Solution Sheet 5

For the exercise class on the 11.05.2023.
Hand in your solutions by 12:00 in the exercise on Thursday 11.05.2023.

Exercise 1 (Conditional Expectation)**.** **(2 Points)**

Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $\mathcal{F}$ a subalgebra of $\mathcal{A}$ and X, Y random vectors. Prove for $\mathcal{F}$-measurable $X \in \mathbb{R}^d$ that we have

$$\mathbb{E}[\langle X, Y\rangle \mid \mathcal{F}] = \langle X, \mathbb{E}[Y \mid \mathcal{F}]\rangle$$

Solution. We simply use linearity of the conditional expectation

$$\mathbb{E}[\langle X, Y\rangle \mid \mathcal{F}] = \mathbb{E}\left[\sum_{i=1}^d X_i Y_i \mid \mathcal{F}\right] = \sum_{i=1}^d \mathbb{E}\big[X_i Y_i \mid \mathcal{F}\big] = \sum_{i=1}^d X_i \mathbb{E}\big[Y_i \mid \mathcal{F}\big] = \langle X, \mathbb{E}[Y \mid \mathcal{F}]\rangle. \quad \square$$

Exercise 2 (Convexity and Expectation)**.** **(2 Points)**

Let Z be a random variable. Let $f(x) := f(x, Z)$ be a random function ($f(x, \omega) = f(x, Z(\omega))$ if you want) and its expectation

$$F(x) = \mathbb{E}[f(x)]$$

Is f almost surely convex if and only if F is convex? Prove or disprove both directions.

Solution. Let us first assume f was almost surely convex. Then due to monotonicity of expectation we have

$$\begin{aligned}
F(\lambda x + (1-\lambda) y) &= \mathbb{E}[f(\lambda x + (1-\lambda) y)] \\
&\le \mathbb{E}[\lambda f(x) + (1-\lambda) f(y)] \\
&= \lambda \mathbb{E}[f(x)] + (1-\lambda)\mathbb{E}[f(y)] = \lambda F(x) + (1-\lambda) F(y)
\end{aligned}$$

so F is convex. The other direction is false. For this let $\mathbb{P}(Z = -1) = (1-p)$ and $\mathbb{P}(Z = 1) = p$ with $p > 0.5$, and $f(x, z) = zx^2$. Then

$$F(x) = \mathbb{E}[f(x, Z)] = \mathbb{E}[Z]x^2 = (2p-1)x^2$$

is convex, but f is not convex, i.e. $f(x) = -x^2$ with probability $(1-p)$. $\square$

Exercise 3 (Convergence of SGD on Strongly Convex Functions)**.** **(2 Points)**

In the lecture we proved for L-smooth functions F and X_n generated by Algorithm 6 (SGD)

$$\|\nabla F(X_n)\|^2 \to 0 \quad \text{a.s.}$$

If we additionally have strong convexity of F, prove $\|X_n - x_*\| \to 0$ almost surely.

Solution. On sheet 3 we proved the PL inequality for L-smooth, strongly convex functions. This together with strong convexity implies

$$\frac{\mu}{2}\|X_n - x_*\|^2 \le F(X_n) - F(x_*) - \langle \underbrace{\nabla F(x_*)}_{=0}, X_n - x_* \rangle \overset{\text{PL}}{\le} \frac{L}{2\mu}\|\nabla F(X_n)\|^2 \to 0.$$

From part in the middle we get $F(X_n) \to F(x_*)$ for free. □

Exercise 4 (Swap Integration with Differentiation)**.** **(9 Points)**

(i) What formal requirements on $f : V \times \Omega \to \mathbb{R}$ with $V \subseteq \mathbb{R}$ and measure μ on Ω are needed, for the following argument using the fundamental theorem of calculus (FTC) to work?

