Lecture 12

12.1 Problem Formulation

We consider the following convex optimization problem:

$$\min_{x \in Q} f(x), \tag{12.1}$$

for a convex set $Q \subseteq \mathbb{R}^n$, and a convex (possibly non-differentiable) function $f : \mathbb{R}^n \to \mathbb{R}$, which, for simplicity of the presentation, we define over the whole space. In modern applications, the dimension n in problem (12.1) is large: $n \to \infty$ (*large-scale optimization*), so we cannot directly apply the heavy machinery of the ellipsoid method. Instead, we will analyze a non-smooth analog of the gradient method.

We assume that for any point $x \in Q$ we can compute a subgradient vector $f'(x) \in \partial f(x)$ that satisfies:

$$f(y) \quad \geq \quad f(x) + \langle f'(x), y - x \rangle, \qquad x, y \in Q.$$

For the set Q we assume a possibility of computing the projection, in the Euclidean norm:

$$\pi_Q(x) \quad := \quad \arg\min_{y \in Q} \|y - x\|. \tag{12.2}$$

This is a different and more expensive operation than a separation oracle for Q. The possibility of computing projections (12.2) typically means that the set Q is *simple*. Thus, as we did when discussing the fully composite problems in Lecture 9, we assume that the main difficulty of solving problem (12.1) lies in the objective function f, but not in the constraints. However, it is possible to generalize the subgradient method to the case of using only a separation oracle for Q, when the set is specified by a number of black-box functional inequalities.

12.1.1 Optimality Condition

Let us discuss an optimality condition for a point $x^\star$ to be a global minimum of (12.1). We have the following generalization of the first-order condition from unconstrained minimization. Let $Df(x^\star)[h]$ denote the directional derivative of f along the direction h:

$$Df(x)[h] \quad := \quad \lim_{\alpha \to 0+} \tfrac{1}{\alpha} \big[f(x + \alpha h) - f(x) \big].$$

- For differentiable functions: $Df(x)[h] \equiv \langle \nabla f(x), h \rangle$.
- For convex functions, the directional derivative along any direction exists at all interior points of the domain. In this case, we have the following interesting relationship:

$$Df(x)[h] \quad = \quad \max \big\{ \langle g, h \rangle \quad : \quad g \in \partial f(x) \big\}.$$

Proposition 12.1.1. *Let $x^\star$ be a constrained minimum of f over set Q. Then*

$$Df(x^\star)[x - x^\star] \;\geq\; 0, \qquad x \in Q. \tag{12.3}$$

Proof. Indeed, by the definition of the directional derivative, we have for a sufficiently small $\alpha > 0$:

$$Df(x^\star)[x - x^\star] \;=\; \tfrac{1}{\alpha}[f(x + \alpha h) - f(x)] + o(1) \;\geq\; o(1),$$

where $o(1)$ goes to zero when $\alpha \to 0$. Taking the limit complete the proof. □

Corollary 12.1.2. *For a constrained minimum of a differentiable function f, we have:*

$$\langle \nabla f(x^\star), x - x^\star \rangle \;\geq\; 0, \qquad x \in Q. \tag{12.4}$$

Remark 12.1.3. See also Theorem 10.1.5 in Lecture 10 for a more general optimality condition suitable for additive composite optimization, that generalizes (12.4).

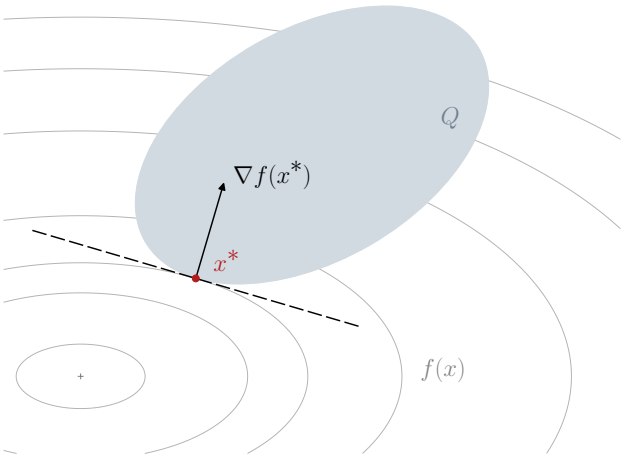

Figure 12.1: The constrained minimum of a function f over a set Q.

