Optimization in Machine Learning | **Universität Mannheim**
HWS 2024 | Prof. Simon Weißmann, Felix Benning

Sheet 6

For the exercise class on the **05.12.2024**.
Hand in your solutions by 10:15 in the lecture on Tuesday 03.12.2024.

Exercise 1 (Conditional expectation)**.** **(8 Points)**

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an underlying probability space.

(i) Let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra and let $X, Y \in L^1(\Omega, \mathcal{F}, \mathbb{P})$. Prove that

(a) for any $\lambda \in \mathbb{R}$ there holds $\mathbb{E}[\lambda X + Y \mid \mathcal{G}] = \lambda \mathbb{E}[X \mid \mathcal{G}] + \mathbb{E}[Y \mid \mathcal{G}]$.

(b) if $X \geq Y$ $\mathbb{P}$-almost surely, then $\mathbb{E}[X \mid \mathcal{G}] \geq \mathbb{E}[Y \mid \mathcal{G}]$ $\mathbb{P}$-a.s..

(c) $|\mathbb{E}[X \mid \mathcal{G}]| \leq \mathbb{E}[|X| \mid \mathcal{G}]$.

(3 pts)

(ii) Let Y_1,Y_2 be iid. random variables with

$$\mathbb{P}(Y_1 = 2) = \mathbb{P}(Y_1 = 0.5) = \frac{1}{2},$$

and set

$$S_0 = 2, \quad S_k = S_0 \cdot \prod_{i=1}^{k} Y_i\,,\ k = 1, 2\,.$$

Compute $\mathbb{E}[S_2 \mid \mathcal{F}_1]$, where $\mathcal{F}_1 = \sigma(S_1)$. (2 pts)

(iii) Let X, Y be independent Bernoulli distributed random variables with parameter $p \in [0, 1]$ and define $Z := \mathbb{1}_{\{X+Y=0\}}$.

(a) Compute $\mathbb{E}[X \mid \sigma(Z)]$ and $\mathbb{E}[Y \mid \sigma(Z)]$.

(b) When are these random variables independent? **Hint:** You may use the fact that a real valued random variable is independent from itself if and only if it is a constant.

(3 pts)

Exercise 2 (Martingales)**.** **(4 Points)**

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be an underlying probability space.

(i) Let $Y_1, \dots, Y_N$ be iid. random variables with $\mathbb{E}[Y_1] = 1$ and $\mathbb{E}[|Y_1|] < \infty$, let $\mathcal{F}_k := \sigma(Y_1, \dots, Y_k)$ and define

$$S_0 = Y_0 := 1, \quad S_k = Y_1 \cdot \dots \cdot Y_k = \prod_{i=1}^{k} Y_i\,,\ k = 1, \dots, N\,.$$

Prove that $(S_k)_{k=1,\dots,N}$ is a Martingale with respect to the filtration $(\mathcal{F}_k)_{k=1,\dots,N}$. (2 pts)

(ii) Let $(X_k)_{k\in\mathbb{N}}$ and $(Y_k)_{k\in\mathbb{N}}$ be two martingales. Prove that $(aX_k + bX_k)_{k\in\mathbb{N}}$ is a martingale for any $a, b \in \mathbb{R}$. (1 pt)

(iii) Let $(X_k)_{k\in\mathbb{N}}$ and $(Y_k)_{k\in\mathbb{N}}$ be two super-martingales. Prove that $(\min(X_k, Y_k))_{k\in\mathbb{N}}$ is a super-martingale. (1 pt)

Exercise 3 (SGD with random batches)**.** **(12 Points)**

Consider the expected risk minimization problem

$$\min_{x\in\mathbb{R}^d} F(x), \quad F(x) := \mathbb{E}_{Z\sim\mu_Z}[f(x, Z)],$$

where $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}$ is measurable and $Z : \Omega \to \mathbb{R}^p$ is a random vector. We assume that the "usual" conditions (conditions for Lemma 4.1.2 such as measurability, integrability, etc.) from the lecture are satisfied.

(i) Assume that $\mathbb{E}[\|\nabla_x f(x, Z) - \nabla_x F(x)\|^2] \le c$ for some $c > 0$ and all $x \in \mathbb{R}^d$, and let $Z^{(1)}, \ldots, Z^{(B)}$, $B \ge 2$, be iid. random variables with $Z^{(1)} \sim \mu_Z$. Prove that

$$\mathbb{E}[\|\frac{1}{B}\sum_{m=1}^{B} \nabla_x f(x, Z^{(m)}) - \nabla_x F(x)\|^2] \le \frac{c}{B}$$

for all $x \in \mathbb{R}^d$. (4 pts)

Let $X_0 : \Omega \to \mathbb{R}^d$ be the initial random variable, $(\alpha_k)_{k\in\mathbb{N}}$ be a sequence of positive deterministic step sizes and for batch sizes $(B_k)_{k\in\mathbb{N}}$ let $(Z_k^{(m)})_{k\in\mathbb{N},\, m=1,\ldots,B_{k-1}}$ be a sequence of iid. random variables with $Z_1^{(1)} \sim \mu_Z$.

(ii) Formulate the stochastic gradient descent (SGD) scheme using the stochastic gradient estimator with batch size B_k:

$$G_k(x) := \frac{1}{B_k}\sum_{m=1}^{B_k} \nabla_x f(x, Z_{k+1}^{(m)}), \quad x \in \mathbb{R}^d.$$

(2 pts)

(iii) Assume that F is L-smooth and μ-strongly convex, and let $\alpha_k \in (0, \frac{1}{L}]$ for all $k \in \mathbb{N}$. Prove that

$$\mathbb{E}[\|X_{k+1} - x_*\|^2] \le (1 - \alpha_k\mu)\mathbb{E}[\|X_k - x_*\|^2] + \tilde{c}\frac{\alpha_k^2}{B_k}$$

for some $\tilde{c} > 0$, where $(X_k)_{k\in\mathbb{N}}$ is generated by SGD with batch-sizes $(B_k)_{k\in\mathbb{N}}$ and $x_* = \arg\min_{x\in\mathbb{R}^d} F(x)$. (3 pts)

(iv) Determine sequences of step sizes $(\alpha_k)_{k\in\mathbb{N}}$ and batch-sizes $(B_k)_{k\in\mathbb{N}}$ to deduce convergence $\lim_{k\to\infty} \mathbb{E}[\|X_{k+1} - x_*\|^2] = 0$. (3 pts)