# Author Response for Paper #7112: Minimum Stein Discrepancy Estimators

We would like to start by thanking the three reviewers for their careful consideration of our paper, as well as their insightful comments which will certainly help further strengthen it. Our response below first addresses comments shared across reviewers then addresses more specific questions from individual reviewers.

## Shared Comments

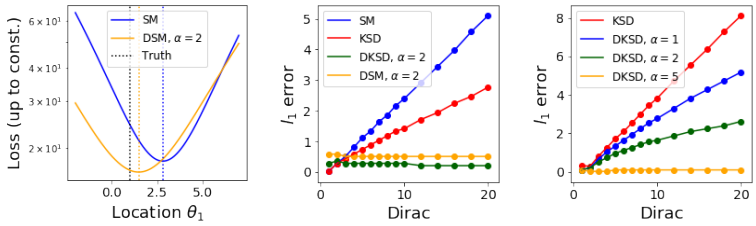
1. **R2:** *"...the paper should be more focused on giving a clear idea of what is being demonstrated by expanding certain sections, and perhaps omitting or making more vague the particularly dense parts"*, **R4:** *"The structure is very clear and the paper is pleasant to read."*, **R6:** *"The presentation of the paper is basically clear"*.

   While there is some disagreement regarding clarity (with the more confident reviewers expressing satisfaction), we agree that some aspects of the paper stand to be more accessible. We propose the following changes:

   - We will make the statement of Thm 1 more concise, as requested. Furthermore, we will shorten Thm 3, and unify Thm 4 & 5 under simpler assumptions. In each case, implications of the theorems will be discussed at greater length, and the full technical details will be relegated to the supplementary information.
   - We will expand Sec. 2.1, adding a number of clarifications on DSM and DKSD which will help the reader develop intuition. This will include greater discussion of how to make choices of $m$ and $K$ to improve performance.

2. **R2:** *"More extensive experiments would have been nice"*, **R4:** *"An illustrative example of the failure of the robustness of SM and how DSM can improve it"*., **R6:** *"compare DSM with SM to demonstrate the benefit."*

   We propose to add an experiment demonstrating the robustness of DSM & DKSD for generalised gamma location models: $p(x|\theta) \propto \exp(-(x-\theta)^c)$. We set $n = 300$ and corrupt 80 points with the value of $x = 8$; a robust estimator should obtain a good approximation of $\theta^*$ even under corruption. *Left plot:* The model is Gaussian for $c = 2$; we see that SM is not robust for this very simple model whereas DSM with $m(x) = 1/(1 + \|x\|^\alpha), \alpha = 2$ is robust. *Middle plot:* The same $m$ also leads to a robust DKSD. *Right plot:* we take $c = 5$ and see that $\alpha$ can be chosen as a function of $c$ to guarantee robustness (our theory requires $\alpha \geq c$).




   We also propose to add an experiment showing the benefits of DKSDs for product-of-expert models. This will demonstrate the advantages of our methodology for applications where SM is typically applied.

## Reviewer 4

- *"... m was picked to match somehow the expression of the parametric model. What guides such choice? Can one think of a general guideline to pick such function to improve the convergence properties of the loss [...]?"*

  Thank you for the opportunity to elaborate on this central point. The novelty of this approach is that $m$ can be chosen so that DKSD satisfies various desirable properties: sample efficiency (e.g., for a location parameter, select $m$ so that $m(x)\nabla_x \log p_\theta(x) \approx -(x-\theta)$), robustness (ensure $\|K(x,y)m(y)\nabla_y \log p_\theta(y)\| \to 0$ as $\|y\| \to \infty$) and convexity, with different choices of $m$ in each case. We will elaborate on general guidelines in the paper.

## Reviewer 6

- *"In Eq. (2), what does it mean that m is a diffusion matrix? Does it have to be symmetric positive definite?"*

  The terminology "diffusion matrix" comes from the fact that the Stein operator arises from the generator of a diffusion process, and is in general an arbitrary matrix function. It needs to be invertible for both DKSD and DSM to be statistical divergences (see Prop 1 and Thm 2). It therefore does not necessarily need to be positive definite.

- *"...discuss the relation between DKSD and the Riemannian kernel Stein discrepancy..."*

  Thank you for pointing this out; we will elaborate on the connection in the paper.

- *"I noted that the identity for DKSD is provided in Line 120, but did not find the one for DSM"*

  The identity holds under the assumptions on $p$ and $q$ stated in Thm 2, where DSM is defined. This can be seen from the proof of the theorem, but we will also clarify it in the main text.

- *"Why the matrix-valued kernel considered in DKSD has to be in the two forms?"*

  Outside of Thm. 1 and Prop. 2, our stated results all hold for general matrix-valued kernels. We restricted the kernel choice for Thm. 1 to make the derivation easier to follow, but this result also extends to all matrix valued kernels, and we will include the general form in the revision.