12.2 Subgradient Method with Normalized Stepsizes

12.2.1 Subgradient Step

We will use the optimality condition to analyze one step of the subgradient method. The method is very similar to the basic gradient method. At point $x \in Q$, we compute a subgradient $f'(x) \in \partial f(x)$ and perform the following update, for some $\eta > 0$:

$$x^+ \;=\; \arg\min_{y \in Q} \Big[\, m(y) \;:=\; \eta \langle f'(x), y - x \rangle + \tfrac{1}{2}\|y - x\|^2 \,\Big]. \tag{12.5}$$

Note that the function $m(\cdot)$ is differentiable and 1-strongly convex. Hence, we have

$$\begin{aligned} m(y) \;&\geq\; m(x^+) + \langle \nabla m(x^+), y - x^+ \rangle + \tfrac{1}{2}\|y - x^+\|^2 \\ &\overset{(12.4)}{\geq}\; m(x^+) + \tfrac{1}{2}\|y - x^+\|^2. \end{aligned}$$

Expanding the definition of the model gives:

$$\begin{aligned} \tfrac{1}{2}\|y-x\|^2 + \eta\langle f'(x), y-x\rangle &\geq \tfrac{1}{2}\|y-x^+\|^2 + \left[\eta\langle f'(x), x^+-x\rangle + \tfrac{1}{2}\|x^+-x\|^2\right] \\ &\geq \tfrac{1}{2}\|y-x^+\|^2 - \tfrac{\eta^2}{2}\|f'(x)\|^2. \end{aligned}$$

Thus, we have proved the following important lemma, which is in the core of analysis of all subgradient methods.

Lemma 12.2.1. *For any $y \in Q$, it holds*

$$\tfrac{1}{2}\|y-x\|^2 + \tfrac{\eta^2}{2}\|f'(x)\|^2 \geq \tfrac{1}{2}\|y-x^+\|^2 + \eta\langle f'(x), x-y\rangle. \tag{12.6}$$

Consequently, substituting $y := x^\star \in Q$ (any minimizer), we get:

$$\tfrac{1}{2}\|x^\star-x\|^2 + \tfrac{\eta^2}{2}\|f'(x)\|^2 \geq \tfrac{1}{2}\|x^\star-x^+\|^2 + \eta\langle f'(x), x-x^\star\rangle. \tag{12.7}$$

Note that for convex functions, the inner product in the right hand side of (12.7) is non-negative, and it is bounded by the functional residual:

$$\langle f'(x), x-x^\star\rangle \geq f(x) - f^\star \geq 0.$$

This reasoning can immediately lead to a convergence rate for the method. But first, let us choose stepsizes in a smart way.

12.2.2 Algorithm

In non-smooth optimization, the main information that is provided by a subgradient $f'(x) \in \partial f(x)$ is of *geometric nature*, as it gives us a certain *separation of the sublevel set* of f at x. At the same time, in contrast to smooth optimization, the magnitude $\|f'(x)\|$ does not reveal much information.

Example 12.2.2. Consider $f(x) = |x|$, $x \in \mathbb{R}$. Then, for any $x \neq 0$, we have $|f'(x)| = 1$, no matter how close we are to the optimum $x^\star = 0$.

Therefore, it is natural to *normalize* the subgradient direction: $\frac{f'(x)}{\|f'(x)\|}$, which happens also to equip our method with *universal* (problem-class independent) convergence rates. We consider the following algorithm.

