}_Y \left[B_\phi \left(Y_{0/1}, p \right) \right] = \mathbb{E}_Y \left[\phi(Y_{0/1}) \right] - \phi(p)$, which we established as an intermediate step in (23):

$$\begin{aligned}
L(p, f) &= \phi(p) - \phi(q) - \phi'(q)(p - q) + \mathbb{E}_Y \left[\phi(Y_{0/1}) \right] - \phi(p) + \underline{L}(0) \\
&= -\phi(q) - \phi'(q)(p - q) + \mathbb{E}_Y \left[\phi(Y_{0/1}) \right] + \underline{L}(0) \\
&= \mathbb{E}_Y \left[\phi(Y_{0/1}) - \phi(q) - \phi'(q)(Y_{0/1} - q) \right] + \underline{L}(0) \\
&= \mathbb{E}_Y \left[B_\phi(Y_{0/1}, q) \right] + \underline{L}(0).
\end{aligned}$$

$\square$

For a central model f^*, the bias and the noise can be separated using this result. Since the minimum risk can be written $\underline{L}(p) = \mathbb{E}_Y \left[B_{-\underline{L}} \left(Y_{0/1}, p \right) \right] + \underline{L}(0)$ and we already have an expression for the excess risk as a Bregman divergence, we can write

$$\begin{aligned}
\mathbb{E}_{\mathbf{X}Y} \left[\ell(Y f^*) \right] &= \mathbb{E}_{\mathbf{X}} \left[L(p, f^*) \right] \\
&= \mathbb{E}_{\mathbf{X}} \left[\underline{L}(p) + L(p, f^*) - \underline{L}(p) \right] \\
&= \underline{L}(0) + \underbrace{\mathbb{E}_{\mathbf{X}Y} \left[B_{-\underline{L}} \left(Y_{0/1}, p \right) \right]}_{\text{noise}} + \underbrace{\mathbb{E}_{\mathbf{X}} \left[B_{-\underline{L}} \left(p, \psi^{-1}(f^*) \right) \right]}_{\text{bias}}.
\end{aligned}$$

Note that $\underline{L}(0)$ is a constant determined by the loss function, and *fair losses*, including all those considered in this paper, is $\underline{L}(0) = 0$. For gradient-symmetric losses, it is also possible to express this decomposition in terms of Bregman divergences with ℓ as the generator.

Theorem 14. *For a gradient-symmetric margin loss ℓ that satisfies $\underline{L}(0) = 0$, the noise can be written*

$$\underline{L}(p) = \mathbb{E}_Y \left[\lim_{g \to g_Y} B_\ell \left(-\frac{L'(p)}{c}, -\frac{g_Y}{c} \right) \right]$$

Proof. From Theorem 6, we have for all $u, v \in \mathbb{R}$

$$B_\ell (u, v) = B_{-\underline{L}} \left([-L']^{-1}(cv), [-L']^{-1}(cu) \right),$$

or equivalently, for all $u, v \in \mathbb{R}$

$$B_\ell \left(\frac{u}{c}, \frac{v}{c} \right) = B_{-\underline{L}} \left([-L']^{-1}(v), [-L']^{-1}(u) \right)$$

Starting with the noise term, if we define the random variable Y_ϵ such that $Y_{0/1} = 1 \implies Y_\epsilon = 1 - \epsilon$ and $Y_{0/1} = 0 \implies Y_\epsilon = \epsilon$,

$$\begin{aligned}
\mathbb{E}_Y \left[B_{-\underline{L}} \left(Y_{0/1}, p \right) \right] &= \mathbb{E}_Y \left[\lim_{\epsilon \to 0} B_{-\underline{L}} \left(Y_\epsilon, p \right) \right] \\
&= \mathbb{E}_Y \left[\lim_{\epsilon \to 0} B_\ell \left(\frac{-L'(p)}{c}, \frac{-L'(Y_\epsilon)}{c} \right) \right],
\end{aligned}$$

$\square$

E   Relationship Between the Decompositions for Proper Scoring Rules and Gradient-Symmetric Losses

In this appendix, we provide additional details about the relationship between the decompositions we present in Section 3.1 and the decomposition in Theorem 4. In Appendix E.1, we show that the centroids in Theorem 1 and Theorem 4 coincide if and only if the loss ℓ is gradient-symmetric. In Appendix E.2, we show that when both decompositions are defined, the bias and variance terms in Theorem 2 and Theorem 4 are equivalent.

