Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

16 Information Measures for Continuous Distributions, and Exponential Families Revisited

Our recent developments have emphasized the case of discrete random variables, and thus the information measures that have arisen have been defined for such discrete distributions. We now show that our development naturally extends to continuous random variables, and leads to similar information measures. In generalizing our development, we also show some important invariance properties of these measures.

Finally, as a simple application, we use the resulting tools to establish that for distributions over continuous alphabets one can always construct an exponential family such that every sufficiently light-tailed distribution in the interior of the simplex is arbitrarily close, in the sense of divergence, to a member of the family.

16.1 Differential Entropy

When working with continuous random variables, entropy is replaced with differential entropy, defined via

$$h(\mathsf{x}) \triangleq -\int_{-\infty}^{\infty} p_{\mathsf{x}}(x) \log p_{\mathsf{x}}(x)\, \mathrm{d}x. \tag{1}$$

Unlike the discrete case, differential entropy is defined only for numeric random variables. It is no longer nonnegative in general, as the following example illustrates.

Example 1. Let x be uniformly distributed over the interval $(0, \Delta)$, for some $\Delta > 0$. Then $h(\mathsf{x}) = \log \Delta$, which is negative if $\Delta < 1$.

In the discrete case, entropy is completely invariant to the labeling of the alphabet. By contrast, differential entropy is *not* invariant to coordinate transformations. To see this, let $\mathsf{x} = g(\mathsf{s})$, where g is some monotonically increasing (and differentiable) mapping.[1] Then $p_{\mathsf{x}}(x) = p_{\mathsf{s}}(s)/g'(s)$ with $s = g^{-1}(x)$, where $g'(s)$ is the Jacobian of the transformation. However, in this case, we have

$$\begin{aligned} h(\mathsf{x}) &= -\int p_{\mathsf{x}}(x) \log p_{\mathsf{x}}(x)\, \mathrm{d}x && (2)\\ &= -\int p_{\mathsf{s}}(s) \frac{1}{g'(s)} \log \left[\frac{p_{\mathsf{s}}(s)}{g'(s)}\right] g'(s)\, \mathrm{d}s && (3)\\ &= h(\mathsf{s}) + \mathbb{E}\left[\log g'(\mathsf{s})\right]. && (4) \end{aligned}$$

[1] While monotonicity is imposed so that the mapping is invertible, the constraint that the mapping be increasing is just for simplicity of exposition. A monotonically decreasing mapping is equally straightforward to analyze, and results in the same conclusion.

Hence, if, for example, $\mathsf{x} = \rho\mathsf{s}$ for some scaling parameter ρ, then (4) implies $h(\mathsf{x}) = h(\mathsf{s}) + \log\rho$.

Hence, while differential entropy is certainly a convenient quantity in our development of inference with continuous distributions, the lack of invariance makes it ultimately a somewhat less fundamental quantity than it might seem at first glance.

Conditional differential entropy is defined in an analogous manner. In particular, (1) implies that

$$h(\mathsf{x}|\mathsf{y}=y) = -\int_{-\infty}^{+\infty} p_{\mathsf{x}|\mathsf{y}}(x|y)\log p_{\mathsf{x}|\mathsf{y}}(x|y)\,\mathrm{d}x,$$

from which we obtain

$$h(\mathsf{x}|\mathsf{y}) \triangleq \sum_{y} p_{\mathsf{y}}(y)\,h(\mathsf{x}|\mathsf{y}=y) = -\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} p_{\mathsf{x},\mathsf{y}}(x,y)\log p_{\mathsf{x}|\mathsf{y}}(x|y)\,\mathrm{d}x\,\mathrm{d}y. \tag{5}$$

16.2 Mutual Information

In the case of continuous random variables, mutual information can be defined in terms of differential entropies via

$$I(\mathsf{x};\mathsf{y}) = h(\mathsf{x}) - h(\mathsf{x}|\mathsf{y}), \tag{6}$$

which, unlike differential entropy, remains nonnegative, as we will discuss shortly.

