Short Notes on Divergence Measures

Danilo Jimenez Rezende

1. Intro

The problem of density estimation is at the foundation of basic sciences, statistics and information theory.

Not surprisingly, this problem receives substantial attention from the ML and DL communities, where it is a hot topic, particually in the context of high-dimensional data such as images, audio and text.

Density estimation requires introducing relevant measures of divergence between probability densities that we can compute efficiently.

Historically many divergences have been proposed, but only a small subset of them may be viable to use on a specific problem due to a combination of computational and other practical reasons which we try to expand below.

In what follows, we are interested in constructing a divergence $D(Q;P)$ between two smooth probability densities $Q\colon \mathbb{R}^d \to \mathbb{R}^+$ and $P\colon \mathbb{R}^d \to \mathbb{R}^+$, where Q is defined by a parametric generative model with parameters $\theta \in \mathbb{R}^l$ and P is a target density. The goal is to both have a consistent interpretation of $D(Q;P)$ as well as use it to modify the parameters θ so that Q aproaches P.

1.1. Divergences, Metrics and Measures of Relative-Information

A pragmatic view common in ML is that all we need is *some* measure $D(Q;P)$ of divergence between two probability densities P and Q and, as long as it has the properties that $D(Q;P) = 0 \Leftrightarrow Q = P$ and $D(Q;P) \geqslant 0$ it should be sufficient for optimization.

Although this view is not strictly mathematically wrong, there is much more nuance to add to the discussion. Importantly, we should note a distiction between an arbitrary divergence measure, a metric and a relative-information measure.

A metric is a divergence $D(Q,P)$ that also satisfies an additional set of assumptions: Symmetry, $D(Q,P) = D(P,Q)$ and Triangle inequality, $D(A,C) \leqslant D(A,B) + D(B,C)$. These supplementary assumptions can have profound implications for optimization and interpretability.

1.1.1. Measures of Relative-Information

One question one might ask is why is the Kullback-Leibler divergence, KL (also known as relative entropy or relative information), so popular in ML given that there are many other divergences? Is that merely a historical bias or there is more to it?

The most important reason for wanting a divergence that is also a measure of relative information is that we want to *quantify surprise or changes in a belief state in a consistent manner*, that is exactly what is achieved by the KL-divergence.

A general reasoning for deriving the expression of the Kullback-Leibler divergence $\mathrm{KL}(Q;P)$ as a way of measuring relative information is by addressing the following question: What is the expected number of bits that we have to transmit to a receiver in order to communicate the density Q given that the receiver already knows the density P?

From coding theory [14, 10, 16, 15] we know that minimum number of bits (codelength) that we need to compress a data point $x \sim P$ is $\lfloor \log_2 p(x) \rfloor$. In order to communicate the density Q to someone who already knows the density P we have to comunicate the difference in codelength $\lfloor \log_2 q(x) \rfloor - \lfloor \log_2 p(x) \rfloor$ bits for every datapoint. On expectation, we will be transmitting $\mathbb{E}_q[\lfloor \log_2 q(x) \rfloor - \lfloor \log_2 p(x) \rfloor] \approx \mathrm{KL}(Q;P)$ bits per datapoint. This shows that the KL divergence has a very concrete meaning: it is the average number of bits that a source must transmit to a receiver in order to communicate the density Q.

Another view of relative information (nicely summarized in [3]) is based on the observation that one can axiomatize how probabilities representing the belief about the state of a physical system should change in face of new evidence and these axioms can be constructed on the basis of universal principles of consistency [17, 6, 19, 18, 8].

