Solution 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 $\mathbb{E}[\lambda X + Y \,|\, \mathcal{G}] = \lambda \mathbb{E}[X \,|\, \mathcal{G}] + \mathbb{E}[Y \,|\, \mathcal{G}]$.

Solution. Let $A \in \mathcal{G}$. Then

$$\begin{aligned}\mathbb{E}\big[\mathbb{1}_A\big(\lambda \mathbb{E}[X \,|\, \mathcal{G}] + \mathbb{E}[Y \,|\, \mathcal{G}]\big)\big] &= \lambda \mathbb{E}\big[\mathbb{1}_A \mathbb{E}[X \,|\, \mathcal{G}]\big] + \mathbb{E}\big[\mathbb{1}_A \mathbb{E}[Y \,|\, \mathcal{G}]\big] \\ &= \lambda \mathbb{E}\big[\mathbb{1}_A X\big] + \mathbb{E}\big[\mathbb{1}_A Y\big] \\ &= \mathbb{E}\big[\mathbb{1}_A(\lambda X + Y)\big]\end{aligned}$$

Thus, by uniqueness of the conditional expectation, $\mathbb{E}[\lambda X + Y \,|\, \mathcal{G}] = \lambda \mathbb{E}[X \,|\, \mathcal{G}] + \mathbb{E}[Y \,|\, \mathcal{G}]$ almost surely. □

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

Solution. Let $A := \{\mathbb{E}[X \,|\, \mathcal{G}] < \mathbb{E}[Y \,|\, \mathcal{G}]\}$. Then A is $\mathcal{G}$ measurable and it follows from the definition of the conditional expectation that

$$0 \overset{\text{def. } A}{\geq} \mathbb{E}[\mathbb{1}_A\big(\mathbb{E}[X \,|\, \mathcal{G}] - \mathbb{E}[Y \,|\, \mathcal{G}]\big)] = \mathbb{E}[\mathbb{1}_A(X - Y)] \overset{X \geq Y}{\geq} 0.$$

Thus, $\mathbb{1}_A(X - Y) = 0$ almost surely and by definition of A this implies $\mathbb{1}_A = 0$ almost surely, i.e. $\mathbb{E}[X \,|\, \mathcal{G}] \geq \mathbb{E}[Y \,|\, \mathcal{G}]$ almost surely. □

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

Solution. One possibility is to apply Jensen's inequality for the conditional expectation to the convex function $x \mapsto |x|$. Another way is to split the random variable X into its positive and negative part $X = X^+ - X^-$ and to use the linearity of the conditional expectation and the previous result, i.e. $X^+ \geq 0$ almost surely implies $\mathbb{E}[X^+ \,|\, \mathcal{G}] \geq 0$ almost surely. □

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

Solution. We have

$$\mathbb{E}[S_2 \,|\, \mathcal{F}_1] = \mathbb{E}[S_1 Y_2 \,|\, \mathcal{F}_1] = S_1 \mathbb{E}[Y_2 \,|\, \mathcal{F}_1] = S_1 \mathbb{E}[Y_2] = S_1\big(\tfrac{2}{2} + \tfrac{1/2}{2}\big) = \tfrac{5}{2} S_1. \qquad \square$$

(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 $\mathbb{E}[X \,|\, \sigma(Z)]$ and $\mathbb{E}[Y \,|\, \sigma(Z)]$.

Solution. We have

$$\mathbb{P}(Z=1) = \mathbb{P}(X+Y=0) = \mathbb{P}(X=0, Y=0) = (1-p)^2. \tag{1}$$

Thus,

$$\begin{aligned}
\mathbb{E}[X \,|\, \sigma(Z)] &= \mathbb{E}[X \,|\, Z=0]\mathbb{1}_{Z=0} + \underbrace{\mathbb{E}[X \,|\, Z=1]}_{=0} \mathbb{1}_{Z=1} \\
&= \frac{\mathbb{E}[X \mathbb{1}_{Z=0}]}{\mathbb{P}(Z=0)} \mathbb{1}_{Z=0} \\
&= \frac{0\mathbb{P}(X=0,Y=1) + 1[\mathbb{P}(X=1,Y=0) + \mathbb{P}(X=1,Y=1)]}{1-(1-p)^2} \mathbb{1}_{Z=0} \\
&= \frac{p(1-p)+p^2}{1-(1-2p+p^2)} \mathbb{1}_{Z=0} \\
&= \frac{p}{2p-p^2} \mathbb{1}_{Z=0} \\
&= \frac{1}{2-p} \mathbb{1}_{Z=0}.
\end{aligned}$$

By a symmetric argument we have $\mathbb{E}[Y \,|\, \sigma(Z)] = \frac{1}{2-p} \mathbb{1}_{Z=0}$. $\square$

(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.

