Some thoughts on scaling the classic exponential family cases

* Do we need all of these to be at a high dimensionality, or is one example at D=100 good enough?
* Increasing from 1 to 4 GPU’s can give us at best a 4x speedup in training. The model must exist on each GPU, and we can split our training up by K/4 or M/4. I’m not sure which if either is best. This may actually get complicated to implement, and might not be the best use of time for a 4x improvement.
* The Gaussian distribution is hard to scale, because it’s flow is a matrix whose parameters increase with the square of the of the dimensionality
  + I realized that I need to index the upper triangular elements out of eta when I feed eta into the theta(eta network). That will reduce the parameterization by half.
* We can scale the Dirichlet to 100 for example, but can’t use a more expressive flow than 40 planar flows (I feel like this should be sufficient).