The axioms summarized in [3] enforce that a reasonable measure of relative information or divergence between a target density Q and a reference density P should have the properties:

i. Locality (local effects must have local consequences). This enforces that a probability divergence $D(Q;P)$ must be of the form $D(Q;P) = \int f(q(x), p(x), x)dx$, where f is an arbitrary function. That is, it only compares $q(x)$ to $p(x')$ at $x = x'$. Note that, as far as locality is concerned, we could also have included local dependecies on the gradients of p and q, e.g. $D(Q;P) = \int f(q(x), p(x), x, \nabla q(x), \nabla p(x))dx$;

ii. Coordinate invariance (the coordinate system used to express the probability densities containts no information). This enforces that, upon a change of measure induced by an invertible map ϕ, the divergence should be invariant. That is, for divergences of the form $D(Q;\, P) = \int f(q(x),\, p(x),\, x)dx$, we would have that $D(\tilde{Q};\, \tilde{P}) = \int f(\tilde{q}(\tilde{x}),\, \tilde{p}(\tilde{x}),\, \tilde{x})d\tilde{x} = \int f\left(\frac{(q \circ \phi^{-1})(\tilde{x})}{\left|\det \frac{\partial \phi}{\partial x}\right|}, \frac{(p \circ \phi^{-1})(\tilde{x})}{\left|\det \frac{\partial \phi}{\partial x}\right|}, \phi^{-1}(\tilde{x}) \right) d\tilde{x} := D(Q;P)$. This constraint enforces that the function f can no longer be arbitrary, but must be of the form $f(q(x), p(x), x) := f\left(\frac{q(x)}{p(x)}\right)\mu(x)$ where f is a scalar and $\mu(x)$ must transform like a density (e.g. like $p(x)$ or $q(x)$ under a change of coordinates). Therefore such divergences can only be of the form $D(Q;\, P) = \int f\left(\frac{q(x)}{p(x)}\right) q(x)dx$ or $\int f\left(\frac{q(x)}{p(x)}\right) p(x)dx$ in order to satisfy this axiom. This axiom can also be interpreted as the continuous generalization of the permutation invariance axiom from information theory;

iii. Subsystem independence (additivity under independent sub-domains). This uniquely constrains the function f to be $f(x) = \log(x)$.

It can be shown that the relative differential entropy, $\mathrm{KL}(Q;P)$, is the only divergence satisfying axioms (i-iii), [3].

In this sense, the divergence $\mathrm{KL}(Q;\, P)$ is indeed a special divergence and arguably the only resonable measure of relative information.

By relaxing some of the universal consistency axioms, we obtain more general families of divergences. For instance,

- The family of f-divergences $D_f(Q;P) = \int f\Big(\frac{q(x)}{p(x)}\Big)p(x)dx$ [5, 1, 12, 11], where $f(x)$ is an arbitrary strictly convex function, violates axiom (iii) when $f(x) \neq x\,\log(x)$ but it still satisfies axioms (i-ii);

- The Stein divergence $D(Q;P) = \sup_{f\in\mathcal{F}} \mathbb{E}_q[\nabla \log p(x)\, f(x) + \nabla f(x)]^2$, where $\mathcal{F}$ is the space of smooth functions for which $\mathbb{E}_p[\nabla \log p(x)\, f(x) + \nabla f(x)] = 0$ (smooth functions that vanish in the infinity) [20, 21]. Violates axioms (ii and iii);

- The Cramer/energy distance $D(Q;P) = 2\mathbb{E}[\|x-y\|] - \mathbb{E}[\|x-x'\|] - \mathbb{E}[\|y-y'\|]$, where x, $x' \sim P$ and y, $y' \sim Q$, [4], violates all 3 axioms. By replacing the Euclidean norm $\|\|$ with the *geodesic distance*, we would obtain a divergence that satisfies axiom (ii) but this may result in negative "distances" [9];

- The Wassertein distance $D_p(Q;P) = [\inf_\rho \int dx dx' \|x-x'\|^p \rho(x,x')]^{\frac{1}{p}}$, where $\rho: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ is a density with marginals $\rho(x) = q(x)$ and $\rho(x') = p(x')$ [2, 13]. Violates all 3 axioms. By replacing the Euclidean norm $\|\|$ with the *geodesic distance*, we would obtain a divergence that satisfies axiom (ii);

- The Fisher distance $D(Q;P) = \mathbb{E}_p[||\nabla_x \ln p(x) - \nabla_x \ln q(x)||^2]$, [7]. It is local, so it satisfies axiom (i) and it can be generalized to metric spaces to be invariant, satisfying axiom (ii).