$$\begin{aligned}
\frac{\partial}{\partial t}\int_\Omega f(t_0,\omega)d\mu(\omega) &\overset{\text{linear}}{=} \lim_{\epsilon\to 0}\int \frac{f(t_0+\epsilon,\omega) - f(t_0,\omega)}{\epsilon}d\mu(\omega) \\
&\overset{\text{FTC II}}{=} \lim_{\epsilon\to 0}\int \frac{1}{\epsilon}\int_{t_0}^{t_0+\epsilon} \frac{\partial}{\partial t} f(t,\omega)dt d\mu(\omega) \\
&\overset{\text{Fubini}}{=} \lim_{\epsilon\to 0}\frac{1}{\epsilon}\int_{t_0}^{t_0+\epsilon}\int \frac{\partial}{\partial t} f(t,\omega)d\mu(\omega)dt \\
&\overset{\text{def.+lin.}}{=} \frac{d}{dy}\int_{t_0}^{y}\int \frac{\partial}{\partial t} f(t,\omega)d\mu(\omega)dt\bigg|_{y=t_0} \\
&\overset{\text{FTC I}}{=} \int \frac{\partial}{\partial t} f(t_0,\omega)d\mu(\omega).
\end{aligned}$$

Formulate the corresponding theorem. (5 pts)

Solution. (a) Linearity of the integral, requires measurability and either positivity ($f \ge 0$) or μ-integrability of f with regard to ω at t_0.

(b) For Fubini, we need either $\frac{\partial}{\partial t} f(t,\omega) \ge 0$, or

$$\int_{t_0}^{t_0+\epsilon}\int \left|\frac{\partial}{\partial t} f(t,\omega)\right| d\omega dt < \infty.$$

Since we let $\epsilon \to 0$, there only needs to be a small environment around t_0, where this is the case. I.e. we need "local-integrability" around t_0 of $\partial_t f$ with regard to t

$$\exists a,b \in \mathbb{R}: \quad t_0 \in [a,b],\ a < b,\ \int_a^b \int \left|\frac{\partial}{\partial t} f(t,\omega)\right| d\omega dt < \infty.$$

This also covers the second usage of linearity.

(c) For the second fundamental theorem of calculus (FTC II), we do not even need $f(\cdot,\omega)$ to be continuous. It is sufficient, if it is for μ-almost-all ω absolutely continuous (i.e. a density exists).

(d) For the first fundamental theorem of calculus (FTC I), we need continuity. I.e.

$$t \mapsto \int \partial_t f(t,\omega) d\mu(\omega)$$

needs to be continuous.

So we get

Theorem (Swap Integration and Differentiation)**.** *Let $f : U \times \Omega \to \mathbb{R}$ be a measurable function for $U \subseteq \mathbb{R}$ which satisfies*

(a) f μ-integrable over $\omega \in \Omega$ at $t_0 \in U$,

(b) for μ-almost all $\omega \in \Omega$ we have: $t \to f(t,\omega)$ is differentiable (or absolutely continuous) in t.

(c) If $\frac{\partial}{\partial t} f$ is further "locally integrable in t_0", i.e. there exists $a < b \in \mathbb{R}$, such that $t_0 \in [a,b] \subseteq U$ and

$$\exists a < b \in \mathbb{R} : \quad t_0 \in [a,b] \subseteq U, \quad \int_a^b \int_\Omega \left| \frac{d}{dt} f(t,\omega) \right| d\mu(\omega) dt < \infty,$$

or $\frac{\partial}{\partial t} f \geq 0$ in the neighborhood $[a,b]$.

(d) In a similar local neighborhood $[a,b]$ of t_0 assume that

$$t \mapsto \int \frac{\partial}{\partial t} f(t,\omega) d\mu(\omega)$$

is continuous.

then swapping derivative in t_0 and integration over ω is allowed

$$\frac{\partial}{\partial t} \int f(t_0,\omega) d\mu(\omega) = \int \frac{\partial}{\partial t} f(t_0,\omega) d\mu(\omega).$$

□

(ii) We want to find an example for a function, where you can not swap integration with differentiation. So for a function $f(t,\omega)$ we need some t_0 such that

$$\frac{\partial}{\partial t} \int_\Omega f(t_0,\omega) d\omega \neq \int_\Omega \frac{\partial}{\partial t} f(t_0,\omega) d\omega.$$

For this consider $f(t,\omega) = t^3 e^{-t^2\omega}$. Prove the inequality at $t_0 = 0$ and $\Omega = [0,\infty)$. Why is this not a contradiction to (i)? (4 pts)

Solution. We have

$$\int_0^\infty f(t,\omega) d\omega = \int_0^\infty t^3 e^{-t^2\omega} d\omega = \begin{cases} -te^{-t^2\omega}\Big|_{\omega=0}^{\omega=\infty} & t \neq 0 \\ 0 & t = 0 \end{cases}$$
$$= t.$$