In addition, when x is continuous (and whether or not y is discrete or continuous), mutual information $I(\mathsf{x};\mathsf{y})$ *is* invariant to coordinate transformations. To see this, let $\mathsf{x} = g(\mathsf{s})$ as above, so

$$h(\mathsf{x}) = h(\mathsf{s}) + \mathbb{E}\left[\log g'(\mathsf{s})\right]. \tag{7}$$

Analogously, we obtain

$$h(\mathsf{x}|\mathsf{y}=y) = h(\mathsf{s}|\mathsf{y}=y) + \mathbb{E}\left[\log g'(\mathsf{s})|\mathsf{y}=y\right], \tag{8}$$

from which we obtain, using iterated expectation,

$$h(\mathsf{x}|\mathsf{y}) = \int h(\mathsf{x}|\mathsf{y}=y)\,p_{\mathsf{y}}(y)\,\mathrm{d}y = h(\mathsf{s}|\mathsf{y}) + \mathbb{E}\left[\log g'(\mathsf{s})\right]. \tag{9}$$

Finally, combining (7) and (9) we obtain

$$\begin{aligned} I(\mathsf{x};\mathsf{y}) &= h(\mathsf{x}) - h(\mathsf{x}|\mathsf{y}) \\ &= \left(h(\mathsf{s}) + \mathbb{E}\left[\log g'(\mathsf{s})\right]\right) - \left(h(\mathsf{s}|\mathsf{y}) + \mathbb{E}\left[\log g'(\mathsf{s})\right]\right) \\ &= h(\mathsf{s}) - h(\mathsf{s}|\mathsf{y}) \\ &= I(\mathsf{s};\mathsf{y}) \end{aligned} \tag{10}$$

We can interpret (10) as meaning that mutual information is a coordinate-free quantity.

16.3 Information Divergence

Likewise, in the case of continuous distributions, information divergence is defined via

$$D(p\|q) \triangleq \int_{-\infty}^{+\infty} p(x) \log \frac{p(x)}{q(x)} \, \mathrm{d}x. \tag{11}$$

It is straightforward to verify that Gibbs' inequality still holds in the continuous case, via which we obtain that divergence continues to be nonnegative in this case.

In turn, this establishes the preceding claim that $I(\mathsf{x};\mathsf{y}) \geq 0$ even when the variables is continuous, since $I(\mathsf{x};\mathsf{y}) = D(p_{\mathsf{x},\mathsf{y}}\|p_{\mathsf{x}}\, p_{\mathsf{y}}) \geq 0$. Moreover, since $I(\mathsf{x};\mathsf{y}) = h(\mathsf{x}) - h(\mathsf{x}|\mathsf{y})$, an additional consequence is that conditioning can still never increase differential entropy, i.e., $h(\mathsf{x}|\mathsf{y}) \leq h(\mathsf{x})$. even though differential entropy may be positive or negative.

Finally, note that divergence, too, is a coordinate-free transformation. In particular, using the same coordinate transformation g as before we have

$$\begin{aligned} D(p_{\mathsf{x}}\|q_{\mathsf{x}}) &= \int p_{\mathsf{x}}(x) \log \frac{p_{\mathsf{x}}(x)}{q_{\mathsf{x}}(x)} \, \mathrm{d}x && (12)\\ &= \int \frac{p_{\mathsf{s}}(s)}{g'(s)} \log \frac{p_{\mathsf{s}}(s)/g'(s)}{q_{\mathsf{s}}(s)/g'(s)} g'(s) \, \mathrm{d}s && (13)\\ &= D(p_{\mathsf{s}}\|q_{\mathsf{s}}). && (14) \end{aligned}$$

16.4 Inference and Modeling with Continuous Alphabets

When extending our modeling results to the case of continuous x, the mixture model takes the form

$$q_w(y) = \int_{x\in\mathcal{X}} w(x)\, p_{\mathsf{y}}(y;x) \, \mathrm{d}x. \tag{15}$$

In this case the model capacity remains $C = \max_{p_{\mathsf{x}}} I(\mathsf{x};\mathsf{y})$, but now mutual information is most naturally expanded in terms of differential entropies. That mutual information is independent of the parameterizations has an important (and reassuring) implication in our inference application—that model capacity is unaffected by reparameterization. In other words, model capacity is unaffected by how we index the models in the class of candidates.

We conclude with example application of divergence for continuous variables.