Algorithm 12.1: *Subgradient Method.*

Initialization: $x_0 \in Q$. Fix $K \geq 1$ and positive parameters $\{\gamma_k\}_{k\geq 0}$.

For $k = 0 \dots K-1$ **iterate:**

1. Compute a subgradient: $f'(x_k) \in \partial f(x_k)$
2. Perform the normalized subgradient step:

$$x_{k+1} = \arg\min_{y\in Q}\Big[\ \tfrac{\gamma_k}{\|f'(x_k)\|}\langle f'(x_k), y-x_k\rangle + \tfrac{1}{2}\|y-x_k\|^2\ \Big] = \pi_Q\Big(x_k - \tfrac{\gamma_k}{\|f'(x_k)\|}f'(x_k)\Big)$$

Return a point $\bar{x}_K := \arg\min\{f(y) : y \in \{x_0, \dots, x_K\}\}$ or the average $\frac{1}{K}\sum\limits_{i=0}^{K-1} x_i$.

Substituting our stepsize choice in the previous lemma, we obtain, for every $k \geq 0$:

$$\frac{\gamma_k^2}{2} + \frac{1}{2}\|x_k - x^\star\|^2 \;\geq\; \frac{1}{2}\|x_{k+1} - x^\star\|^2 + \gamma_k \Delta_k, \tag{12.8}$$

where

$$\Delta_k \;:=\; \frac{\langle f'(x_k), x_k - x^\star \rangle}{\|f'(x_k)\|}$$

is a certain *measure of optimality*. We also denote:

$$\bar{\Delta}_K \;:=\; \frac{1}{K}\sum_{i=0}^{K-1} \Delta_i \qquad \text{and} \qquad \Delta_K^\star \;:=\; \min_{0 \leq i \leq K-1} \Delta_i.$$

Clearly, we have $\bar{\Delta}_K \geq \Delta_K^\star$.

Telescoping inequality (12.8) for the first $K \geq 1$ iterations of the method, we get the following progress:

$$\frac{1}{2}\|x_0 - x^\star\|^2 + \frac{1}{2}\sum_{i=0}^{K-1} \gamma_i^2 \;\geq\; \sum_{i=0}^{K-1} \gamma_i \Delta_i \;\geq\; \Big(\sum_{i=0}^{K-1} \gamma_i\Big) \Delta_K^\star \tag{12.9}$$

Therefore, we obtain the following bound on our new accuracy measure:

$$\Delta_K^\star \;\leq\; \frac{\|x_0 - x^\star\|^2 + \sum\limits_{i=0}^{K-1} \gamma_i^2}{2\sum\limits_{i=0}^{K-1} \gamma_i} \;=\; \varphi(\gamma_0, \ldots, \gamma_{K-1}). \tag{12.10}$$

In principle, we want to choose $\{\gamma_k\}_{k \geq 0}$ such that the right hand side of (12.10) is as small as possible, i.e. to minimize it: $\varphi(\cdot) \to \min$. Notice that

- $\varphi(\cdot)$ is convex;
- $\varphi(\cdot)$ is symmetric in γ: its value is invariant to any permutation of $\{\gamma_0, \ldots, \gamma_{K-1}\}$.

For a convex symmetric function, there is always exists a solution with all the same arguments,

$$\gamma_0^\star \;=\; \gamma_1^\star \;=\; \ldots \;\gamma_{k-1}^\star \;=\; \gamma.$$

Therefore, a constant choice $\gamma > 0$ is able to give us the best bound in (12.10), when the number of iterations K is fixed. In practice, however, we might want to use a decreasing sequence, e.g. $\gamma_k = O(1/\sqrt{k})$, so not to fix K.