1.2. Building Invariant Divergences

Invariance under a change of measure or coordinate system is a fundamental concept both in information theory and physics. It is widely accepted in both fields that the coordinate system employed to express concepts carries no information. This brings us back to the axiom (ii) of previous section, illustrated in Figure (1).

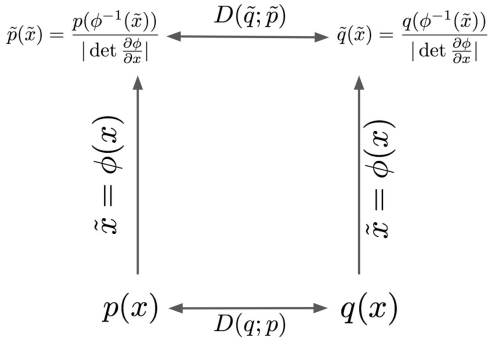

Figure 1. Diagram relating $p(x)$ and $q(x)$ with their twins $\tilde{p}(\tilde{x})$ and $\tilde{q}(\tilde{x})$ in a different coordinate system induced by an invertible smooth map $\phi(x)$. The principle of invariance under change of measures postulates that $D(q;p) = D(\tilde{q};\tilde{p})$ for any ϕ.

To develop a better understanding of how invariant divergences can be constructed, we show in Table (1) how a few relevant quantities transform under a change of coordinates $\tilde{x} = \phi(x)$.

Name	Expression	Expression in new coordinate system
coordinates	x	$\tilde{x} = \phi(x)$
measure	dx	$d\tilde{x} = \left\lvert \det \frac{\partial \phi}{\partial x} \right\rvert dx$
scalar function	$f(x)$	$\tilde{f}(\tilde{x}) = (f \circ \phi^{-1})(\tilde{x})$
scalar density	$p(x)$	$\tilde{p}(\tilde{x}) = \frac{(p \circ \phi^{-1})(\tilde{x})}{\left\lvert \det \frac{\partial \phi}{\partial x} \right\rvert}$
gradient operator	∇_x	$\tilde{\nabla}_{\tilde{x}} = \frac{\partial x}{\partial \tilde{x}} \nabla_x = \left(\frac{\partial \phi}{\partial x} \right)^{-1} \nabla_x$
metric tensor on a tangent space	$G_{ij} = g\left(\frac{\partial}{\partial x_i}, \frac{\partial}{\partial x_j} \right)$	$\tilde{G} = \left(\left(\frac{\partial \phi}{\partial x} \right)^{-1} \right)^T G \left(\frac{\partial \phi}{\partial x} \right)^{-1}$
inverse metric	G^{-1}	$\tilde{G}^{-1} = \frac{\partial \phi}{\partial x} G^{-1} \left(\frac{\partial \phi}{\partial x} \right)^T$
score "function"	$\nabla_x \ln p(x)$	$\tilde{\nabla}_{\tilde{x}} \ln \tilde{p}(\tilde{x}) = \left(\frac{\partial \phi}{\partial x} \right)^{-1} \nabla_x \ln \frac{(p \circ \phi^{-1})(x)}{\left\lvert \det \frac{\partial \phi}{\partial x} \right\rvert}$

Table 1. Examples of how a few relevant quantities change after a change of coordinates induced by an invertible smooth map ϕ.

From Table (1) it is clear why any divergence of the form $D(Q;P) = \int f(q(x), p(x), x)dx$ can only be invariant if $D(Q;P) = \int f(q(x), p(x), x)dx = \int f\left(\frac{q(x)}{p(x)}\right) q(x)dx$: First the measure dx alone is not invariant, but the product of the measure by a scalar density $q(x)dx$ is; Second, since both Q and P transform in the same multiplicative way, we can produce an invariant quantity by taking their ratio.

This explains why probability ratios are so ubiquitous in statistics: its a consequence of the invariance under change of measure.