E.1   Coinciding Centroids

In this section, we prove that $\psi^{-1}(f^*)$ and q^* coincide for a loss ℓ if and only if the loss is gradient-symmetric.

We start with the following lemma, which is built around a result given by copper.hat (2020).

Lemma 6. *Let ℓ be a strictly convex differentiable margin loss, then ψ is both continuous and invertible (with the range being $\mathbb{R}$ or a finite interval thereof).*

Proof. Since $L(p,v) = p\ell(v) + (1-p)\ell(-v)$ and the sum of two strictly convex functions is also strictly convex, $L(p,v)$ is strictly convex in v. Furthermore, since margin losses (up to a constant factor) are upper bounds on the zero-one loss, both terms on the right hand side are positive.

We will show that ψ is continuous around the arbitrary point $p_\infty \in (0,1)$. Define the sequence $\{p_k\}$ with $p_k \in (0,1)$ for all k and such that $p_k \to p_\infty$. Let $\{x_k\}$ be the sequence such that $x_k = \psi(p_k)$ and define $x_\infty = \psi(p_\infty)$.

We will first show that all x_k are contained in a compact set. We do this, as establishing this fact means that there is a convergent subsequence of $\{x_k\}$. We will show that this subsequence necessarily converges to x_∞, and from the continuity of $L(p,v)$, this will give that ψ is also continuous.

Since $\{p_k\}$ converges to $p_\infty \in (0,1)$ we can choose $\epsilon > 0$ such that all p_k, we have $p_k \in (\epsilon, 1-\epsilon)$. For all strictly convex functions on the real line either $\lim_{v\to\infty} \ell(v) = \infty$ or $\lim_{v\to\infty} \ell(-v) = \infty$. We will complete our proof assuming the latter case, though the former can be proved by simple modifications of the same argument. Assuming the latter, we now show for all $p \in (\epsilon, 1-\epsilon)$, that $\psi(p) \in \mathbb{R}$ is in a bounded region around 0. Since $\ell(-v)$ tends to ∞, there exists v_0 such that for all $v > v_0$, $\ell(-v) > \frac{\ell(0)}{\epsilon}$ and therefore

$$
\begin{aligned}
L(p,-v) &= p\ell(-v) + (1-p)\ell(v)\\
&\geq p\ell(-v)\\
&> p\frac{\ell(0)}{\epsilon}\\
&> \ell(0) = L(p,0).
\end{aligned}
$$

This means that for all $v > v_0$, we have $-v$ cannot be the minimizer of $L(p,\cdot)$ with $p \in (\epsilon, 1-\epsilon)$, so by definition of the optimal link function, $\psi(p) > -v_0$. In particular, for all k, $x_k = \psi(p_k) > -v_0$. Similarly for $v > v_0$,

$$
\begin{aligned}
L(p,v) &= p\ell(v) + (1-p)\ell(-v)\\
&\geq (1-p)\ell(-v)\\
&> \epsilon\ell(-v)\\
&> \epsilon\frac{\ell(0)}{\epsilon}\\
&= \ell(0) = L(p,0).
\end{aligned}
$$

Hence, we can conclude that for $p \in (\epsilon, 1-\epsilon)$, $\psi(p) < v_0$.

Combining the two cases, we find that $x_k = \psi(p_k) \in (-v_0, v_0)$ for all k. Since the sequence $\{x_k\}$ is all contained in a compact set, it necessarily has a convergent subsequence. We choose such a subsequence and call its limit x_{acc} (so called because it is an accumulation point of $\{x_k\}$).

Note that by definition of ψ as the minimizer, for all $v \in \mathbb{R}$, we have

$$
L(p_k, x_k) = L(p_k, \psi(p_k)) \leq L(p_k, v).
$$

For our subsequence of $\{x_k\}$, by virtue of $L(\cdot, \cdot)$ being continuous, we have

$$\lim_{k \to \infty} L(p_k, x_k) = L(p_\infty, x_{\text{acc}}). \tag{25}$$

For the subsequence we also have

$$\lim_{k \to \infty} L(p_k, x_k) = \lim_{k \to \infty} L(p_k, \psi(p_k)) \leq \lim_{k \to \infty} L(p_k, v) = L(p_\infty, v), \tag{26}$$

From (26) and (25), we have for all $v \in \mathbb{R}$, that $L(p_\infty, x_{\text{acc}}) \leq L(p_\infty, v)$, so x_{acc} is a minimizer of $L(p_0, \cdot)$. By strict convexity, this minimizer is unique so $x_{\text{acc}} = x_\infty$.

