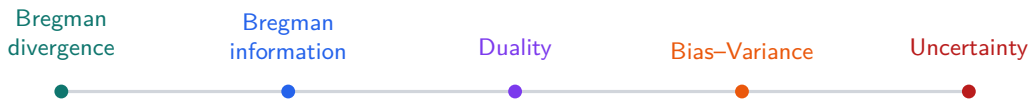
Bregman Divergences, Bregman Information and the Generalised Bias–Variance Decomposition



Mira Jürgens

Department of Data Analysis and Mathematical Modelling
Ghent University

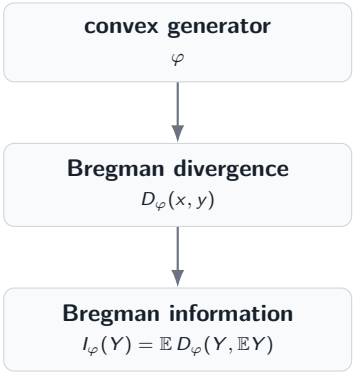
One recurring distance measure in machine learning

Cross entropy and KL-divergence are objectives used everywhere in ML:

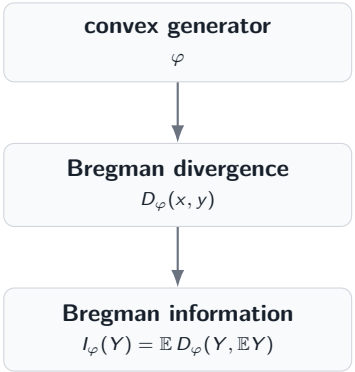
- as training objectives (e.g. cross-entropy loss)
- as variational objectives (e.g. KL regularisation)
- as uncertainty measures (e.g. mutual information)

Last talk by Antonio: How do entropy and KL divergence connect?
What is the underlying structure?

The framework that connects them



The framework that connects them



Bregman divergence	Bregman information
Squared error	Variance
KL divergence	Information entropy
?	?

Bregman divergences

Definition

Let $\varphi : \Omega \to \mathbb{R}$ be strictly convex and continuously differentiable on a convex domain $\Omega \subseteq \mathbb{R}^d$.

The Bregman divergence generated by φ is

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

Bregman (1967) *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming.*

Definition

Let $\varphi : \Omega \to \mathbb{R}$ be strictly convex and continuously differentiable on a convex domain $\Omega \subseteq \mathbb{R}^d$.

The Bregman divergence generated by φ is

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

Bregman (1967) *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming.*

- the *first-order Taylor remainder* of φ at y, evaluated at x

Definition

Let $\varphi : \Omega \to \mathbb{R}$ be strictly convex and continuously differentiable on a convex domain $\Omega \subseteq \mathbb{R}^d$.

The Bregman divergence generated by φ is

$$\boxed{D_\varphi(x, y) \;=\; \varphi(x) - \varphi(y) - \langle \nabla\varphi(y),\, x - y \rangle.}$$

Bregman (1967) *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming.*

- the *first-order Taylor remainder* of φ at y, evaluated at x
- $D_\varphi(x, y) = 0 \iff x = y$

Definition

Let $\varphi : \Omega \to \mathbb{R}$ be strictly convex and continuously differentiable on a convex domain $\Omega \subseteq \mathbb{R}^d$.

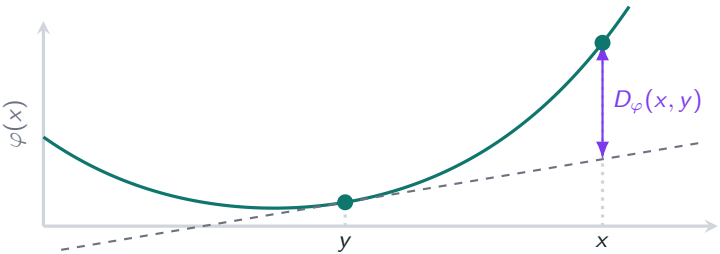
The Bregman divergence generated by φ is

$$\boxed{D_\varphi(x, y) \;=\; \varphi(x) - \varphi(y) - \langle \nabla\varphi(y),\, x - y \rangle.}$$

