MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Technical Report No. 2 27 July 2011

NONCONVEX PROXIMAL SPLITTING: BATCH AND INCREMENTAL ALGORITHMS

Suvrit Sra

Abstract. Within the unmanageably large class of nonconvex optimization, we consider the rich subclass of nonsmooth problems having *composite* objectives (this includes the extensively studied convex, composite objective problems as a special case). For this subclass, we introduce a powerful, new framework that permits asymptotically *non-vanishing* perturbations. In particular, we develop perturbation-based batch and incremental (online like) nonconvex proximal splitting algorithms. To our knowledge, this is the first time that such perturbation-based nonconvex splitting algorithms are being proposed and analyzed. While the main contribution of the paper is the theoretical framework, we complement our results by presenting some empirical results on matrix factorization.

Note. This technical report includes verbatim the authors original writeup that was submitted to Neural Information Processing Systems, 2011 on June 3, 2011.

Nonconvex proximal splitting: batch and incremental algorithms

Anonymous Author(s)
Affiliation
Address
email

Abstract

Within the unmanageably large class of nonconvex optimization, we consider the rich subclass of nonsmooth problems having *composite* objectives (this includes the extensively studied convex, composite objective problems as a special case). For this subclass, we introduce a powerful, new framework that permits asymptotically *non-vanishing* perturbations. In particular, we develop perturbation-based batch and incremental (online like) nonconvex proximal splitting algorithms. To our knowledge, this is the first time that such perturbation-based nonconvex splitting algorithms are being proposed and analyzed. While the main contribution of the paper is the theoretical framework, we complement our results by presenting some empirical results on matrix factorization.

1 Introduction

Within the unmanageably vast class of nonconvex optimization, we consider the rich subclass of problems that have *nonconvex composite objectives*. Specifically, we study problems of the form

$$\text{minimize} \quad F(x) + \psi(x), \quad \text{s.t. } x \in \mathcal{X}, \tag{1}$$

where $\mathcal{X} \subset \mathbb{R}^n$ is a compact convex set, $F : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function, and $\psi : \mathbb{R}^n \to \mathbb{R}$ is a lower semi-continuous (lsc) convex function. We make the common assumption that $F \in C_L^1(\mathcal{X})$, i.e., the gradient ∇F is (locally) *Lipschitz continuous* on $\mathcal{X}$ with constant L,

$$\|\nabla F(x) - \nabla F(y)\| \leq L\|x - y\| \qquad \text{for all} \quad x, y \in \mathcal{X}. \tag{2}$$

Problem (1) is a natural but far-reaching generalization of *composite objective* convex problems that enjoy tremendous importance in machine learning; see [1–4], for example. Although, convex formulations are extremely useful, often for many difficult problems a nonconvex formulation is more natural. Familiar examples include, matrix factorization [5, 6], blind deconvolution [7], dictionary learning [8, 5], and neural networks [9, 10].

The main contribution of our paper is a new framework NOCOPS (an acronym for nonconvex proximal splitting). This framework solves (1) while allowing *computational errors*, a capability that proves key to deriving scalable batch and incremental (online like). A realistic feature of our framework is that it does *not* require the computational errors to vanish in the limit or its stepsizes to shrink to zero; such choices that are often assumed in standard analysis of (convex) incremental-gradient like methods [11] or even in stochastic gradient methods [12].

NOCOPS builds on the remarkable work of Solodov [13], but it is strictly *more general* than [13] (which solves (1) only for $\psi \equiv 0$). Like Solodov [13]'s framework, NOCOPS also allows nonvanishing errors, which is practical, since often one has limited or no true control over computational errors (e.g., fixed noise level in a simulation). To our knowledge, ours is the first work on nonconvex proximal splitting that has both batch and incremental incarnations, even if we disregard the ability to handle nonvanishing errors.

Related Work. Among batch nonconvex splitting methods, an early work is [14]. More recently, in his pioneering paper on composite minimization, Nesterov [15] solved (1) via a splitting-like algorithm. Fukushima and Mine [14] ensured convergence by forcing monotonic descent (using line-search); Nesterov [15] proved convergence (for the nonconvex case) by also ensuring monotonic descent. Even more recently, Attouch et al. [16] introduced a powerful method based on Kurdyka-Łojasiewicz theory, though convergence again hinged on descent. This insistence on monotonic descent makes these methods unsuitable to obtaining incremental, stochastic, or online variants.

But there are some incremental and stochastic methods that do apply to (1), namely the generalized gradient-type algorithms of [17] and stochastic generalized gradient methods of [18, 19]. Both approaches are analogous to subgradient methods from convex optimization, and face similar difficulties. For example, as is well recognized (see [15, 1], e.g.), subgradient-style methods fail to exploit composite objectives. Moreover, they exhibit the effect of the regularizer only in the limit; for example, if $\psi(x) = \|x\|_1$, then the sparse solutions are obtained only in the limit, and intermediate iterates may be dense.