16.5 Approximating Distributions by Exponential Families

We are now equipped to return to a question we raised during our introduction to exponential families. In particular, we claimed that there is a sequence of exponential families such that any sufficiently well-behaved distribution over a continuous alphabet $\mathcal{Y} \subset \mathbb{R}$ is arbitrarily close to a distribution in this sequence of families. From our

development of the role of information measures in inference, it is clear that the right measure of closeness from the perspective of inference is information divergence.

To proceed, we must define a class of suitably well-behaved distributions whose members we seek to approximate. While larger classes are possible, a convenient (convex) class for our purposes are the light-tailed distributions that are bounded, strictly positive, and decay at least as fast as a Gaussian distribution. Formally,[2]

$$\mathcal{Q}_{\mathbb{N}+} = \left\{ q \colon 0 < q(y) < \infty \text{ for all } y, \text{ and } \exists \sigma > 0, y_0 \geq 0 \text{ s.t. } \sigma\, q(\sigma y) \leq e^{-y^2} \text{ for } |y| > y_0 \right\}, \tag{16}$$

where we note that

$$\tilde{q}^{\sigma}(y) = \sigma\, q(\sigma y) \tag{17}$$

is a rescaled version of the distribution q.[3]

For any $q \in \mathcal{Q}_{\mathbb{N}+}$, we have

$$\begin{aligned}
\int e^{-y^2} \, |\ln q(y)|^2 \, \mathrm{d}y &= \int_{|y|>y_0} e^{-y^2} \, |\ln q(y)|^2 \, \mathrm{d}y + \int_{|y|\leq y_0} e^{-y^2} \, |\ln q(y)|^2 \, \mathrm{d}y \\
&\leq \int_{|y|>y_0} e^{-y^2} \left| \frac{y^2}{\sigma^2} + |\ln \sigma| \right|^2 \mathrm{d}y + \int_{|y|\leq y_0} e^{-y^2} \, |\ln q(y)|^2 \, \mathrm{d}y \\
&\leq \int e^{-y^2} \left| \frac{y^2}{\sigma^2} + |\ln \sigma| \right|^2 \mathrm{d}y + 2y_0 \, e^{-y_0^2} \max_{\{y \in \mathbb{R} \colon |y| \leq y_0\}} |\ln q(y)|^2 \\
&< \infty.
\end{aligned}$$

Hence, $\ln q$ for $q \in \mathcal{Q}_{\mathbb{N}+}$ lies in the Hilbert space of functions with inner product

$$\langle f, g \rangle = \int e^{-y^2} \, f(y) \, g(y) \, \mathrm{d}y, \tag{18}$$

and thus $\ln q(y)$ can be expanded in an orthonormal basis of the form

$$\ln q(y) = \sum_{k=0}^{\infty} x_k \, \phi_k^{\mathrm{H}}(y), \qquad y \in \mathcal{Y}, \tag{19}$$

where

$$\phi_k^{\mathrm{H}}(y) = (-1)^k \, e^{y^2} \, \frac{\mathrm{d}^k}{\mathrm{d}y^k} e^{-y^2}, \qquad k = 0, 1, 2, \ldots \tag{20}$$

are the so-called *Hermite polynomials*, and where the coefficients are given by

$$x_k = \frac{1}{\sqrt{\pi}\, 2^k \, k!} \int e^{-y^2} \, \ln q(y) \, \phi_k^{\mathrm{H}}(y) \, \mathrm{d}y.$$

[2] This is a subset of what are sometimes referred to as *sub-Gaussian* distributions.

[3] It is straightforward to verify that distributions with, e.g., rescalings such that $\tilde{q}^{\sigma}(y) \leq e^{-y^2 + o(y^2)}$, which include shifted Gaussians, are contained in $\mathcal{Q}_{\mathbb{N}+}$.

Defining

$$q_K(y) \triangleq \exp\left[\sum_{k=0}^{K} x_k t_k(y)\right]$$

with

$$t_k(y) = \phi_k^{\mathrm{H}}(y), \qquad k = 0, 1, 2, \ldots,$$

this means that $q_K(\cdot)$ for $K = 0, 1, 2, \ldots$ is a sequence of distributions in a corresponding sequence of K-dimensional linear exponential families $\mathtt{E}_K$ defined via

$$p_{\mathsf{y}}(y; \mathbf{x}) = \exp\left[\sum_{k=0}^{K} x_k\, t_k(y) - \alpha_K(\mathbf{x})\right], \qquad K = 0, 1, 2, \ldots \tag{21}$$

such that, as (18) and (19) imply,

$$\int e^{-y^2} \left|\ln \frac{q(y)}{q_K(y)}\right|^2 \mathrm{d}y = \int e^{-y^2} \left|\ln q(y) - \ln q_K(y)\right|^2 \mathrm{d}y \to 0 \quad \text{as } K \to \infty. \tag{22}$$

