Practical Session 5

Optimization

1 Convexity

Problem 1: Show that affine functions of the form $\boldsymbol{w}^T\boldsymbol{x} + b$ are both convex and concave.

A function is convex iff $\lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y}) \geq f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y})$ and concave iff $\lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y}) \leq f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y})$. Hence, for a function to be both convex and concave it must hold that

$$\lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y}) = f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) .$$

We have

$$\begin{aligned}
\lambda f(\boldsymbol{x}) + (1-\lambda)f(\boldsymbol{y}) &= \lambda[\boldsymbol{w}^T\boldsymbol{x} + b] + (1-\lambda)[\boldsymbol{w}^T\boldsymbol{y} + b] \\
&= \lambda\boldsymbol{w}^T\boldsymbol{x} + (1-\lambda)\boldsymbol{w}^T\boldsymbol{y} + \lambda b + (1-\lambda)b \\
&= \lambda\boldsymbol{w}^T\boldsymbol{x} + (1-\lambda)\boldsymbol{w}^T\boldsymbol{y} + b \\
&= \boldsymbol{w}^T[\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}] + b \\
&= f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \qquad \square
\end{aligned}$$

Problem 2: Show that a twice differentiable function $f(\boldsymbol{x})$ with a convex domain is convex if and only if its Hessian or second derivative is positive semidefinite: $\nabla^2 f(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \operatorname{dom}(f)$.

We first assume we have a single dimension ($n = 1$). Suppose $f : \mathbb{R} \to \mathbb{R}$ is convex. Let $x, y \in \operatorname{dom}(f)$.

By the first-order condition,

$$\begin{aligned}
&f'(x)(y-x) \leq f(y) - f(x) , \\
&f'(y)(x-y) \leq f(x) - f(y) \quad \Leftrightarrow \quad f(y) - f(x) \leq f'(y)(y-x) .
\end{aligned}$$

Hence,

$$f'(x)(y-x) \leq f'(y)(y-x) .$$

Subtracting the lefthand side from the righthand side and dividing by $(y-x)^2$ gives:

$$\frac{f'(y) - f'(x)}{y-x} \geq 0$$

Taking the limit for $y \to x$ yields $f''(x) \geq 0$, for any $x \in \operatorname{dom}(f)$.

Conversely, suppose $f''(z) \geq 0$ for all $z \in \mathrm{dom}(f)$. Consider two arbitrary points $x, y \in \mathrm{dom}(f)$. Without loss of generality we assume that $x < y$. We have

$$\begin{aligned} 0 &\leq \int_x^y f''(z)(y-z)\,\mathrm{d}z \\ &= \left(f'(z)(y-z)\right)\Big|_{z=x}^{z=y} + \int_x^y f'(z)\,\mathrm{d}z \\ &= -f'(x)(y-x) + f(y) - f(x) \end{aligned}$$

i.e. $f(y) \geq f(x) + f'(x)(y-x)$, which is the first order convexity condition and shows that f is convex.

To generalize to $n > 1$, we note that a function is convex if and only if it is convex on all lines, i.e. iff the one-dimensional function $g(t) = f(\boldsymbol{x}_0 + t\boldsymbol{v})$ is convex in t for all $\boldsymbol{x}_0 \in \mathrm{dom}(f)$ and all $\boldsymbol{v} \in \mathbb{R}^n$, for values satisfying $\boldsymbol{x}_0 + t\boldsymbol{v} \in \mathrm{dom}(f)$. Therefore, f is convex if and only if

$$g''(t) = \boldsymbol{v}^T \nabla^2 f(\boldsymbol{x}_0 + t\boldsymbol{v})\boldsymbol{v} \geq 0\,.$$

In other words, it is necessary and sufficient that $\boldsymbol{v}^T \nabla^2 f(\boldsymbol{x})\boldsymbol{v} \geq 0$ for all $\boldsymbol{x} = \boldsymbol{x}_0 + t\boldsymbol{v} \in \mathrm{dom}(f)$, which is exactly the definition of the Hessian $\nabla^2 f(\boldsymbol{x})$ being positive semi-definite. □

Note: For the strictly convex case one can show that if $\nabla^2 f(\boldsymbol{x})$ is positive definite then f is strictly convex. The converse, however, is not true: For example, the function $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^4$ is strictly convex but has zero second derivative at $x = 0$.