For convex problems, a powerful alternative to subgradient methods is offered by *proximal splitting* (see [20] for a survey). These methods split (1) into smooth and nonsmooth parts. The smooth part is handled as in gradient-projection while the nonsmooth part is handled via a proximity operator. Owing to their ability to effectively tackle the nonsmooth part, proximal methods become valuable in machine learning and related areas; see [20, 4, 2, 1] and the references therein.

2 The NOCOPS Framework

We begin by defining the function $g : \mathbb{R}^n \to \mathbb{R}$ to be the sum $g(x) := \psi(x) + \delta(x|\mathcal{X})$, where $\delta(x|\mathcal{X})$ is the *indicator function* for the set $\mathcal{X}$. With this notation, the main problem of this paper is

$$\text{minimize}_{x\in\mathbb{R}^n} \quad \phi(x) := F(x) + g(x). \tag{3}$$

Next, we recall a definition central to our analysis.

Definition 1 (Proximity operator)**.** Let $g : \mathbb{R}^n \to \mathbb{R}$ be lsc and convex. The *proximity operator* for g, indexed by scalar $\eta > 0$, is the nonlinear map [see e.g., 21; Def. 1.22]:

$$P^g_\eta \;:\quad y \mapsto \underset{x\in\mathbb{R}^n}{\operatorname{argmin}} \Big(g(x) + \frac{1}{2\eta}\|x - y\|^2\Big). \tag{4}$$

Proximity operators are key to forward-backward splitting [22], which, for convex $F \in C^1_L(\mathcal{X})$ and appropriate stepsizes η_k, optimizes (1) by essentially iterating

$$x^{k+1} = P^g_{\eta_k}(x^k - \eta_k \nabla F(x^k)), \quad k = 0, 1, \ldots. \tag{5}$$

Our new framework NOCOPS introduces two powerful generalizations to (5). First, it allows F to be nonconvex, and second, it allows perturbations. Formally, NOCOPS performs the iteration

$$x^{k+1} = P^g_{\eta_k}(x^k - \eta_k \nabla F(x^k) + \eta_k \vartheta(x^k)), \quad k = 0, 1, \ldots, \tag{6}$$

where the stepsizes η_k satisfy the standard bounds and conditions

$$c \le \liminf_k \eta_k, \quad \limsup_k \eta_k \le \min\{1, 2/L - c\}, \quad 0 < c < 1/L. \tag{7}$$

The *perturbation* term $\eta_k \vartheta(x^k)$ in (6) represents the computational errors, which occur for example when only an approximation to the full gradient $\nabla F(x)$ is available.

To make NOCOPS well-defined, following [13], we also impose a mild restriction on the perturbations. Specifically, we assume that for all η smaller than a fixed value $\bar{\eta}$, it holds that

$$\eta\|\vartheta(x)\| \le \bar{\epsilon}, \quad \text{for some fixed } \bar{\epsilon} \ge 0, \quad \text{and } \forall x \in \mathcal{X}. \tag{8}$$

Condition (8) is weaker than the typical vanishing error requirement $\eta\|\vartheta(x^k)\| \to 0$ imposed by most analyses. Since nonvanishing errors are allowed, exact stationary points cannot not always be obtained, but appropriate *inexact stationary points* can be found. To formalize this, recall that a point $x^* \in \mathbb{R}^n$ is stationary for (3), if and only if it satisfies the inclusion

$$0 \in \partial_C \phi(x^*) := \nabla F(x^*) + \partial g(x^*), \tag{9}$$

where $\partial_C \phi$ denotes the Clarke (generalized) subdifferential [23]. Inclusion (9) may be equivalently recast as the fixed-point condition

$$x^* = P^g_\eta(x^* - \eta\nabla F(x^*)), \quad \text{for } \eta > 0. \tag{10}$$

We use (10) to characterize approximate stationary points. Define thus the *proximal residual*

$$\rho(x) := x - P^g_1(x - \nabla F(x)), \tag{11}$$

so that for stationary x^*, the residual norm $\|\rho(x^*)\| = 0$. At point x, let the level of perturbation be given by $\epsilon(x) \geq 0$. We define a point $\bar{x}$ to be ϵ-*stationary* if the residual norm is bounded satisfies

$$\|\rho(\bar{x})\| \leq \epsilon(\bar{x}). \tag{12}$$

To control overall level of perturbation in the system, we require $\epsilon(x) \geq \eta\|\vartheta(x)\|$. Thus, intuitively, by letting η become small enough, we can obtain a stationary point of any desired accuracy.

