1 Major changes

(i) A short introductory section on statistical manifolds is added. In particular, the notion of Banach manifold is somewhat clarified.

(ii) Al occurrences of Fréchet derivatives are replaced by the weaker notion of directional derivatives. The topology is left unspecified.

(iii) The introductory paragraph of the section "Parametric Models" is modified.

(iv) The formulation of Theorem 2 is improved. The connections $\Gamma^{(\pm 1)}$ are not defined at the moment of stating the theorem. Therefore they are replaced by the corresponding covariant derivatives.

(v) We have added a table, giving specific examples of (ρ, τ) combinations.

(vi) Also the proof of Theorem 2 is clarified.

(vii) We fixed the references, including adding one as suggested by one of the Reviewers.

2 Detailed Response to Referee 1

- "Most of the literature quoted in the introduction considers a manifold of probability densities while the divergences are defined as functions on random variables. If the random variables are non-negative it is a set of unnormalized densities otherwise I do not know. I feel that some word of explanation would improve the clarity of the argument."
  `Response.` *Two sentences are added: "Later on, manifolds of probability densities are considered. Note that these are non-negative-valued random variables with expectation equal to 1."*

- "... the logarithm f a (strictly positive) density function is considered ... In contrast to that, the definition of directional derivative top of page 5 is not really compatible with any positivity assumption."
  `Response.` *We changed to an approach based on taking directional derivatives of P rather than taking directonal derivatives of* $\log(P)$ *or more generally $h(P)$. In this way, the positivity assumption of $P + \epsilon X$ will not be needed.*

- Apparently, Eq. (11) is intended to overcome the difficulty with the introduction of the τ function as a generalized embedding.

`Response.` *No, (11) intends to remind the reader that the vectors in the tangent plane correspond with derivatives of the expectation functional.*

- In the classical case, the embedding $p \mapsto \sqrt{p}$ maps densities to the L^2 sphere, and the push-back of the sphere geometry gives a geometry on densities. The corresponding set-up in the generalization should be specified in order the argument to be clear.
  `Response.` *See answers to the previous two questions related to "positivity assumption" which is not needed in our approach.*

- The statement following Eq. (11) of this section should be clarified, at least to explain the general set-up, before referring the reader to [20]. I can see that formal equivalence of Eq. (11) with the following equation can be derived from Eq. (5).
  `Response.` *By having all derivatives in the tangent plane these difficulties disappear.*

- page 2, line 11: A measure space is usually written with brackets, $(\mathcal{X}, \mu)$.
  `Response.` *Brackets are added*

- page 2, line 12: The notation for the expectation is not consistent through the paper. Sometimes there are brackets (of two types) and in other cases the brackets are missing.
  `Response.` *Indeed, brackets are used only if needed for grouping terms.*

- page 2, Eq. (2): As ρ is differentiable, it holds $\mathrm{d}\rho(x) = \rho'(x)\mathrm{d}x$. What is the advantage of the Stieltjes Integral notation? In fact, both ρ' and τ' are used in Eq. (12).
  `Response.` *Some of our formulas can be used when ρ and τ are not differentiable. This is however not the focus of the present text.*

- page 2, line -3: An "open convex domain of $\mathbb{R}$" is an open interval, correct? The statement as it is makes the reader wonder if it would be possible to define the rho-tau-divergence with ρ and τ defined on $\mathbb{R}^d$, $d > 1$.
  `Response.` *We now replaced 'convex domain' with 'interval'; the generalization of the rho-tau formalism is a relevant problem, however not under consideration here.*

- page 3, line 8: According to the following Eq. (6), $p(\zeta, \eta)$ is joint density function. Is this a special case or the existence of a joint density is a general assumption? As it is stated, it is not clear. However, no special

assumption is needed to prove inequality (6).
`Response.` *The formulation of the sentence is improved. Note that (6) is not needed for the sequel. The statement is made only to mention this remarkable inequality.*

- page 3, Eq. (6): The last inequality is Cauchy–Schwarz inequality. Why is the result relevant? To give sufficient condition for the divergence to be finite? Why not to use the most general Hölder's inequality?
  `Response.` *One can use indeed Hölder's inequality, or other inqualities even weaker (e.g. involving Young functions). The only intention of (6) is to make the reader aware of the point that the divergence of an arbitrary pair of random variables might diverge.*

- page 4, line -14: The statement "The Fréchet derivative of a random variable . . . " is confusing. The random variables are the points in the manifold, while the Fréchet derivative applies to functions on some domain.
  `Response.` *The word 'Fréchet' is omitted. What remains is rather straightforward. The derivative is a limit and limits of random variables are again random variables. Note that these lines of text moved to the new Section 2.*

- page 5, line 5: Please explain "to deform the logarithmic function". It is a deformation or a replacement? If τ is an embedding, the space into which the embedding is done should be specified.
  `Response.` *The sentence is modified to better express the intention.*

- page 5, Eq. (11): Please explain the relation between Eq. (11) and Eq. (10). Also, how the inner product defined in the equation following (11) relates with the metric in Eq. (12)?
  `Response.` *(i) The derivation of (10) starting from (11) has been added.(ii) The use of the inner product in (12) is just a notational issue. The expression is derived from the divergence function in a straightforward manner.*

- page 5, line 17: Please check ". . . in the form of for the non-parametric . . . ".
  `Response.` *Corrected. The missing reference to (12) has been inserted.*

- page 5, line -12: Reference "From (3)" does not seem correct. Is it (12)?
  `Response.` *It is indeed (12). We corrected it.*

3 Detailed Response to Referee 2

- Here, although can be derived, the statement of the generalized Pythagorean equality being satisfied by the rho-tau divergence for any three given points P, Q, R could be linked to a reference.
  `Response.` *Previously, in [20], the rho-tau divergence, denoted D-divergence there, was shown to be the canonical divergence, and as a result, Pythagorean equality would hold. Here, we explicitly write out the Pythagorean equality.*