Bregman (1967) *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming.*

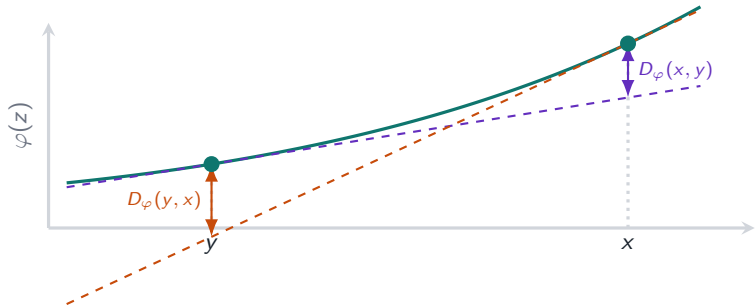
- the *first-order Taylor remainder* of φ at y, evaluated at x
- $D_\varphi(x, y) = 0 \iff x = y$
- $D_\varphi(x, y) \geq 0$

Geometric intuition: the tangent gap



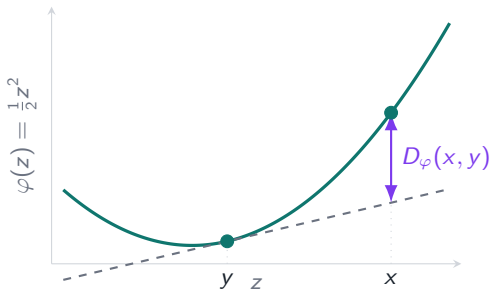
$$D_\varphi(x, y) \;=\; \varphi(x) \;-\; \underbrace{\big[\varphi(y) + \langle \nabla\varphi(y), x - y\rangle\big]}_{\text{tangent at } y\text{, evaluated at } x}.$$

Asymmetry



$$D_\varphi(x, y) \neq D_\varphi(y, x) \qquad \text{unless } \varphi \text{ is quadratic.}$$

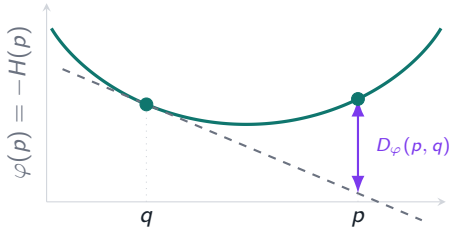
Example 1: quadratic generator $\varphi(x) = \frac{1}{2}\|x\|^2$



Here $\nabla\varphi(y) = y$. Hence

$$\begin{aligned} D_\varphi(x, y) &= \tfrac{1}{2}\|x\|^2 - \tfrac{1}{2}\|y\|^2 - \langle y, x - y\rangle \\ &= \tfrac{1}{2}\|x\|^2 - \langle x, y\rangle + \tfrac{1}{2}\|y\|^2 \\ &= \tfrac{1}{2}\|x - y\|^2. \end{aligned}$$

Example 2: negative entropy



$$\varphi(p) = \sum_{k=1}^{K} p_k \log p_k = -H(p),$$

$$\nabla\varphi(q)_k = \log q_k + 1.$$

$$\begin{aligned} D_\varphi(p, q) &= \sum_k p_k \log p_k - \sum_k q_k \log q_k - \sum_k (\log q_k + 1)(p_k - q_k) \\ &= \sum_k p_k \log p_k - \sum_k p_k \log q_k - \sum_k (p_k - q_k) \\ &= \sum_k p_k \log \frac{p_k}{q_k} = \mathrm{KL}(p \,\|\, q). \end{aligned}$$

Bregman divergences in context

Metric vs. divergence

A metric d satisfies four axioms:

1. Non-negativity: $d(x, y) \geq 0$.
2. Identity: $d(x, y) = 0 \iff x = y$.
3. Symmetry: $d(x, y) = d(y, x)$.
4. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

A divergence retains only (1) and (2).

Metric vs. divergence

A metric d satisfies four axioms:

1. Non-negativity: $d(x, y) \geq 0$.
2. Identity: $d(x, y) = 0 \iff x = y$.
3. Symmetry: $d(x, y) = d(y, x)$.
4. Triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

A divergence retains only (1) and (2).

Why do we often use divergences in ML?

- Maximum likelihood estimation is a divergence operation (truth $\to$ model).

