Mathematics for Intelligent Systems
Lecture 7 Homework
(Gradients and Hessians)

Marc Toussaint, Andrea Baisero

1 Gradient and Hessian of a GP posterior

We are given a fixed set of data points $\{x_i\}_{i=1}^n, x_i \in \mathbb{R}^d$, a fixed vector $y \in \mathbb{R}^n$, a fixed matrix $G \in \mathbb{R}^{n\times n}$, and a symmetric kernel function

$$k:\ \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R},\ k(x,x') = k(x',x)\ . \tag{1}$$

We assume that we know the gradient and hessian

$$\nabla_x k(x,x')\ ,\ \nabla_x^2 k(x,x')\ . \tag{2}$$

We also define the vector-valued function $\kappa:\ \mathbb{R}^d \to \mathbb{R}^n$,

$$\kappa(x) = \begin{pmatrix} k(x,x_1) \\ k(x,x_2) \\ \vdots \\ k(x,x_n) \end{pmatrix}\ . \tag{3}$$

In machine learning, the so-called Gaussian Process is given by the posterior mean and covariance functions,

$$\mu(x) = \kappa(x)^\intercal G y\ , \tag{4}$$

$$\sigma^2(x) = k(x,x) - \kappa(x)^\intercal G \kappa(x)\ . \tag{5}$$

(a) Compute the gradient $\nabla_x \mu(x)$ and hessian $\nabla_x^2 \mu(x)$ of the posterior mean.

(b) Compute the gradient $\nabla_x \sigma(x)$ and hessian $\nabla_x^2 \sigma(x)$ of the standard deviation function $\sigma(x) = \sqrt{\sigma^2(x)}$.

2 Gradient vs Direction of Steepest Descent

Recall that the derivative of a function (aka differential, aka directional derivative) $f : \mathcal{V} \to \mathbb{R}$ at a point x is defined as $df_x \in \mathcal{V}^*$,

$$df_x(v) = \lim_{t\to 0} \frac{f(x+tv) - f(x)}{t} \tag{6}$$

We have also seen that the gradient $\nabla f(x)$ is defined as the "column vector" of partial derivatives evaluated at x,

$$\nabla f(x) \equiv \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix} \tag{7}$$

Notice that the notion of derivative df_x is a coordinate-less one; it exists and is the same whatever the choice of a basis is. On the other hand, the notion of gradient $\nabla f(x)$ is by definition dependent on an algebraeic expression (i.e. on a coordinate representation). These two notion are nonetheless related by $df_x(v) = \nabla f(x) \cdot v$, which holds in any basis.

We define the vector of steepest descent $\delta^* \in \mathcal{V}$ at the point x as

$$\delta^* \equiv \arg\min_{\delta \in \mathcal{V}} df_x(\delta) \text{ s.t. } \delta^2 = 1 \tag{8}$$

The negative gradient $-\nabla f(x)$ is typically confused to be the vector of steepest descent. However, we will see that it is not necessarily the case.

(a) Find the vector of steepest descent δ^* in the case of Euclidean metric space, $\langle x, y \rangle = x^\intercal y$. Graphical argument is allowed.

(b) Find the vector of steepest descent δ^* in the case of non-Euclidean metric space, with metric tensor $A\mathbb{R}^{d \times d}$, i.e. $\langle x, y \rangle = x^\intercal A y$. Graphical argument is allowed, although some algebraeic manipulation is required. Hint: Use the Cholesky decomposition of the metric tensor $A = B^\intercal B$.