Математички Билтен
49 (LXXV) 1 (2025), 1–15

UDC: 519.853:519.863
DOI: 10.37560/matbil2549101n

ON THE CONVERGENCE OF THE PROXIMAL GRADIENT METHOD WITH VARIABLE STEP SIZES

FILIP NIKOLOVSKI [1] AND IRENA STOJKOVSKA [2]

Abstract. Composite optimization problems arise frequently in modeling, since the objective function might contain components that do not possess some "nice" properties like differentiability; the case of ℓ_1 (LASSO) regularization is one such example. The proximal gradient methods are designed to handle this kind of optimization problems, and can solve them efficiently when the proximal mapping has a closed-form solution. Theoretical analyses of the convergence properties of the proximal gradient method with constant step size have showed sublinear and linear convergence for convex and strongly convex objective functions respectively. In this paper we show that under standard assumptions the same kind of convergence result can be established for the proximal gradient method with variable step sizes in the general setting of bounded step sizes. Further, a recently proposed step size selection for the proximal gradient method with variable step sizes is considered, and the above convergence analysis is implemented for this method.

1. INTRODUCTION

Optimization models in data science, signal processing, and statistical learning frequently combine a smooth *data fidelity* term with a (possibly) nonsmooth structural term. This leads to the *composite optimization problem*:

$$\min_{x\in\mathbb{R}^d} \; F(x) \equiv f(x) + g(x), \tag{1.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable with Lipschitz continuous gradient and $g : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is proper convex and lower semicontinuous, allowing for nonsmooth penalties, and even constraints using indicator functions. The split is useful in the modeling: f measures agreement with the data (e.g. in a mean squared error sense), while g encodes prior structure such as sparsity,

2010 *Mathematics Subject Classification.* Primary: 90C25, 90C30 Secondary: 65K05.

Key words and phrases. convex optimization, first-order methods, proximal gradient method, variable step size, regularization, composite optimization problem.

group structure, or boundedness, thus improving stability and interpretability [10, 7]. We also assume that an optimal solution x^* of the problem (1.1) exists.

The composite structure of the objective function is directly connected to *regularization*. Classical ℓ_2 (Tikhonov, ridge) regularization stabilizes ill-conditioned problems, whereas ℓ_1 (LASSO) regularization promotes sparsity and interpretability in linear regression and inverse problems; both easily fit as choices of g [2]. In practice, these penalties enable scalable solvers whose per-iteration cost is dominated by matrix–vector products and simple proximity steps [2, 10].

An extensive algorithmic base exists for the composite problem class (1.1), with or without a convergence analysis of the proposed method. When $g \equiv 0$, one recovers the usual gradient descent method; when g is an indicator of a convex set, one recovers the projected gradient method [12]. For general convex g, *proximal gradient* (forward-backward) updates relate a gradient step on f with a proximity operator of g, leading to straightforward descent analyses and with an $\mathcal{O}(1/k)$ complexity in the convex setting [10, 14]. Acceleration via Nesterov-type extrapolation improves the worst-case rate to $\mathcal{O}(1/k^2)$ while preserving the simplicity of first-order steps [2]. Beyond fixed (Euclidean) metrics, variable-metric variants (e.g. diagonal Barzilai–Borwein scaling) adapt step sizes to local curvature with a low per-iteration cost, and can be combined with line search for robustness [11]. In problems that are nonsmooth and potentially nonconvex, nonmonotone and inertial extensions of proximal gradient methods provide convergence to critical points under the usual assumptions (e.g. *Kurdyka-Łojasiewicz* or *Polyak-Łojasiewicz* inequalities), and often result in a notable increase of the speed in practice [5]. To the best of the authors' knowledge, a convergence analysis for the proximal gradient method with variable step sizes in its general form, assuming only bounded step sizes, has not yet been presented. Sources we analyzed do contain a reference to bounded step sizes, however this is always coupled with a backtracking/extrapolation and/or another similar line search-type procedure; see [3, 4, 6].

We adopt the composite perspective and in this paper we analyze the convergence of the proximal gradient method with variable step sizes in a general setting assuming bounded step sizes, aiming to balance theoretical guarantees with practical adaptability, as per the discussions in [7, 10]. Our results indicate that, under standard assumptions, the proximal gradient method achieves a sublinear convergence rate for a convex component function f, whereas it achieves a linear convergence rate when the component f is μ-strongly convex.

The rest of the paper is organized as follows. In section 2 some preliminaries about the proximal gradient method with constant step sizes are given. Section 3 gives a convergence analysis of the proximal gradient method with variable step sizes in a general setting of bounded step sizes. In section 4, a recently proposed step size selection is considered jointly with a convergence analysis of the method. The conclusions are presented in the last section.

2. Preliminaries: Proximal gradient method with constant step sizes

We are solving the composite optimization problem (1.1). The *proximal mapping* is a key concept in the method we analyze.

Definition 2.1. *For a function g with a domain in $\mathbb{R}^d$ and scalar $\lambda > 0$ we define a* ***proximal mapping*** *as:*

$$\operatorname{prox}_{\lambda g}(z) = \operatorname*{argmin}_{y \in \mathbb{R}^d} \left\{ g(y) + \frac{1}{2\lambda} \|y - z\|^2 \right\}.$$

When g is the indicator of a closed convex set, the mapping $\operatorname{prox}_{\lambda g}(\cdot)$ reduces to a (Euclidean) projection [10].

While conceptually simple, the practical computation of proximal mappings faces several obstacles:

- **No closed form solutions.** For an arbitrary convex function g, the minimization problem in the definition has no explicit general solution. This requires an additional iterative solver within each proximal step [1, 10].
- **High-dimensional coupling.** Functions that couple variables (e.g. total variation, matrix nuclear norm) result in non-separable proximal mappings, often demanding large-scale decompositions like singular value decompositions [1].
- **Computational cost.** Even when calculable, proximal mappings may be much more expensive to obtain than a gradient step (e.g. spectral regularizers require full eigenvalue or singular value computations [10]).
- **Nonconvexity.** If g is nonconvex, the proximal mapping can be set-valued or ill-defined, complicating its calculation [1].

Despite the described challenges, several important cases do have closed-form solutions:

(1) ℓ_1-regularization (LASSO): for $g(x) = \alpha\|x\|_1$,

$$(\operatorname{prox}_{\alpha g}(z))_i = \operatorname{sign}(z_i)\, \max\{|z_i| - \alpha, 0\},$$

which is the *soft-threshold operator* [2].