Other divergence families

f-divergences
For convex $f : \mathbb{R}_{>0} \to \mathbb{R}$ with $f(1) = 0$:

$$\boxed{D_f(P \,\|\, Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right].}$$

Other divergence families

f-divergences
For convex $f : \mathbb{R}_{>0} \to \mathbb{R}$ with $f(1) = 0$:

$$\boxed{D_f(P \,\|\, Q) = \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right].}$$

Examples:

$f(t) = t \log t$	KL
$f(t) = \frac{1}{2}\lvert t-1\rvert$	Total Variation
$f(t) = (\sqrt{t}-1)^2$	Hellinger^2
$f(t) = (t-1)^2$	χ^2

Integral Probability Metrics

For a class of test functions $\mathcal{F}$:

$$\boxed{\gamma_{\mathcal{F}}(P, Q) \;=\; \sup_{f \in \mathcal{F}} \left| \mathbb{E}_P f - \mathbb{E}_Q f \right|.}$$

Integral Probability Metrics

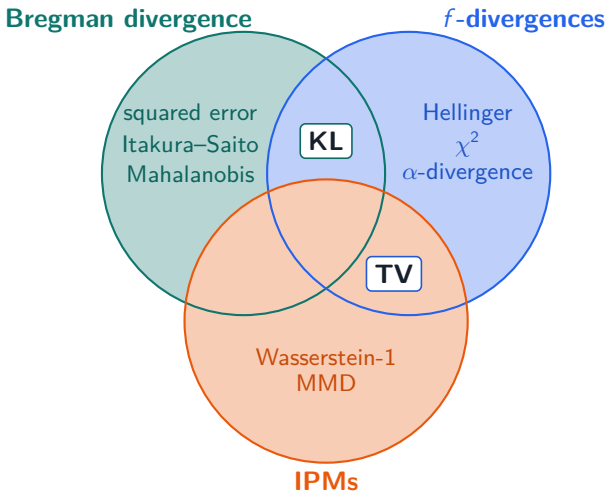
For a class of test functions $\mathcal{F}$:

$$\boxed{\gamma_{\mathcal{F}}(P, Q) \;=\; \sup_{f \in \mathcal{F}} \left| \mathbb{E}_P f - \mathbb{E}_Q f \right|.}$$

Examples:

$\|f\|_\infty \leq 1$	Total Variation
f 1-Lipschitz	Wasserstein-1
$\|f\|_{\mathcal{H}} \leq 1$	Maximum Mean Discrepancy

Three families, two intersections



Why Bregman divergences matter in machine learning

proper scoring rules, likelihood, optimisation

Proper scoring rules induce Bregman divergences

Let $\ell(q, y)$ be a loss for predicting the distribution of $Y \sim p^*$.

ℓ is a proper scoring rule if:

$$p^* \in \arg\min_{q \in \mathcal{P}} \mathbb{E}_{Y \sim p^*} \ell(q, Y).$$

Proper scoring rules induce Bregman divergences

Let $\ell(q, y)$ be a loss for predicting the distribution of $Y \sim p^*$.

ℓ is a proper scoring rule if:

$$p^* \in \arg\min_{q \in \mathcal{P}} \mathbb{E}_{Y \sim p^*} \ell(q, Y).$$

The *excess risk* is a Bregman divergence:

$$\boxed{\mathbb{E}_{Y \sim p^*} \ell(q, Y) - \mathbb{E}_{Y \sim p^*} \ell(p^*, Y) = D_\varphi(p^*, q),}$$

with generator $\varphi(p) = -\mathbb{E}_{Y \sim p} \ell(p, Y)$

Gruber & Buettner (2023) *Uncertainty Estimates of Predictions via a General Bias-Variance Decomposition.*

Maximum likelihood is KL minimisation

Let $P_\star$ be the data-generating distribution and $\{Q_\theta : \theta \in \Theta\}$ a model class.

$$R(\theta) := \mathbb{E}_{Y \sim P_\star}[-\log q_\theta(Y)].$$

$$R(\theta) = H(P_\star) + \mathrm{KL}(P_\star \,\|\, Q_\theta).$$

Maximum likelihood is KL minimisation

Let $P_\star$ be the data-generating distribution and $\{Q_\theta : \theta \in \Theta\}$ a model class.

$$R(\theta) := \mathbb{E}_{Y \sim P_\star}[-\log q_\theta(Y)].$$

$$R(\theta) = H(P_\star) + \mathrm{KL}(P_\star \,\|\, Q_\theta).$$

Since $H(P_\star)$ does not depend on θ,

$$\boxed{\arg\min_\theta R(\theta) = \arg\min_\theta \mathrm{KL}(P_\star \,\|\, Q_\theta).}$$

MLE is the empirical version

The mean is the (primal) Bregman centroid

Let Y be a random variable, and let $\bar{y} = \mathbb{E}Y$.