We denote by $R \geq \|x_0 - x^\star\|$ *any* upper bound for the distance from the initial point to any of the solutions. Using this bound in (12.9) along with the constant choice, $\gamma_i \equiv \gamma > 0$, leads to

$$\bar{\Delta}_K \overset{(12.9)}{\leq} \frac{\|x_0 - x^\star\|^2}{2K\gamma} + \frac{K\gamma}{2} \;\leq\; \frac{R^2}{2K} + \frac{K\gamma}{2}.$$

Minimizing the right-hand side in $\gamma > 0$, we obtain the optimal choice

$$\boxed{\gamma \;:=\; \frac{R}{K^{1/2}}.} \tag{12.11}$$

Thus, we have proved the following result.

Theorem 12.2.3. *Let all γ_k in Algorithm 12.1 be chosen according to* (12.11)*. Then,*

$$\Delta_K^\star \;\leq\; \bar{\Delta}_K \;\leq\; \frac{R}{K^{1/2}}. \tag{12.12}$$

12.2.3 Lipschitz Functions

The question is how to relate our quantities Δ_k with a standard accuracy measure, the functional residual $f(x_k) - f^\star$. The simplest reasoning is as follows.

Assume that our subgradients are bounded:

$$\|f'(x)\| \;\leq\; M, \qquad \forall x \in Q, \tag{12.13}$$

which means that the function f is Lipschitz:

$$|f(y) - f(x)| \;\leq\; M\|y - x\|, \qquad x, y \in Q.$$

Then, by convexity we immediately obtain the following relationship:

$$\Delta_k \;:=\; \tfrac{\langle f'(x_k), x_k - x^\star\rangle}{\|f'(x_k)\|} \overset{(12.13)}{\geq} \tfrac{\langle f'(x_k), x_k - x^\star\rangle}{M} \;\geq\; \tfrac{f(x_k) - f^\star}{M}. \tag{12.14}$$

Corollary 12.2.4. *We have,*

$$f\Big(\tfrac{1}{K}\sum_{i=0}^{K-1} x_i\Big) - f^\star \;\leq\; \tfrac{1}{K}\sum_{i=0}^{K-1}[f(x_i) - f^\star] \overset{(12.14)}{\leq} M \cdot \bar{\Delta}_K \;\leq\; \tfrac{MR}{K^{1/2}}. \tag{12.15}$$

Note that we used a conservative choice of stepsizes that requires fixing the number of iterations. Therefore, technically, (12.15) is not a "rate" of convergence, as this bound is achieved only once, for the final output of the algorithm. If we wanted to achieve higher precision, we would need to rerun the method with a smaller stepsize γ. In Lecture 14, we discuss a more advanced approach using *adaptive stepsizes* that solves this issue, as they do not require fixing the number of iterations in advance, and they also work for stochastic problems.

We also assume we know a bound $R \geq \|x_0 - x^\star\|$. At the same time, it is easy to see that any choice of $R > 0$ in (12.11) will give us a similar bound. For example, in practice, we can set $R := 1$ if no other information is available. However, the closer R is to $\|x_0 - x^\star\|$, the faster the method converges.

In the next lecture, we prove that the rate $O\Big(\tfrac{MR}{K^{1/2}}\Big)$ is *optimal* for this problem class.

12.3 Functional Growth

Note that to prove our bound (12.15) for the subgradient method on convex Lipschitz functions, it is enough to take much more conservative steps, using the same constant for each iteration in (12.5):

$$\eta \;:=\; \tfrac{\gamma}{M} \overset{(12.11)}{=} \tfrac{R}{MK^{1/2}}, \tag{12.16}$$

in contrast to using normalized stepsizes.

Exercise 12.3.1. Check that subgradient steps (12.5) with η given by (12.16) achieve the same bound (12.15) for the functional residual.

In this section, we provide a more advanced analysis that reveals the true power of normalized stepsizes: they enable us to prove convergence rates for problems well beyond the Lipschitz assumption.

