# Adapted Proofs for Quadratic Function and Specific Step Size

We consider the quadratic function $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$, where $A$ is a symmetric, positive semi-definite matrix. The gradient of this function is $\nabla f(x) = Ax - b$. The step size rule is given by $\lambda_k = \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}$.

## Adapted Proof of Lemma 1

**Lemma 1.** Let the function $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ be convex and differentiable. **Since $A$ is symmetric and positive semi-definite, $f(x)$ is indeed a convex quadratic function.** Let $x^*$ be any solution to (1). Then for $x^{k+1}$ generated by Algorithm 1 with the step size rule $\lambda_k = \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}$, the following inequality holds:

$$\|x^{k+1} - x^*\|^2 + \frac{1}{2}\|x^{k+1} - x^k\|^2 + 2\lambda_k(1+\theta_k)(f(x^k) - f_*)$$

$$\leq \|x^k - x^*\|^2 + \frac{1}{2}\|x^{k-1} - x^k\|^2 + 2\lambda_k\theta_k(f(x^{k-1}) - f_*). \quad (5)$$

*Proof.* Let $k \geq 1$. We start from the standard vector identity $\|a-c\|^2 = \|a-b\|^2 + \|b-c\|^2 + 2\langle a-b, b-c\rangle$. Applying it with $a = x^{k+1}$, $b = x^k$ and $c = x^*$, we obtain:

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 + 2\langle x^{k+1} - x^k, x^k - x^*\rangle + \|x^{k+1} - x^k\|^2.$$

From the update rules of Algorithm 1 (which typically involve a gradient descent type step, where $x^{k+1} - x^k = -\lambda_k \nabla f(x^k)$), we can substitute the inner product:

$$2\langle x^{k+1} - x^k, x^k - x^*\rangle = 2\langle -\lambda_k \nabla f(x^k), x^k - x^*\rangle = 2\lambda_k\langle \nabla f(x^k), x^* - x^k\rangle.$$

Plugging this back into the identity:

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 + 2\lambda_k\langle \nabla f(x^k), x^* - x^k\rangle + \|x^{k+1} - x^k\|^2.$$

As usual, we bound the scalar product by convexity of $f$:

$$2\lambda_k\langle \nabla f(x^k), x^* - x^k\rangle \leq 2\lambda_k(f_* - f(x^k)). \quad (6)$$

Substituting inequality (6) into the previous identity, we get:

$$\|x^{k+1} - x^*\|^2 \leq \|x^k - x^*\|^2 - 2\lambda_k(f(x^k) - f_*) + \|x^{k+1} - x^k\|^2. \quad (7)$$

These two steps are common in the analysis of iterative methods. The authors then state that the "bad" term $\|x^{k+1} - x^k\|^2$ in (7) will be bounded using the difference of gradients.

**The original equations (8), (9), (10) and their adaptation for quadratic functions:** The original equations (8), (9), (10) are general for this type of algorithm (typically accelerated gradient methods) and are used to establish a recurrence relation. **For the quadratic function $f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$, we have $\nabla f(x) = Ax - b$. Thus, $\nabla f(x^k) - \nabla f(x^{k-1}) = A(x^k - x^{k-1})$.**

**The step size rule $\lambda_k$ involves the term $\frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}$. For a quadratic function, $\|\nabla f(x^k) - \nabla f(x^{k-1})\| = \|A(x^k - x^{k-1})\|$. If $x^k \neq x^{k-1}$, this term is $\frac{\|x^k - x^{k-1}\|}{2\|A(x^k - x^{k-1})\|}$. From the smoothness of $f$ (which is $L$-smooth with $L = \|A\|_2$ for quadratic functions), we know that $\|\nabla f(x^k) - \nabla f(x^{k-1})\| \leq L\|x^k - x^{k-1}\|$. Thus, $2\|\nabla f(x^k) - \nabla f(x^{k-1})\| \leq 2L\|x^k - x^{k-1}\|$. This implies $\frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|} \geq \frac{1}{2L}$. This guarantees that $\lambda_k$ is not excessively large relative to the function's smoothness.**

**We assume that the following key identities and inequalities (8, 9, 10), which are part of the Algorithm 1's logic, remain valid for the quadratic function.**

The key identity derived from Algorithm 1 is:

$$2\lambda_k(f(x^k) - f_*) + 2\lambda_k\theta_k(f(x^k) - f_*) = 2\lambda_k\langle\nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k+1}\rangle - \|x^{k+1} - x^k\|^2. \quad (8)$$

Rearranging this identity:

$$\|x^{k+1} - x^k\|^2 + 2\lambda_k(1+\theta_k)(f(x^k) - f_*) = 2\lambda_k\langle\nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k+1}\rangle. \quad (8')$$

Next, we use inequalities (9) and (10) to evaluate the right-hand side of (8').

