We would like to thank all of the reviewers for their valuable time and their constructive comments. In what follows, we would like to address all concerns raised.

Reviewer 1: We will incorporate the proposed minor corrections in the final version of the paper. We thank the reviewer for proposing the interesting comparisons, and we will include more comprehensive discussion in the final paper. Below please find some details regarding the proposed methods.

- The two-stage approach, i.e., $i)$ running gradient descent to convergence, and then $ii)$ projection onto sparsity set, is known to be sub-optimal even for simple problems such as least-squares with sparsity constraints. In fact, the results when using such approach on ℓ_2-norm minimization are the same as the Greedy baseline: It $i)$ converges to the global optimum $q(\cdot)$, and then $ii)$ use greedy projection to try to minimize $F[p(\cdot)] = \|p(\cdot) - q(\cdot)\|_2^2$ subject to sparsity constraint, which has been shown to be inferior than IHT (subsection 4.1).
- In the general case, taking the k-heaviest coordinates of q (without assuming any structure) would result into a non-valid putative solution; recall, by definition of the discrete setting, we have n coordinates, each of which takes m points, leading to a m^n sample space. Simply taking the k-heaviest coordinates of that long vector would result into an intermediate representation of non-zero positions that does not correspond to a probability distribution. A variation of this approach is fine for the "vector-sparsity" special case.

On whether support set changes during iterations, we observe that in experiments (subsection 4.1) IHT changes support, on average, 36.1 times for ℓ_2-norm minimization and 6.9 times for KL minimization. We also conduct experiments on fixing the support after some iterations: IHT on ℓ_2-norm minimization (subsection 4.1) after 400 iterations, fixing supports after 1, 5, 10 and 15 support sets change, give average results of 0.0026, 0.0020, 0.0018, 0.0016, respectively.

Regarding motivation, model compression is an exciting immediate application of our proposed approach, especially since our general approach may be flexibly applied to specialized problem-specific losses. We are currently investigating extensions of the current work to model/policy compression for reinforcement learning, where the loss can be constructed to preserve post-compression expected reward; but the details of that approach deserve a different publication.

Reviewer 2: We thank the reviewer for the supportive and constructive review. Regarding the comment in lines 211-214, we chose not to include this detail as extending convergence analyses from global to local strong convexity is considered fairly standard; see e.g., [Agarwal2010].

Regarding the comment in lines 198-202, we apologize for any confusion. This argument is a standard one in NP-hardness proofs: proving the existence of corner cases is sufficient to show the hardness of the problem. Note that even if *empirically* we observe that common cases are easy to solve, this does not guarantee this is the norm. Showing that with high-probability the corner cases rarely appear is an interesting question by itself.

Regarding variance in experiments, we have observed high variance is not enough for the algorithm to get "lucky". In fact, we observe that the **best** performance of Greedy is often worse than the **worst** performance of IHT. Low variance is especially desirable when the algorithm is only applied a few times to save computation, as in large discrete optimization problems. We believe that this makes IHT preferable in practice.

We also thank the reviewer for the suggestions on free energy / ELBO using IHT; we will consider these as future work.

Reviewer 3: We thank the reviewer for the constructive comments. We are happy to fully describe the structure nesting i.e. why $\mathcal{D}_k$ is included in $\mathcal{D}'_k$ in general. QM-AM stands for Quadratic Mean-Arithmetic Mean inequality.

Unfortunately, we are unaware of methods that are strictly better than greedy ℓ_2-norm projection. However, trading time for performance may be an interesting topic for future work. Regarding the novelty, although we unveil a relationship between the functional derivative and the gradient (l. 408), they are different in many aspects. The ℓ_2-norm projection for vector sparsity can be done optimally, but we have shown that ℓ_2-norm projection for the general case is computationally hard in general (Theorems 1 & 2). Despite the provable hardness, we provide some intuition for why greedy projection is effective in our main Algorithm (Theorem 3). These are challenges of extending vector optimization to general distribution optimization. We also note that the convergence analysis is for the functional setting, not the vector setting (Theorem 4), which is not only more general, but also paves the way for future work on continuous distributions.

Regarding **why** greedy seems effective in practice, we have provided the intuition and its supporting theorem in section 3.4. Regarding the greedy projection (Algorithm 2), it is also possible to use KL divergence as distance metric, but our convergence analysis (Theorem 4) suggests that a good projection in ℓ_2-norm (Definition 5) is preferable.

We will improve the presentation as suggested in the final version of the paper.

[Agarwal2010] Agarwal, Alekh and Negahban, Sahand and Wainwright, Martin J, "*Fast global convergence rates of gradient methods for high-dimensional statistical recovery*", Advances in Neural Information Processing Systems, 2010.