Consequently, we have that as $p \to p_\infty$, $\psi(p) \to \psi(p_\infty)$, so ψ is continuous at p_∞. Since p_∞ is an arbitrary point on the interval $(0, 1)$, ψ is continuous on the entire interval.

We now show that ψ is strictly monotonic. For v to be the minimizer of $L(p, v)$ for a specific p, it must be the case that

$$\frac{\partial}{\partial v} L(p, v) = 0.$$

Expanding and rearranging this, we find

$$\frac{\partial}{\partial v} L(p, v) = 0$$
$$\implies \frac{\partial}{\partial v} \left[p\ell(v) + (1 - p)\ell(-v) \right] = 0$$
$$\implies p\ell'(v) - (1 - p)\ell'(-v) = 0$$
$$\implies p = \frac{\ell(-v)}{\ell'(v) + \ell'(-v)}.$$

From this, we can see that ψ must be strictly monotonic. If it were not, then by continuity, there would necessarily be $p_1 \neq p_2$ such that $\psi(p_1) = \psi(p_2)$, but this would imply

$$p_1 = \frac{\ell(-\psi(p_1))}{\ell'(\psi(p_1)) + \ell'(-\psi(p_1))} = \frac{\ell(-\psi(p_2))}{\ell'(\psi(p_2)) + \ell'(-\psi(p_2))} = p_2,$$

contradicting the assumption that p_1 and p_2 are distinct. Therefore ψ is continuous and strictly monotonic, and is therefore invertible (on some potentially restricted interval of the real line). $\qquad \square$

Note that for some losses, such as the squared margin loss $\ell(v) = (1 - v)^2$, the optimal link function is not invertible on $\mathbb{R}$ and the result only holds if we restrict our model f to have values only in some compact interval determined by ℓ.

In Section 3.2, we claimed that the central model and f^* and the model $\psi(q^*)$ coincide if and only the loss is gradient-symmetric. Here we formally state and prove the claim.

Theorem 15. *For a strictly convex differentiable margin loss ℓ where both Theorem 2 and Theorem 4 hold, define the central model as*

$$f^* = \mathbb{E}_D \left[f \right]$$

and define the model corresponding to the central probability estimate as $\psi(q^*)$ where

$$q^* = [-\underline{L}']^{-1} \left(\mathbb{E}_D \left[-\underline{L}'(q) \right] \right)$$

the two models coincide, i.e., $f^* = \psi(q^*)$ if and only if ℓ is gradient-symmetric.

Proof. ($\implies$) Assume that $f^* = \psi(q^*)$, then by the definitions of f^* and $\psi(q^*)$

$$\mathbb{E}_D \left[f \right] = \psi \circ [-\underline{L}']^{-1} \left(\mathbb{E}_D \left[-\underline{L}' \circ \psi^{-1}(f) \right] \right)$$

Since ψ and $-\underline{L}'$ are both monotonic and continuous (the former using Lemma 6), so is their composition[3]. By the application of Thereom 12, this means that $\psi \circ [-\underline{L}']^{-1}$ must be a linear function. Defining $l = \psi \circ [-\underline{L}]'^{-1}$ where $\ell(v) = av + b$, we have $-\underline{L}' = l^{-1} \circ \psi$ and by Lemma 3, $b = 0$. We therefore have that $\psi(p) = -a\underline{L}'(p)$ for a constant a, and therefore by Proposition 2, ℓ is gradient-symmetric.