2.1 Convergence analysis

Our analysis builds on the pioneering works of Nesterov [15] and Solodov [13]. But as mentioned, our problem and analysis are *strictly* more general. Specifically, in contrast to [15], we permit perturbations and *do not* rely on strict descent, and unlike [13], we consider nonsmooth objective functions. Our analysis yields, to our knowledge, the first nonconvex proximal splitting algorithm with nonvanishing noise, and also the first nonconvex incremental proximal splitting algorithm, regardless of vanishing or nonvanishing nature of the noise.

We begin by recalling two simple facts without proof; the first is a classical descent lemma.

Lemma 1 (Descent lemma [see e.g., 24; Lemma 2.1.3])**.** *Let $F \in C^1_L(\mathcal{X})$. Then,*

$$|F(x) - F(y) - \langle \nabla F(y),\, x - y\rangle| \leq \tfrac{L}{2}\|x - y\|^2, \quad \forall\, x, y \in \mathcal{X}. \tag{13}$$

Lemma 2 (Nonexpansivity [see e.g., 22; Lemma 2.4])**.** *The operator P^g_η is* nonexpansive*, that is,*

$$\|P^g_\eta x - P^g_\eta y\| \leq \|x - y\|, \quad \forall\, x, y \in \mathbb{R}^n. \tag{14}$$

Next, we prove a useful monotonicity result about proximity operators.

Lemma 3 (Monotonicity)**.** *Let $P_\eta \equiv P^g_\eta$; let $y, z \in \mathbb{R}^n$, and $\eta > 0$, and define*

$$p(\eta) := \quad \eta^{-1}\|P_\eta(y - \eta z) - y\|, \tag{15}$$

$$q(\eta) := \quad \|P_\eta(y - \eta z) - y\|. \tag{16}$$

Then, $p(\eta)$ is a decreasing function of η, while $q(\eta)$ is an increasing function of η.

Proof. Both (15) and (16) follow as corollaries to well-known properties of Moreau-envelopes [21, 22] (also see [15]). To set our notation, we provide a proof in the language of proximity operators. Consider thus the "deflected" proximity objective

$$m_g(x, \eta; y, z) := \langle z,\, x - y\rangle + \tfrac{1}{2}\eta^{-1}\|x - y\|^2 + g(x), \tag{17}$$

to which we associate the (deflected) *Moreau-envelope*

$$E_g(\eta) := \inf_{x \in \mathcal{X}}\ m_g(x, \eta; y, z). \tag{18}$$

The infinum in (18) is attained at the unique point $P^g_\eta(y - \eta z)$. Thus, $E_g(\eta)$ is differentiable, and

$$E'_g(\eta) = -\tfrac{1}{2}\eta^{-2}\|P^g_\eta(y - \eta z) - y\|^2 = -\tfrac{1}{2}p(\eta)^2.$$

Since $E_g(\eta)$ is convex ([21; Theorem 2.26]), E'_g is increasing; equivalently $p(\eta)$ is decreasing. Similarly, note that $\hat{E}_g(\gamma) := E_g(1/\gamma)$ is concave in γ, as it is a pointwise (indexed by x) infimum of functions linear in γ [25; §3.2.3]; thus differentiate $\hat{E}_g$ and conclude (16). □

Remark 1. The monotonicity results (15) and (16) complement similar monotonicity results for projection operators derived in [26; Lemma 1].

Now we analyze the difference $\phi(x^k) - \phi(x^{k+1})$; specifically, we derive an inequality

$$\phi(x^k) - \phi(x^{k+1}) \geq h(x^k), \tag{19}$$

where the potential function $h(x)$ depends on $\|\rho(x)\|$ and $\epsilon(x)$. Note that the potential $h(x)$ is allowed to take on negative values because we do not insist on monotonic descent. To simplify notation, let $u \equiv x^{k+1}$, $x \equiv x^k$, and $\eta \equiv \eta_k$, so that update (6) becomes

$$u = P_\eta(x - \eta \nabla F(x) + \eta \vartheta(x)). \tag{20}$$

With this notation, we now have the following descent-like theorem.

Theorem 1. *Let $x \in \mathcal{X}$, u, η be as in* (20)*, and $\epsilon(x) \geq \eta\|\vartheta(x)\|$. Then we have the bound*

$$\phi(x) - \phi(u) \quad \geq \quad \tfrac{2-L\eta}{2\eta}\|u-x\|^2 - \tfrac{1}{\eta}\epsilon(x)\|u-x\|. \tag{21}$$

Proof. Consider the directional derivative $\mathrm{d}m_g$ (of m_g, with respect to x, and in the direction w), which satisfies at $x = u$ the optimality condition

$$\mathrm{d}m_g(u, \eta; y, z)(w) = \langle z + \eta^{-1}(u-y) + s,\, w\rangle \geq 0, \quad s \in \partial g(u). \tag{22}$$