2 Logistic Regression

Problem 3: Prove that the objective function of logistic regression

$$E(\boldsymbol{w}) = -\ln p\big(\boldsymbol{y} \mid \boldsymbol{w}, \boldsymbol{X}\big) = -\sum_{i=1}^{N} \big(y_i \ln \sigma(\boldsymbol{w}^T\boldsymbol{x}_i) + (1-y_i)\ln(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))\big) \tag{1}$$

is convex. What is the benefit of having a convex function for optimization?

First, notice that if we can prove that the following two functions

$$-\ln \sigma(\boldsymbol{w}^T\boldsymbol{x}_i) \qquad \text{and} \qquad -\ln(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))$$

are convex, our objective function as given in Eq.1 must also be convex since any linear combination (with positive constants) of two or more convex combinations is also convex. Since y_i and $1 - y_i$ are positive this holds.

To prove that the first function is convex we will use the second-order condition of convexity.

Reminder: A function $f(x)$ which is twice-differentiable is convex if and only if its Hessian matrix (matrix of second-order partial derivatives) is positive semi-definite.

To compute the Hessian matrix we first calculate the derivative of the sigmoid function:

$$\begin{aligned}\frac{\partial}{\partial x}\sigma(x) &= \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x)\frac{e^{-x}}{1+e^{-x}} = \sigma(x)\left(1+\frac{e^{-x}-1-e^{-x}}{1+e^{-x}}\right) = \sigma(x)\left(1-\frac{1}{1+e^{-x}}\right)\\ &= \sigma(x)(1-\sigma(x))\,.\end{aligned}$$

Using this, we can derive the Hessian:

$$\begin{aligned}\nabla^2_{\boldsymbol{w}}\left[-\ln\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)\right] &= \nabla_{\boldsymbol{w}}\left[\nabla_{\boldsymbol{w}}(-\ln\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))\right]\\ &= \nabla_{\boldsymbol{w}}\left[-\boldsymbol{x}_i\frac{\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))}{\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)}\right]\\ &= \nabla_{\boldsymbol{w}}\left[\boldsymbol{x}_i(\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)-1)\right]\\ &= \sigma(\boldsymbol{w}^T\boldsymbol{x}_i)(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))\boldsymbol{x}_i\boldsymbol{x}_i^T\,.\end{aligned}$$

Next, we show that this Hessian matrix is positive semi-definite:

$$\begin{aligned}\forall\boldsymbol{z}:\quad & \boldsymbol{z}^T\nabla^2_{\boldsymbol{w}}\left[-\ln\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)\right]\boldsymbol{z}\\ &= \boldsymbol{z}^T\left[\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))\boldsymbol{x}_i\boldsymbol{x}_i^T\right]\boldsymbol{z}\\ &= \sigma(\boldsymbol{w}^T\boldsymbol{x}_i)(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))(\boldsymbol{x}_i^T\boldsymbol{z})^2 \geq 0\end{aligned}$$

To prove that the second function is convex, we first notice:

$$\begin{aligned}-\ln(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)) &= -\ln\left(1-\frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}_i}}\right) = -\ln\left(\frac{e^{-\boldsymbol{w}^T\boldsymbol{x}_i}}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}_i}}\right)\\ &= \boldsymbol{w}^T\boldsymbol{x}_i - \ln\left(\frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}_i}}\right) = \boldsymbol{w}^T\boldsymbol{x}_i - \ln\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)\end{aligned}$$

This is a sum of two convex functions, since the affine function $\boldsymbol{w}^T\boldsymbol{x}_i$ is convex and we just showed that $-\ln\sigma(\boldsymbol{w}^T\boldsymbol{x}_i)$ is convex. Hence, $-\ln(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}_i))$ is convex as well. □

The benefit of having a convex function for optimization is that, subject to relatively mild assumptions, stochastic gradient descent converges almost surely to a global minimum.

3 Optimization methods

Problem 4: Discuss the following topics:

- Condition number
- Consistency, convergence, stability
- Stiffness

- Condition number: How much does output vary depending on the input, more precisely the maximum ratio of the relative error in the output $\hat{y}$ due to the relative error in the input x. Ill-conditioned: High condition number. This is a property of the problem itself, not of the algorithm.
- Consistency: Local discretization error (error due to a single step) $l(\delta t) \to 0$ for $\delta t \to 0$
- Convergence: Global discretization error (overall error) $e(\delta t) \to 0$ for $\delta t \to 0$
- Stability: Algorithms that do not magnify approximation errors. Instabilities can be caused e.g. by nearby singularities (e.g. very small eigenvalues), truncation errors, or loss of significance. Stability + consistency = convergence
- Stiffness: Local property of the algorithm's solution imposes extremely high resolution.