* Same thing for GANs? Or for matrix factorizations? **Nicolas**
* How do different optimization variants affect generalization (test error)? **Arthur**
* Second-order methods: Do (Quasi-)Newton methods go to differently shaped local minima in neural net- works? Or: Is the secant method a viable alternative training method? **Arthur**
* Quantized SGD: As quantization of gradients or iterates becomes coarser, what happens to the optimization algorithms, and to generalization? Is it different for DL as compared to linear ML models? **Robin**
* AdaGrad / Adam / signSGD: Can you suggest/try different data-dependent coordinate-wise learning rate schemes and compare them? **Arthur**
* For training deep nets with very large SGD mini-batches: when does the scalability gain disappear? Is it influenced by other properties of the optimizer? For example, what is the effect of first slowly growing the learning rate and then later decreasing it? **Nicolas**
* How well do zero-order optimization methods do for ML applications, compared to standard first-order methods? **Arthur**
* Asynchronous SGD: How do different delays affect convergence? How does it interplay with momentum? Does it act as a regularizer, like drop-out? **Robin**
* Implicit regularization effect of SGD, as in flattening the landscape, or encouraging different trajectories or converging to solutions with different properties **Robin Nicolas**