Solution. Since we have $\mathbb{E}[X \,|\, \sigma(Z)] = \mathbb{E}[Y \,|\, \sigma(Z)]$ almost surely by the previous result, $\mathbb{E}[X \,|\, \sigma(Z)]$ is independent from $\mathbb{E}[Y \,|\, \sigma(Z)]$ if and only if it is a constant almost surely. For this to be the case, we need $\mathbb{1}_{Z=0}$ to be constant almost surely, i.e. $\mathbb{P}(Z=0) = 0$ or $\mathbb{P}(Z=0) = 1$. This is only the case for $p \in \{0,1\}$ by (1).
In summary, $\mathbb{E}[X \,|\, \sigma(Z)]$ is independent from $\mathbb{E}[Y \,|\, \sigma(Z)]$ if and only if $p \in \{0,1\}$. $\square$

(3 pts)

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

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

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

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

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

Solution. We first check $\mathbb{E}[|S_k|] < \infty$ for all k,

$$\mathbb{E}[|S_k|] = \mathbb{E}\left[\left|\prod_{i=1}^{k} Y_i\right|\right] = \mathbb{E}\left[\prod_{i=1}^{k} |Y_i|\right] \overset{\text{iid}}{=} \prod_{i=1}^{k} \mathbb{E}[|Y_i|] < \infty\,.$$

For the martingale property note that S_k is $\mathcal{F}_k$-measurable by definition, thus we have

$$\begin{aligned}\mathbb{E}[S_{k+1}|\mathcal{F}_k] &= \mathbb{E}\Big[\prod_{i=1}^{k+1} Y_i \,|\, \mathcal{F}_k\Big] = \mathbb{E}\Big[Y_{k+1}S_k \,|\, \mathcal{F}_k\Big] = \mathbb{E}\Big[Y_{k+1} \,|\, \mathcal{F}_k\Big]S_k = \mathbb{E}[Y_{k+1}]S_k \\ &= S_k.\end{aligned}$$

□

(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)

Solution. We have

$$\mathbb{E}[|aX_k + bY_k|] < |a|\mathbb{E}[|X_k|] + |b|\mathbb{E}[|Y_k|] < \infty$$

for all k and the martingale property follows from linearity of the conditional expectation

$$\mathbb{E}[aX_{k+1} + bY_{k+1}|\mathcal{F}_k] = a\mathbb{E}[X_{k+1}|\mathcal{F}_k] + b\mathbb{E}[Y_{k+1}|\mathcal{F}_k] = aX_k + bY_k.$$

□

(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)

Solution. Since X_k, Y_k are super-martingales, we have $\mathbb{E}[|\min\{X_k, Y_k\}|] \leq \mathbb{E}[|X_k|] < \infty$. Since $\min$ is a measurable function, $Z_k := \min\{X_k, Y_k\}$ is $\mathcal{F}_k$-measurable and we obtain the super-martingale property from

$$\mathbb{E}[Z_{k+1} \,|\, \mathcal{F}_k] = \mathbb{E}[\min\{X_{k+1}, Y_{k+1}\} \,|\, \mathcal{F}_k] \leq \mathbb{E}[X_{k+1} \,|\, \mathcal{F}_k] \leq X_k$$

and similarly $\mathbb{E}[Z_{k+1} \,|\, \mathcal{F}_k] \leq Y_k$ which together imply

$$\mathbb{E}[Z_{k+1} \,|\, \mathcal{F}_k] \leq \min\{X_k, Y_k\}.$$

□

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

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

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

Solution. From $F(x) = \mathbb{E}[f(x,Z)]$ follows $\nabla_x F(x) = \mathbb{E}[\nabla_x f(x,Z)]$. Using

$$\Big\|\sum_k v_k\Big\|^2 = \Big\langle \sum_k v_k, \sum_n v_n \Big\rangle = \sum_{k,n} \langle v_k, v_n\rangle \tag{2}$$

we have, since the $Z^{(i)}$ are iid

$$\begin{aligned}
&\mathbb{E}\Big[\Big\|\frac{1}{B}\sum_{m=1}^{B} \nabla_x f(x, Z^{(m)}) - \nabla_x F(x)\Big\|^2\Big] \\
&\overset{(2)}{=} \mathbb{E}\Big[\frac{1}{B^2}\sum_{m,n=1}^{B} \Big\langle \nabla_x f(x, Z^{(m)}) - \nabla_x F(x), \nabla_x f(x, Z^{(n)}) - \nabla_x F(x)\Big\rangle^2\Big] \\
&= \frac{1}{B^2}\sum_{m,n=1}^{B} \mathbb{E}\Big[\Big\langle \nabla_x f(x, Z^{(m)}) - \nabla_x F(x), \nabla_x f(x, Z^{(n)}) - \nabla_x F(x)\Big\rangle^2\Big] \\
&\overset{\text{iid}}{=} \frac{1}{B^2}\sum_{m=1}^{B} \underbrace{\mathbb{E}\Big[\big\|\nabla_x f(x, Z^{(m)}) - \nabla_x F(x)\big\|^2\Big]}_{\leq c} \\
&\leq \frac{c}{B}.
\end{aligned}$$