(2) ℓ_2-regularization (Tikhonov/ridge penalty): for $g(x) = \frac{\alpha}{2}\|x\|_2^2$,

$$\operatorname{prox}_{\alpha g}(z) = \frac{1}{1+\alpha}\, z.$$

(3) Indicator of a convex set: for $g(x) = \iota_C(x)$, where $\iota_C(\cdot)$ is the indicator of a closed convex set C,

$$\operatorname{prox}_{\alpha g}(z) = \operatorname{proj}_C(z),$$

i.e. the (Euclidean) projection of z onto C.

(4) ℓ_2-norm regularization: for $g(x) = \alpha\|x\|_2$,

$$\operatorname{prox}_{\alpha g}(z) = \begin{cases} \left(1 - \dfrac{\alpha}{\|z\|_2}\right) z & \text{if } \|z\|_2 > \alpha, \\ 0 & \text{otherwise,} \end{cases}$$

Algorithm 1 Proximal gradient method with constant step size

input: $x_0 \in \mathbb{R}^d$, $\lambda > 0$, $N \in \mathbb{N}$
set: $k = 0$
while $k < N$ **do**
    $x_{k+1} = \text{prox}_{\lambda g}(x_k - \lambda \nabla f(x_k))$
    $k \leftarrow k + 1$

which corresponds to the vector soft-threshold or shrinkage operator [10].

The properties of the proximal mapping are well-understood and studied. For detailed treatise on this topic see [1]. We give one property which we use further in the paper.

Proposition 2.1. [1] *Let $g : \mathbb{R}^d \to (-\infty, +\infty]$ be a proper, closed and convex function. Then for all $x, y \in \mathbb{R}^d$:*

(i) $\|\text{prox}_{\lambda g}(x) - \text{prox}_{\lambda g}(y)\|^2 \leqslant (x - y)^\top (\text{prox}_{\lambda g}(x) - \text{prox}_{\lambda g}(y))$.
(ii) $\|\text{prox}_{\lambda g}(x) - \text{prox}_{\lambda g}(y)\| \leqslant \|x - y\|$.

Property (i) in Proposition 2.1 is called *firm nonexpansivity*, while property (ii) is called *nonexpansivity*.

The proximal gradient method with constant step size solves problem (1.1) by generating the iterative sequence of approximate solutions as follows. Select a fixed, constant step size $\lambda > 0$ and an initial approximation of the solution x_0. Generate subsequent iterates as:

$$x_{k+1} = \text{prox}_{\lambda g}(x_k - \lambda \nabla f(x_k)),$$

for $k = 0, 1, \ldots$ until some convergence criterion is not satisfied (e.g. a predefined number of iterations). Algorithm 1 describes the outlined process formally; see also [2, 9, 10].

With a slight rearrangement of the update rule in the proximal gradient method, we can rewrite it in a form structurally similar to ordinary the gradient method. We define a *generalized gradient* $G_\lambda(\cdot)$ as:

$$G_\lambda(x) = \frac{1}{\lambda} \left(\text{prox}_{\lambda g}(x - \lambda \nabla f(x)) - x \right).$$

Then the proximal gradient method update can be represented as:

$$x_{k+1} = x_k + \lambda G_\lambda(x_k).$$

This implies that the optimality condition in this case is, per [10]:

$$G_\lambda(x^*) = 0 \iff 0 \in \nabla f(x^*) + \partial g(x^*),$$

where $\partial g(x) = \{\gamma \in \mathbb{R}^d | g(y) \geqslant g(x) + \gamma^T (y - x)\}$ is the subdifferential (i.e. the set of all subgradients) of g at x.

To analyze the convergence of the proximal gradient method with constant step sizes, we make the following assumptions [7, 9, 10].

Assumption 2.1: For the component functions f and g in the composite optimization problem (1.1) we assume:

(i) $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable and L-smooth, i.e. there exists a constant $L > 0$ such that for all $x, y \in \mathbb{R}^d$:

$$f(y) \leqslant f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2}\|y - x\|^2.$$

Equvalently, the gradient of f is Lipschitz continuous with constant L:

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|, \quad \text{for all} \quad x, y \in \mathbb{R}^d.$$

(ii) $g : \mathbb{R}^d \to (-\infty, +\infty]$ is proper, closed, convex, and possibly nonsmooth.

Under Assumption 2.1, straightforward convergence results can be established for the proximal gradient method with constant step size. Detailed analysis with appropriate proofs can be found in [2, 9, 10, 12, 14]. Based on these we formulate the following convergence results for the proximal gradient method with constant step size.

Theorem 1. [9] *Let $F(x) = f(x) + g(x)$ with $f : \mathbb{R}^d \to \mathbb{R}$ continuously differentiable, convex and L-smooth, and $g : \mathbb{R}^d \to \mathbb{R}$ continuous, convex and possibly non-smooth. Let $0 < \lambda \leqslant \frac{1}{L}$ and let $\{x_k\}$ be the sequence generated by Algorithm 1. Then for any $k \geqslant 1$ we have:*

$$F(x_k) - F(x^*) \leqslant \frac{1}{2k\lambda}\|x_0 - x^*\|^2,$$

where x^* is a solution to (1.1).

If, in addition, we assume that f is μ-strongly convex, then we have the following result.

Theorem 2. [9] *Let $F(x) = f(x) + g(x)$ with $f : \mathbb{R}^d \to \mathbb{R}$ continuously differentiable, L-smooth and μ-strongly convex, and $g : \mathbb{R}^d \to \mathbb{R}$ continuous, convex and possibly non-smooth. Let $\{x_k\}$ be the sequence generated by Algorithm 1. Then for $0 < \lambda \leqslant \frac{1}{L}$ the following holds for all $k \geqslant 1$:*

$$F(x_k) - F(x^*) \leqslant \left(1 + \frac{\lambda\mu}{4}\right)^{-k} (F(x_0) - F(x^*)),$$

where x^* is a solution to (1.1).

Obviously, Theorems 1 and 2 hold for the frequent choice $\lambda = \frac{1}{L}$. As noted in [10] the proximal gradient method will still converge for step sizes smaller than $2/L$, not just $1/L$.

