SYS 6003: Optimization	Fall 2016
Lecture 10	
Instructor: Quanquan Gu	*Date: Sep 26^{th}*

In the following, we will introduce another convexity preserving operation: **composition with scalar function**.

Theorem 1 *Let $g : \mathbb{R}^d \to \mathbb{R}$ and $h : \mathbb{R} \to \mathbb{R}$, define $f(\mathbf{x}) = h(g(\mathbf{x}))$, where*

$$\mathbf{dom}\, f = \{\mathbf{x} \in \mathbf{dom}\, g \quad | \quad g(\mathbf{x}) \in \mathbf{dom}\, h\}.$$

we have:

1) f is convex if g is convex, h is convex and $\widetilde{h}$ is nondecreasing.

2) f is convex if g is concave, h is convex and $\widetilde{h}$ is nonincreasing.

Proof: f is convex if g is convex, h is convex and $\widetilde{h}$ is nondecreasing. For any $\mathbf{x}, \mathbf{y} \in \mathbf{dom}\, f, \quad \alpha \in [0, 1]$
Since g is convex and h is nondecreasing,

$$\begin{aligned} f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= \tilde{h}(g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y})) \\ &\leq \tilde{h}(\alpha g(\mathbf{x}) + (1-\alpha)g(\mathbf{y})). \end{aligned}$$

Since h is convex, $\tilde{h}$ is also convex. We have

$$\begin{aligned} \tilde{h}(\alpha g(\mathbf{x}) + (1-\alpha)g(\mathbf{y})) &\leq \alpha\tilde{h}(g(\mathbf{x})) + (1-\alpha)\tilde{h}(g(\mathbf{y})) \\ &= \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y}). \end{aligned}$$

Therefore, $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$. By the definition of convex function, we conclude that f is a convex function. ■

Remark 1 *In the above statements, the monotonic conditions on $\widetilde{h}$ rather than h are necessary. In the proof of this argument, we cannot write: for any $\mathbf{x}, \mathbf{y} \in \mathbf{dom}\, f$, and any $\alpha \in [0, 1]$, $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) = h(g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}))$, because this actually may not be true, since even though $g(\mathbf{x}) \in \mathbf{dom}\, h, g(\mathbf{y}) \in \mathbf{dom}\, h$, we may have $g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \notin \mathbf{dom}\, h$. However, $g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y})$ must belong to $\mathbf{dom}\, \tilde{h}$. Thus, we can write $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) = \tilde{h}(g(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}))$.*

Remark 2 *To give a specific example, consider: $f(x) = h(g(x))$ where $g(x) = x^2$ with $\mathbf{dom}\, g = \mathbb{R}$ and $h(x) = x$ with $\mathbf{dom}\, h = [1, 4]$. In this example, $\widetilde{h}$ is **not** nondecreasing, so $f(x)$ is **not** convex. In fact, we can show that here $f(x) = x^2$ with $\mathbf{dom}\, f = [-2, -1] \cup [1, 2]$.*

Now let us see some examples of convex functions, which can be verified by the rule of composition with scalar function.

Example 1 $f(\mathbf{x}) = \exp(g(\mathbf{x}))$ *is convex if* $g : \mathbb{R}^d \to \mathbb{R}$ *is convex.*

Example 2 $f(\mathbf{x}) = 1/g(\mathbf{x})$ *is convex if* $g : \mathbb{R}^d \to \mathbb{R}$ *is concave and positive. To see that, notice that* $h(x) = 1/x$ *is convex where* **dom** $h = [0, +\infty]$ *and* $\widetilde{h}$ *is nonincreasing.*

Next, we generalize the scalar composition to **Vector Composition**. The proof for the following theorem is similar to the proof in the scalar composition case and is therefore omitted.

Theorem 2 *Let* $g : R^d \to R^m$ *and* $h : R^m \to R$*, define*

$$f(\mathbf{x}) = h(g(\mathbf{x})) = h(g_1(\mathbf{x}), g_2(\mathbf{x}), \ldots, g_m(\mathbf{x}))$$

we have:

1) f *is convex if* $g_i(\mathbf{x})$ *is convex,* h *is convex and* $\widetilde{h}$ *is nondecreasing in each augment.*

2) f *is convex if* $g_i(\mathbf{x})$ *is concave,* h *is convex and* $\widetilde{h}$ *is nonincreasing in each augment.*

Example 3 $f(\mathbf{x}) = \sum_{i=1}^{m} \log(g_i(\mathbf{x}))$ *is concave if* $g_i : \mathbb{R}^d \to \mathbb{R}$*'s are concave and positive.*

Example 4 $f(\mathbf{x}) = \log \sum_{i=1}^{m} \exp(g_i(\mathbf{x}))$ *is convex if* $g_i : \mathbb{R}^d \to \mathbb{R}$*'s are convex.*

So far, we have discussed the following concepts:

1) general form of optimization problem;

2) local and global minimum;

3) convex sets;

4) convex functions.

Staring from here, we will begin to study some important algorithms in convex optimization.