□

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 $(Z_k^{(m)})_{k\in\mathbb{N},\, m=1,\dots,B_{k-1}}$ 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)

Solution. Until convergence select

$$X_{k+1} := X_k - \alpha_k G_k(X_k)$$

and increase k (cf. Algorithm 8 in the lecture notes). □

(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

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

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)

Solution. The proof is analogous to the proof of Theorem 4.1.16, i.e. let the filtration be given by $\mathcal{F}_n := \sigma(X_0, (Z_k^{(m)})_{m=1,\dots,B_k}, k \leq n)$ such that by induction X_n is $\mathcal{F}_n$-measurable with induction step $n \to n+1$

$$X_{n+1} = X_n - \alpha_n G_n(X_n) = \underbrace{X_n}_{\mathcal{F}_n\text{-meas.}} - \alpha_n \frac{1}{B_n}\sum_{m=1}^{B_m} \underbrace{\nabla f(X_n, Z_{n+1}^{(m)})}_{\mathcal{F}_{n+1}\text{-meas.}}.$$

Let $M_{k+1} := \nabla F(X_k) - G_k(X_k)$, then we have

$$\begin{aligned}
&\big\|X_{k+1} - x_*\big\|^2 \\
&= \big\|X_k - \alpha_k G_k(X_k) - x_*\big\|^2 \\
&= \big\|X_k - x_*\big\|^2 - 2\alpha_k\big\langle \underbrace{G_k(X_k)}_{=\nabla F(X_k) - M_{k+1}}, X_k - x_*\big\rangle + \alpha_k^2\big\|G_k(X_k)\big\|^2 \\
&= \big\|X_k - x_*\big\|^2 - 2\alpha_k\big\langle \nabla F(X_k), X_k - x_*\big\rangle + \alpha_k^2\big\|\nabla F(X_k)\big\|^2 \\
&\qquad + 2\alpha_k\big\langle M_{k+1}, X_k - x_*\big\rangle - 2\alpha_k^2\big\langle M_{k+1}, \nabla F(X_k)\big\rangle + \alpha_k^2\big\|M_{k+1}\big\|^2
\end{aligned}$$

Since X_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[G_k(X_k) \,|\, \mathcal{F}_k] = \nabla F(X_k)$, we have

$$\mathbb{E}[M_{k+1} \,|\, \mathcal{F}_k] = 0 \tag{3}$$

$$\mathbb{E}\big[\big\|M_{k+1}\big\|^2 \,|\, \mathcal{F}_k\big] \overset{(i)}{\leq} \frac{c}{B_k}. \tag{4}$$

This implies

$$\begin{aligned}
&\mathbb{E}\big[\big\|X_{k+1} - x_*\big\|^2 \,|\, \mathcal{F}_k\big] \\
&\overset{(3)}{=} \big\|X_k - x_*\big\|^2 - 2\alpha_k\big\langle \nabla F(X_k), X_k - x_*\big\rangle + \alpha_k^2\big\|\nabla F(X_k)\big\|^2 + \alpha_k^2\mathbb{E}\big[\big\|M_{k+1}\big\|^2 \,|\, \mathcal{F}_k\big] \\
&\overset{(4)}{\leq} \big\|X_k - x_*\big\|^2 - 2\alpha_k\big\langle \nabla F(X_k), X_k - x_*\big\rangle + \alpha_k^2\big\|\nabla F(X_k)\big\|^2 + \alpha_k^2\frac{c}{B_k}.
\end{aligned} \tag{5}$$

Recall that μ-strong convexity implies

$$\langle \nabla F(X_k), x_* - X_k\rangle + \tfrac{\mu}{2}\big\|x_* - X_k\big\|^2 \leq F(x_*) - F(X_k)$$

which, rearranged, gives

$$-\langle \nabla F(X_k), X_k - x_*\rangle \leq F(x_*) - F(X_k) - \tfrac{\mu}{2}\big\|x_* - X_k\big\|^2. \tag{6}$$