3. Proximal gradient method with variable step size

Once again we are looking into the solution of the composite optimization problem (1.1), and we assume that an optimal solution x^* exists. Instead of taking a constant step size $\lambda > 0$ at each iteration of the proximal gradient method, we now allow the step sizes to vary from one iteration to the next, within some closed interval $[\lambda_{\min}, \lambda_{\max}]$ with $\lambda_{\min} > 0$. We give the general form of the *proximal*

Algorithm 2 General form of proximal gradient method with variable step size

1: **input:** $x_0 \in \mathbb{R}^d$, $0 < \lambda_{\min} < \lambda_{\max}$, $\lambda_0 \in [\lambda_{\min}, \lambda_{\max}]$, $N \in \mathbb{N}$
2: **set:** $k = 0$
3: **while** $k < N$ **do**
4: $\quad x_{k+1} = \text{prox}_{\lambda_k g}(x_k - \lambda_k \nabla f(x_k))$
5: $\quad$ **select** $\lambda_{k+1} \in [\lambda_{\min}, \lambda_{\max}]$
6: $\quad k \leftarrow k + 1$

gradient method with variable step size in Algorithm 2. The details about selection of the step sizes λ_{k+1} in line 5 of Algorithm 2 will be discussed in the next section.

In this section, under standard assumptions, we establish the same theoretical convergence results for the proximal gradient method with variable step sizes, as done with its constant step size counterpart.

In the following propositions we denote, for brevity: $f(x_k) \equiv f_k$, $\nabla f(x_k) \equiv \nabla f_k$, $f(x^*) \equiv f^*$, $F(x_k) \equiv F_k$ and $F(x^*) \equiv F^*$.

Theorem 3. *Let Assumption 2.1 hold and let $\{x_k\}$ be the sequence of iterates generated by Algorithm 2. Then for all $z \in \mathbb{R}^d$ and $k \geqslant 0$:*

$$F_{k+1} \leqslant f_k + g(z) + \nabla f_k^\top (z - x_k) - \frac{\lambda_k^{-1} - L}{2}\|x_{k+1} - x_k\|^2 + \\ + \frac{\lambda_k^{-1}}{2}\left(\|x_k - z\|^2 - \|x_{k+1} - z\|^2\right).$$

Proof. By Algorithm 2 we have $x_{k+1} = \text{prox}_{\lambda_k g}(x_k - \lambda_k \nabla f_k)$, so it follows from the definition of the proximal mapping that there exists a subgradient $\gamma_{k+1} \in \partial g(x_{k+1})$ such that $\gamma_{k+1} + \lambda_k^{-1}(x_{k+1} - (x_k - \lambda_k \nabla f_k)) = 0$. By definition, for any subgradient and any $z \in \mathbb{R}^d$, $g(z) \geqslant g_{k+1} + \gamma_{k+1}^\top (z - x_{k+1})$. Now, for any $z \in \mathbb{R}^d$ and $k \geqslant 0$:

$$\begin{aligned} F_{k+1} = f_{k+1} + g_{k+1} &\leqslant f_k + \nabla f_k^\top (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 + g_{k+1} \leqslant \\ &\leqslant f_k + \nabla f_k^\top (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 + g(z) - \gamma_{k+1}^\top (z - x_{k+1}) = \\ &= f_k + \nabla f_k^\top (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 + g(z) + \\ &\qquad + \lambda_k^{-1}(x_{k+1} - x_k + \lambda_k \nabla f_k)^\top (z - x_{k+1}) = \\ &= f_k + \nabla f_k^\top (z - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 + g(z) + \lambda_k^{-1}(x_{k+1} - x_k)^\top (z - x_{k+1}). \end{aligned}$$

From $a^\top b = \frac{1}{2}\left(\|a + b\|^2 - \|a\|^2 - \|b^2\|\right)$, for $a = x_{k+1} - x_k$ and $b = z - x_{k+1}$ we obtain:

$$(x_{k+1} - x_k)^\top (z - x_{k+1}) = \frac{1}{2}\left(\|x_k - z\|^2 - \|x_{k+1} - z\|^2 - \|x_{k+1} - x_k\|^2\right),$$

so, then:

$$F_{k+1} \leqslant f_k + \nabla f_k^\top (z - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2 + g(z) + \\ + \frac{\lambda_k^{-1}}{2}\left(\|x_k - z\|^2 - \|x_{k+1} - z\|^2 - \|x_{k+1} - x_k\|^2\right) = \\ = f_k + g(z) + \nabla f_k^\top (z - x_k) - \frac{\lambda_k^{-1} - L}{2}\|x_{k+1} - x_k\|^2 + \\ + \frac{\lambda_k^{-1}}{2}\left(\|x_k - z\|^2 - \|x_{k+1} - z\|^2\right). \qquad \square$$

We prove the following lemma before proceeding with the analysis of the convergence of the iterates generated by Algorithm 2.

Lemma 1. *Let Assumption 2.1 hold and let $\{\lambda_k\}$ be the sequence of step sizes generated by Algorithm 2. If $\lambda_{\max} < \frac{2}{L}$, then there exists $\varepsilon > 0$ such that for all $k \geqslant 0$, $\lambda_k \leqslant \frac{2}{L} - \varepsilon$.*

Proof. If $\lambda_{\max} < \frac{2}{L}$, then for $\varepsilon = \frac{1}{2}\left(\frac{2}{L} - \lambda_{\max}\right) > 0$ and any $k \geqslant 0$:

$$\lambda_k \leqslant \lambda_{\max} < \lambda_{\max} + \varepsilon = \frac{2}{L} - \varepsilon.$$

Note that for $\varepsilon > 0$ constructed as above we have $\varepsilon < \frac{1}{L} < \frac{2}{L}$. $\square$

We are now ready to prove a convergence proposition about sequence of iterates $\{x_k\}$ generated by Algorithm 2.

Theorem 4. *Let Assumption 2.1 hold and let $\{x_k\}$ be the sequence of iterates generated by Algorithm 2 with $\lambda_{\max} < \frac{2}{L}$. Then:*

(i) $\sum_{k=0}^{\infty}\|x_{k+1} - x_k\|^2 < \infty$.

(ii) $\|x_{k+1} - x_k\| \to 0 \quad as \quad k \to \infty$.

