2.1. The idea of the methods can be demonstrated when considering the problem
\[ f(x) \rightarrow \min \mid x \in E, \]
where \( f \) is a convex function belonging to \( C^{1,1}(1) \) [i.e., with Lipschitz continuous, with constant 1, gradient] which attains its minimum on Hilbert space \( E, \| \cdot \| \). In this situation we want to design a method with accuracy of \( i \)-th iterate \( O(1/i^2) \).

2.1.1. Were \( f \) quadratic, for our purposes we could use the conjugate gradient algorithm (it is well known). Let us try to understand, which geometry ensures in this case the desired rate of convergence and how could we arrive at the same geometry in the general non-quadratic case.

Let \( V(x) = f(x) - \min_E f \), let \( x^* \) be the minimizer of \( f \), let \( f \) be quadratic, and let \( x_0 = \bar{x}, x_1, x_2, \ldots \) be the trajectory of the conjugate gradients on \( f \), and \( p_i = f'(x_i) \). It turns out that the desired convergence rate can be extracted from the following facts:

(A) \( \langle p_i, x_i - \bar{x} \rangle = 0 \)
(B) \( \langle p_i, p_j \rangle = 0, \ i \neq j \),
(C) \( f(x_i) \leq f(x_{i-1}) - \frac{1}{2} \| p_i \|^2 \)

(which, of course, immediately follows from the definition of \( x_i \) as the minimizer of \( f \) on the affine plane \( \bar{x} + \mathbf{R}p_1 + \ldots + \mathbf{R}p_{i-1} \)). Indeed, by convexity of \( f \) we have
\[ \langle p_i, x^* - x_i \rangle \leq -V(x_i) \equiv -V_i \quad (2.1) \]
whence by (A)
\[ \langle p_i, x^* - \bar{x} \rangle \leq -V_i \quad (2.2) \]
Let \( \lambda_i \) be positive reals; multiplying (2.2) by \( \lambda_i \) and summing the results over \( i = 1, \ldots, N \), we get
\[ \sum_{i=1}^{N} \lambda_i V_i \leq \langle \sum_{i=1}^{N} \lambda_i p_i, \bar{x} - x^* \rangle \leq \frac{1}{2} \| \sum_{i=1}^{N} \lambda_i p_i \|^2 + \frac{1}{2} \| \bar{x} - x^* \|^2. \quad (2.3) \]
By (B) \[ \| \sum_1^N \lambda_i p_i \|^2 = \sum_1^N \lambda_i^2 \| p_i \|^2, \] and by (C) the latter quantity is at most \[ 2 \sum_1^N \lambda_i^2 (V_i - V_{i+1}). \] Thus, we arrive at the inequality

\[ \sum_1^N \lambda_i V_i \leq \sum_1^N \lambda_i^2 (V_i - V_{i+1}) + \frac{1}{2} \| \bar{x} - x^* \|^2. \] (2.4)

From this inequality by appropriate choice of \( \lambda_i \) we can extract the required estimate \( V_{N+1} \leq O(N^{-2}); \) to this end it suffices to set

\[ \lambda_i = \frac{R^2}{2N(V_i - V_{i+1})} \] (2.5)

\( (R > 0 \) is a parameter). Indeed, with this choice of \( \lambda_i \) (2.4) implies that

\[ \sum_1^N \frac{V_i}{\sqrt{V_i - V_{i+1}}} \leq \sqrt{2N} \frac{R^2 + R_i^2}{2R} \] (2.6)

where \( R_* = \| \bar{x} - x^* \|. \) From (2.6) it is easy to conclude that

\[ V_{N+1} \leq \frac{(R^2 + R_i^2)^2}{2R^2} N^{-2}. \] (2.7)

2.1.2. Looking at our reasoning, we observe that it did not use (B) “as is;” what was sufficient for us is the weaker condition

(B') \[ \langle p_i, \sum_1^{i-1} \lambda_j p_j \rangle = 0, \ i \geq 1, \]

(with \( \lambda_i \) defined by (2.5)), since this is the condition which allows to rewrite \( \| \sum_1^N \lambda_i p_i \|^2 \) as \( \sum_1^N \lambda_i^2 \| p_i \|^2. \)

The conditions (A), (B'), (C) can be ensured without much effort for every convex function \( f \) belonging to \( C^{1,1}(1), \) not necessary a quadratic one. To this end it suffices, given \( x_0 = \bar{x}, x_1, ..., x_{i-1}, \) to set \( \hat{x}_i = x_{i-1} - f'(x_{i-1}), \) to minimize \( f \) on the affine plane \( E_i \) passing through \( \bar{x}, \hat{x}_i \) and parallel to the vector \( q_{i-1} = \sum_{j=1}^{i-1} \lambda_j f'(x_j), \) and to define \( x_i \) as the corresponding minimizer of \( f. \) With this approach, \( p_i = f'(x_i) \) will be orthogonal to \( x_i - \bar{x} \) (since \( \bar{x} \in E_i \)) and to \( q_{i-1} \) (since \( E_i \) is parallel to \( q_{i-1} \)), ensuring (A) and (B'); besides, \( f(x_i) \leq f(\hat{x}_i) \leq f(x_{i-1}) - \frac{1}{2} \| p_{i-1} \|^2 \) due to \( f \in C^{1,1}(1), \) so that (C) also takes place.
Thus, we can get a method with accuracy estimate $O(N^{-2})$, assuming we are able to carry out two-dimensional minimization at a step; thus construction was first proposed (in the context of minimizing strongly convex function on Hilbert space) by the author in [164] and extended to minimizing smooth convex function on $(\kappa, r)$-smooth space in [88]. It was shown that the 2D minimization can be solved approximately with explicitly given accuracy; “price” for the 2D minimization results in the total number of computations of $f$ and $f'$ in course of $N$ iterations becoming $O(N \ln N)$.

