# AN INTRODUCTION TO CONVEX OPTIMIZATION

GIDEON DRESDNER, HADI DANESHMAND

## 1. Smooth optimization

1.1. **Gradient free optimization and curse of dimensionality.** We highlight the computational complexity of grid search of the space for optimization. Consider the objective function $f : \mathbb{R}^d \to \mathbb{R}$ that is $L$-Lipschitz, namely it is differentiable with a continuous derivative and

$$|f(\mathbf{w}) - f(\mathbf{v})| \leq \ell \|\mathbf{w} - \mathbf{v}\| \tag{1.1}$$

holds for all $\mathbf{w}$ and $\mathbf{v}$ in $\mathbb{R}^d$. We aim at approximating the solution of the following minimization

$$\arg \min_{0<\mathbf{w}_i<1} f(\mathbf{w}). \tag{1.2}$$

Let $f^*$ denotes the solution of the above optimization problem. $\mathbf{w} \in \mathbb{R}^d$ is an $\epsilon$-accurate solution of the above objective, if

$$f(\mathbf{w}) - f^* \leq \epsilon \tag{1.3}$$

holds. The smoothness of $f$ allows us to grid-search over $\mathbb{R}^d$ to find an $\epsilon$-accurate solution[2]. More precisely, we search over set $\mathcal{S} = \{\mathbf{w} = [n_1/p, n_2/p, \ldots, n_d/p]\}$ where $n_i \in \{0, \ldots, p\}$ as

$$\widehat{\mathbf{w}} = \arg \min_{\mathbf{w} \in \mathcal{S}} f(\mathbf{w}). \tag{1.4}$$

One can readily check that

$$f(\widehat{\mathbf{w}}) - f^* \leq \ell/p \tag{1.5}$$

Therefore, we need $|\mathcal{S}| = \mathrm{O}((\ell/\epsilon)^d)$ trial to find an $\epsilon$-accurate solution. The exponential blow-up of the computational complexity in $d$ is the main challenge of optimization. During this lecture we will show how (and when) we can get ride of this dependency using gradient descent method.

1.2. **Steepest descent direction.** Let's assume that $f$ is $L$-smooth, namely the following holds for all $\mathbf{w} \in \mathbb{R}^d$

$$\|\nabla^2 f(\mathbf{w})\|_2 \leq L. \tag{1.6}$$

In this section, we focus on unconstraint optimization of $f$, i.e.

$$\arg \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}). \tag{1.7}$$

A local search algorithm looks for a $\mathbf{v} \in \mathbb{R}^d$ such $f(\mathbf{w}+\mathbf{v}) \leq f(\mathbf{w})$. Taylor expansion can lead us to find such a direction:

$$\begin{aligned} f(\mathbf{w}+\mathbf{v}) &= f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} \rangle + \frac{1}{2} \int_{\tau=0}^{1} \mathbf{v}^\top \nabla^2 f(\mathbf{w}_1 + \tau(\mathbf{v}))\mathbf{v} \\ &\overset{(1.6)}{\leq} f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} \rangle + \frac{L}{2}\|\mathbf{v}\|^2 \end{aligned}$$

Let's rewrite the established upperbound in the last equation as a function of $\mathbf{v}$:

$$g(\mathbf{v}) := f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} \rangle + \frac{L}{2}\|\mathbf{v}\|^2 \tag{1.8}$$

The above function is the 1st order Taylor expansion of $f$ at $\mathbf{w}$ regularized by $\|\mathbf{v}\|^2$. We can find the minimum of the quadratic function $g(\mathbf{v})$ by setting gradient to zero:

$$\nabla g(\mathbf{v}^*) = 0 \Leftrightarrow \mathbf{v}^* = -\nabla f(\mathbf{w})/L \tag{1.9}$$

In this way, we find a distance direction $\mathbf{v}^*$ that guarantees the following decrease in $f$,

$$f(\mathbf{w} - \nabla f(\mathbf{w})/L) - f(\mathbf{w}) \leq -\|\nabla f(\mathbf{w})\|^2/(2L) \tag{1.10}$$

Since the gradient descent direction $-\nabla f(\mathbf{w})/L$ is steepest direction over the established upperbound $g(\mathbf{w})$, it is often called steepest descent direction [1].

1.3. **Gradient Descent (GD) method.** GD is a recurrence over the steepest descent approach as

$$\mathbf{w}_n = \mathbf{w}_{n-1} - \nabla f(\mathbf{w}_{n-1})/L. \tag{1.11}$$

The above recurrence keep descending the function value, since

$$f(\mathbf{w}_n) - f(\mathbf{w}_0) = \sum_{k=1}^{n} f(\mathbf{w}_k) - f(\mathbf{w}_{k-1}) \overset{(1.10)}{\leq} -(2L)^{-1} \sum_{k=1}^{n-1} \|\nabla f(\mathbf{w}_k)\|^2. \tag{1.12}$$

Using the above result, we can conclude

$$\min_{k \leq n} \|\nabla f(\mathbf{w}_k)\|^2 \leq \sum_{k=1}^{n-1} \|\nabla f(\mathbf{w}_k)\|^2/n \overset{(1.12)}{\leq} 2L(f(\mathbf{w}_0) - f^*)/n \tag{1.13}$$

Suppose that $f$ has a unique minimum (i.e. $f(\mathbf{w}_0) - f^*$ is finite), then GD yields an $\epsilon$-accurate solution (in terms of the squared norm of gradient) in $\mathrm{O}(1/\epsilon)$. On the contrary to the zero-order optimization scheme, the complexity of GD does not scale with the dimensionality. This is the main advantage of the gradient.