Proof. For $z = x_k$, by Theorem 3 we get:

$$F_{k+1} \leqslant f_k + g_k - \frac{\lambda_k^{-1} - L}{2}\|x_{k+1} - x_k\|^2 - \frac{\lambda_k^{-1}}{2}\|x_{k+1} - x_k\|^2 = \\ = F_k - \left(\lambda_k^{-1} - \frac{L}{2}\right)\|x_{k+1} - x_k\|^2. \tag{3.1}$$

By Algorithm 2 and Lemma 1, $\lambda_{\min} \leqslant \lambda_k \leqslant \frac{2}{L} - \varepsilon$ for some $0 < \varepsilon < \frac{2}{L}$, so we have:

$$\frac{L}{2 - \varepsilon L} \leqslant \lambda_k^{-1} \leqslant \frac{1}{\lambda_{\min}},$$

from where:

$$\lambda_k^{-1} - \frac{L}{2} \geqslant \frac{L}{2 - \varepsilon L} - \frac{L}{2} = \underbrace{\frac{\varepsilon L^2}{2(2 - \varepsilon L)}}_{=\delta} > 0,$$

since $\varepsilon < \frac{2}{L}$ implies $2 - \varepsilon L > 0$. Let us denote $\delta = \frac{\varepsilon L^2}{2(2-\varepsilon L)} > 0$. Now, from (3.1) we get that for any $k \geqslant 0$:

$$F_{k+1} \leqslant F_k - \delta \|x_{k+1} - x_k\|^2. \tag{3.2}$$

As $\delta > 0$, it follows that $\{F_k\}$ is a decreasing sequence:

$$F_0 \geqslant F_1 \geqslant \ldots \geqslant F^*,$$

and by recursively applying the inequality (3.2) we obtain:

$$F_{k+1} \leqslant F_0 - \delta \sum_{i=0}^{k} \|x_{i+1} - x_i\|^2$$

$$F_{k+1} - F^* \leqslant (F_0 - F^*) - \delta \sum_{i=0}^{k} \|x_{i+1} - x_i\|^2,$$

from where it follows that:

$$\sum_{i=0}^{k} \|x_{i+1} - x_i\|^2 \leqslant \frac{1}{\delta}((F_0 - F^*) - (F_{k+1} - F^*)) \leqslant \frac{F_0 - F^*}{\delta} < \infty.$$

The last inequality holds for any $k \geqslant 0$, so we conclude that the series is convergent, i.e. that $\sum_{i=0}^{\infty} \|x_{i+1} - x_i\|^2 < \infty$. This proves (i). The claim in (ii) follows directly from the convergence of the series. □

Let us note that expecting the upper bound of the step sizes to not surpass a value of $2/L$ is reasonable, and stems from the standard convergence analysis of the gradient method – the same one which establishes the constant step size $\lambda = 1/L$ as *optimal* in the case when the objective function is L-smooth in a sense that it results in largest per-iteration function value decrease; see [8, 13] for details.

Also, in the formulation and proofs of Theorems 3 and 4 it is not assumed that f is convex. This additional assumption leads to the following result.

Theorem 5. *Let Assumption 2.1 hold and let f be convex. Let $\{x_k\}$ be the sequence of iterates generated by Algorithm 2 with $\lambda_{\max} < \frac{2}{L}$. Then for any $k \geqslant 0$:*

$$F_{k+1} - F^* \leqslant \frac{\|x_0 - x^*\|^2 + \max\{0, L\lambda_{\max} - 1\}\overline{S}}{2\lambda_{\min}(k+1)},$$

where:

$$\overline{S} = \lim_{k \to \infty} \sum_{i=0}^{k} \|x_{i+1} - x_i\|^2 < \infty.$$

Proof. Since f satisfies the appropriate conditions, Theorems 3 and 4 hold. From the convexity of f we have:

$$f(x_k) + \nabla f_k^\top (z - x_k) \leqslant f(z),$$

where for $z = x^*$ we obtain:

$$f(x_k) + \nabla f_k^\top (x^* - x_k) \leqslant f(x^*)$$

From the result of Theorem 3, for $z = x^*$ we get:

$$F_{k+1} \leqslant f_k + g^* + \nabla f_k^\top (x^* - x_k) - \frac{\lambda_k^{-1} - L}{2}\|x_{k+1} - x_k\|^2 + \\ + \frac{\lambda_k^{-1}}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right) \leqslant \\ \leqslant f^* + g^* - \frac{\lambda_k^{-1} - L}{2}\|x_{k+1} - x_k\|^2 + \frac{\lambda_k^{-1}}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right).$$

Multiplying by $2\lambda_k$ now we get:

$$2\lambda_k F_{k+1} \leqslant 2\lambda_k F^* - (1 - L\lambda_k)\|x_{k+1} - x_k\|^2 + \|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2,$$

or, equivalently:

$$\|x_{k+1} - x^*\|^2 \leqslant \|x_k - x^*\|^2 + (L\lambda_k - 1)\|x_{k+1} - x_k\|^2 - 2\lambda_k(F_{k+1} - F^*), \tag{3.3}$$

so, by recursively applying (3.3):

$$\|x_{k+1} - x^*\|^2 \leqslant \|x_0 - x^*\|^2 + \sum_{i=0}^{k}(L\lambda_i - 1)\|x_{i+1} - x_i\|^2 - 2\sum_{i=0}^{k}\lambda_i(F_{i+1} - F^*) \tag{3.4}$$

For $\lambda_i \in [\lambda_{\min}, \frac{1}{L}]$ it follows that $L\lambda_i - 1 \leqslant 0$, while for $\lambda_i \in [\frac{1}{L}, \lambda_{\max}]$ it follows that $L\lambda_i - 1 \leqslant L\lambda_{\max} - 1$, which in turn implies that $L\lambda_i - 1 \leqslant \max\{0, L\lambda_{\max} - 1\}$, for all i. By Theorem 3, the series $\sum_{i=0}^{\infty}\|x_{i+1} - x_i\|^2$ is convergent, so let its sum be denoted by $\overline{S} < \infty$. Then, since $\{F_k\}$ is a decreasing sequence, i.e. $F_{i+1} - F^* \geqslant F_{k+1} - F^*$, for all $0 \leqslant i \leqslant k$ in (3.4) we get:

$$0 \leqslant \|x_{k+1} - x^*\|^2 \leqslant \|x_0 - x^*\|^2 + \max\{0, L\lambda_{\max} - 1\}\overline{S} - 2(F_{k+1} - F^*)\sum_{i=0}^{k}\lambda_i \\ \iff F_{k+1} - F^* \leqslant \frac{\|x_0 - x^*\|^2 + \max\{0, L\lambda_{\max} - 1\}\overline{S}}{2\sum_{i=0}^{k}\lambda_i}.$$

Now, $\lambda_k \geqslant \lambda_{\min}$ implies $\sum_{i=0}^{k}\lambda_i \geqslant \lambda_{\min}(k+1)$, from where we finally conclude:

$$F_{k+1} - F^* \leqslant \frac{\|x_0 - x^*\|^2 + \max\{0, L\lambda_{\max} - 1\}\overline{S}}{2\lambda_{\min}(k+1)}. \qquad \square$$

Next we analyze the convergence properties under the additional assumption that f is strongly convex with coefficient μ. Before proving the main result, we formulate some auxiliary propositions about strongly convex functions.