We have for any Bregman divergence

$$\boxed{\arg\min_m \mathbb{E}\big[D_\varphi(Y, m)\big] = \mathbb{E}Y.}$$

The mean is the (primal) Bregman centroid

Let Y be a random variable, and let $\bar{y} = \mathbb{E}Y$.

We have for any Bregman divergence

$$\boxed{\arg\min_m \mathbb{E}\big[D_\varphi(Y, m)\big] = \mathbb{E}Y.}$$

Holds since

$$\mathbb{E}D_\varphi(Y, m) = \mathbb{E}D_\varphi(Y, \mathbb{E}Y) + D_\varphi(\mathbb{E}Y, m).$$

Banerjee et al. (2005), *Clustering with Bregman divergences*.

Bregman information: generalised form of variance

Using the Bregman centroid,

$$I_\varphi(Y) := \min_m \mathbb{E}\big[D_\varphi(Y, m)\big] = \mathbb{E}\big[D_\varphi(Y, \mathbb{E}Y)\big] = \mathbb{E}[\varphi(Y)] - \varphi(\mathbb{E}Y) \geq 0.$$

Gap of the Jensen inequality of φ

Banerjee et al. (2005), Definition 2; Gruber & Buettner (2023), Definition 2.2.

Duality and centroids

slopes become coordinates

Primal and dual means

Because D_φ is asymmetric, there are two centroids!

$$\boxed{\bar{y} = \arg\min_m \mathbb{E}[D_\varphi(Y, m)] = \mathbb{E}Y} \qquad \text{primal mean}$$

$$\boxed{\tilde{y} = \arg\min_m \mathbb{E}[D_\varphi(m, Y)] = \nabla\varphi^*(\mathbb{E}[\nabla\varphi(Y)])} \qquad \text{dual mean}$$

Primal and dual means

Because D_φ is asymmetric, there are two centroids!

$$\boxed{\bar{y} = \arg\min_m \mathbb{E}[D_\varphi(Y, m)] = \mathbb{E}Y} \qquad \text{primal mean}$$

$$\boxed{\tilde{y} = \arg\min_m \mathbb{E}[D_\varphi(m, Y)] = \nabla\varphi^*(\mathbb{E}[\nabla\varphi(Y)])} \qquad \text{dual mean}$$

Generator	primal mean $\bar{y}$	dual mean $\tilde{y}$
$\frac{1}{2}\|x\|^2$	arithmetic mean	arithmetic mean
negative entropy	arithmetic mean of probabilities	softmax(mean logits)

Generalised bias–variance decomposition

Generalised bias–variance decomposition

Let A index a random training procedure (seed, bootstrap, posterior sample), so q_A is a random predictor.

$$\bar{y} := \mathbb{E}[Y], \qquad \tilde{q} := \nabla\varphi^*\big(\mathbb{E}_A \nabla\varphi(q_A)\big).$$

Generalised bias–variance decomposition

Let A index a random training procedure (seed, bootstrap, posterior sample), so q_A is a random predictor.

$$\bar{y} := \mathbb{E}[Y], \qquad \tilde{q} := \nabla\varphi^*\big(\mathbb{E}_A \nabla\varphi(q_A)\big).$$

For any Bregman divergence D_φ, the expected risk decomposes as

$$\mathbb{E}_{A,Y}\, D_\varphi\big(Y, q_A\big) = \underbrace{\mathbb{E}_Y\, D_\varphi\big(Y, \bar{y}\big)}_{\text{aleatoric}} + \underbrace{D_\varphi\big(\bar{y}, \tilde{q}\big)}_{\text{bias}} + \underbrace{\mathbb{E}_A\, D_\varphi\big(\tilde{q}, q_A\big)}_{\text{variance}}.$$

D. Pfau (2013/2025) *A generalized bias-variance decomposition for Bregman divergences.*

Geometric interpretation

aleatoric (noise): $Y \to \bar{y}$, bias: $\bar{y} \to \tilde{q}$, variance: $\tilde{q} \to q_A$.

