A Generalized Bias-Variance Decomposition for Bregman Divergences

David Pfau

June 11, 2013

Definition 0.1 (Bregman Divergence). Let $F : S \to \mathbb{R}$ be a strictly convex differentiable function, then the Bregman Divergence derived from $F$ is a function $D_F : S \times S \to \mathbb{R}^+$ defined as

$$D_F[x||y] \triangleq F(x) - F(y) - \langle \nabla F(y), x - y \rangle.$$ 

Lemma 0.1 (Minimum Expected Bregman Divergence). Let $F : S \to \mathbb{R}$ be a strictly convex differentiable function, and $X$ be a random variable on $S$. Then $x^* = \arg\min_z \mathbb{E}[D_F[z||X]] \iff \nabla F(x^*) = \mathbb{E}[\nabla F(X)]$ and $\mathbb{E}[X] = \arg\min_z \mathbb{E}[D_F[X||z]]$.

Proof. A necessary condition for $x^*$ to minimize the expected divergence is that its gradient should be zero. The gradient of the expected Bregman divergence when the expectation is taken over the second argument is given by

$$\nabla_z \mathbb{E}[D_F[z||X]] = \nabla_z \mathbb{E}[F(z) - F(X) - \langle \nabla F(X), z - X \rangle]$$

$$= \nabla F(z) - \nabla z \langle \mathbb{E}[\nabla F(X)], z \rangle$$

$$= \nabla F(z) - \mathbb{E}[\nabla F(X)] = 0$$

$$\Rightarrow \nabla F(z) = \mathbb{E}[\nabla F(X)]$$

by the linearity of expectations and the independence of $z$ from $X$. Since $F$ is convex, if an $x^*$ exists that satisfies this condition then it is unique, and therefore the minimum.

When the expectation is taken over the first argument, the gradient is then

$$\nabla_z \mathbb{E}[D_F[X||z]] = \nabla_z \mathbb{E}[F(X) - F(z) - \langle \nabla F(z), X - z \rangle]$$

$$= -\nabla F(z) - \nabla \langle \nabla F(z), \mathbb{E}[X] \rangle + \nabla \langle \nabla F(z), z \rangle$$

$$= -\nabla F(z) - \nabla^2 F(z) \mathbb{E}[X] + \nabla^2 F(z) z + \nabla F(z)$$

$$= -\nabla^2 F(z) \mathbb{E}[X] + \nabla^2 F(z) z = 0$$

$$\Rightarrow \nabla^2 F(z) z = \nabla^2 F(z) \mathbb{E}[X]$$

$$\Rightarrow z = \mathbb{E}[X]$$

where the last step follow from the fact that the Hessian of a strictly convex function is positive definite and therefore invertible. \qed
\textbf{Theorem 0.1} (Decomposition of Expected Bregman Divergence). Let $F : S \to \mathbb{R}$ be a strictly convex differentiable function, and $X$ be a random variable on $S$. Then for any point $s \in S$, the expected Bregman divergences have the following exact decomposition:

$$
\mathbb{E}[D_F[s||X]] = D_F[s||x^*] + \mathbb{E}[D_F[x^*||X]],
$$

where $x^* = \arg\min_x \mathbb{E}[D_F[z||X]]$

$$
\mathbb{E}[D_F[X||s]] = D_F[x^*||s] + \mathbb{E}[D_F[X||x^*]],
$$

where $x^* = \arg\min_x \mathbb{E}[D_F[X||z]] = \mathbb{E}[X]$.

