Convex Programming
Problem Set 5 – CS 6515/4540 (Fall 2025)

This problem set is due on **Tuesday November 4th**. Submission is via Gradescope. Your solution must be a typed pdf (e.g. via LaTeX) – no handwritten solutions.

18 Convex Functions

1. Given two convex functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, is the sum $f + g$ also convex? Either prove it or give a counterexample.

> Yes, because
>
> $$(f + g)(\lambda x + (1 - \lambda)y) = f(\lambda x + (1 - \lambda)y) + g(\lambda x + (1 - \lambda)y))$$
>
> $$\leq \lambda f(x) + (1 - \lambda)f(x) + \lambda g(x) + (1 - \lambda)g(y) = \lambda(f(x) + g(x)) + (1 - \lambda)(f(x) + g(x))$$
>
> and so we're done.

2. Given two convex functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, is the product fg also convex? Either prove it or give a counterexample.

> No, consider $f(x) = -1$ and $g(x) = x^2$. Both are convex but $fg = -x^2$ is not convex.

3. What about the convexity of the function $h : \mathbb{R} \to \mathbb{R}$ defined by $h(x) = \max(f(x), g(x))$? Prove it or give a counterexample.

> Yes,
>
> $$h(\lambda x + (1 - \lambda)y) = \max(f(\lambda x + (1 - \lambda)y), g(\lambda x + (1 - \lambda y)))$$
>
> $$\leq \max(\lambda f(x) + (1 - \lambda)f(y), \lambda g(x) + (1 - \lambda)g(y)) \leq \lambda h(x) + (1 - \lambda)h(y)$$
>
> where the last step follows from the fact that $f(x) \leq h(x)$, $g(x) \leq h(x)$ for all x.

4. Show an α-strongly convex function (defined in Pb 20) $f : \mathbb{R} \to \mathbb{R}$ satisfies $f(x) = \Omega(x^2)$

> Simply plug in $x = x^*$, where x^* is the unique minimum of f where $\nabla f(x^*) = 0$, and then it is quite obvious we have a quadratic via the definition of strong convexity:
>
> $$f(y) \geq f(x^*) + \frac{\alpha}{2}(y - x^*)^2$$
>
> Since x^* is a constant, the above equation is a quadratic in y, showing that f grows at least as fast as a quadratic.

19 Gradient Descent Failure

Suppose we run an unconstrained Gradient Descent on $f(x) = \frac{1}{2}x^2$ with some arbitrary step size. Give (and justify) an example consisting of

1. a step size $\eta > 0$
2. an initial point $x_0 \in \mathbb{R}$

such that t-th **average** $\overline{x}_t = \frac{1}{t}\sum_{i\leq t} x_i$ of an unconstrained Gradient Descent with the above parameters does not converge (e.g it diverges) to the optimum as $t \to \infty$.

Example

We provide the following example:

$$\eta = 3 \tag{1}$$
$$x_0 = 1 \tag{2}$$

More generally, any step size $\eta > 2$ with any initial point $x_0 \neq 0$ will cause the average to diverge.

Setup and Gradient Descent Update

For the function $f(x) = \frac{1}{2}x^2$, we have:

- Gradient: $\nabla f(x) = x$
- Optimum: $x^* = 0$ with $f(x^*) = 0$
- GD update rule: $x_{t+1} = x_t - \eta\nabla f(x_t) = x_t - \eta x_t = (1-\eta)x_t$

Computing the Average

The t-th average is:

$$\overline{x}_t = \frac{1}{t}\sum_{i=0}^{t-1} x_i = \frac{1}{t}\sum_{i=0}^{t-1}(1-\eta)^i x_0 = \frac{x_0}{t}\sum_{i=0}^{t-1}(1-\eta)^i \tag{3}$$