Proposition 3.1. [8] *Let $f : \mathbb{R}^d \to \mathbb{R}$ be μ-strongly convex function. Then for any $x, y \in \mathbb{R}^d$:*

$$\mu\|x - y\| \leqslant \|\nabla f(x) - \nabla f(y)\|.$$

Proposition 3.2. [8] *Let $f : \mathbb{R}^d \to \mathbb{R}$ be L-smooth and μ-strongly convex function. Then for any $x, y \in \mathbb{R}^d$:*

$$(\nabla f(x) - \nabla f(y))^\top (x - y) \geqslant \frac{\mu L}{\mu + L}\|x - y\|^2 + \frac{1}{\mu + L}\|\nabla f(x) - \nabla f(y)\|^2.$$

Theorem 6. *Let Assumption 2.1 hold and let f be μ-strongly convex. Let $\{x_k\}$ be the sequence of iterates generated by Algorithm 2 with $\lambda_{\max} < \frac{2}{L}$. Then for any $k \geqslant 0$:*

$$\|x_{k+1} - x^*\|^2 \leqslant \rho^{k+1}\|x_0 - x^*\|^2,$$

for some $\rho \in (0,1)$.

Proof. From $x^* = \text{prox}_{\lambda_k g}(x^*) = \text{prox}_{\lambda_k g}(x^* - \lambda_k \nabla f^*)$, proposition 2.1, and proposition 3.2 we have:

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &= \|\text{prox}_{\lambda_k g}(x_k - \lambda_k \nabla f_k) - \text{prox}_{\lambda_k g}(x^* - \lambda_k \nabla f^*)\|^2 \leqslant \\
&\leqslant \|x_k - \lambda_k \nabla f_k - x^* + \lambda_k \nabla f^*\|^2 = \\
&= \|x_k - x^*\|^2 - 2\lambda_k (\nabla f_k - \nabla f^*)^\top (x_k - x^*) + \lambda_k^2 \|\nabla f_k - \nabla f^*\|^2 \leqslant \\
&\leqslant \|x_k - x^*\| - \frac{2\lambda_k}{\mu + L}\left(\|\nabla f_k - \nabla f^*\|^2 + \mu L \|x_k - x^*\|^2\right) + \\
&\qquad + \lambda_k^2 \|\nabla f_k - \nabla f^*\|^2 = \\
&= \left(1 - \frac{2\lambda_k \mu L}{\mu + L}\right)\|x_k - x^*\|^2 - \lambda_k \left(\frac{2}{\mu + L} - \lambda_k\right)\|\nabla f_k - \nabla f^*\|^2.
\end{aligned} \tag{3.5}$$

Let $\tilde{\mu} = \min\{\mu, L, \frac{1}{2\lambda_{\min}}\} \leqslant \mu$. If f is μ-strongly convex, then f is also $\tilde{\mu}$-strongly convex and the inequality (3.5) holds with $\mu = \tilde{\mu}$. We now bound the last term on the right-hand side of (3.5). We discuss two cases.

Case 1: $\lambda_k \in [\lambda_{\min}, \frac{2}{\tilde{\mu}+L}]$. Then, $\frac{2}{\tilde{\mu}+L} - \lambda_k \geqslant 0$, and since $\|\nabla f_k - \nabla f^*\| \geqslant \tilde{\mu}\|x_k - x^*\|$, by Proposition 3.1, applying this in (3.5) with $\mu = \tilde{\mu}$ we obtain:

$$\begin{aligned}
\|x_{k+1} - x^*\|^2 &\leqslant \left(1 - \frac{2\lambda_k \tilde{\mu} L}{\tilde{\mu} + L}\right)\|x_k - x^*\|^2 - \lambda_k \tilde{\mu}^2 \left(\frac{2}{\tilde{\mu} + L} - \lambda_k\right)\|x_k - x^*\|^2 = \\
&= \left(1 - \frac{2\lambda_k \tilde{\mu} L}{\tilde{\mu} + L} - \frac{2\lambda_k \tilde{\mu}^2}{\tilde{\mu} + L} + \lambda_k^2 \tilde{\mu}^2\right)\|x_k - x^*\|^2 = \\
&= \left(1 - \frac{2\lambda_k \tilde{\mu}(\tilde{\mu} + L)}{\tilde{\mu} + L} + \lambda_k^2 \tilde{\mu}^2\right)\|x_k - x^*\|^2 = \\
&= (1 - \lambda_k \tilde{\mu})^2 \|x_k - x^*\|^2.
\end{aligned}$$

As $L \geqslant \min\{\mu, L, \frac{1}{2\lambda_{\min}}\} = \tilde{\mu}$, we get $\lambda_k \leqslant \frac{2}{\tilde{\mu}+L} \leqslant \frac{2}{\tilde{\mu}+\tilde{\mu}} = \frac{1}{\tilde{\mu}}$, from where $\lambda_k \tilde{\mu} \leqslant 1$ i.e. $1 - \lambda_k \tilde{\mu} \geqslant 0$. Obviously $1 - \lambda_k \tilde{\mu} < 1$. This implies $1 - \lambda_k \tilde{\mu} \in [0, 1)$, and hence $(1 - \lambda_k \tilde{\mu})^2 \in [0, 1)$.

Similarly, $\frac{1}{\lambda_{\min}} > \frac{1}{2\lambda_{\min}} \geqslant \min\{\mu, L, \frac{1}{2\lambda_{\min}}\} = \tilde{\mu}$, so it follows that $1 - \lambda_{\min}\tilde{\mu} > 0$, and from $1 - \lambda_{\min}\tilde{\mu} < 1$, it follows that $1 - \lambda_{\min}\tilde{\mu} \in (0, 1)$, from where $(1 - \lambda_{\min}\tilde{\mu})^2 \in (0, 1)$.