What about divergences that also contain gradient terms such as the score function? As shown in Table (1), gradients transform as *covariant tensors*.

The only way to build invariants from covariant tensors is to contract them with *contravariant tensors* (such as the inverse metric) to form scalar functions.

For example, the quantity $(\nabla_x \ln p(x))^T \nabla_x \ln p(x)$ is not an invariant as we are contracting two covariant tensors together. In order to make it invariant, we need to introduce a metric tensor g, with components G, and sandwich its inverse in between the gradients: $(\nabla_x \ln p(x))^T G^{-1} \nabla_x \ln p(x)$. Similarly for terms involving cross score functions such as $(\nabla_x \ln p(x))^T G^{-1} \nabla_x \ln q(x)$.

This reasoning demonstrates that the Fisher divergence, $D(Q;\ P) = \mathbb{E}_p[||\nabla_x \ln\ p(x) - \nabla_x \ln\ q(x)||^2]$, will only be invariant if we replace the L^2 norm by a contraction with the inverse metric tensor.

2. Case-by-case Analysis

Here we assume to have a target density P: $\mathbb{R}^d \to \mathbb{R}^+$ and a model density Q_θ: $\mathbb{R}^d \to \mathbb{R}^+$ parametrized by a vector $\theta \in \mathbb{R}^l$, both belonging to the set of smooth probability densties $\mathcal{F}$.

We are concerned in characterizing a divergence D: $\mathcal{F}^2 \to \mathbb{R}^+$ by its ability to be used in real-world applications. Assuming that we will use a gradient-based optimization algorithm with gradients of the form

$$\nabla_\theta D(Q_\theta; P) \ \propto \mathbb{E}[G_\theta(x)],$$

where $x \in \mathbb{R}^d$ is a datapoint and the expectation is either with respect to P or Q or both (in this later case the gradient estimator G_θ will depend on multiple points $x_1, ..., x_K$).

We can write a general estimator of $\nabla_\theta D(Q_\theta; P)$ as an expectation under Q_θ,

$$\begin{aligned}\nabla_\theta D(Q_\theta; P) &= \int dx \frac{\delta D(Q_\theta; P)}{\delta Q_\theta(x)} \nabla_\theta Q_\theta(x) \\ &= \mathbb{E}_{Q_\theta}\left[\frac{\delta D(Q_\theta; P)}{\delta Q_\theta(x)} \nabla_\theta \ln Q_\theta(x) \right],\end{aligned}$$

where $\frac{\delta D(Q_\theta; P)}{\delta Q_\theta(x)}$ is the functional derivative of the functional $D(Q_\theta; P)$ with respect to Q_θ at the point x. In this derivation, the first line is the chain-rule for functional derivatives while the second line is just an elementary manipulation, $\nabla_\theta Q_\theta(x) = Q_\theta(x) \nabla_\theta \ln Q_\theta(x)$. Although unbiased, this estimator is not always helpful for computational reasons.

In practice, each real-world application will require different assumptions about what can be known about P, Q and D ***at an acceptable computational or physical budget***.

We will characterize the available knowledge about P and Q using the definitions below:

- $\mathbb{S}$: The set of probability densities that we can cheaply sample from;
- $\mathbb{N}$: The set of probability densities for which the normalized likelihood is known;
- $\mathbb{U}$: The set of probability densities for which only the unnormalized likelihood is known.

We will characterize the available knowledge about a divergence $D(Q; P)$ using the definitions below:

- UBG: The set of divergences for which unbiased gradient estimators are known (gradients with respect to Q);
- BG: The set of divergences for which only biased gradient estimators are known (gradients with respect to Q);
- ICM: The set of divergences that is invariant under a change of measure;
- NICM: The set of divergences that is not invariant under a change of measure;
- M: The set of divergences that are also metrics.

For example, if we can cheaply sample from P and compute its likelihood, we will indicate this by $P \in \mathbb{N} \bigcap \mathbb{S}$, where $\bigcap$ stands for the intersection operator. Conversely, if all we can know about P with our given computational budget is its likelihood up to a normalization constant, we will indicate this by $P \in \mathbb{U}$.

