Labs
Optimization for Machine Learning
Spring 2019

EPFL
School of Computer and Communication Sciences
Martin Jaggi
github.com/epfml/OptML_course

Problem Set 8 — Solutions (Quasi-Newton Methods)

Exercise 48. *Consider a step of the secant method:*

$$x_{t+1} = x_t - f(x_t)\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}, \quad t \geq 1.$$

Assuming that $x_t \neq x_{t-1}$ and $f(x_t) \neq f(x_{t-1})$, prove that the line through the two points $(x_{t-1}, f(x_{t-1}))$ and $(x_t, f(x_t))$ intersects the x-axis at the point $x = x_{t+1}$.

Solution: Let the line be $y = ax + b$. Then we have

$$\begin{aligned} f(x_t) &= ax_t + b, \\ f(x_{t-1}) &= ax_{t-1} + b. \end{aligned}$$

Subtracting the two equations yields

$$a = \frac{f(x_t) - f(x_{t-1})}{x_t - x_{t-1}}.$$

To compute the intersection with the x-axis, we need to solve

$$0 = ax + b.$$

Subtracting from this the first of the previous two equations yields

$$-f(x_t) = a(x - x_t) \quad \Leftrightarrow \quad x = x_t - f(x_t)a^{-1} = x_t - f(x_t)\frac{x_t - x_{t-1}}{f(x_t) - f(x_{t-1})}.$$

By definition of the secant method, $x = x_{t+1}$.

Exercise 49. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a twice differentiable function with nonzero Hessians everywhere. Prove that the following two statements are equivalent.*

(i) f is a nondegenerate quadratic function, meaning that

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top M\mathbf{x} - \mathbf{q}^\top\mathbf{x} + c,$$

where $M \in \mathbb{R}^{d\times d}$ is an invertible symmetric matrix, $\mathbf{q} \in \mathbb{R}^d, c \in R$ (see also Lemma 7.1).

(ii) Applied to f, Newton's update step

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \nabla^2 f(\mathbf{x}_t)^{-1}\nabla f(\mathbf{x}_t), \quad t \geq 1$$

defines a Quasi-Newton method for all $\mathbf{x}_0, \mathbf{x}_1 \in \mathbb{R}^d$.

Solution: If f is a nondegenerate quadratic function then

$$\nabla f(\mathbf{x}) = M\mathbf{x} - \mathbf{q}, \quad \nabla^2 f(\mathbf{x}) = M,$$

so

$$\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_{t-1}) = M(\mathbf{x}_t - \mathbf{x}_{t-1}),$$

hence the secant condition (8.7) is satisfied with $H_t = M = \nabla^2 f(\mathbf{x})$, so Newton's method is a Quasi-Newton method.

Conversely, if for all $\mathbf{x}_0, \mathbf{x}_1$,

$$\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_0) = \nabla^2 f(\mathbf{x}_1)(\mathbf{x}_1 - \mathbf{x}_0),$$

then substituting $\mathbf{x}_0 = \mathbf{x}, \mathbf{x}_1 = \mathbf{0}$, we obtain

$$\nabla f(\mathbf{x}) - \nabla f(\mathbf{0}) = \nabla^2 f(\mathbf{0})\mathbf{x}, \quad \mathbf{x} \in \mathbb{R}^d.$$

Now let $g(t) = f(t\mathbf{x})$. We have $g'(t) = \nabla f(t\mathbf{x})^\top \mathbf{x}$. Using the previous equation in the fourth line, we get

$$\begin{aligned}
f(\mathbf{x}) - f(\mathbf{0}) &= g(1) - g(0) \\
&= \int_0^1 \nabla f(t\mathbf{x})^\top \mathbf{x} \, dt \\
&= \int_0^1 \left((\nabla f(t\mathbf{x}) - \nabla f(\mathbf{0}))^\top \mathbf{x} + \nabla f(\mathbf{0})^\top \mathbf{x}\right) dt \\
&= \int_0^1 \left((\nabla^2 f(\mathbf{0}) t\mathbf{x})^\top \mathbf{x} + \nabla f(\mathbf{0})^\top \mathbf{x}\right) dt \\
&= \mathbf{x}^\top \nabla^2 f(\mathbf{0})\mathbf{x} \int_0^1 t dt + \nabla f(\mathbf{0})^\top \mathbf{x} \\
&= \frac{1}{2}\mathbf{x}^\top \nabla^2 f(\mathbf{0})\mathbf{x} + \nabla f(\mathbf{0})^\top \mathbf{x}.
\end{aligned}$$

Hence, f is a nondegenerate quadratic function.