By the discussion above and the fact that $\lambda_k \geqslant \lambda_{\min}$, we conclude that $(1 - \lambda_k \tilde{\mu})^2 \leqslant (1 - \lambda_{\min}\tilde{\mu})^2$ i.e. for all k such that $\lambda_k \in [\lambda_{\min}, \frac{2}{\tilde{\mu}+L}]$ we have:

$$\|x_{k+1} - x^*\|^2 \leqslant (1 - \lambda_{\min}\tilde{\mu})^2 \|x_k - x^*\|^2. \tag{3.6}$$

Case 2: $\lambda_k \in [\frac{2}{\tilde{\mu}+L}, \lambda_{\max}]$. Then, $\frac{2}{\tilde{\mu}+L} - \lambda_k \leqslant 0$, and since $\|\nabla f_k - \nabla f^*\| \leqslant L\|x_k - x^*\|$, applying this in (3.5) with $\mu = \tilde{\mu}$ we obtain:

$$\begin{aligned}\|x_{k+1} - x^*\|^2 &\leqslant \left(1 - \frac{2\lambda_k\tilde{\mu}L}{\tilde{\mu}+L}\right)\|x_k - x^*\|^2 - \lambda_k L^2\left(\frac{2}{\tilde{\mu}+L} - \lambda_k\right)\|x_k - x^*\|^2 = \\ &= \left(1 - \frac{2\lambda_k\tilde{\mu}L}{\tilde{\mu}+L} - \frac{2\lambda_k L^2}{\tilde{\mu}+L} + \lambda_k^2 L^2\right)\|x_k - x^*\|^2 = \\ &= \left(1 - \frac{2\lambda_k L(\tilde{\mu}+L)}{\tilde{\mu}+L} + \lambda_k^2 L^2\right)\|x_k - x^*\|^2 = \\ &= (1 - \lambda_k L)^2\|x_k - x^*\|^2.\end{aligned}$$

Now, by Lemma 1 there exists $\varepsilon = \frac{1}{2}\left(\frac{2}{L} - \lambda_{\max}\right) > 0$ such that $\lambda_k \leqslant \frac{2}{L} - \varepsilon$, which implies that $\lambda_k L - 1 \leqslant 1 - L\varepsilon = 1 - \frac{L}{2}\left(\frac{2}{L} - \lambda_{\max}\right) = \frac{L}{2}\cdot\lambda_{\max} < \frac{L}{2}\cdot\frac{2}{L} = 1$. On the other hand, $\lambda_k L - 1 \geqslant \frac{2L}{\tilde{\mu}+L} - 1 = \frac{L-\tilde{\mu}}{\tilde{\mu}+L} \geqslant 0$ since $L \geqslant \min\{\mu, L, \frac{1}{\lambda_{\min}}\} = \tilde{\mu}$. This means that $\lambda_k L - 1 \in [0, 1)$ and $(\lambda_k L - 1)^2 \in [0, 1)$. Also, $1 - L\varepsilon = \frac{\lambda_{\max} L}{2} \in (0, 1)$, so $(1 - L\varepsilon)^2 = (\frac{\lambda_{\max} L}{2})^2 \in (0, 1)$.

By the discussion above and the fact that $\lambda_k L - 1 \leqslant 1 - L\varepsilon$, we conclude that $(\lambda_k L - 1)^2 \leqslant (1 - L\varepsilon)^2$, i.e. for all k such that $\lambda_k \in [\frac{2}{\tilde{\mu}+L}, \lambda_{\max}]$ we have:

$$\|x_{k+1} - x^*\|^2 \leqslant (1 - L\varepsilon)^2\|x_k - x^*\|^2 = \left(\frac{\lambda_{\max} L}{2}\right)^2\|x_k - x^*\|^2. \tag{3.7}$$

Let now $\rho = \max\{(1 - \lambda_{\min}\tilde{\mu})^2, (\frac{\lambda_{\max} L}{2})^2\} \in (0, 1)$. Then from (3.6) and (3.7) we finally conclude that for all $k \geqslant 0$:

$$\|x_{k+1} - x^*\|^2 \leqslant \rho\|x_k - x^*\|^2 \leqslant \left(\prod_{i=0}^{k}\rho\right)\|x_0 - x^*\|^2 = \rho^{k+1}\|x_0 - x^*\|^2. \qquad \square$$

In Theorem 4 we proved that the $\{x_k\}$, the sequence of iterates generated by Algorithm 2, is convergent as a Cauchy sequence in the complete metric space $\mathbb{R}^d$; however, the proof does not give an insight into the nature of the limit point. In Theorems 5 and 6 we proved more essential convergence results which depend on some additional properties of the "nice" component f in the composite optimization problem (1.1) we are solving.

In this sense, if f is convex then, as per Theorem 5, the Algorithm 2 generates sequences of iterates $\{x_k\}$ and corresponding function values $\{F_k\}$ such that $F_k \to F^*$, and the convergence speed is *sublinear*, i.e. of order $\mathcal{O}(1/k)$. On the other hand, if f is μ-strongly convex then, as per Theorem 6, for the sequence of iterates it holds that $x_k \to x^*$, and the convergence speed is *linear*, i.e. of order $\mathcal{O}(\rho^k)$ for some $\rho \in (0, 1)$.

All in all, under the same standard assumptions, we showed that the convergence speed of the proximal gradient method with variable step size is, theoretically, of the same order as the convergence speed of its constant step counterpart. In practice, however, the variable step sizes result in faster convergence of the method, [9].