Using the geometric series formula (assuming $\eta \neq 0$):

$$\sum_{i=0}^{t-1} r^i = \frac{1-r^t}{1-r} \tag{4}$$

we obtain:

$$\overline{x}_t = \frac{x_0}{t} \cdot \frac{1-(1-\eta)^t}{1-(1-\eta)} = \frac{x_0}{t\eta}\left[1-(1-\eta)^t\right] \tag{5}$$

Analysis of Convergence

Case 1: $0 < \eta < 2$ **(Standard Convergent Regime)**

In this case, $|1-\eta| < 1$, so $(1-\eta)^t \to 0$ as $t \to \infty$. Therefore:

$$\overline{x}_t = \frac{x_0}{t\eta}\left[1-(1-\eta)^t\right] \to \frac{x_0}{t\eta} \to 0 \tag{6}$$

The average converges to the optimum $x^* = 0$.

Case 2: $\eta = 2$ (Boundary Case)

Here, $1 - \eta = -1$, so the iterates oscillate:

$$x_t = (-1)^t x_0 \tag{7}$$

The sequence is $x_0, -x_0, x_0, -x_0, \ldots$
For the average:

- If t is even: $\overline{x}_t = 0$ (equal numbers of $+x_0$ and $-x_0$)
- If t is odd: $\overline{x}_t = \frac{x_0}{t} \to 0$

Therefore, $\overline{x}_t \to 0$. The average still converges despite the oscillation of individual iterates.

Case 3: $\eta > 2$ (Divergent Regime)

This is the regime where the average diverges.
For $\eta > 2$, we have $1 - \eta < -1$, so $|1 - \eta| > 1$. Let $r = 1 - \eta$ with $|r| > 1$. Then:

$$\overline{x}_t = \frac{x_0}{t\eta}(1 - r^t) = \frac{x_0}{t\eta} - \frac{x_0 r^t}{t\eta} \tag{8}$$

As $t \to \infty$, the term $\frac{x_0 r^t}{t\eta}$ behaves as:

$$\frac{|r|^t}{t} \to \infty \tag{9}$$

This is because exponential growth $|r|^t$ dominates polynomial growth t. More formally, for $|r| > 1$:

$$\lim_{t\to\infty} \frac{|r|^t}{t} = \lim_{t\to\infty} \frac{e^{t \ln |r|}}{t} = \infty \tag{10}$$

Therefore, $|\overline{x}_t| \to \infty$, and the average diverges. For this specific example:

- $1 - \eta = 1 - 3 = -2$
- $x_t = (-2)^t$
- Iterates: $1, -2, 4, -8, 16, -32, \ldots$

The average is:

$$\overline{x}_t = \frac{1 - (-2)^t}{3t} = \frac{1}{3t} - \frac{(-2)^t}{3t} \tag{11}$$

For large t:

$$|\overline{x}_t| \approx \frac{2^t}{3t} \to \infty \tag{12}$$

The average diverges to $\pm\infty$ (oscillating in sign). The step size $\eta = 3$ causes the algorithm to overshoot dramatically:

- Each iteration, we move by $\eta x_t = 3x_t$
- This takes us to $x_{t+1} = x_t - 3x_t = -2x_t$
- We overshoot the optimum and end up twice as far on the opposite side
- The magnitude grows exponentially: $|x_t| = 2^t$

Even though we're averaging, the exponential growth overwhelms the $1/t$ decay from averaging, causing divergence.
Note that if $x_0 = 0$ (starting at the optimum), then $x_t = 0$ for all t regardless of η, and the average trivially remains at the optimum. Therefore, any valid example must have $x_0 \neq 0$.

20 Gradient Descent for Strongly-Convex Functions

A differentiable function f is α-strongly convex for $\alpha > 0$ if for all $x, y \in \mathbb{R}^n$ we have

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\alpha}{2}\|y - x\|_2^2.$$

Consider an α-strongly convex differentiable function f with the 2-norm of its gradient always bounded by G. The goal is to minimize f and let x^* denote its minimum.