Exercise 52. *Consider the BFGS method (Definition 8.5).*

(i) Prove that $\mathbf{y}^\top \sigma > 0$, unless $\mathbf{x}_t = \mathbf{x}_{t-1}$, or $f(\lambda \mathbf{x}_t + (1-\lambda)\mathbf{x}_{t+1}) = \lambda f(\mathbf{x}_t) + (1-\lambda) f(\mathbf{x}_{t-1})$ for all $\lambda \in (0,1)$.

(ii) Prove that if H is positive definite and $\mathbf{y}^\top \sigma > 0$, then also H' is positive definite. You may want to use the product form of the BFGS update as developed in Observation 8.6.

Solution: (i) We first prove that for any differentiable convex function f and any $\mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$, we have

$$(\nabla f(\mathbf{y}) - \nabla f(\mathbf{x}))^\top (\mathbf{y} - \mathbf{x}) \geq 0.$$

This property is also known as *monotonicity* of the gradient. It easily follows from the first order charatcerization of convexity, according to which we have

$$\begin{aligned}
f(\mathbf{y}) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}), \\
f(\mathbf{x}) &\geq f(\mathbf{y}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}).
\end{aligned}$$

Adding the two equations, we get

$$0 \geq \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \nabla f(\mathbf{y})^\top (\mathbf{x} - \mathbf{y}) = (\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))^\top (\mathbf{y} - \mathbf{x}).$$

We now also claim that the inequality in monotonicity of the gradient is strict unless $\mathbf{x} = \mathbf{y}$ or $f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) = \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y})$ for all $\lambda \in (0,1)$. So suppose that $\mathbf{x} \neq \mathbf{y}$ and there is some $\lambda \in (0,1)$ such that (by convexity)

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) < \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}).$$

Set $\mathbf{z} = \lambda\mathbf{x} + (1-\lambda)\mathbf{y}$. Hence, using the first-order charatcerization for $\mathbf{x}, \mathbf{z}$,

$$\lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}) > f(\mathbf{z}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{z} - \mathbf{x}) = f(\mathbf{x}) + (1-\lambda)\nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).$$

Rewriting this yields

$$(1-\lambda) f(\mathbf{y}) > (1-\lambda) f(\mathbf{x}) + (1-\lambda)\nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}),$$

and dividing by $1-\lambda$, we see that

$$f(\mathbf{y}) > f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}).$$

Using this strict inequality in the proof of monotonicity of the gradient, we get a strict inequality also in the latter.

(ii) According to the product form, we have

$$H' = \left(I - \frac{\boldsymbol{\sigma}\mathbf{y}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right) H \left(I - \frac{\mathbf{y}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right) + \frac{\boldsymbol{\sigma}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}} =: Q^T H Q + P.$$

For any $\mathbf{x} \in \mathbb{R}^d$, H being positive definite yields

$$\mathbf{x}^\top Q^T H Q \mathbf{x} =: \mathbf{z}^\top H \mathbf{z} \geq 0.$$

Furthermore, $\mathbf{y}^\top\boldsymbol{\sigma} > 0$ yields

$$\mathbf{x}^\top P \mathbf{x} = \mathbf{x}^\top \frac{\boldsymbol{\sigma}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}} \mathbf{x} = \frac{(\mathbf{x}^\top\boldsymbol{\sigma})^2}{\mathbf{y}^\top\boldsymbol{\sigma}} \geq 0.$$

To ensure that $\mathbf{x}^\top H' \mathbf{x} > 0$ for $\mathbf{x} \neq \mathbf{0}$, it remains to prove that for $\mathbf{x} \neq \mathbf{0}$, one of the previous two inequalities is strict.

Indeed, if $\mathbf{x}^\top\boldsymbol{\sigma} \neq \mathbf{0}$, the second inequality is strict. And if $\mathbf{x}^\top\boldsymbol{\sigma} = \mathbf{0}$, we get

$$Q\mathbf{x} = \left(I - \frac{\mathbf{y}\boldsymbol{\sigma}^\top}{\mathbf{y}^\top\boldsymbol{\sigma}}\right)\mathbf{x} = \mathbf{x} \neq \mathbf{0}.$$

Since H is positive definite, this yields $\mathbf{x}^\top Q^T H Q \mathbf{x} = \mathbf{x}^\top H \mathbf{x} > 0$.