Since the natural measure of closeness in the approximation for the purposes of inference is the information divergence of q_K from q, it remains to show that (22) implies

$$D(q\|q_K) \to 0 \quad \text{as } K \to \infty. \tag{23}$$

Proceeding, since, as developed in Section 16.3, divergence is a coordinate-free measure and thus, e.g., invariant to rescalings, we have

$$D(q\|q_K) = D(\tilde{q}^{\sigma}\|\tilde{q}_K^{\sigma}), \tag{24}$$

for any $\sigma > 0$, where $\tilde{q}$ is as defined in (17) and $\tilde{q}_K^{\sigma}(y) \triangleq \sigma\, q_K(\sigma y)$. In turn, since scaled exponential families are themselves exponential familes, we can restrict our attention to verifying (23) for the rescaled distributions, and so without loss of generality we focus on q in $\mathcal{Q}_{\mathtt{N}+}$ such that $\sigma = 1$.

Proceeding, for $q \in \mathcal{Q}_{\mathtt{N}+}$ with $\sigma = 1$ and some $y_0 \geq 0$, we write

$$\begin{aligned}
D(q\|q_K) &\leq \int_{|y|>y_0} q(y) \left|\ln \frac{q(y)}{q_K(y)}\right| \mathrm{d}y + \int_{|y|\leq y_0} q(y) \left|\ln \frac{q(y)}{q_K(y)}\right| \mathrm{d}y && (25)\\
&\leq \int_{|y|>y_0} e^{-y^2} \left|\ln \frac{q(y)}{q_K(y)}\right| \mathrm{d}y + \int_{|y|\leq y_0} q(y) \left|\ln \frac{q(y)}{q_K(y)}\right| \mathrm{d}y && (26)\\
&\leq \underbrace{\int e^{-y^2} \left|\ln \frac{q(y)}{q_K(y)}\right| \mathrm{d}y}_{D_+} + \left(\max_{|y|\leq y_0} q(y)\right) \underbrace{\int_{|y|\leq y_0} \left|\ln \frac{q(y)}{q_K(y)}\right| \mathrm{d}y}_{D_-}, && (27)
\end{aligned}$$

where to obtain (25) we have used the triangle inequality, and where to obtain (26) we have used (16).

Using the Cauchy-Schwarz inequality, the first integral on the right-hand side of (27) is bounded according to

$$D_+^2 \leq \int e^{-y^2}\,\mathrm{d}y \int e^{-y^2} \left|\ln \frac{q(y)}{q_K(y)}\right|^2 \mathrm{d}y = \sqrt{\pi} \int e^{-y^2} \left|\ln \frac{q(y)}{q_K(y)}\right|^2 \mathrm{d}y, \tag{28}$$

so using (22) we see $D_+ \to 0$ as $K \to \infty$.

In turn, the second integral on the right-hand side of (27) is bounded according to

$$\begin{aligned} D_-^2 &\leq 2y_0 \int_{|y|<y_0} \left|\ln \frac{q(y)}{q_K(y)}\right|^2 \mathrm{d}y && (29)\\ &\leq 2y_0\, e^{y_0^2} \int_{|y|<y_0} e^{-y^2} \left|\ln \frac{q(y)}{q_K(y)}\right|^2 \mathrm{d}y && (30)\\ &\leq 2y_0\, e^{y_0^2} \int e^{-y^2} \left|\ln \frac{q(y)}{q_K(y)}\right|^2 \mathrm{d}y, && (31) \end{aligned}$$

where to obtain (29) we have also used the Cauchy-Schwarz inequality. Using (31) with (22) we see $D_- \to 0$ as $K \to \infty$.

Since both D_+ and D_- both converge to zero, (23) follows, i.e., for any $q \in \mathcal{Q}_{\mathtt{N}+}$ we can construct an exponential family with a member arbitrarily close in divergence from q.