4. STEP SIZE SELECTION

In this section we illustrate the preceding convergence analysis on an existing proximal gradient method with variable step sizes, recently proposed in [9]. The main idea of the method in [9] is that the L-smoothness as a geometric property is not necessarily global and it is natural to explore ways of constructing approximations of the local smoothness coefficient in neighborhoods of the iterates. In the k-th iteration of the method, once x_{k+1} is calculated, it is assumed that f is locally smooth with coefficient L_k not necessarily equal to L, i.e. that $\|\nabla f(x_{k+1}) - \nabla f(x_k)\| \leqslant L_k \|x_{k+1} - x_k\|$. Since locally the optimal choice of the step size is still $\lambda_k = 1/L_k$, one can put:

$$\lambda_k = \frac{1}{L_k} \leqslant \frac{\|x_{k+1} - x_k\|}{\|\nabla f(x_{k+1}) - \nabla f(x_k)\|}. \tag{4.1}$$

Now, starting with an initial step size $\lambda_0 > 0$, the method in [9] iteratively generates step sizes ensuring that their magnitude remains smaller than the ratio of the norms on the right-hand side of (4.1), taking care not to generate too small step sizes which would slow down the process. To achieve this, two constants $0 < \mu_1 < \mu_0 < 1$ and a sequence $\{\eta_k\}$ such that $\eta_k > 0$ and $\sum \eta_k < \infty$ are fixed. To update λ_k to λ_{k+1} in the k-th iteration it is tested if:

$$\lambda_k > \frac{\mu_0 \|x_{k+1} - x_k\|}{\|\nabla f(x_{k+1}) - \nabla f(x_k)\|}.$$

If the test is true, then λ_k is either greater than the norm ratio, or it is smaller than, but too close to it, i.e its value is greater then $\mu_0 \cdot 100\%$ of the value of the norm ratio. In such case the step size should be decreased to an acceptable size, as per (4.1), so the next step size is selected as:

$$\lambda_{k+1} = \frac{\mu_1 \|x_{k+1} - x_k\|}{\|\nabla f(x_{k+1}) - \nabla f(x_k)\|},$$

i.e. it is set to $\mu_1 \cdot 100\%$ of the norm ratio, which ensures $\lambda_{k+1} < \lambda_k$. In case the test is false, the current step size λ_k might be too small, so a controlled increase of its magnitude using the sequence $\{\eta_k\}$ is performed by setting:

$$\lambda_{k+1} = \lambda_k + \min\{\lambda_k, 1\} \cdot \eta_k.$$

The concrete choices of the constants μ_0 and μ_1 allow for a control over how large departures from the norm ratio from (4.1) one is willing to tolerate in the step size generation process. The full description of the method is given in Algorithm 3.

The numerical results given in [9] demonstrate almost the same complexity of the proposed method compared with its constant step size counterpart; the paper, however, lacks any theoretical analyses and results. We now formulate theoretical results about the performance of this method based on the convergence analysis developed in the previous section.

First, we show that step sizes $\{\lambda_k\}$ generated by Algorithm 3 are bounded.

Lemma 2. *Let Assumption 2.1 hold and let $\{\lambda_k\}$ be the sequence of step sizes generated by the Algorithm 3. Then the following holds:*

Algorithm 3 Proximal gradient method with variable step sizes from [9]

1: **input:** $x_0 \in \mathbb{R}^d$, $\lambda_0 > 0$, $0 < \mu_1 < \mu_0 < 1$, $N \in \mathbb{N}$, $\{\eta_k\}$ s.t. $\sum \eta_k < \infty$
2: **set:** $k = 0$
3: **while** $k < N$ **do**
4: $\quad x_{k+1} = \operatorname{prox}_{\lambda_k}(x_k - \lambda_k \nabla f(x_k))$
5: $\quad$ **if** $\|\nabla f(x_{k+1}) - \nabla f(x_k)\| > \frac{\mu_0}{\lambda_k}\|x_{k+1} - x_k\|$ **then**
6: $\quad\quad \lambda_{k+1} = \dfrac{\mu_1 \|x_{k+1} - x_k\|}{\|\nabla f(x_{k+1}) - \nabla f(x_k)\|}$
7: $\quad$ **else**
8: $\quad\quad \lambda_{k+1} = \lambda_k + \min\{\lambda_k, 1\} \cdot \eta_k$
9: $\quad k \leftarrow k + 1$

(i) For all $k \geqslant 0$,

$$\lambda_k \geqslant \min\left\{\lambda_0, \frac{\mu_1}{L}\right\} = \lambda_{\min}.$$

(ii) For all $k \geqslant 0$,

$$\lambda_k \leqslant \lambda_0 + \eta^* = \lambda_{\max},$$

where $\sum_{k=0}^{\infty} \eta_k = \eta^* < \infty$.

Proof. From the L-smoothness of f we have $\|\nabla f(x_{k+1}) - \nabla f(x_k)\| \leqslant L\|x_{k+1} - x_k\|$ for any $k \geqslant 0$. If at the k-th iteration λ_{k+1} is selected by the update rule on line 6 in Algorithm 3, then the condition on line 5 is satisfied and we have:

$$\lambda_{k+1} > \frac{\mu_0 \|x_{k+1} - x_k\|}{\|\nabla f(x_{k+1}) - \nabla f(x_k)\|} > \frac{\mu_1 \|x_{k+1} - x_k\|}{\|\nabla f(x_{k+1}) - \nabla f(x_k)\|} \geqslant \frac{\mu_1}{L}.$$

If λ_{k+1} is selected by the update rule on line 8 in Algorithm 3, then:

$$\lambda_{k+1} = \lambda_k + \min\{\lambda_k, 1\} \cdot \eta_k > \lambda_k.$$

This means that for $k = 0$, we have $\lambda_1 > \frac{\mu_1}{L}$ or $\lambda_1 > \lambda_0$, which implies $\lambda_1 \geqslant \min\{\frac{\mu_1}{L}, \lambda_0\}$. By induction it can be shown that $\lambda_k \geqslant \min\{\lambda_0, \frac{\mu_1}{L}\} = \lambda_{\min}$ for all $k \geqslant 0$. This proves (i).

To prove (ii), we reason in a similar way. If at the k-th iteration λ_{k+1} is selected by the update rule on line 6 in Algorithm 3, then the condition on line 5 is satisfied and we have:

$$\lambda_{k+1} = \frac{\mu_1 \|x_{k+1} - x_k\|}{\|\nabla f(x_{k+1}) - \nabla f(x_k)\|} < \frac{\mu_1 \lambda_k}{\mu_0} < \lambda_k.$$

If λ_{k+1} is selected by the update rule on line 8 in Algorithm 3, then:

$$\lambda_{k+1} = \lambda_k + \min\{\lambda_k, 1\} \cdot \eta_k \leqslant \lambda_k + \eta_k.$$

This means that for $k = 0$, we have $\lambda_1 < \lambda_0$ or $\lambda_1 \leqslant \lambda_0 + \eta_0$, which implies $\lambda_1 \leqslant \lambda_0 + \eta_0$. By induction it can be shown that $\lambda_k \leqslant \lambda_0 + \sum_{i=0}^{k} \eta_i \leqslant \lambda_0 + \eta^* = \lambda_{\max}$ for all $k \geqslant 0$, which proves (ii). □

Now, as a consequence of Lemma 2, Theorem 5 and Theorem 6, we can prove the following convergence theorem for the method given in [9].