L-smoothness on the other hand implies via the descent lemma

$$F(x_*) \leq F(X_k) - \tfrac{1}{2L}\|\nabla F(X_k)\|^2$$

which, rearranged, gives

$$F(x_*) - F(X_k) \leq -\tfrac{1}{2L}\|\nabla F(X_k)\|^2 \tag{7}$$

Starting with (5) and applying (6) and then (7) we get

$$\begin{aligned}
&\mathbb{E}\big[\big\|X_{k+1} - x_*\big\|^2 \,|\, \mathcal{F}_k\big] \\
&\overset{(6)}{\leq} \big\|X_k - x_*\big\|^2 + 2\alpha_k\Big[F(x_*) - F(X_k) - \tfrac{\mu}{2}\big\|x_* - X_k\big\|^2\Big] + \alpha_k^2\big\|\nabla F(X_k)\big\|^2 + \alpha_k^2\frac{c}{B_k} \\
&\overset{(7)}{\leq} \big\|X_k - x_*\big\|^2 + 2\alpha_k\Big[-\tfrac{1}{2L}\|\nabla F(X_k)\|^2 - \tfrac{\mu}{2}\big\|x_* - X_k\big\|^2\Big] + \alpha_k^2\big\|\nabla F(X_k)\big\|^2 + \alpha_k^2\frac{c}{B_k} \\
&= (1-\alpha_k\mu)\big\|X_k - x_*\big\|^2 + \alpha_k\underbrace{\big(\alpha_k - \tfrac{1}{L}\big)}_{\leq 0}\big\|\nabla F(X_k)\big\|^2 + \alpha_k^2\frac{c}{B_k} \\
&\leq (1-\alpha_k\mu)\big\|X_k - x_*\big\|^2 + c\frac{\alpha_k^2}{B_k}.
\end{aligned}$$

□

(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)

Solution. There are many possible combinations of sequences, one possibility is to set the batchsize to be constant $B_k = B$. In this case, our upper bound is equal to the one without minibatching up to a constant and the step size from Corollary 4.1.17 works.

Another possibility is to select a constant step size $\alpha < \frac{1}{L}$ and only change the batch size. In this case we have for $\Delta_k = \mathbb{E}[\|X_k - x_*\|^2]$ by induction

$$\Delta_{n+1} \leq (1-\alpha\mu)^{n+1}\Delta_0 + c\alpha^2\sum_{k=0}^{n}\frac{(1-\alpha\mu)^{n-k}}{B_k}.$$

The induction step $(n-1) \to n$ follows from the previous exercise, i.e.

$$\begin{aligned}
\Delta_{n+1} &\leq (1-\alpha\mu)\Delta_n + c\alpha^2\frac{1}{B_n} \\
&\overset{\text{ind.}}{\leq} (1-\alpha\mu)\Big[(1-\alpha\mu)^n\Delta_0 + c\alpha^2\sum_{k=0}^{n-1}\frac{(1-\alpha\mu)^{n-1-k}}{B_k}\Big] + c\alpha^2\frac{1}{B_n} \\
&\leq (1-\alpha\mu)^{n+1}\Delta_0 + c\alpha^2\sum_{k=0}^{n}\frac{(1-\alpha\mu)^{n-k}}{B_k}.
\end{aligned}$$

To prove that $0 < \alpha < \frac{1}{L}$ and $\sum_{k=0}^{\infty}\frac{1}{B_k} < \infty$ are sufficient conditions for convergence we use

$\frac{1}{L} \leq \frac{1}{\mu}$, which implies $1 > (1 - \alpha\mu) > 0$ and split the sum into two parts

$$\begin{aligned}
\Delta_{n+1} &\leq (1-\alpha\mu)^{n+1}\Delta_0 + c\alpha^2\Big[\sum_{k=0}^{\lceil \frac{n}{2} \rceil} \frac{(1-\alpha\mu)^{n-\lceil \frac{n}{2} \rceil}}{B_k} + \sum_{k=\lceil \frac{n}{2} \rceil+1}^{n} \frac{(1-\alpha\mu)^{n-k}}{B_k}\Big] \\
&\leq \underbrace{(1-\alpha\mu)^{n+1}}_{\to 0}\Delta_0 + c\alpha^2\Big[\underbrace{(1-\alpha\mu)^{\frac{n}{2}}}_{\to 0}\underbrace{\sum_{k=0}^{\infty}\frac{1}{B_k}}_{<\infty} + \underbrace{\sum_{k=\lceil \frac{n}{2} \rceil+1}^{\infty}\frac{1}{B_k}}_{\to 0}\Big] \\
&\to 0 \qquad (n \to \infty).
\end{aligned}$$

Many other options are available and it is not clear what the optimal choice is. □