The points marked for each exercise give an indication of the relative importance. All answers must be justified.

Ex. 1 — (2 points) Let $F(w) = \frac{1}{n}\sum_{i=1}^{n} f(w,i)$, be the function we want to minimize. To this purpose, we have considered the following general iterative algorithm

$$w_{k+1} = w_k - \alpha_k g(w_k, \xi_k),$$

where ξ_k denotes in general a subset of the indices $\{1, 2, \dots n\}$.

1. How is α_k called?
2. What are ξ_k and the function $g(.)$ for the stochastic gradient?
3. For the mini-batch?
4. For the full-gradient?

Ex. 2 — (3 points) Consider the following assumptions:

1) the function $F : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and L-smooth,

2) the sequence of iterates $\{w_k\}$ is contained in an open set over which $F(.)$ is bounded below by a scalar F_{inf},

3) there exist scalar $\mu_G \geq \mu \geq 0$ such that, for all $k \in \mathbb{N}$,

$$\nabla F(w_k)^T \mathbb{E}_{\xi_k}[g(w_k, \xi_k)] \geq \mu ||\nabla F(w_k)||^2$$

and

$$||\mathbb{E}_{\xi_k}[g(w_k, \xi_k)]|| \leq \mu_G ||\nabla F(w_k)||,$$

4) there exist scalars $M \geq 0$ and $M_G \geq 0$ such that, for all $k \in \mathbb{N}$,

$$\mathbb{E}_{\xi_k}[||g(w_k, \xi_k)||^2] \leq M + M_G ||\nabla F(w_k)||^2.$$

In class we proved:
Lemma The iterates of the algorithm in Exercise 1 satisfy the following inequality for all $k \in \mathbb{N}$

$$E_{\xi_k}[F(w_{k+1})] - F(w_k) \leq -\alpha_k \left(\mu - \frac{1}{2}\alpha_k L M_G\right) ||\nabla F(w_k)||^2 + \frac{1}{2}\alpha_k^2 L M.$$

1. What does the lemma imply?
2. How do the different quantities $(\alpha_k, L, M, \dots)$ affect the result? Is it what you expect from their definition?

Ex. 3 — (2 points)

1. Write the equations of a gradient method using momentum.
2. Write the equations of a gradient method using Nesterov momentum.

Ex. 4 — (2 points) What does it mean that an optimization problem is ill-conditioned?

Ex. 5 — (3 points) Describe two tricks of the trade one can use to speed up ML training.

Ex. 6 — (2 points) What is data parallelism? And model parallelism?

Ex. 7 — (2 points) Are MapReduce-based frameworks like Hadoop and Spark suited for ML?

Ex. 8 — (3 points) Are second-order methods used for neural network training? Why?

Ex. 9 — (2 points) What are consensus-based optimization methods? How do they differ from the traditional parameter server?