Theorem 7. *Let Assumption 2.1 hold and let $\{x_k\}$ be the sequence of iterates generated by Algorithm 3 with $\lambda_0 + \eta^* < \frac{2}{L}$, where $\sum_{k=0}^{\infty} \eta_k = \eta^* < \infty$. Then:*

(i) $\sum_{k=0}^{\infty} \|x_{k+1} - x_k\|^2 < \infty$.

(ii) $\|x_{k+1} - x_k\| \to 0 \quad as \quad k \to \infty$.

(iii) *If f is convex, then for any $k \geqslant 0$:*

$$F_{k+1} - F^* \leqslant \frac{\|x_0 - x^*\|^2 + \max\{0, L\lambda_{\max} - 1\}\overline{S}}{2\lambda_{\min}(k+1)},$$

where $\lambda_{\min} = \min\left\{\lambda_0, \frac{\mu_1}{L}\right\}$, $\lambda_{\max} = \lambda_0 + \eta^*$, and $\overline{S} = \lim_{k\to\infty} \sum_{i=0}^{k} \|x_{i+1} - x_i\|^2 < \infty$.

(iv) *If f is μ-strongly convex, then for any $k \geqslant 0$:*

$$\|x_{k+1} - x^*\|^2 \leqslant \rho^{k+1} \|x_0 - x^*\|^2,$$

for some $\rho \in (0, 1)$.

As shown here, any proximal gradient method with variable step sizes that are bounded can be treated in a similar way, and same convergence results can be established.

5. Conclusions

Proximal gradient methods are designed for composite optimization problems, in particular for cases when components of the composite objective are potentially not differentiable. They can be a very efficient optimization tool when the proximal mapping induced in the method has a closed-form solution. Theoretical analyses of the convergence of the proximal gradient method with constant step size have shown sublinear convergence in the case when the "nice" component of the composite objective is convex, and linear convergence in the case when this component is strongly convex.

In this paper we showed that under standard assumptions the same kind of convergence result can be established for the proximal gradient method with variable step sizes in a general setting of bounded step sizes. We implemented this analysis to a recently proposed proximal gradient method with variable step sizes and established its convergence. In fact, any proximal gradient method with variable step sizes that are bounded, as outlined in our analysis, can be treated in a similar way.

It will be challenging for future research to weaken some of the assumptions like convexity or Lipschitz continuity and analyze convergence in such setting.

References

[1] A. Beck, *First-Order Methods in Optimization*, MOS–SIAM Series on Optimization, SIAM, Philadelphia, 2017.

[2] A. Beck, M. Teboulle, *A fast iterative shrinkage-thresholding algorithm for linear inverse problems*, SIAM Journal on Imaging Sciences, vol. 2, no. 1, pp. 183–202, 2009.
[3] P. L. Combettes, J.-C. Pesquet, *Proximal splitting methods in signal processing*, in: H. H. Bauschke, R. S. Burachik, P. L. Combettes, V. Elser, D. R. Luke, H. Wolkowicz (eds.), *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, Springer, New York, pp. 185–212, 2011.
[4] C. Kanzow, P. Mehlitz, *Convergence properties of monotone and nonmonotone proximal gradient methods revisited*, Journal of Optimization Theory and Applications, vol. 195, no. 2, pp. 624–646, 2022.
[5] H. W. Liu and T. Wang, A Nonmonotone Accelerated Proximal Gradient Method with Variable Stepsize Strategy for Nonsmooth and Nonconvex Minimization Problems. *Journal of Global Optimization*, vol. 89, no.4, pp. 863–897, 2024.
[6] A. De Marchi, *Proximal gradient methods beyond monotony*, Journal of Nonsmooth Analysis and Optimization, vol. 4, 2023.
[7] Y. Nesterov, *Gradient methods for minimizing composite functions*, Mathematical Programming, vol. 140, no. 1, pp. 125–161, 2013.
[8] Y. Nesterov, *Lectures on Convex Optimization*, 2nd edition, Springer Optimization and Its Applications, vol. 137, Springer, Cham, 2018.
[9] F. Nikolovski, I. Stojkovska, K. Hadzi-Velkova Saneva, Z. Hadzi-Velkov, *Gradient descent methods for regularized optimization*, Contributions, Section of Natural, Mathematical and Biotechnical Sciences, MASA, submitted, 2024. Available at: `https://arxiv.org/pdf/2412.20115v1`.
[10] N. Parikh, S. Boyd, *Proximal algorithms*, Foundations and Trends in Optimization, vol. 1, no. 3. pp. 127–239, 2014.
[11] Y. Park, S. Dhar, S. Boyd, S. Shah, *Variable metric proximal gradient method with diagonal Barzilai–Borwein stepsize*, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3597–3601, 2020.
[12] B. Recht, *Projected gradient methods*, lecture notes, 2012. Available at: `https://pages.cs.wisc.edu/~brecht/cs726docs/ProjectedGradientMethods.pdf`
[13] A. B. Taylor, J. M. Hendrickx, F. Glineur, *Exact worst-case convergence rates of the proximal gradient method for composite convex minimization*, Journal of Optimization Theory and Applications, vol. 178, no. 2, pp. 455–476, 2018.
[14] N. D. Vanlı, M. Gürbüzbalaban, A. Özdaglar, *A simple proof for the iteration complexity of the proximal gradient algorithm*, in *NIPS Workshop on Optimization for Machine Learning*, 2020.

FILIP NIKOLOVSKI
SS. CYRIL AND METHODIUS UNIVERSITY IN SKOPJE,
FACULTY OF MECHANICAL ENGINEERING,
RUGJER BOSHKOVIKJ 18, 1000 SKOPJE, N. MACEDONIA
Email address: `filip.nikolovski@mf.edu.mk`

IRENA STOJKOVSKA
SS. CYRIL AND METHODIUS UNIVERSITY IN SKOPJE,
FACULTY OF NATURAL SCIENCES AND MATHEMATICS,
ARHIMEDOVA 4, 1000 SKOPJE, N. MACEDONIA
Email address: `irenatra@pmf.ukim.mk`

Received 26.09.2025
Revised 26.10.2025
Accepted 21.11.2025