2.1.3. The next step was made by the author in [90]; it turned out that instead of two-dimensional minimization of $f$ on the plane $\hat{x}_i + R(\hat{x}_i - \bar{x}) + R \sum_{i=1}^{i-1} \lambda_j f'(x_j)$ one can minimize $f$ on the line $\bar{x}_i + R (\hat{x}_i - \bar{x} + \sum_{i=1}^{i-1} \lambda_j f'(x_j))$.

Indeed, in this case

$$\langle f'(x_i), \hat{x}_i - \bar{x} \rangle = -\langle f'(x_i), \sum_{i=1}^{i-1} \lambda_j f'(x_j) \rangle$$

and for some $t_i$,

$$x_i = \hat{x}_i + t_i \left( \hat{x}_i - \bar{x} + \sum_{i=1}^{i-1} \lambda_j f'(x_j) \right),$$

or, which is the same,

$$\bar{x} - x_i = (1 + t_i)(\bar{x} - \hat{x}_i) - t_i \sum_{i=1}^{i-1} \lambda_j f'(x_j),$$

so that (2.1) yields (below $p_j = f'(x_j)$):

$$-V_i \geq \langle p_i, x^* - x_i \rangle = \langle p_i, x^* - \bar{x} \rangle + \langle p_i, \bar{x} - x_i \rangle = \langle p_i, x^* - \bar{x} \rangle + (1 + t_i) \langle p_i, \bar{x} - \hat{x}_i \rangle - t_i \langle p_i, \sum_{i=1}^{i-1} \lambda_j p_j \rangle,$$

whence, by (2.9),

$$-V_i \geq \langle p_i, x^* - \bar{x} \rangle + \langle p_i, \sum_{i=1}^{i-1} \lambda_j p_j \rangle,$$

or

$$\lambda_i V_i + \langle \lambda_i p_i, \sum_{i=1}^{i-1} \lambda_j p_j \rangle + \langle \lambda_i p_i, x^* - \bar{x} \rangle \leq 0.$$
The second term in the left hand side of (2.10) is
\[
\frac{1}{2} \| \sum_{i=1}^{i} \lambda_j p_j \|^2 - \frac{1}{2} \| \sum_{i=1}^{i-1} \lambda_j p_j \|^2 - \frac{1}{2} \| \lambda_i p_i \|^2,
\]
and summing up (2.10) over \(i=1,2,...,N\), we get the inequality
\[
\sum_{i=1}^{N} \lambda_i V_i + \langle \sum_{i=1}^{N} \lambda_i p_i, x^* - \bar{x} \rangle + \frac{1}{2} \| \sum_{i=1}^{N} \lambda_i p_i \|^2 \leq \frac{1}{2} \| \sum_{i=1}^{N} \lambda_i^2 p_i \|^2,
\]
implying that
\[
\sum_{i=1}^{N} \lambda_i V_i \leq \frac{1}{2} \sum_{i=1}^{N} \lambda_i^2 \| p_i \|^2 + \frac{1}{2} \| x^* - \bar{x} \|^2,
\]
which, along with (C), was used to arrive at (2.4).

Our method now takes the following form:
\[
\begin{align*}
x_0 &= \bar{x} \\
\hat{x}_i &= x_{i-1} - f'(x_{i-1}) \\
x_i &= \hat{x}_i + t_i (\hat{x}_i - \bar{x} + \sum_{j=1}^{i-1} \lambda_j f'(x_j))
\end{align*}
\]
with
\[
t_i \in \text{Argmin}_{t \in \mathbb{R}} \left\{ f(\hat{x}_i + t(\hat{x}_i - \bar{x} + \sum_{j=1}^{i-1} \lambda_j p_j)) \right\}
\]
(2.11)

One-dimensional minimizations also can be carried out approximately; \(N\) iterations still require \(O(N \ln N)\) computations of \(f\) and \(f'\). Note that the geometric nature of the method (2.11), (2.12) is not as transparent as for the method with two-dimensional minimization; we now are using certain “analytical trick.”

2.1.4. The concluding stem in this avenue of research was carried out by Yu. E. Nesterov [99,100]; he noticed that with appropriate rule for generating \(\lambda_j\) one can get rid of one-dimensional minimization; it turns out that in (2.11) one can specify \(t_i\) by an explicit formula, resulting in method with convergence rate \(O(N^{-2})\) and \(2N\) computations of \(f\) and \(f'\) per \(N\) iterations. The brilliant idea of Nesterov is also of analytical nature, and in our opinion, the geometry of the optimal method for smooth convex minimization discovered by him is not clear yet.