($\Longleftarrow$) If ℓ is gradient-symmetric, then by Proposition 2, $-\underline{L}'(p) = c\psi(p)$, and therefore taking the inverse functions of both sides $[-\underline{L}']^{-1}(f) = \psi^{-1}(f/c)$. Applying these two facts to $\psi(q^*)$, we find

$$
\begin{aligned}
\psi(q^*) = \psi \circ [-\underline{L}]^{-1} &\left(\mathbb{E}_D \left[[-\underline{L}'] \circ \psi^{-1}(f) \right] \right) \\
&= \psi \circ [-\underline{L}]^{-1} \left(\mathbb{E}_D \left[cf \right] \right) \\
&= \frac{\mathbb{E}_D \left[cf \right]}{c} = \mathbb{E}_D \left[f \right],
\end{aligned}
$$

so gradient-symmetry implies that $\psi(q^*) = f^*$ $\qquad\square$

E.2 Equivalence of Terms Between Decompositions

Proposition 3 (Equivalence of Bias Terms). *For a strictly convex gradient-symmetric loss ℓ with invertible differentiable optimal link,*

$$
\mathbb{E}_Y \left[\ell(Yf^*) \right] = B_{-\underline{L}} \left(p, \psi^{-1}(f^*) \right) + \underline{L}(p)
$$

Proof. The result can be derived immediately from Theorem 13 and the definition of $L(p, f)$:

$$
\begin{aligned}
\mathbb{E}_Y \left[\ell(Yf^*) \right] &= L(p, f^*) \\
&= L(p, f^*) - \underline{L}(p) + \underline{L}(p) \\
&= B_{-\underline{L}} \left(p, \psi^{-1}(f^*) \right) + \underline{L}(p).
\end{aligned}
$$

$\qquad\square$

Proposition 4 (Equivalence of Variance Terms). *For a strictly convex gradient-symmetric loss ℓ with invertible differentiable optimal link,*

$$
\mathbb{E}_D \left[B_\ell \left(f, f^* \right) \right] = \mathbb{E}_D \left[B_{-\underline{L}} \left(q^*, q \right) \right]
$$

Proof. We could use the Proposition 3 and the two bias-variance decompositions to show that the two variance terms are equal, but it is perhaps more instructive to show that the two terms are equivalent using facts that we know about gradient-symmetric losses.

Let $h(v) = -cv$, then by Proposition 2, we have $-\underline{L}' = h \circ \psi$. Developing this gives

$$
\begin{aligned}
-\underline{L}' &= h \circ \psi \\
\Longleftrightarrow h^{-1} \circ [-\underline{L}'] &= \psi \\
\Longleftrightarrow \psi^{-1} &= [-\underline{L}']^{-1} \circ h \\
\Longleftrightarrow [-\underline{L}'] \circ \psi^{-1} &= h.
\end{aligned}
$$

Remember that for gradient-symmetric losses, $\psi(q^*) = f^*$, so $q^* = \psi^{-1}(f^*)$, and we have

$$
\begin{aligned}
\mathbb{E}_D \left[B_{-\underline{L}} \left(q^*, q \right) \right] &= \mathbb{E}_D \left[B_{-\underline{L}^*} \left([-\underline{L}'(q)], -\underline{L}'(q^*) \right) \right] \\
&= \mathbb{E}_D \left[B_{-\underline{L}^*} \left([-\underline{L}'(\psi^{-1}(f))], -\underline{L}'(\psi^{-1}(f^*)) \right) \right] \\
&= \mathbb{E}_D \left[B_{-\underline{L}^*} \left(-cf, -cf^* \right) \right] \\
&= \mathbb{E}_D \left[[-\underline{L}^*](-cf) - [-\underline{L}^*](-cf^*) - [-\underline{L}^*]'(-cf^*)(f - f^*) \right] \\
&= \mathbb{E}_D \left[[-\underline{L}^*](-cf) - [-\underline{L}^*](-cf^*) \right],
\end{aligned}
$$

---

[3]We have not proved that $-\underline{L}$ is differentiable, but note that if it is not $B_{-\underline{L}}$ is not uniquely defined, and neither is the bias-variance decomposition of Theorem 4

where the last line uses that $f^* = \mathbb{E}_D[f]$. Now, using Theorem 5, $[-\underline{L}]^*(cv) = \ell(v)$, so

$$\begin{aligned}
\mathbb{E}_D\left[B_{-\underline{L}}(q^*, q)\right] &= \mathbb{E}_D\left[\ell(-f) - \ell(-f^*)\right] \\
&= \mathbb{E}_D\left[\ell(-f) - \ell(-f^*) - \ell'(f^*)(f - f^*)\right] \\
&= \mathbb{E}_D\left[B_\ell(-f, -f^*)\right] \\
&= \mathbb{E}_D\left[B_\ell(f, f^*)\right],
\end{aligned}$$

where the last line uses the fact that ℓ is gradient-symmetric, along with Proposition 1. $\qquad\square$