A SIMPLE PROOF OF NESTEROV CONVERGENCE

OMAR HIJAB

Let $f(w)$ be a scalar function of a point w in euclidean space. A basic problem is to minimize $f(w)$, that is, to find or compute a minimizer w^*,

$$f(w) \geq f(w^*), \qquad \text{for every } w.$$

A *descent sequence* is a sequence $w_0, w_1, w_2, \ldots$ satisfying

$$f(w_0) \geq f(w_1) \geq f(w_2) \geq \ldots.$$

In a descent sequence, the point after $w = w_n$ is $w^+ = w_{n+1}$, and the point before w is $w^- = w_{n-1}$. Then $(w^-)^+ = w = (w^+)^-$.

We assume $f(w)$ is smooth and strictly convex: There are positive constants $m < L$ with

$$\frac{m}{2}|x-a|^2 \leq f(x) - f(a) - \nabla f(a) \cdot (x-a) \leq \frac{L}{2}|x-a|^2. \tag{1}$$

Then there is a unique global minimizer w^*.

Theorem (Nesterov [1, 2, 3])**.** *Let $r = m/L$, $E(w) = f(w) - f(w^*)$, and*

$$t = \frac{1}{L}, \qquad s = \frac{1-\sqrt{r}}{1+\sqrt{r}}, \qquad \rho = 1 - \sqrt{r}.$$

Starting from any initial w_0, the sequence $w_{-1} = w_0, w_1, w_2, \ldots$ given by

$$\begin{aligned} w^\circ &= w + s(w - w^-), \\ w^+ &= w^\circ - t\nabla f(w^\circ). \end{aligned} \tag{2}$$

converges to w^* at the rate

$$E(w_n) \leq 2\rho^n E(w_0), \qquad n = 1, 2, \ldots. \tag{3}$$

The proof presented here is a rearrangement of the proof in the book of Wright and Recht [3]. A consequence of the current proof is the natural emergence of the expressions for s and ρ.

Proof. Starting from w_0, and setting $w_{-1} = w_0$, the loss sequence $f(w_0)$, $f(w_1)$, $f(w_2), \ldots$ is not always decreasing. Because of this, we seek another function $V(w)$ where the corresponding sequence $V(w_0)$, $V(w_1)$, $V(w_2), \ldots$ is decreasing.

To explain this, it's best to assume $w^* = 0$ and $f(w^*) = 0$. This can always be arranged by translating the coordinate system. Then it turns out

$$V(w) = f(w) + \frac{L}{2}|w - \rho w^-|^2, \tag{4}$$

with a suitable choice of ρ, does the job. With the right choices for ρ and s, we will show

$$V(w^+) \leq \rho V(w). \tag{5}$$

Date: January 2024.

We first show how (5) implies the result (3), assuming $\rho = 1 - \sqrt{r}$. Insert $x = w_0$ and $a = w^* = 0$ in (1). Then

$$V(w_0) = f(w_0) + \frac{L}{2}|w_0 - \rho w_0|^2 = f(w_0) + \frac{m}{2}|w_0|^2 \leq 2f(w_0).$$

Moreover $f(w) \leq V(w)$. Iterating (5), we obtain

$$f(w_n) \leq V(w_n) \leq \rho^n V(w_0) \leq 2\rho^n f(w_0),$$

which is (3). We now derive (5).

Since $t = 1/L$ is the standard short-step learning rate, the second half of (2), together with (1), implies

$$f(w^+) \leq f(w^\circ) - \frac{t}{2}|g^\circ|^2, \qquad g^\circ = \nabla f(w^\circ). \tag{6}$$

By (1) with $x = w$ and $a = w^\circ$,

$$f(w^\circ) \leq f(w) - g^\circ \cdot (w - w^\circ) - \frac{m}{2}|w - w^\circ|^2. \tag{7}$$

By (1) with $x = w^* = 0$ and $a = w^\circ$,

$$f(w^\circ) \leq g^\circ \cdot w^\circ - \frac{m}{2}|w^\circ|^2. \tag{8}$$

Multiply (7) by ρ and (8) by $1 - \rho$ and add, then insert the sum into (6). After some simplification,

$$f(w^+) \leq \rho f(w) + g^\circ \cdot (w^\circ - \rho w) - \frac{r}{2t}\left(\rho|w - w^\circ|^2 + (1-\rho)|w^\circ|^2\right) - \frac{t}{2}|g^\circ|^2. \tag{9}$$

Since $(w^\circ - \rho w) - tg^\circ = w^+ - \rho w$,

$$\frac{1}{2t}|w^+ - \rho w|^2 = \frac{1}{2t}|w^\circ - \rho w|^2 - g^\circ \cdot (w^\circ - \rho w) + \frac{t}{2}|g^\circ|^2.$$

Adding this to (9) leads to

$$V(w^+) \leq \rho f(w) - \frac{r}{2t}\left(\rho|w - w^\circ|^2 + (1-\rho)|w^\circ|^2\right) + \frac{1}{2t}|w^\circ - \rho w|^2. \tag{10}$$

Let

$$R(a,b) = r\left(\rho s^2|b|^2 + (1-\rho)|a + sb|^2\right) - |(1-\rho)a + sb|^2 + \rho|(1-\rho)a + \rho b|^2.$$

Solving for $f(w)$ in (4) and inserting into (10) leads to

$$V(w^+) \leq \rho V(w) - \frac{1}{2t}R(w, w - w^-). \tag{11}$$

If we can choose s and ρ so that $R(a,b)$ is a *positive* scalar multiple of $|b|^2$, then, by (11), (5) follows, completing the proof. Based on this, we choose s, ρ to make $R(a,b)$ independent of a. But

$$\nabla_a R = 2(1-\rho)\left(\left(r - (1-\rho)^2\right)a + \left(\rho^2 - s(1-r)\right)b\right),$$

and $\nabla_a R = 0$ is two equations in two unknowns s, ρ. This leads to the choices for s and ρ made above. Once these choices are made, $s(1-r) = \rho^2$ and $\rho > s$. From this,

$$R(a,b) = R(0,b) = (rs^2 - s^2 + \rho^3)|b|^2 = \rho^2(\rho - s)|b|^2, \tag{12}$$

which is positive. □

Note: Since the proof is dimension-independent, a version of this result should hold in Hilbert space.

References

[1] Sébastien Bubeck, *Convex Optimization: Algorithms and Complexity*, Now Publishers (2015).
[2] Yurii Nesterov, *Lectures on Convex Optimization*, Springer (2018).
[3] Stephen J. Wright and Benjamin Recht, *Optimization for Data Analysis*, Cambridge University (2022).

Temple University
Email address: hijab@temple.edu