12.3.1 Geometric Interpretation

First, let us understand the geometric meaning of the quantity $\Delta_k := \frac{\langle f'(x_k), x_k - x^\star \rangle}{\|f'(x_k)\|}$. For simplicity, consider the case of unconstrained optimization, thus

$$Q \equiv \mathbb{R}^n.$$

Staying at point x_k, the subgradient $f'(x_k) \neq 0$ provides us with the supporting hyperplane:

$$L_k = \left\{ y \in \mathbb{R}^n : \langle f'(x_k), x_k - y \rangle = 0 \right\}.$$

Let us look at the optimal solution $x^\star \in \mathbb{R}^n$ and find the projection of it onto L_k:

$$\min_{y \in L_k} \|y - x^\star\|. \tag{12.17}$$

We take vector $h := \Delta_k \frac{f'(x_k)}{\|f'(x_k)\|}$ and the perturbed solution $y^\star := x^\star + h$. We have

$$\begin{aligned} \langle f'(x_k), x_k - y^\star \rangle &= \langle f'(x_k), x_k - x^\star \rangle + \langle f'(x_k), h \rangle \\ &= \Delta_k \|f'(x_k)\| - \Delta_k \frac{\langle f'(x_k), f'(x_k) \rangle}{\|f'(x_k)\|} = 0, \end{aligned}$$

which concludes that $y^\star \in L_k$ is the solution to (12.17). Note that

$$\|y^\star - x^\star\| = \|h\| = \Delta_k.$$

Corollary 12.3.1. *Δ_k is the distance from $x^\star$ to the hyperplane L_k.*

Consider the localizing polyhedron:

$$G_{k+1} = \Big\{ y \in \mathbb{R}^n : \langle f'(x_0), x_0 - y \rangle \geq 0, \ldots, \langle f'(x_k), x_k - y \rangle \geq 0 \Big\}.$$

By convexity $x^\star \in G_{k+1}$. Note that $\Delta_k^\star$ is the minimal distance from $x^\star$ to one of the hyperplanes defining the polyhedron. Hence, $\Delta_k^\star$ is *the maximal radius of the Euclidan ball that is contained in the localizer G_{k+1}*:

$$\Delta_k^\star = \max \Big\{ r \geq 0 : B_r(x^\star) \subset G_{k+1} \Big\}$$

and the subgradient method manages to $\Delta_k^\star \to 0$. Note that such quantity can be used to define an approprite size$(\cdot)$ of a set (see previous lecture).

12.3.2 Convergence Rate with Functional Growth

Define the following quantity, called *the growth of f at point $x^\star$*:

$$\omega_f(r) := \max \Big\{ f(x) - f(x^\star) : \|x - x^\star\| \leq r \Big\}.$$

Clearly, $\omega_f(\cdot)$ is a nondecreasing function of r.

We can relate our measure of optimality Δ_k with the functional residual. By convexity, we know that

$$\Delta_k = \frac{\langle f'(x_k), x_k - x^\star \rangle}{\|f'(x_k)\|} \geq \frac{f(x_k) - f(x^\star)}{\|f'(x_k)\|}.$$

Now we present a more advanced reasoning. Consider the perturbation, as we fixed before:

$$h \;:=\; \tfrac{\Delta_k}{\|f'(x_k)\|} f'(x_k).$$

Then,

$$\begin{aligned}
\langle f'(x_k), x_k - x^\star \rangle &= \langle f'(x_k), x_k - x^\star - h \rangle + \langle f'(x_k), h \rangle \\
&\geq f(x_k) - f(x^\star + h) + \|f'(x_k)\| \Delta_k \\
&= f(x_k) - f^\star + (f^\star - f(x^\star + h)) + \|f'(x_k)\| \Delta_k.
\end{aligned}$$

Note that $\langle f'(x_k), x_k - x^\star \rangle = \|f'(x_k)\| \Delta_k$. Rearranging the terms, we obtain

$$f(x_k) - f^\star \;\leq\; f(x^\star + h) - f^\star.$$

This inequality has a clear geometric meaning: due to convexity, the function value at $f(x_k)$ is better than that at $f(x^\star + h)$, as shown on Fig. 12.2. We established the following result.

