Plan of Study

Clayton W. Seitz *February 2, 2022*

1 Introduction

The central dogma of molecular biology has recently stimulated the development of a rich set of computational techniques for cellular phenotyping at the molecular scale. The intracellular environment harbors a diverse set of biomolecules which undergo complex interactions in order to carry out cellular functions. Classical biological techniques have yielded great insights into biochemical and biophysical mechanisms of their interactions. However, the development of even simple models of the interaction scheme of DNA, RNA, and protein remains a challenge, which in turn precludes the abstraction of these fundamental interactions to the scale of the phenotype. Recent advances in single-cell sequencing and fluorescent labeling of nucleic acids and proteins holds great promise if coupled with the appropriate analysis paradigm. High-throughput variants of these methods e.g., genome wide or transcriptome wide analyses, output data volumes that enable discovery-driven research. This has important implications for both fundamental biological research as well as the discovery of the molecular drivers of disease.

2 Deep generative models

Generative models have emerged as an invaluable tool in a range of supervised learning tasks, such as assigning labels to images, as well as unsupervised learning tasks, such as dimensionality reduction and out-of-sample generation. Recently, generative models have found applications in molecular biology research such as protein design, analyzing genomic mutations, and dissecting transcriptional variability between single cells. In supervised generative learning, we try to explicity learn the joint distribution which produced that data - a task which is generally a more difficult task than discriminative learning e.g., clustering. Overcoming the barrier to generative modeling carries significant benefits; f fitting complete multivariate distributions goes beyond correlation-based or clustering approaches as it permits the discovery of partial correlations, outlier detection, as well as sampling based inference. Implementing generative models as deep networks, as in the variational autoencoder (VAE), can be a favorable choice in genomic, transcriptomic, and proteomic research as they are known to be flexible and expressive. However, the flexibility of deep models can simulataneously make them difficult to interpret in a biologically meaningful way.

2.1 The curse of dimensionality

When predicting $p(\mathbf{x})$, the dimension N of $\mathbf{x}$ is often very large such as in an image or a genome wide expression profile. Typically, we are given a finite number of empirical observations of $\mathbf{x}$ with the hope of constructing $p(\mathbf{x})$ from those observations. In high dimensions, i.e. $N >> 1$, the volume of the state space grows exponentially in the number of dimensions while we generally do not have a comparable number of training examples to accurately "populate" this volume. Even if we did have a sufficient number of samples to obtain a good estimate of $p(\mathbf{x})$, simply *counting* the number of times each $\mathbf{x}$ to construction $p(\mathbf{x})$ is generally intractable. This can be seen from the fact that if $\mathbf{x_i} \in \Omega$ where Ω is the state space of each element of $\mathbf{x}$, normalizing the counts requires an summation over $|\Omega|^N$ possible states. This issue arises in many contexts, for example Bayesian inference and a large class of problems in statistical physics.

2.2 Thermodynamics of deep learning and stochastic gradient descent

In classical gradient descent, we attempt to find a global minimum a generally non-convex function $\mathcal{L}(\Phi)$ by compute the gradient $\nabla_\Phi \mathcal{L}$ with respect to the parameters or "position vector" Φ. By iteratively computing the gradient and updating the parameters, we can arrive at a local minimum of $\mathcal{L}$. This method works well when we can explicitly evaluate $\nabla_\Phi \mathcal{L}$ but this is rarely the case in real-world optimization problems. We typically can only *estimate* the value of the loss $\mathcal{L}(\Phi)$ based on a batch of training examples. Those training examples are sampled from $p(\mathbf{x})$ at random, which means the evolution of the parameters Φ is a stochastic process. In this scenario, conventional gradient descent becomes stochastic gradient descent or SGD.

$$\frac{d\Phi}{dt} = -\eta\hat{g}$$

which is analogous to a "velocity". It follows that

$$\Delta\Phi \approx -\eta\Delta t\hat{g}$$

which can be decomposed into

$$\Delta\Phi \approx -\eta\Delta t g(\Phi) + \eta\sum_i(g(\Phi) - \hat{g}_i(\Phi))$$

Applying the central limit theorem to the last term gives the Langevin equation

$$\Delta\Phi \approx -\eta\Delta t g(\Phi) + \eta\epsilon\sqrt{N}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. We can now see that $\Phi(t)$ follows a stochastic differential equation reminiscent of the Ornstein-Uhlenbeck (OU) process. The OU process has a corresponding Fokker-Planck equation. We would like the stationary distribution of that stochastic process to center on the point at which $\mathcal{L}$ is a minimum.

3 Specific Aims

3.1 Specific Aim 1: Develop an interpretable deep generative model for single-cell expression data

3.2 Specific Aim 2: Discover latent variables in single-cell expression profiles