2.1. When $P \in \mathbb{S}$ and $Q \in \mathbb{N} \bigcap \mathbb{S}$

Historically, this is the typical case for density estimation where only samples $x \in \mathbb{R}^d$ from a data distribution P are known and our model defines a parametric density Q_θ with parameters $\theta \in \mathbb{R}^l$.

Given what is known about P and Q_θ, we would like to use a divergence D with the following properties:

1. Its value can be approximated up to a constant from samples of P only. The estimator is allowed to use the numerical values of the likelihood of Q_θ;
2. There is an unbiased estimator of its gradients with respect to θ. This gradient estimator can be approximated from samples of P and the numerical values of the likelihood of Q_θ. This is a strong constraint, it essentially forces the divergence D to not contain any non-linear function of an expectation under Q_θ;

3. Both (1) and (2) can be estimated from samples in $O(n)$ algorithmic complexity in the number of samples.

- Among the family of f-divergences, the moment-matching $\mathrm{KL}(P;Q)$, is the only f-divergence satisfying properties (1) and (2);
- Jensen-Shannon divergence is not applicable;
- The Cramer/energy distance is applicable satisfying (1) and (2);
- The Wasserstein distance is applicable but has biased gradients (violates 2);
- MMD distance is applicable but is expensive (violates 3);
- The Stein divergence is applicable but has biased gradients (violates 2).

2.2. When $P \in \mathbb{U}$ and $Q \in \mathbb{N} \bigcap \mathbb{S}$

This is the case we are typically interested during the inference phase when training latent-variable models. Here Q_θ is the inference model or approximating posterior density (we can sample from and is cheap to evaluate) and P is a density only known up to a normalization constant and is expensive to sample from.

In this case, the ideal divergence D and its gradients with respect to Q can rely on the unnormalized value of P:

- Both the mode-seeking $\mathrm{KL}(Q;P)$ and moment-matching $\mathrm{KL}(P;Q)$ can be used in this case;
- Jensen-Shannon divergence is applicable;
- Cramer/energy distance not viable;
- MMD distance not viable;
- Wasserstein distance not viable;
- The Stein divergence is applicable but has biased gradients (violates 2).

2.3. When $P \in \mathbb{S}$ and $Q \in \mathbb{S}$

This case emerges when evaluating and training implicit models. That is, models for which we also don't know the numerical values of the likelihood (even up to a constant) but we can still cheaply sample from.

Implicit models are typically latent-variable models defining a joint density $q(x,z)$ where the likelihood term $q(x|z)$ is a Dirac-delta,

$$q(x) = \int dz \delta(x - g(z))\pi(z).$$

Since we cannot solve the integral above in general, any divergence relying explicitly on the value of the likelihood $q(x)$ or on the conditional densities $q(x|z)$ and $q(z|x)$ will not be applicable.

However, we can easily see that divergences depending on Q only through an expectation can be used. Since any expectation of the form $\mathbb{E}_q[f(x,z)]$ can be simplified to an expectation only with respect to z:

$$\begin{aligned}\mathbb{E}_q[f(x)] &= \int dx\, dz\, f(x)\delta(x - g(z))\pi(z) \\ &= \int dx\, f(g(z))\pi(z) \\ &= \mathbb{E}_\pi[f(g(z))].\end{aligned}$$

- Any of the f-divergences is not viable in this case as they require the numerical value of the model's likelihood;
- Cramer distance can still be used;
- The Wasserstein distance is applicable but has biased gradients (violates 2);
- MMD distance is applicable but is expensive (violates 3);
- The Stein divergence is applicable but has biased gradients (violates 2).

2.4. When $P \in \mathbb{S} \bigcap N$ and $Q \in \mathbb{S}$

This case appears when we are trying to approximate a density P that we can sample from and know its likelihood with an implicit model Q.

3. Conclusion

In these notes we have discussed divergences, metrics, measures of relative information between two probability densities and how to build coordianate-invariant divergences.