Show that the gradient descent algorithm with step size $\frac{1}{\alpha(t+1)}$ satisfies

$$f\Big(\frac{\sum_t x_t}{T}\Big) - f(x^*) \leq \frac{G^2(1 + \log T)}{2\alpha T}.$$

Thus, strong-convexity allows us to get $1/T$ dependency in regret instead of $1/\sqrt{T}$ dependency for general convex functions.

(Hint: Change the potential function in the analysis from class to $\Phi(t) = \frac{t\alpha}{2}\|x_t - x^*\|^2$. Also, use that $\sum_{t \in \{1,\ldots,T\}} \frac{1}{t} \leq 1 + \log T$.)

Let n be the step size and f be our feedback function. Consider the potential function $\Phi(t) = \frac{\alpha t}{2}||x_t - x^*||^2$. Note that $||y_{t+1} - x^*||^2 \geq ||x_{t+1} - x^*||^2$ by the lecture notes (projection on K). Using this, as well as our step size $n = \frac{1}{\alpha(t+1)}$ and the upper bound on our gradient (G) along with the definition of α-strong convexity,

$$\Phi(t+1) = \frac{\alpha(t+1)}{2}(||x_{t+1} - x^*||^2) \leq \frac{\alpha(t+1)}{2}(||y_{t+1} - x^*||^2) = \frac{\alpha(t+1)}{2}||x_t - x^* - n\nabla f(x_t)||^2$$

$$= \frac{\alpha(t+1)}{2}\left(||x_t - x^*||^2 + (n\nabla f(x_t))^2 - 2n\langle x_t - x^*, \nabla f(x_t)\rangle\right)$$

$$= \Phi(t) + \frac{\alpha}{2}||x_t - x^*||^2 + \frac{(\nabla f(x_t))^2}{2\alpha(t+1)} - \langle x_t - x^*, \nabla f(x_t)\rangle$$

$$\leq \Phi(t) + f(x^*) - f(x_t) + \frac{G^2}{2\alpha(t+1)}$$

$$\implies f(x_t) - f(x^*) \leq \Phi(t) - \Phi(t+1) + \frac{G^2}{2\alpha(t+1)}$$

$$\implies \sum_t f(x_t) - f(x^*) \leq \Phi(0) - \Phi(T+1) + \frac{G^2}{2\alpha}\sum_t \frac{1}{t+1} \leq \frac{G^2(1 + \log(T))}{2\alpha}$$

where the last step utilizes the observations that $\Phi(0) - \Phi(t+1) \leq 0$, and $\sum_t \frac{1}{t} \leq 1 + \log(T)$. Applying convexity to the left hand side, we have $f(\frac{\sum_t x_t}{T}) - f(x^*) \leq \frac{G^2(1+\log(T))}{2\alpha T}$ and we are done.

21 Non-Convex Function

In class we assumed function f is convex. We now want to consider the non-convex case. We want to show that for L-smooth f, after t iterations with step size $\eta \leq 1/L$ we can find a point x' with

$$\|\nabla f(x')\| \leq \sqrt{\frac{2}{\eta \cdot t}(f(x^0) - f(x^*))}.$$

(Note that for a local optimum we have $\nabla f(x) = 0$, so a small norm $\|\nabla f(x')\|$ indicates that we are close to a local optimum or saddle point.)

Proving this from scratch is a bit tricky, so we provide the following subproblems to guide you to a proof. Each subproblem can be solved in a few lines of calculation/algebra.

Problem:

1. Show $f(x^{t+1}) \leq f(x^t) - \frac{\eta}{2}\|\nabla f(x^t)\|^2$ (Hint: Check the proof from class for convex functions. Does it work for non-convex functions?)
2. Show $\sum_{k=0}^{t} \|\nabla f(x^k)\|^2 \leq \frac{2}{\eta}(f(x^0) - f(x^*))$ (Hint: 1. implies $\frac{\eta}{2}\|\nabla f(x^t)\|^2 \leq$ ...)
3. Show $\min_{k=0\ldots t} \|\nabla f(x^k)\| \leq \sqrt{\frac{2}{\eta \cdot t}(f(x^0) - f(x^*))}$.

where x^* is the global optimum $f(x^*) = \min_x f(x)$.