Example: squared loss

For

$$\varphi(y) = \tfrac{1}{2}\|y\|^2, \qquad D_\varphi(y, q_A) = \tfrac{1}{2}\|y - q_A\|^2,$$

$q_A := \hat{y}$. The dual Bregman centroid of q_A is the ordinary mean:

$$\tilde{q} = \mathbb{E}_A[q_A] =: \bar{q}.$$

Example: squared loss

For

$$\varphi(y) = \tfrac{1}{2}\|y\|^2, \qquad D_\varphi(y, q_A) = \tfrac{1}{2}\|y - q_A\|^2,$$

$q_A := \hat{y}$. The dual Bregman centroid of q_A is the ordinary mean:

$$\tilde{q} = \mathbb{E}_A[q_A] =: \bar{q}.$$

The general decomposition becomes

$$\mathbb{E}_{A,Y}\, \tfrac{1}{2}\big(Y - q_A\big)^2 = \underbrace{\tfrac{1}{2}\,\mathrm{Var}(Y)}_{\text{aleatoric}} + \underbrace{\tfrac{1}{2}\big(\mathbb{E}[Y] - \bar{q}\big)^2}_{\text{bias}^2} + \underbrace{\tfrac{1}{2}\,\mathbb{E}_A\big(\bar{q} - q_A\big)^2}_{\text{variance}}.$$

classical bias–variance identity.

Bias-variance as uncertainty decomposition

In the Bregman decomposition,

$$\mathbb{E}_{A,Y}\, D_\varphi(Y, q_A) = \underbrace{\mathbb{E}_Y\, D_\varphi(Y, \bar{y})}_{\text{aleatoric}} + \underbrace{D_\varphi(\bar{y}, \tilde{q}) + \mathbb{E}_A\, D_\varphi(\tilde{q}, q_A)}_{\text{epistemic}}.$$

Bias-variance as uncertainty decomposition

In the Bregman decomposition,

$$\mathbb{E}_{A,Y}\, D_\varphi(Y, q_A) = \underbrace{\mathbb{E}_Y\, D_\varphi(Y, \bar{y})}_{\text{aleatoric}} + \underbrace{D_\varphi(\bar{y}, \tilde{q}) + \mathbb{E}_A\, D_\varphi(\tilde{q}, q_A)}_{\text{epistemic}}.$$

So epistemic uncertainty has two components:

$$\boxed{\text{epistemic} = \text{bias} + \text{variance}.}$$

Standard estimators capture only the variance term!

Estimated aleatoric and epistemic uncertainty

Bregman information as uncertainty estimate

Recall

$$I_\varphi(Y) := \mathbb{E}\big[D_\varphi(Z, \mathbb{E}Z)\big].$$

Using the definition of D_φ, one obtains

$$\boxed{I_\varphi(Z) = \mathbb{E}[\varphi(Z)] - \varphi(\mathbb{E}Z).}$$

$$\text{epistemic estimate} \;=\; \text{total} \;-\; \text{aleatoric estimate}$$

Example: squared loss

$$I_\varphi(Z) = \mathbb{E}\big[\tfrac{1}{2}(Z - \mathbb{E}Z)^2\big] = \tfrac{1}{2}\ \mathrm{Var}(Z).$$

Assume each ensemble member outputs a Gaussian predictive distribution:

$$\hat{Y}|A \sim \mathcal{N}\big(\mu_A, \sigma_A^2\big).$$

Then the total predictive variance decomposes as

$$\boxed{\mathrm{Var}(\hat{Y}) = \underbrace{\mathbb{E}_A[\sigma_A^2]}_{\text{aleatoric}} + \underbrace{\mathrm{Var}_A(\mu_A)}_{\text{epistemic}}.}$$

Example: log loss

Let q_A be the predictive distribution of a random model A.
Then the generalised variance is

$$I_\varphi(q_A) = \mathbb{E}_A \,\mathrm{KL}\big(q_A \,\|\, \bar{q}\big) = H(\bar{q}) - \mathbb{E}_A[H(q_A)].$$

Hence

$$\underbrace{H(\bar{q})}_{\text{total predictive uncertainty}} = \underbrace{\mathbb{E}_A[H(q_A)]}_{\text{expected conditional entropy}} + \underbrace{I(\hat{Y}; A)}_{\text{mutual information}} .$$

Thank you.

