Labs
Optimization for Machine Learning
Spring 2024

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

Problem Set 7 — Solutions (Non-Convex Optimization and Newton's Method)

Non-Convex Optimization

Non-convex

Exercise 40. *Prove Lemma 6.3 (gradient descent does not overshoot on smooth functions).*

Solution: On the one hand, we have sufficient decrease, since f is also smooth with parameter $L' > L$:

$$f(\mathbf{x}') \leq f(\mathbf{x}) - \frac{1}{2L'} \|\nabla f(\mathbf{x})\|^2 .$$

Now assume for contradiction that $\mathbf{x}'$ is a critical point, meaning that $\nabla f(\mathbf{x}') = \mathbf{0}$. Then, by smoothness with parameter L, and because $\mathbf{x}' = \mathbf{x} - \gamma \nabla f(\mathbf{x})$, we get that

$$\begin{aligned}
f(\mathbf{x}) &\leq f(\mathbf{x}') + \nabla f(\mathbf{x}')(\mathbf{x} - \mathbf{x}') + \frac{L}{2} \|\mathbf{x} - \mathbf{x}'\|^2 \\
&= f(\mathbf{x}') + \frac{L}{2} \|\mathbf{x} - \mathbf{x}'\|^2 \\
&= f(\mathbf{x}') + \frac{L}{2} \frac{1}{(L')^2} \|\nabla f(\mathbf{x})\|^2 < f(\mathbf{x}') + \frac{L'}{2} \frac{1}{(L')^2} \|\nabla f(\mathbf{x})\|^2 \\
&= f(\mathbf{x}') + \frac{1}{2L'} \|\nabla f(\mathbf{x})\|^2 ,
\end{aligned}$$

where the strict inequality in the second-to-last line uses $\nabla f(\mathbf{x}) \neq \mathbf{0}$. Hence,

$$f(\mathbf{x}') > f(\mathbf{x}) - \frac{1}{2L'} \|\nabla f(\mathbf{x})\|^2 ,$$

which contradicts sufficient decrease.

Exercise 41. *Consider the function* $f(\mathbf{x}) = \frac{1}{2}\left(\prod_{k=1}^{d} x_k - 1\right)^2$. *Prove that for any starting point* $\mathbf{x}_0 \in X = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{x} > \mathbf{0}, \prod_k \mathbf{x}_k \geq 1\}$ *and any* $\varepsilon > 0$, *gradient descent attains* $f(\mathbf{x}_T) \leq \varepsilon$ *for some iteration* T.

Solution: We first prove smoothness along the trajectory. With C being the maximum value of $\prod_{k \neq I} (\mathbf{x}_0)_k$ over all sets I of size at most 2, we bound the squared Frobenius norm of the Hessian by bounding each of the d^2 entries:

$$\left\|\nabla^2 f(\mathbf{x}_0)\right\|_F^2 = \sum_{i,j} |\nabla^2 f(\mathbf{x}_0)_{i,j}|^2 \leq 9d^2C^4 ,$$

where we used that the diagonal term are smaller than $C \leq 3C^2$ and the off-diagonal terms are smaller than $3C^2$ by the triangle inequality. Note that if we start from $\mathbf{x} \in X$, then each gradient descent step can only *de*crease the values x_k as long as we do not overshoot. We therefore get as in Lemma 6.7 that $\left\|\nabla^2 f(\mathbf{x})\right\| \leq \left\|\nabla^2 f(\mathbf{x})\right\|_F \leq 3dC^2$ along the trajectory. Up to the first point of overshooting, f is therefore smooth with parameter $3dC^2$ over the trajectory (Lemma 6.1), and then the smooth step size $1/3dC^2$ guarantees that we actually never overshoot (Lemma 6.3). Hence, Lemma **??** yields sufficient decrease:

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{6dC^2} \|\nabla f(\mathbf{x}_t)\|^2 .$$

We still need a lower bound for

$$\|\nabla f(\mathbf{x})\|^2 = 2f(\mathbf{x})\sum_{i=1}^{d}\left(\prod_{k\neq i} x_k\right)^2,$$

for $\mathbf{x}\in X$. We claim that for some i,

$$\prod_{k\neq i} x_k \geq 1.$$

If not, we would have

$$1 > \prod_{i=1}^{d}\prod_{k\neq i} x_k = \left(\prod_{k} x_k\right)^{d-1},$$

which would mean that $\prod_k x_k < 1$, contradiction. Hence, we have

$$\|\nabla f(\mathbf{x}_t)\|^2 \geq 2f(\mathbf{x}_t),$$

and hence

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{3dC^2}f(\mathbf{x}_t) = \left(1-\frac{1}{3dC^2}\right)f(\mathbf{x}_t).$$

Convergence follows.