We start with the following general unconstrained convex optimization problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} f(\mathbf{x}),$$

where $f(\mathbf{x})$ is convex.

The gradient descent algorithm is probably one of the most widely used algorithms for convex optimization. It is shown in Algorithm 1. In the algorithm, η_t is the step size, which is indexed by t, and may or may not depend on t. In other words, the step size η_t could be either fixed or time varying.

The following theorem gives the convergence rate of the gradient descent algorithms for convex and differential functions with bounded gradient and with fixed step size.

Algorithm 1 Gradient Descent

Input: η_t, T
Initialization: $\mathbf{x}_1$
for $t = 1$ to T **do**
  $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$
end for
Output: $\mathbf{x}_{T+1}$

Theorem 3 *If $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable, and f has bounded gradient, i.e., $\|\nabla f(\mathbf{x})\|_2 \le G$ for all $\mathbf{x} \in \mathbb{R}^d$ and some $G > 0$. then the gradient descent with $\eta_t = \eta = R/(G\sqrt{T})$ satisfies*

$$f\Big(\frac{1}{T}\textstyle\sum_{t=1}^{T} \mathbf{x}_t\Big) - f(\mathbf{x}^*) \le \frac{RG}{\sqrt{T}},$$

where $\mathbf{x}^* = \arg\min f(\mathbf{x})$ is the global minimum point and $R = \|\mathbf{x}_1 - \mathbf{x}^*\|_2$.

Proof: Since $f(\mathbf{x})$ is convex, by the first order condition, we have

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\le \nabla f(\mathbf{x}_t)^T(\mathbf{x}_t - \mathbf{x}^*) \\ &= \frac{1}{\eta}(\mathbf{x}_t - \mathbf{x}_{t+1})^T(\mathbf{x}_t - \mathbf{x}^*) \\ &= \frac{1}{2\eta}(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|_2^2 + \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2), \end{aligned}$$

where the first equality follows from the updating rule of the gradient descent algorithm and the second one is from the identity: $2\mathbf{x}^\top\mathbf{y} = \|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - \|\mathbf{x} - \mathbf{y}\|_2^2$. By definition, $\mathbf{x}_t - \mathbf{x}_{t+1} = -\eta\nabla f(\mathbf{x}_t)$, it then follows that

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}^*) &\le \frac{1}{2\eta}(\|\eta\nabla f(\mathbf{x}_t)\|_2^2 + \|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2) \\ &= \frac{1}{2\eta}(\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2) + \frac{\eta}{2}\|\nabla f(\mathbf{x}_t)\|_2^2 \\ &\le \frac{1}{2\eta}(\|\mathbf{x}_t - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^*\|_2^2) + \frac{\eta}{2}G^2, \end{aligned}$$

where the last inequality follows from the fact that f is G-Lipschitz and Lemma **??**.

Note that this inequality holds for any positive integer t. Specifically, we have

$$\begin{aligned} f(\mathbf{x}_1) - f(\mathbf{x}^*) &\le \frac{1}{2\eta}(\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_2 - \mathbf{x}^*\|_2^2) + \frac{\eta}{2}G^2, \\ f(\mathbf{x}_2) - f(\mathbf{x}^*) &\le \frac{1}{2\eta}(\|\mathbf{x}_2 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_3 - \mathbf{x}^*\|_2^2) + \frac{\eta}{2}G^2, \\ &\vdots \\ f(\mathbf{x}_T) - f(\mathbf{x}^*) &\le \frac{1}{2\eta}(\|\mathbf{x}_T - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2) + \frac{\eta}{2}G^2. \end{aligned}$$

Adding these inequalities gives

$$\begin{aligned}
\sum_{t=1}^{T} f(\mathbf{x}_t) - Tf(\mathbf{x}^*) &\le \frac{1}{2\eta}(\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{T+1} - \mathbf{x}^*\|_2^2) + \frac{T\eta}{2}G^2 \\
&\le \frac{1}{2\eta}\|\mathbf{x}_1 - \mathbf{x}^*\|_2^2 + \frac{T\eta}{2}G^2 \\
&= \frac{1}{2\eta}R^2 + \frac{T\eta}{2}G^2,
\end{aligned}$$

which can be rewritten as

$$\frac{1}{T}\sum_{i=1}^{T} f(\mathbf{x}_t) - f(\mathbf{x}^*) \le \frac{1}{2\eta T}R^2 + \frac{\eta}{2}G^2.$$

Since $f(\mathbf{x})$ is convex, we have $f\big(\sum_{t=1}^{T}\mathbf{x}_t/T\big) \le \sum_{t=1}^{T} f(\mathbf{x}_t)/T$. Then,

$$f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\right) - f(\mathbf{x}^*) \le \frac{1}{2\eta T}R^2 + \frac{\eta}{2}G^2.$$

Since the above inequality holds for any $\eta > 0$ and the right-hand side is minimized when $\eta = R/G\sqrt{T}$, it follows that

$$f\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{x}_t\right) - f(\mathbf{x}^*) \le \frac{RG}{\sqrt{T}}.$$

■