Differentiable inference

We already know how to specify some expressive and flexible generative models, including entire languages of models that can express arbitarily complicated structure. However, until recently such models were hard to apply to real datasets, because inference methods (such as Markov chain Monte Carlo methods) were not usually fast or scalable enough to run on large models or even medium-sized datasets.

The past few years have seen major progress in methods to train and do inference in generative models, loosely following four strands:

- **Variational autoencoders** - Latent-variable models that use a neural network to do approximate inference. The *recognition network* looks at each datapoint x and outputs an approximate posterior on the latents q(z | x) for that datapoint.
- **Generative adversarial networks** - A way to train generative models by optimizing them to fool a classifier, the *discriminator network*, that tries to distinguish between real data and data generated by the model.
- **Invertible density estimation** - A way to specify complex generative models by transforming a simple latent distribution with a series of invertible functions. These approaches are restricted to a more limited set of possible operations, but sidestep the difficult integrals required to train standard latent variable models.
- **Autoregressive models** - Another way to model p(x) is to break the model into a series of conditional distributions: $p(x) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1) \ldots$ This is the approach used, for example, by recurrent neural networks. These models are also realitvely easy to train, but the downside is that they don't support all of the same queries we can make of latent-variable models.

The common thread among these approaches that lets them scale to high-dimensional models is that their loss functions are *end-to-end differentiable*. This is in contrast to previous inference strategies such as MCMC or early variational inference strategies, which required alternating inference and optimization steps and didn't allow gradient-based tuning of the inference procedure.

These new inference schemes are allowing great progress in generative models of images and text.