Exercise 42. *Consider the function $f(\mathbf{x}) = \frac{1}{2}\left(\prod_{k=1}^{d} x_k - 1\right)^2$. Prove that for even dimension $d \geq 2$, there is a point $\mathbf{x}_0$ (not a critical point) such that gradient descent does not converge to a global minimum when started at $\mathbf{x}_0$, regardless of step size(s).*

Solution: Throughout, let $\mathbf{x}_0$ be such that all entries have the same absolute value. We first prove that gradient descent maintains this property in all iterations. Recall that with $\Delta := -\gamma(\prod_k x_k - 1)(\prod_k x_k)$, the gradient descent step is

$$x_k' = x_k + \frac{\Delta}{x_k}, \quad k = 1,\ldots,d.$$

Suppose that $|x_k| = \alpha$ for all k. Then $x_k' \in \{\alpha+\Delta/\alpha, -\alpha-\Delta/\alpha\}$, hence $|x_k'| = |\alpha+\Delta/\alpha|$ for all k. We also see that either all entries in $\mathbf{x}'$ have the *same* sign as in $\mathbf{x}$ (if $\alpha+\Delta/\alpha > 0$), or all entries in $\mathbf{x}'$ have the *opposite* sign as in $\mathbf{x}$ (if $\alpha+\Delta/\alpha < 0$). (The special case where $\alpha+\Delta/\alpha = 0$ leads to $\mathbf{x}' = \mathbf{0}$ in which case we have already converged to a saddle point, so we do not consider this case further.)

If d is even, any starting point with an odd number of negative signs will lead to *all* iterates having an odd number of negative signs. This means that – regardless of stepsize – we will always have $\prod_k x_k \leq 0$, so we can never converge to an optimal point where $\prod_k x_k = 1$.

Newton's Method

Exercise 48. *Prove Lemma 7.6!*

Solution: We use that for any two matrices, $\|AB\| \leq \|A\|\,\|B\|$. Indeed,

$$\|AB\| = \max_{\mathbf{v}\neq\mathbf{0}}\frac{\|AB\mathbf{v}\|}{\|\mathbf{v}\|} \leq \max_{\mathbf{v}\neq\mathbf{0}}\frac{\|A\|\,\|B\mathbf{v}\|}{\|\mathbf{v}\|} = \|A\|\,\|B\|.$$

Hence,

$$1 = \left\|\nabla^2 f(\mathbf{x}^\star)\nabla^2 f(\mathbf{x}^\star)^{-1}\right\| \leq \left\|\nabla^2 f(\mathbf{x}^\star)\right\|\left\|\nabla^2 f(\mathbf{x}^\star)^{-1}\right\| \leq \left\|\nabla^2 f(\mathbf{x}^\star)\right\|\frac{1}{\mu},$$

so, $\left\|\nabla^2 f(\mathbf{x}^\star)\right\| \geq \mu$.

Now, by the Lipschitz assumption and Corollary 7.5,

$$\left\|\nabla^2 f(\mathbf{x}_T) - \nabla^2 f(\mathbf{x}^\star)\right\| \leq B\,\|\mathbf{x}_T - \mathbf{x}^\star\| \leq \mu\left(\frac{1}{2}\right)^{2^T-1}.$$

Together with $\left\|\nabla^2 f(\mathbf{x}^\star)\right\| \geq \mu$, the statement follows.

Exercise 50. *Let $\delta > 0$ be any real number. Find an example of a convex function $f : \mathbb{R} \to \mathbb{R}$ such that (i) the unique global minimum $x^\star$ has a vanishing second derivative $f''(x^\star) = 0$, and (ii) Newton's method satisfies*

$$|x_{t+1} - x^\star| \geq (1 - \delta)|x_t - x^\star|,$$

for all $x_t \neq x^\star$.

Solution: We take $f(x) = x^k$ for some even natural number k satisfying $k \geq 4$ and $1/(k-1) \leq \delta$. We have

$$\begin{aligned} f'(x) &= kx^{k-1}, \\ f''(x) &= k(k-1)x^{k-2} \geq 0, \end{aligned}$$

hence f is convex by the second-order characterization of convexity (Lemma **??**), and we have $x^\star = 0$ as well as $f''(x^\star) = 0$. Suppose w.l.o.g. that $x_t > 0$. The Newton step (7.1) is

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)} = x_t - \frac{kx_t^{k-1}}{k(k-1)x_t^{k-2}} = x_t - \frac{1}{k-1}x_t \geq (1-\delta)x_t.$$

Quasi-Newton Methods

Exercise 53. *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, $x = x_{t+1}$.

Fixed Point Iteration

Solutions are provided in `solution/`.