So its derivative is constant

$$\frac{\partial}{\partial t}\int_0^\infty f(t,\omega)d\omega = 1.$$

In particular this is also the case at $t_0 = 0$. On the other hand we have

$$\frac{\partial}{\partial t}f(t,\omega) = 3t^2e^{-t^2\omega} - 2t^4\omega e^{-t^2\omega} = t^2e^{-t^2\omega}(3-2t^2\omega).$$

In particular $\frac{\partial}{\partial t}f(t_0,\omega) = 0$. Therefore

$$\int_0^\infty \frac{\partial}{\partial t}f(t_0,\omega)d\omega = 0 \neq 1 = \frac{\partial}{\partial t}\int_0^\infty f(t_0,\omega)d\omega.$$

We also have for $t \neq 0$

$$\begin{aligned}
\int_0^\infty \frac{\partial}{\partial t}f(t,\omega)d\omega &= -3e^{-t^2\omega}\Big|_{\omega=0}^{\omega=\infty} - 2t^2\int_0^\infty t^2\omega e^{-t^2\omega}d\omega \\
&= 3 - 2t^2\left[\omega e^{-t^2\omega}\Big|_{\omega=0}^{\omega=\infty} - \int_0^\infty \underbrace{\left(\frac{d}{d\omega}\omega\right)}_{=1} e^{-t^2\omega}d\omega\right] \\
&= 3 - 2\int_0^\infty t^2e^{-t^2\omega}d\omega \\
&= 1
\end{aligned}$$

So in total

$$\int_0^\infty \frac{\partial}{\partial t}f(t,\omega)d\omega = \begin{cases} 0 & t = 0 \\ 1 & t \neq 0. \end{cases}$$

In particular

$$t \mapsto \int_0^\infty \frac{\partial}{\partial t}f(t,\omega)d\omega$$

is not continuous. But this was a requirement for (i) so this is not a contradiction. □

Exercise 5 (SGD on quadratic functions)**.** **(9 Points)**

Throughout we use the notation for SGD

$$X_{n+1} = X_n - \alpha_n \nabla f_{n+1}(X_n)$$

using $X_0 = x_0 \in \mathbb{R}^d$ with sample errors $\epsilon_n = \nabla f_n(x) - \nabla F(x)$ for stochastic gradients

$$f_n(x) := f(x, Z_n)$$

for sample data $(Z_n)_n$ with $Z_n \overset{\text{iid}}{\sim} \mu$ random vectors in $\mathbb{R}^d$. Additionally we write for GD

$$x_{n+1} = x_n - \alpha_n F(x_n).$$

(i) Prove for any y_0 and the recursion

$$y_{n+1} := y_n - \tfrac{1}{n+1}(y_n - z_{n+1})$$

that y_n is a running mean

$$y_n = \frac{1}{n}\sum_{k=1}^n z_k =: \bar{z}_n, \quad \forall n \in \mathbb{N} \qquad \text{(1 pt)}$$

Solution. This is simply induction with induction start

$$y_1 = x_0 - (y_0 - z_1) = z_1$$

and induction step

$$y_{n+1} = \left(1 - \tfrac{1}{n+1}\right)\frac{1}{n}\sum_{k=1}^{n} z_n + \tfrac{1}{n+1} z_{n+1} = \frac{1}{n+1}\sum_{k=1}^{n+1} z_k. \qquad \square$$

(ii) Let $Z \in \mathbb{R}^d$ be a random vector and consider the sample loss

$$f(x, Z) := \tfrac{1}{2}\|x - Z\|_H^2 \overset{\text{recall}}{=} \tfrac{1}{2}\langle x - Z, H(x - Z)\rangle$$

Prove that

$$F(x) = \mathbb{E}[f(x, Z)] = \tfrac{1}{2}\|x - x_*\|_H^2 + \text{const.}$$

with

$$x_* = \operatorname*{argmin}_x \mathbb{E}[\|x - Z\|_H^2].$$

What is x_*? What is the (in the L^2 sense) optimal step size for SGD in the case $H = \mathbb{I}$? (4 pts)