As observed in section 1.1, the main reason for introducing the concept of relative information is to consistenly quantify surprise and updates in belief states. If our goal is only to optimize the parameters of a density Q so that it approaches another density P, it is sufficient to minimize any divergence $D(Q, P)$ with respect to the parameters of Q.

On practice, different divergences will emphasize different aspects of the density while de-emphasizing others. So there is no unique divergence for all applications.

Still, some of the axioms defining relative information measures are sensible properties to be expected from usefull divergences, such as invariance under a change of measure.

Depending on what is known about the target density P and about our model Q, different divergences can be used, we summarize this analysis in Table 1 below.

Q/P	$\mathbb{S} \bigcap \mathbb{N}$			$\mathbb{U}$			$\mathbb{S}$		
$\mathbb{S} \bigcap \mathbb{N}$		UBG	BG		UBG	BG		UBG	BG
	ICM	JS-divergence All f-divergences		ICM	KL(Q;P)		ICM	KL(P;Q)	
	NICM	MMD Cramer/energy	Wasserstein Stein	NICM		Stein	NICM	MMD Cramer/energy	Wasserstein
$\mathbb{S}$		UBG	BG		UBG	BG		UBG	BG
	ICM			ICM			ICM		
	NICM	MMD Cramer/energy	Wasserstein Stein	NICM		Stein	NICM	MMD Cramer/energy	Wasserstein

Table 2. Summary of viable divergences $D(Q; P)$ based on what is known about P and Q.

Bibliography

[1] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 131–142, 1966.

[2] Vladimir I Bogachev and Aleksandr V Kolesnikov. The monge-kantorovich problem: achievements, connections, and perspectives. *Russian Mathematical Surveys*, 67(5):785, 2012.

[3] Ariel Caticha. Relative entropy and inductive inference. In *AIP Conference Proceedings*, volume 707, pages 75–96. AIP, 2004.

[4] Harald Cramér. On the composition of elementary errors: first paper: mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1):13–74, 1928.

[5] I Csisz et al. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

[6] Imre Csiszar. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The annals of statistics*, pages 2032–2066, 1991.

[7] Oliver Thomas Johnson. *Information theory and the central limit theorem*. World Scientific, 2004.

[8] SN Karbelkar. On the axiomatic approach to the maximum entropy principle of inference. *Pramana*, 26(4):301–310, 1986.

[9] Lev Borisovich Klebanov, Viktor Beneš, and Ivan Saxl. *N-distances and their applications*. Charles University in Prague, the Karolinum Press, 2005.

[10] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[11] Pranesh Kumar and S Chhina. A symmetric information divergence measure of the csiszár's f-divergence class and its bounds. *Computers & Mathematics with Applications*, 49(4):575–588, 2005.

[12] Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006.

[13] Ingram Olkin and Friedrich Pukelsheim. The distance between two random vectors with given dispersion matrices. *Linear Algebra and its Applications*, 48:257–263, 1982.

[14] MA RA Fisher. On the mathematical foundations of theoretical statistics. *Phil. Trans. R. Soc. Lond. A*, 222(594-604):309–368, 1922.

[15] Alfréd Rényi. On the foundations of information theory. *Revue de l'Institut International de Statistique*, pages 1–14, 1965.

[16] Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423, 1948.

[17] John Shore and Rodney Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory*, 26(1):26–37, 1980.

[18] John Skilling. The axioms of maximum entropy. In *Maximum-Entropy and Bayesian Methods in Science and Engineering*, pages 173–187. Springer, 1988.

[19] John Skilling. Classic maximum entropy. In *Maximum entropy and Bayesian methods*, pages 45–52. Springer, 1989.

[20] Charles Stein, Persi Diaconis, Susan Holmes, Gesine Reinert et al. Use of exchangeable pairs in the analysis of simulations. In *Stein's Method*, pages 1–25. Institute of Mathematical Statistics, 2004.

[21] Charles Stein et al. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*. The Regents of the University of California, 1972.