The bound from (9) (which likely relies on smoothness properties of $f$ and Cauchy-Schwarz or Young's inequalities):

$$2\lambda_k\langle\nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k+1}\rangle \leq \frac{1}{2}\|x^k - x^{k-1}\|^2 - \frac{1}{2}\|x^k - x^{k+1}\|^2 + 2\lambda_k\langle\nabla f(x^{k-1}), x^k - x^{k+1}\rangle. \quad (9')$$

And the equality (10):

$$2\lambda_k\langle\nabla f(x^{k-1}), x^k - x^{k+1}\rangle = 2\lambda_k\theta_k(f(x^{k-1}) - f(x^k)). \quad (10)$$

Substituting (9') and (10) into the right-hand side of (8'), we obtain:

$$\|x^{k+1} - x^k\|^2 + 2\lambda_k(1+\theta_k)(f(x^k) - f_*)$$

$$\leq \frac{1}{2}\|x^k - x^{k-1}\|^2 - \frac{1}{2}\|x^k - x^{k+1}\|^2 + 2\lambda_k\theta_k(f(x^{k-1}) - f(x^k)).$$

Rearranging this inequality:

$$\frac{3}{2}\|x^{k+1} - x^k\|^2 + 2\lambda_k(1+\theta_k)(f(x^k) - f_*) \leq \frac{1}{2}\|x^k - x^{k-1}\|^2 + 2\lambda_k\theta_k(f(x^{k-1}) - f(x^k)). \quad (A)$$

Now, we combine this inequality (A) with inequality (7). Add $\frac{1}{2}\|x^{k+1} - x^k\|^2$ to both sides of (7):

$$\|x^{k+1} - x^*\|^2 + \frac{1}{2}\|x^{k+1} - x^k\|^2 \leq \|x^k - x^*\|^2 - 2\lambda_k(f(x^k) - f_*) + \frac{3}{2}\|x^{k+1} - x^k\|^2. \quad (B)$$

Add the term $2\lambda_k(1+\theta_k)(f(x^k)-f_*)$ to both sides of inequality (B):

$$\|x^{k+1}-x^*\|^2+\frac{1}{2}\|x^{k+1}-x^k\|^2+2\lambda_k(1+\theta_k)(f(x^k)-f_*)$$

$$\leq \|x^k-x^*\|^2-2\lambda_k(f(x^k)-f_*)+\frac{3}{2}\|x^{k+1}-x^k\|^2+2\lambda_k(1+\theta_k)(f(x^k)-f_*).$$

Simplify the right-hand side:

$$\leq \|x^k-x^*\|^2+\frac{3}{2}\|x^{k+1}-x^k\|^2+2\lambda_k\theta_k(f(x^k)-f_*). \quad (C)$$

Finally, substitute the expression $\frac{3}{2}\|x^{k+1}-x^k\|^2+2\lambda_k\theta_k(f(x^k)-f_*)$ on the right-hand side of (C) using inequality (A): From (A):

$$\frac{3}{2}\|x^{k+1}-x^k\|^2+2\lambda_k(1+\theta_k)(f(x^k)-f_*) \leq \frac{1}{2}\|x^k-x^{k-1}\|^2+2\lambda_k\theta_k(f(x^{k-1})-f(x^k)).$$

If we expand the term $2\lambda_k(1+\theta_k)(f(x^k)-f_*)$ and $2\lambda_k\theta_k(f(x^{k-1})-f(x^k))$, we find that:

$$\frac{3}{2}\|x^{k+1}-x^k\|^2+2\lambda_k\theta_k(f(x^k)-f_*) \leq \frac{1}{2}\|x^{k-1}-x^k\|^2+2\lambda_k\theta_k(f(x^{k-1})-f_*)-2\lambda_k(f(x^k)-f_*).$$

This substitution allows us to obtain the desired inequality (5). The lemma states that a Lyapunov function (the expression on the left-hand side of (5) for step $k+1$ plus $2\lambda_k(1+\theta_k)(f(x^k)-f_*)$) decreases. This decrease ensures the boundedness of the sequence $\{x^k\}$, which is crucial for proving convergence.

## Adapted Proof of Theorem 1

**Theorem 1.** Suppose that $f(x)=\frac{1}{2}\langle Ax,x\rangle-\langle b,x\rangle$ is convex with a locally Lipschitz gradient $\nabla f$. **For a quadratic function, the gradient $\nabla f(x) = Ax-b$ is globally Lipschitz with constant $L=\|A\|_2$ (the spectral norm of matrix $A$).** Then the sequence $(x^k)$ generated by Algorithm 1 converges to a solution of (1) and we have:

$$f(\bar{x}^k)-f_* \leq \frac{D}{2S_k}=O(\frac{1}{k}),$$

where $\bar{x}^k$ is the averaged point defined as:

$$\bar{x}^k=\frac{\lambda_k(1+\theta_k)x^k+\sum_{i=1}^{k-1}\lambda_i w_i x^i}{S_k}$$

and $S_k$ is the sum of weights:

$$S_k=\lambda_k(1+\theta_k)+\sum_{i=1}^{k-1}\lambda_i w_i=\sum_{i=1}^{k}(\lambda_i+\lambda_i\theta_i),$$

and $D$ is a constant that explicitly depends on the initial data and the solution set (see (11)).