References I

C.E. Shannon (1948). *A mathematical theory of communication.* Bell System Technical Journal 27:379–423.

L.M. Bregman (1967). *The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming.* USSR Comp. Math. Math. Phys. 7(3):200–217.

A. Banerjee, S. Merugu, I.S. Dhillon, J. Ghosh (2005). *Clustering with Bregman divergences.* JMLR 6:1705–1749.

T. Gneiting, A.E. Raftery (2007). *Strictly proper scoring rules, prediction, and estimation.* JASA 102(477):359–378.

D. Pfau (2013/2025). *A generalized bias-variance decomposition for Bregman divergences.* arXiv:2511.08789.

E. Hüllermeier, W. Waegeman (2021). *Aleatoric and epistemic uncertainty in machine learning: an introduction.* Machine Learning 110:457–506.

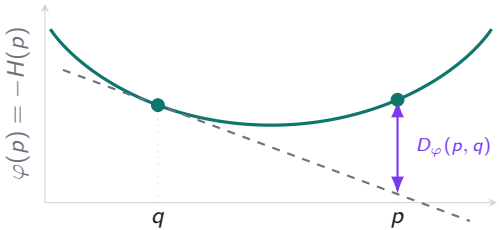
References II

S. Depeweg et al. (2018). *Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning*. ICML.

S. Jiménez, M. Jürgens, W. Waegeman (2025). *Position: Epistemic uncertainty estimation methods are fundamentally incomplete.* arXiv:2505.23506.

Appendix

Example 2: binary negative entropy



$p, q \in (0, 1)$.

$$\varphi(p) = p \log p + (1 - p) \log(1 - p),$$

so

$$\varphi'(q) = \log \frac{q}{1 - q}.$$

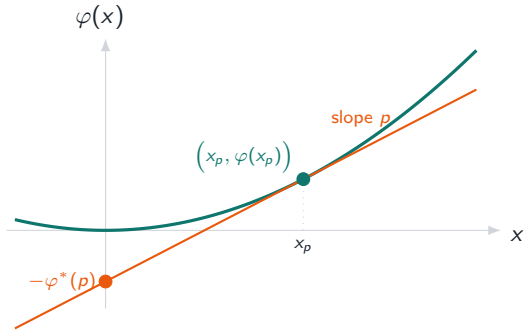
Hence

$$\begin{aligned} D_\varphi(p, q) &= \varphi(p) - \varphi(q) - \varphi'(q)(p - q) \\ &= p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} \\ &= \mathrm{KL}(p \,\|\, q). \end{aligned}$$

A small zoo of generators

Generator φ	Divergence $D_\varphi(x, y)$	Typical use
$\frac{1}{2}\|x\|^2$	$\frac{1}{2}\|x - y\|^2$	regression
$\frac{1}{2}x^\top Ax$	$\frac{1}{2}(x - y)^\top A(x - y)$	Mahalanobis geometry
$\sum_i p_i \log p_i$	$\sum_i p_i \log(p_i/q_i)$	classification
$\sum_i x_i \log x_i$	$\sum_i x_i \log(x_i/y_i) - x_i + y_i$	Poisson/count models
$-\sum_i \log x_i$	$\sum_i \left(x_i/y_i - \log(x_i/y_i) - 1\right)$	spectra, NMF, audio

Legendre transform: from points to slopes



For a convex function φ,

$$\varphi^*(p) = \sup_x \{\langle p, x\rangle - \varphi(x)\}.$$

Example: negative entropy / log loss

For categorical prediction,

$$\varphi(p) = \sum_k p_k \log p_k, \qquad D_\varphi(p, q) = \mathrm{KL}(p \,\|\, q).$$

Let $Y \sim p^*$. The dual centroid of q_A is

$$\tilde{q}_A \propto \exp\big(\mathbb{E}_A \log q_{A,k}\big),$$

i.e. the normalised geometric mean.

Example: negative entropy / log loss

For categorical prediction,

$$\varphi(p) = \sum_k p_k \log p_k, \qquad D_\varphi(p, q) = \mathrm{KL}(p \,\|\, q).$$

Let $Y \sim p^*$. The dual centroid of q_A is

$$\tilde{q}_A \propto \exp\big(\mathbb{E}_A \log q_{A,k}\big),$$

i.e. the normalised geometric mean. Then

$$\mathbb{E}_{A,Y}\big[-\log q_A(Y)\big] = \underbrace{H(p^\star)}_{\text{aleatoric}} + \underbrace{\mathrm{KL}\big(p^\star \,\|\, \tilde{q}_A\big)}_{\text{bias}} + \underbrace{\mathbb{E}_A\,\mathrm{KL}\big(\tilde{q}_A \,\|\, q_A\big)}_{\text{variance}}.$$

For log loss, the natural ensemble centroid is geometric, not arithmetic.