Solution. We have

$$\begin{aligned} 2\mathbb{E}[f(x, Z)] = \mathbb{E}\|x - Z\|_H^2 &= \|x\|_H^2 - 2\mathbb{E}\langle x, Z\rangle_H + \mathbb{E}\|Z\|_H^2 \\ &= \|x\|_H^2 - 2x^T H\mathbb{E}[Z] + \|\mathbb{E}[Z]\|_H^2 + (\mathbb{E}\|Z\|_H^2 - \|\mathbb{E}[Z]\|_H^2) \\ &= \|x - \mathbb{E}[Z]\|_H^2 + \underbrace{(\mathbb{E}\|Z\|_H^2 - \|\mathbb{E}[Z]\|_H^2)}_{=\text{const.}}. \end{aligned}$$

So $x_* = \mathbb{E}[Z]$ does the job. We know that $\bar{Z}_n$ is the minimum variance estimator of $\mathbb{E}[Z]$, so it would be perfect if $X_n = \bar{Z}_n$. And with $\alpha_n = \frac{1}{n+1}$ this is in fact possible if we recall

$$\nabla f_n(x) = \nabla_x \tfrac{1}{2}\|x - Z_n\|_H^2 = H(x - Z_n) = (x - Z_{n+1}).$$

Because then we get

$$X_{n+1} = X_n - \alpha_n \nabla f_{n+1}(X_n) = X_n - \tfrac{1}{n+1}(X_n - Z_{n+1}).$$

By the previous exercise, this is therefore the optimal step size. $\square$

(iii) Prove for this quadratic loss, that SGD can be written as GD plus accumulated error

$$X_n - x_* = (x_n - x_*) - \sum_{k=0}^{n-1} \alpha_k \left(\prod_{i=k+1}^{n-1} (1 - \alpha_i H)\right)\epsilon_{k+1}. \qquad \text{(2 pts)}$$

Solution. Recall

$$\nabla F(x) = \nabla_x \tfrac{1}{2}\|x - x_*\|_H^2 = H(x - x_*).$$

Then by induction with induction start $n = 0$ (clear) and induction step

$$
\begin{aligned}
X_{n+1} - x_* &= X_n - \alpha_n(\nabla F(X_n) + \epsilon_{n+1}) - x_* \\
&= X_n - x_* - \alpha_n H(X_n - x_*) - \alpha_n \epsilon_{n+1} \\
&= (1 - \alpha_n H)(X_n - x_*) - \alpha_n \epsilon_{n+1} \\
&\overset{\text{ind.}}{=} (1 - \alpha_n H)(x_n - x_*) - (1 - \alpha_n H) \sum_{k=0}^{n-1} \alpha_k \left(\prod_{i=k+1}^{n-1} (1 - \alpha_i H) \right) \epsilon_{k+1} - \alpha_n \epsilon_{n+1} \\
&= (x_{n+1} - x_*) - \sum_{k=0}^{n-1} \alpha_k \left(\prod_{i=k+1}^{n} (1 - \alpha_i H) \right) \epsilon_{k+1} - \alpha_n \epsilon_{n+1} \\
&= (x_{n+1} - x_*) - \sum_{k=0}^{n} \alpha_k \left(\prod_{i=k+1}^{n} (1 - \alpha_i H) \right) \epsilon_{k+1}. \qquad \square
\end{aligned}
$$

(iv) Consider the previous setting with constant step sizes $\alpha_n = \alpha$. Additionally we are going to assume f is a quadratic loss with $H = \mathbb{I}$. Prove

$$X_n = (1 - \alpha)^n x_0 + \sum_{k=1}^{n} \alpha (1 - \alpha)^{n-k} Z_k.$$

Compare the estimate X_n to the mean $\bar{Z}_n$. (2 pts)

Solution. Proof by induction with induction start $X_0 = x_0$ and induction step

$$
\begin{aligned}
X_{n+1} &= X_n - \alpha \nabla f_{n+1}(X_n) \\
&= X_n - \alpha[X_n - Z_{n+1}] = (1 - \alpha) X_n - \alpha Z_{n+1} \\
&\overset{\text{ind.}}{=} (1 - \alpha)\Big((1 - \alpha)^n x_0 + \sum_{k=1}^{n} \alpha (1 - \alpha)^{n-k} Z_k\Big) - \alpha Z_{n+1} \\
&= (1 - \alpha)^{n+1} x_0 + \sum_{k=1}^{n+1} \alpha (1 - \alpha)^{n+1-k} Z_k. \qquad \square
\end{aligned}
$$