- Also, the definition of the rho-tau entropy in Equation (7) is not well justified. I believe it is a direct extension from the Shannon definition and usage of general f-divergence, but it would worth mentioning the origin of it so one can easily get the result from Equation (8).
  `Response.` *A reference is added and one sentence which makes clear that this is a very general definition of entropy.*

- One question that come to my mind is that in such general proposed model one does not need to assume the function (divergence) is Gateaux differentiable in order to ensure the convergence conditions for the expectations?
  `Response.` *For the derivatives of the divergence function it suffices that the functions ρ and τ are differentiable. However, it was implicitly assumed that the manifold itself is differentiable. Therefore one sentence is added in the new introductory section, stating that "Throughout the text it is assumed that the manifold is differentiable and that for each X in $\mathbb{M}$ the tangent plane $T_X\mathbb{M}$ is well-defined."*

- Regarding the gauge freedom, for the case of $\rho(u) = 1/\tau'(u)$ and the corresponding deformed logarithm the divergence becomes a generalization of the Kullback-Leibler one. Would it be possible to select different $\rho(u)$ such that the divergence becomes a general model for the Rényi's divergence? If so, could you comment on that?
  `Response.` *Rényi's divergence is not of the Bregman type. For $\alpha \neq 1$ it is not the expectation of a random variable. In particular, it is not a rho-tau divergence. Hence, we cannot say much about it.*

- With respect to the deformed exponential family (Section 7) if we select a general model, as the one proposed in [17], does the rho-tau divergence can be reduced or extended to the model proposed in [17]? If so, which conditions should be imposed over the deformed family for such situation?

`Response.` *Section 7 (now Section 8) is about the parametric case whereas in [17], but also [14-16], the manifold has maximal extent. As said in the paper, we use concepts of the latter although it is left for future research to to clarify the interplay between both approaches.*

4 Detailed Response to Referee 3

- Is faithful in Theo 1 a usual term (add ref then)
  `Response.` *The notion of a 'faithful state' is common in the context of C*-algebras. Since probability distributions are special cases of states on a C*-algebra it is justified, although not common, to use the term here. Because it is not common, the meaning of the term is explained in the text.*

- In Eq 7, index S by ρ, τ?
  `Response.` *These have been added at 4 places.*

- why index proba density with superscript theta instead of usual subscript? add footnote?
  `Response.` *The reason is not very deep. In the case of a discrete probabilities p_i^θ is somewhat easier than $p_{\theta,i}$. Also $p^\theta(x)$ is better than $p(x|\theta)$ and p_i^θ is better than $p(i|\theta)$. But again, it is more a question of taste.*

- page 6, a table with the ρ, τ, f, f^* would be nice with various examples
  `Response.` *Such a table has been added.*

- Sec 5 "by taking two derivatives" "by taking its second-order partial derivatives"?
  `Response.` *No, one takes one derivative of each of the two arguments to get the metric tensor.*

- Sec 7 specify that you use regular exp fam
  `Response.` *No, the statements in this section is NOT specific the regular (or for that matter, deformed) exponential family. They are generic. We added a sentence to that effect.*

- I also recommend to add those citations: ...
  `Response.` *We have added a reference to the book of Shima in Section 6, now 7. We did not include the other reference, as the Reviewer did not explain why voronoi diagram is relevant to the current investigation.*

5 Detailed Response to Referee 4

- `Response.` *After (6) the sentence "To obtain the latter the Cauchy-Schwarz inequality is used." has been added.*

- `Response.` *"Pythagorian" ⇒ "Pythagorean".*

- `Response.` *Various typos have been corrected.*

- Line after (13): "hermitian" with a capital?
  `Response.` *We have changed to "adjoint" as the correct terminology.*

- ∇^{-1} instead of ∇_Z^{-1}
  `Response.` *It has been corrected.*

- `Response.` *The statement and the proof of Theorem 2 are both clarified.*

- The probability distributions of the deformed exponential family belong to statistics and are not random variables.
  `Response.` *We consider them as special cases of random variables. Usually it is the other way round: random variables are often associated with measures.*

- The function $\phi(v)$ 'must be strictly monotone'.
  `Response.` *One can drop this condition but then the deformed logarithm is not concave.*

6 Detailed Response to Referee 5

- Furthermore, it would be excellent if any physical notion associated with the gauge freedom is given.
  `Response.` *A paragraph is added to clarify the meaning in Physics.*

- It would be nice to suggest any relevant perspectives connecting the parametric with the nonparametric framework.
  `Response.` *The connection between both has been worked out a little bit further. The long term goal of developing a common framework goes beyond the present paper.*

- The deformed exponential model authors discuss can be connected with the idea of maximum entropy if the rho-tau divergence is decomposed into the cross and diagonal entropies? For this it might be necessary to consider a constrain by the escort expectation.

`Response.` *The deformed exponential model can indeed be derived using a maximum entropy approach, both with and without escort constraints. There is a link with the gauge freedom of the rho-tau formalism. See for instance [12]. However, it would take an additional Section to work this out in the present paper, while not contributing to the central topic.*

- In the notation of the paper the arguments P and Q are used as random variables, for example, as in equation (1). However, in the standard notation P and Q are used to express probability measures, and so it would be very confusing for readers who work in probabilistic paradigms and applications.
  `Response.` *We tried a systematic use of capitals for random variables, not only P and Q, but also X and Y for tangent vectors, small characters for probabilities and greek symbols for scalar functions and parameters. In functional analysis it would be highly unusual to use $p(x)$ for the value of a probability distribution P at the point x. So yes, in many places we deviate from habits found in statistics.*