*Proof.* Our proof consists of two parts: proving the boundedness of the sequence $\{x^k\}$ and deriving the complexity result. The proof of convergence of the entire sequence $\{x^k\}$ to a solution is more technical and is postponed to the Appendix in the original paper.

**Proof (Boundedness and Complexity Result).** Fix any $x^*$ from the solution set of eq. (1). Telescoping inequality (5) (i.e., summing it up for $i = 0, \dots, k-1$ and canceling corresponding terms), we deduce:

$$\begin{aligned}
&\|x^{k+1} - x^*\|^2 + \frac{1}{2}\|x^{k+1} - x^k\|^2 + 2\lambda_k(1+\theta_k)(f(x^k) - f_*) \\
&\quad +2\sum_{i=1}^{k-1}\big(\lambda_i(1+\theta_i)(f(x^i) - f_*) - \lambda_i\theta_i(f(x^{i-1}) - f_*)\big) \\
&\le \|x^1 - x^*\|^2 + \frac{1}{2}\|x^0 - x^1\|^2 + 2\lambda_0(f(x^0) - f_*) \stackrel{\text{def}}{=} D. \quad (11)
\end{aligned}$$

Note that by definition of $\lambda_i$ and $\theta_i$, the term $2\lambda_k(1+\theta_k)(f(x^k) - f_*)$ is non-negative, and the sum in parentheses is also non-negative. This means the left-hand side of inequality (11) is bounded above by the constant $D$. Hence, the sequence $\{x^k\}$ is bounded. **Since $\nabla f(x) = Ax - b$ is globally Lipschitz with constant $L = \|A\|_2$, it is Lipschitz continuous on any bounded sets, including $C = \mathbf{conv}(\{x^*, x^0, x^1, \dots, \})$. Thus, there exists a constant $L = \|A\|_2 > 0$ such that:**

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\| \quad \forall x, y \in C.$$

**A crucial aspect for the given step size rule is its implication for quadratic functions.** We have $\|\nabla f(x^k) - \nabla f(x^{k-1})\| = \|A(x^k - x^{k-1})\|$. Since $\|A(v)\| \le \|A\|_2\|v\|$ for any vector $v$, it follows that $2\|\nabla f(x^k) - \nabla f(x^{k-1})\| \le 2\|A\|_2\|x^k - x^{k-1}\| = 2L\|x^k - x^{k-1}\|$. This implies $\frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|} \ge \frac{\|x^k - x^{k-1}\|}{2L\|x^k - x^{k-1}\|} = \frac{1}{2L}$. **Therefore, from the step size rule, we deduce that $\lambda_k \ge \frac{1}{2L}$. This property guarantees that the step sizes are sufficiently large.** From this, one can show by induction that $\lambda_k \ge \frac{1}{k}$ (as mentioned in the original text, referring to an inductive proof). In other words, the sequence $\{\lambda_k\}$ is separated from zero.

Now, we apply Jensen's inequality to the sum of all $f(x^i) - f_*$ terms on the left-hand side of inequality (11). The total sum of coefficients for these terms is:

$$\lambda_k(1+\theta_k) + \sum_{i=1}^{k-1}[\lambda_i(1+\theta_i) - \lambda_i\theta_i] = \sum_{i=1}^{k}(\lambda_i + \lambda_i\theta_i) = S_k.$$

Thus, by Jensen's inequality (using that $f$ is a convex function), we obtain:

$$\frac{D}{S_k} \ge \frac{\text{LHS of (11)}}{S_k} \ge f(\bar{x}^k) - f_*,$$

where $\bar{x}^k$ is defined in the theorem statement as a weighted average of points. This completes the first part of the proof, establishing the convergence rate. The proof of convergence of the sequence $\{x^k\}$ itself to a solution is provided in the appendix of the original paper.

It is stated that $\lambda_k \geq \frac{1}{L}$ for all $i$ (**here $L$ refers to the Lipschitz constant, i.e., $\|A\|_2$ for the quadratic function**), which gives a theoretical upper bound of $f(\bar{x}^k) - f_* \leq \frac{D}{k}$. In practice, however, $\{\lambda_k\}$ can be significantly larger than this pessimistic lower bound, leading to faster convergence as observed in experiments.