Connection to exponential families

A regular exponential family has the form

$$q_\theta(x) = h(x) \exp(\langle \theta, T(x) \rangle - A(\theta)), \qquad \mu(\theta) = \nabla A(\theta).$$

We have

$$\boxed{-\log q_\theta(x) = D_\varphi(T(x), \mu(\theta)) + c(x).}$$

Banerjee, Merugu, Dhillon, Ghosh (2005), Theorem 4 and Theorem 6.

Two familiar variances

squared generator

$$\varphi(z) = \tfrac{1}{2}\|z\|^2$$

$$I_\varphi(Y) = \tfrac{1}{2}\,\mathbb{E}\|Y - \mathbb{E}Y\|^2$$

$$= \tfrac{1}{2}\,\mathrm{Var}(Y)$$

negative entropy

$$\varphi(p) = \sum_k p_k \log p_k$$

$$I_\varphi(P) = H(\mathbb{E}P) - \mathbb{E}H(P)$$

$$= I(Y; A)$$

Entropy: one of many

every Bregman generator has an entropy

Cross-entropy decomposition — for any generator

Define

$$H_\varphi(p) \;:=\; -\varphi(p), \qquad H_\varphi(p,q) \;:=\; -\varphi(q) - \langle \nabla\varphi(q),\, p-q \rangle.$$

Cross-entropy decomposition — for any generator

Define

$$H_\varphi(p) \;:=\; -\varphi(p), \qquad H_\varphi(p,q) \;:=\; -\varphi(q) - \langle \nabla\varphi(q),\, p-q\rangle.$$

Then directly from the definition of D_φ:

$$D_\varphi(p,q) \;=\; H_\varphi(p,q) \;-\; H_\varphi(p).$$

Cross-entropy decomposition — for any generator

Define

$$H_\varphi(p) \;:=\; -\varphi(p), \qquad H_\varphi(p,q) \;:=\; -\varphi(q) - \langle \nabla\varphi(q),\, p-q \rangle.$$

Then directly from the definition of D_φ:

$$D_\varphi(p,q) \;=\; H_\varphi(p,q) \;-\; H_\varphi(p).$$

For Shannon's φ this is the familiar

$$\mathrm{KL}(p \,\|\, q) \;=\; \underbrace{H(p,q)}_{\text{cross-entropy}} \;-\; \underbrace{H(p)}_{\text{entropy}}.$$

Each Bregman divergence has its own entropy, cross-entropy, and likelihood.

Many entropies

Entropy	Formula	Yields divergence
Shannon	$-\sum_i p_i \log p_i$	KL
Rényi-α	$\frac{1}{1-\alpha} \log \sum_i p_i^\alpha$	Rényi-α
Tsallis-q	$\frac{1}{q-1}\left(1 - \sum_i p_i^q\right)$	β-divergence
Min-entropy	$-\log \max_i p_i$	—
Differential	$-\int p \log p \, d\mu$	continuous KL
Von Neumann	$-\mathrm{tr}(\rho \log \rho)$	quantum relative entropy

Each strictly convex generator $\Rightarrow$ entropy, divergence, exponential family.

Why Shannon entropy in particular?

Shannon entropy is picked out by three independent characterisations:

- Coding (Shannon, 1948): the expected code-length to communicate X is bounded below by $H(X)$.
- Locality (Bernardo, 1979): $-\log q(y)$ is the unique strictly proper scoring rule that depends on q only through its value at the realised outcome.
- Chain rule (Csiszár, 1991): KL is the unique f-divergence satisfying

$$\mathrm{KL}(p_{XY} \,\|\, q_{XY}) = \mathrm{KL}(p_X \,\|\, q_X) + \mathbb{E}_X[\mathrm{KL}(p_{Y\,|\,X} \,\|\, q_{Y\,|\,X})].$$

Other entropies (Rényi, Tsallis, min-entropy) drop one of these axioms.