Assignment – Stochastic Gradient Descent and Online Learning

Questions:

1) How does optimization in machine learning differ from pure optimization?
2) How does 'on-line learning' differ from 'deterministic' (also called 'batch) approaches to machine learning? Can you describe this in a way a non-statistician would understand?
3) How do batch gradient descent, stochastic gradient descent, and mini-batch gradient descent differ?
4) There is a new source of variation introduced when we implement stochastic gradient descent (compared with batch gradient descent). What is it that new source of variation? Does it decrease as the algorithm converges?
5) What does it mean for a function to be 'ill-conditioned'? How does this manifest in ML optimization (think of Jacobian and Hessian)?
6) What is the advantage of using a learning rate schedule (dynamic learning rate) instead of a fixed learning rate? When do we need to use a dynamic learning rate instead of a fixed learning rate? Why?