You are allowed to use subproblems to solve later subproblems (e.g., use 1+2 to solve 3), even if you did not prove them.

1. This property has already been proven in class, and actually follows from L-smoothness alone; it does not require convexity of function f. The proof is repeated here for completeness:

   For an L-smooth function, from the lemma in lectures, we have:

   $$|f(y) - (f(x) + \nabla f(x)^\top (y-x))| \leq \frac{L}{2}\|y-x\|^2 \Rightarrow f(y) \leq f(x) + \nabla f(x)^\top (y-x) + \frac{L}{2}\|y-x\|^2$$

   Substituting in $y = x^{t+1} = x^t - \eta \nabla f(x)$ (for $\eta \leq 1/L$), and $x = x^t$ in the above, we get:

   $$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \nabla f(x^t)^\top (x^t - \eta \nabla f(x) - x^t) + \frac{L}{2}\|x^t - \eta \nabla f(x) - x^t\|^2 \\ &= f(x^t) - \eta \nabla f(x^t)^\top \nabla f(x) + \frac{L}{2}\|\eta \nabla f(x^t)\|^2 \\ &= f(x^t) - \eta \|\nabla f(x^t)\|^2 + \frac{L\eta^2}{2}\|\nabla f(x^t)\|^2 \\ &\leq f(x^t) - \eta \|\nabla f(x^t)\|^2 + \frac{\eta}{2}\|\nabla f(x^t)\|^2 \quad \text{(since } L \leq 1/\eta) \\ &= f(x^t) - \frac{\eta}{2}\|\nabla f(x^t)\|^2 \end{aligned}$$

2. Rearranging the result of part 1, we get:

   $$\|\nabla f(x^k)\|^2 \leq \frac{2}{\eta}[f(x^k) - f(x^{k+1})]$$

   The above inequality holds for all t. So, summing over all $k = 0, \ldots, t$, we get:

   $$\sum_{k=0}^{t} \|\nabla f(x^k)\|^2 \leq \sum_{k=0}^{t} \frac{2}{\eta}[f(x^k) - f(x^{k+1})] = \frac{2}{\eta}[f(x^0) - f(x^{t+1}))]$$

   In the above, the equality follows from the fact that the summation telescopes (part of previous terms in the summation cancel with part of future terms). Now, by definition of $x^* = \text{argmin}_x f(x)$, we have $f(x^*) \leq f(x^{t+1}) \Leftrightarrow -f(x^*) \leq -f(x^{t+1})$. As a result, we get:

   $$\sum_{k=0}^{t} \|\nabla f(x^k)\|^2 \leq \frac{2}{\eta}[f(x^0) - f(x^*))]$$

3. Dividing the inequality obtained from part 2 by 2, we have:

   $$\frac{1}{t}\sum_{k=0}^{t} \|\nabla f(x^k)\|^2 \leq \frac{2}{\eta \cdot t}[f(x^0) - f(x^*))]$$

   Since the average of a set of numbers is at least the minimum among the set, we have the relation:

   $$\min_{k=0\ldots t} \|\nabla f(x^k)\|^2 \leq \frac{1}{t}\sum_{k=0}^{t} \|\nabla f(x^k)\|^2$$

   Therefore, we conclude:

   $$\min_{k=0\ldots t} \|\nabla f(x^k)\|^2 \leq \frac{2}{\eta \cdot t}[f(x^0) - f(x^*))] \Leftrightarrow \min_{k=0\ldots t} \|\nabla f(x^k)\| \leq \sqrt{\frac{2}{\eta \cdot t}[f(x^0) - f(x^*))]}$$