\textit{Proof.}

\begin{align*}
D_F[s||x^*] + \mathbb{E}[D_F[x^*||X]] &= F(s) - F(x^*) - \langle \nabla F(x^*), s - x^* \rangle \\
&+ \mathbb{E}[F(x^*) - F(X) - \langle \nabla F(X), x^* - X \rangle] \\
&= F(s) - \langle \mathbb{E}[\nabla F(X)], s - x^* \rangle \\
&+ \mathbb{E}[-F(X) - \langle \nabla F(X), x^* - X \rangle] \\
&= \mathbb{E}[F(s) - F(X) - \langle \nabla F(X), s - x^* + x^* - X \rangle] \\
&= \mathbb{E}[F(s) - F(X) - \langle \nabla F(X), s - X \rangle] \\
&= \mathbb{E}[D_F[s||X]]
\end{align*}

\begin{align*}
D_F[\mathbb{E}[X]||s] + \mathbb{E}[D_F[X||\mathbb{E}[X]]] &= F(\mathbb{E}[X]) - F(s) - \langle \nabla F(s), \mathbb{E}[X] - s \rangle \\
&+ \mathbb{E}[F(s) - F(\mathbb{E}[X]) - \langle \nabla F(\mathbb{E}[X]), X - \mathbb{E}[X] \rangle] \\
&= -F(s) - \langle \nabla F(s), \mathbb{E}[X] - s \rangle \\
&+ \mathbb{E}[F(X) - \langle \nabla F(\mathbb{E}[X]), \mathbb{E}[X] - \mathbb{E}[X] \rangle] \\
&= \mathbb{E}[F(X) - F(s) - \langle \nabla F(s), X - s \rangle] \\
&= \mathbb{E}[D_F[X||s]]
\end{align*}

Suppose we wish to predict some random variable $Y \in S$ that is dependent on another variable $X \in \mathcal{R}$. We are given a training set $D = \{\{x_1, y_1\}, \ldots, \{x_n, y_n\}\}$ of input/output pairs sampled iid from the joint distribution of $X$ and $Y$, and have an algorithm that learns a deterministic prediction function from the data $f_D : \mathcal{R} \to S$. If the loss function for evaluating the quality of prediction is the Bregman divergence derived from $F$, $L(y, f_D(x)) = D_F[y||f_D(x)]$ then the expected loss can be decomposed exactly.

\textbf{Theorem 0.2} (Generalized Bias-Variance Decomposition). Let $F : S \to \mathbb{R}$ be a strictly convex differentiable function, $f_D : \mathcal{R} \to S$ be the prediction function trained on data $D = \{\{x_1, y_1\}, \ldots, \{x_n, y_n\}\}$, and $Y$ be the random variable we are trying to predict from $X$. Then the expected Bregman divergence of the data obeys a generalized bias-variance decomposition:

\begin{align*}
\mathbb{E}_{D,Y}[D_F[Y||f_D(X)]] &= \mathbb{E}_Y[D_F[Y||f^*(X)]] + D_F[f^*(X)||\hat{f}(X)] + \mathbb{E}_D[D_F[\hat{f}(X)||f_D(X)]]
\end{align*}
where $f^*(X) = \arg \min_z \mathbb{E}_Y[D_F[Y||z]] = \mathbb{E}_Y[Y]$, $\bar{f}(X) = \arg \min_z \mathbb{E}_D[D_F[z||f_D(X)]]$, and all expectations are implicitly conditioned on $X$.

**Proof.** The proof is a straightforward consequence of Theorem 0.1.

\[
\begin{align*}
\mathbb{E}_{D,Y}[D_F[Y||f_D(X)]] & = \mathbb{E}_D[\mathbb{E}_Y[D_F[Y||f_D(X)]|D]] \\
& = \mathbb{E}_D[\mathbb{E}_Y[D_F[Y||f^*(X)]|D] + D_F[f^*(X)||f_D(X)]] \\
& = \mathbb{E}_D[D_F[Y||f^*(X)] + \mathbb{E}_D[D_F[f^*(X)||f_D(X)]] \\
& = \mathbb{E}_D[D_F[Y||f^*(X)] + D_F[f^*(X)||\bar{f}(X)] + \mathbb{E}_D[\bar{f}(X)||f_D(X)]]
\end{align*}
\]