1.4. **Global optimization with GD and convexity.** In the last section, we observed that GD obtains a poly-time complexity to reach an approximate stationary point –where the gradient is zero. Here, we determines the class of functions $\mathcal{F}$ for which the convergence to a stationary point leads to global optimization of $f$ (i.e. reaching to the global optimum of $f$). Our proposal for the function class $\mathcal{F}$ is

1 Let $f \in \mathcal{F}$. If $\nabla f(\mathbf{w}^*) = 0$, then $f(\mathbf{w}^*) \leq f(\mathbf{w})$ holds for all $\mathbf{w} \in \mathbb{R}^d$.
2 If $f_1, f_2 \in \mathcal{F}$, then $\alpha f_1 + \beta f_2 \in \mathcal{F}$ for all $\alpha, \beta > 0$.
3 Any linear function belongs to $\mathcal{F}$

Suppose that $f \in \mathcal{F}$. Assumption 3 implies that $h(\mathbf{v}) = \langle -\nabla f(\mathbf{w}), \mathbf{v} \rangle$, which is a linear function in $\mathbf{v}$, belongs to $\mathcal{F}$; hence, $f + h \in \mathcal{F}$ according to the property 3:

$$\varphi(\mathbf{v}) = f(\mathbf{v}) - \langle \nabla f(\mathbf{w}), \mathbf{v} \rangle \in \mathcal{F} \tag{1.14}$$

Since $\nabla \varphi(\mathbf{w}) = 0$, the assumption 2 implies that $\mathbf{w}$ is the global optimum of $\varphi$, namely

$$\varphi(\mathbf{v}) \geq \varphi(\mathbf{w}) \tag{1.15}$$

Replacing the closed form of $\varphi$ in the above Eq. yields

$$f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle \tag{1.16}$$

The above inequality is definition of convexity for differentiable function. Indeed, the convex functions are a class of function on which GD reaches the global minimum.

1.5. **Convergence of GD on convex functions.** The convexity of $f$ implies that

$$f(\mathbf{w}) - f^* \leq \langle \nabla f(\mathbf{w}), \mathbf{w} - \mathbf{w}^* \rangle \leq \|\nabla f(\mathbf{w})\| \|\mathbf{w} - \mathbf{w}^*\| \tag{1.17}$$

Therefore the convergence in $\|\nabla f(\mathbf{w})\|$ implies the convergence in suboptimality $f(\mathbf{w}) - f^*$. Since $\|\nabla f(\mathbf{w}_n)\|$ deceases in $\mathrm{O}(1/\sqrt{n})$, the above bound leads to a slow convergence rate $1/\sqrt{n}$ in $f(\mathbf{w}_n) - f^*$. Here, we improve the convergence rate to $\mathrm{O}(1/n)$[1]. Consider the compact notations $\Delta_n := f(\mathbf{w}_n) - f^*$, $\gamma = 1/L$ and $r_n := \|\mathbf{w}_n - \mathbf{w}^*\|^2$. According to Eq. (1.10), the following holds:

$$\Delta_{n+1} \leq \Delta_n - \gamma \|\nabla f(\mathbf{w}_n)\|^2 \tag{1.18}$$

$$\overset{(1.17)}{\leq} \Delta_n - \gamma \Delta_n^2 / (2 r_n) \tag{1.19}$$

We can show that $r_n \leq r_0$ (see Theorem 2.1.14 of [2]). Then

$$\Delta_{n+1} \leq \Delta_n - \gamma \Delta_n^2 / (2 r_0) \tag{1.20}$$

holds. Dividing both sides by $\Delta_{n+1} \Delta_n$ and rearranging the terms yields

$$\frac{\gamma \Delta_n}{r_0 \Delta_{n+1}} \leq \frac{1}{\Delta_{n+1}} - \frac{1}{\Delta_n} \tag{1.21}$$

Since $\Delta_{n+1} \leq \Delta_n$, we have

$$\frac{\gamma}{r_0} \leq \frac{1}{\Delta_{n+1}} - \frac{1}{\Delta_n} \tag{1.22}$$

Summing up over $n$ obtains

$$\frac{n\gamma}{r_0} \leq \frac{1}{\Delta_n} - \frac{1}{\Delta_0} \leq \frac{1}{\Delta_n} \tag{1.23}$$

The above inequality concludes a $\mathrm{O}(1/n)$-convergence rate in $\Delta_n$:

$$\Delta_n \leq r_0 / (n\gamma) \tag{1.24}$$

[1]The established improved convergence of GD on convex functions is not necessary for the exam and exercises. You can skip this part of the note.

1.6. **Different definitions of convexity.** So far, we have established the definition of convexity for differentiable functions [2]. Yet, the convexity notion is extendable to non-differentiable functions. It is useful to keep the following three definitions of the convexity in mind:

**General definition:** $f$ is convex over $\mathbb{R}^d$ if for all $\mathbf{w}, \mathbf{v} \in \mathbb{R}^d$ and $\theta \in [0, 1]$ the following holds

$$f(\theta \mathbf{w} + (1-\theta)\mathbf{v}) \leq \theta f(\mathbf{w}) + (1-\theta) f(\mathbf{v}). \tag{1.25}$$

- **Definition for differentiable functions:** The continuously differentiable function $f$ is convex if and only if

$$f(\mathbf{v}) \geq f(\mathbf{w}) + \langle \nabla f(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle \tag{1.26}$$

- **Definition for twice differentiable functions:** $f \in C^2$ is convex if and only if its Hessian $\nabla^2 f(\mathbf{w})$ is semi-positive definite uniformly in $\mathbf{w}$.

## References

1. Stephen Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge university press, 2004.
2. Yurii Nesterov, *Introductory lectures on convex optimization*, vol. 8, Springer Science & Business Media, 2004.

[2] We follow the algorithmic approach of [2].