Theorem 12.3.2. *For the result of the subgradient method, it holds:*

$$f(\bar{x}_K) - f^\star \;\leq\; \omega_f(\Delta_K^\star), \qquad \text{and} \qquad \Delta_K^\star \;\leq\; \tfrac{R}{K^{1/2}}. \tag{12.18}$$

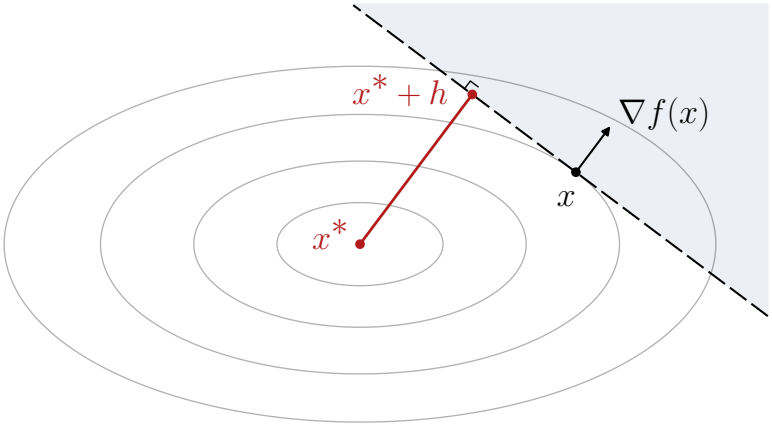

Figure 12.2: Point $x^\star + h$ is the projection of the unconstrained minimum $x^\star$ to the supporting hyperplane defined by a subgradient at x.

12.3.3 Example: Lipschitz Functions

Proposition 12.3.3. *Consider Lipschitz functions: $f(x + h) - f(x) \leq M\|h\|$. Then, we have*

$$\omega_f(r) \;\leq\; Mr. \tag{12.19}$$

Hence, we immediately recover the bound from the previous reasoning:

$$f(\bar{x}_K) - f^\star \overset{(12.18)}{\leq} \omega_f\Big(\tfrac{R}{K^{1/2}}\Big) \overset{(12.19)}{\leq} \tfrac{MR}{K^{1/2}}. \tag{12.20}$$

To obtain an ε-solution we need $\left[\frac{MR}{\varepsilon}\right]^2$ oracle calls.

Let us consider the following objective:

$$f(x) \;=\; \max_{1\le i\le m}\Big[\langle a_i, x\rangle - b_i\Big].$$

Then

$$\begin{aligned} f(y)-f(x) &= \max_{1\le i\le m}\Big[\langle a_i, x\rangle - b_i\Big] - \max_{1\le j\le m}\Big[\langle a_j, y\rangle - b_j\Big] \\ &= \max_{1\le i\le m}\Big[\langle a_i, x\rangle - b_i - \max_{1\le j\le m}\Big[\langle a_j, y\rangle - b_j\Big]\Big] \\ &\le \max_{1\le i\le m}\Big[\langle a_i, x-y\rangle\Big] \;\le\; \max_{1\le i\le m}\|a_i\|\cdot\|x-y\|. \end{aligned}$$

Hence, $M = \max\limits_{1\le i\le m}\|a_i\|$.

At every point x, a subgradient can be computed as $f'(x) := a_i$, where i it the active index.

12.3.4 Example: Smooth Functions

Proposition 12.3.4. *Let f be differentiable and smooth. Then,*

$$\begin{aligned} \omega_f(r) &= \max\{f(x^\star+h)-f(x^\star) \;:\; \|h\|\le r\} \\ &\le \max\{\langle\nabla f(x^\star),h\rangle + \tfrac{L}{2}\|h\|^2 \;:\; \|h\|\le r\} \qquad (12.21) \\ &\le \|\nabla f(x^\star)\|r + \tfrac{L}{2}r^2. \end{aligned}$$

Substituting this value into our bound, we get:

$$f(\bar x_K)-f^\star \overset{(12.18)}{\le} \omega_f\Big(\tfrac{R}{K^{1/2}}\Big) \overset{(12.21)}{\le} \tfrac{\|\nabla f(x^\star)\|R}{K^{1/2}} + \tfrac{LR^2}{2K}.$$

Hence, in case of smooth functions and small $\|\nabla f(x^\star)\|$, our method automatically receives a faster rate of convergence than that one from (12.20). For unconstrained minimization, $\nabla f(x^\star)=0$.

12.3.5 Example: Maximum of Smooth Functions

Let

$$f(x) \;=\; \max_{1\le i\le m}\Big[f_i(x)\Big], \qquad (12.22)$$

where each $f_i(x) = \frac{1}{2}\langle A_i x, x\rangle - \langle b_i, x\rangle$ is a convex quadratic function, or, more generally, each f_i is convex and has the Lipschitz gradient with constant L_i.

Then,

$$\begin{aligned} f_i(x) &\le f_i(x^\star) + \langle\nabla f_i(x^\star), x-x^\star\rangle + \tfrac{L_i}{2}\|x-x^\star\|^2 \\ &\le f_i(x^\star) + \|\nabla f_i(x^\star)\|\cdot\|x-x^\star\| + \tfrac{L_i}{2}\|x-x^\star\|^2. \end{aligned}$$

And we obtain that

$$\omega_f(r) \;\le\; \max_{1\le i\le m}\|\nabla f_i(x^\star)\|\cdot r + \max_{1\le i\le m} L_i\cdot\tfrac{r^2}{2}. \qquad (12.23)$$

Corollary 12.3.5. *It holds:*

$$f(\bar x_K)-f^\star \overset{(12.18)}{\le} \omega_f\Big(\tfrac{R}{K^{1/2}}\Big) \overset{(12.23)}{\le} \tfrac{\max_{1\le i\le m}\|\nabla f_i(x^\star)\|R}{K^{1/2}} + \tfrac{\max_{1\le i\le m} L_i R^2}{2K}.$$

We see that exactly the same subgradient method with normalized steps will work on different subclasses of non-smooth convex problems, even though the objective in (12.22) is not Lipschitz.

Literature

The subgradient method was discovered by Naum Shor in 1962 [Sho12].

Another powerful selection of stepsizes for the subgradient method includes Polyak's stepsizes [Pol87, HK19] and their recent modification suitable for constrained and composite optimization [Nes24].

See also Section 7 in [Nem95] and Section 3.2 in [Nes18] for the design of subgradient methods for constrained problems with functional inequalities.

An advanced analysis of the convergence rate for the last iterate x_k of the subgradient method, instead of $\bar{x}_k$, was developed recently in [ZG25].

[HK19] Elad Hazan and Sham Kakade. Revisiting the polyak step size. *arXiv preprint arXiv:1905.00313*, 2019.

[Nem95] Arkadi Nemirovski. *Information-based complexity of convex programming*. Lecture notes, 1995.

[Nes18] Yurii Nesterov. *Lectures on convex optimization*. Springer, 2018.

[Nes24] Yurii Nesterov. Primal subgradient methods with predefined step sizes. *Journal of Optimization Theory and Applications*, 203(3):2083–2115, 2024.

[Pol87] Boris T Polyak. Introduction to optimization. 1987.

[Sho12] Naum Zuselevich Shor. *Minimization methods for non-differentiable functions*. Springer Science & Business Media, 2012.

[ZG25] Moslem Zamani and François Glineur. Exact convergence rate of the last iterate in subgradient methods. *SIAM Journal on Optimization*, 35(3):2182–2201, 2025.