In (22), set $z = \nabla F(x) - \vartheta(x)$, $y = x$, and $w = x - u$; then, rearrange to obtain

$$\langle \nabla F(x) - \vartheta(x),\, u - x\rangle \quad \leq \quad \langle \eta^{-1}(u-x) + s,\, x - u\rangle. \tag{23}$$

By Lemma 1 we have

$$\phi(u) \leq F(x) + \langle \nabla F(x),\, u - x\rangle + \tfrac{L}{2}\|u-x\|^2 + g(u). \tag{24}$$

Adding and subtracting $\vartheta(x)$ in (24), and then combining with (23) we further obtain

$$\begin{aligned}
\phi(u) &\leq F(x) + \langle \nabla F(x) - \vartheta(x),\, u-x\rangle + \tfrac{L}{2}\|u-x\|^2 + g(u) + \langle \vartheta(x),\, u-x\rangle\\
&\leq F(x) + \langle \eta^{-1}(u-x) + s,\, x-u\rangle + \tfrac{L}{2}\|u-x\|^2 + g(u) + \langle \vartheta(x),\, u-x\rangle\\
&= F(x) + g(u) + \langle s,\, x-u\rangle + \big(\tfrac{L}{2} - \tfrac{1}{\eta}\big)\|u-x\|^2 + \langle \vartheta(x),\, u-x\rangle\\
&\leq F(x) + g(x) - \tfrac{2-L\eta}{2\eta}\|u-x\|^2 + \langle \vartheta(x),\, u-x\rangle\\
&\leq \phi(x) - \tfrac{2-L\eta}{2\eta}\|u-x\|^2 + \|\vartheta(x)\|\|u-x\|\\
&\leq \phi(x) - \tfrac{2-L\eta}{2\eta}\|u-x\|^2 + \tfrac{1}{\eta}\epsilon(x)\|u-x\|,
\end{aligned}$$

where the third inequality follows from convexity of g, the fourth one from Cauchy-Schwarz, and the last one from the definition of $\epsilon(x)$. Flipping signs, we immediately obtain (22). □

Next, we bound the right-hand side terms in (21), for which we derive two-sided bounds on $\|x-u\|$.

Lemma 4. *Let x, u, and η be as in Theorem 1, and c be as defined in* (7)*. Then,*

$$c\|\rho(x)\| - \eta^{-1}\epsilon(x) \leq \|x-u\| \leq \|\rho(x)\| + \eta^{-1}\epsilon(x). \tag{25}$$

Proof. From Lemma 3 we see that for $\eta > 0$ it holds that

$$1 \leq \eta \implies q(1) \leq q(\eta), \quad \text{and} \quad 1 \geq \eta \implies p(1) \leq p(\eta) = \eta^{-1}q(\eta) \tag{26}$$

Using (26), the triangle inequality, and nonexpansivity (Lemma 2), we see that

$$\begin{aligned}
\min\{1,\eta\}\, q(1) = \min\{1,\eta\}\, \|\rho(x)\| \quad &\leq \quad \|P_\eta(x - \eta\nabla f(x)) - x\|\\
&\leq \quad \|x-u\| + \|u - P_\eta(x - \eta\nabla f(x))\|\\
&\leq \quad \|x-u\| + \|\vartheta(x)\| \quad \leq \quad \|x-u\| + \eta^{-2}\epsilon(x).
\end{aligned}$$

Since $c \leq \liminf_k \eta_k$, we have $\|x-u\| \geq c\|\rho(x)\| - \epsilon(x)$. An upper bound follows upon noting

$$\begin{aligned}
\|x-u\| &\leq \|x - P_\eta(x - \eta\nabla f(x))\| + \|P_\eta(x - \eta\nabla f(x)) - u\|\\
&\leq \max\{1,\eta\}\, \|\rho(x)\| + \|\vartheta(x)\| \quad \leq \quad \|\rho(x)\| + \eta^{-1}\epsilon(x).
\end{aligned}$$

□

Theorem 1 and Lemma 4 have done the hard work; they imply the following crucial corollary.

Corollary 1. *Let x, u, η, and c be as in Lemma 4 and Theorem 1. Then, we have*

$$\phi(x) - \phi(u) \;\geq\; h(x),$$

where the function h is given by

$$h(x) := \tfrac{2-L\eta}{2\eta}c^2\|\rho(x)\|^2 - \big(c\tfrac{2-L\eta}{\eta^2} + \tfrac{1}{\eta}\big)\|\rho(x)\|\epsilon(x) - \big(\tfrac{2-L\eta}{2\eta^3} + \tfrac{1}{\eta^2}\big)\epsilon(x)^2. \tag{27}$$

Proof. Plug in the bounds (25) into (21) and simplify. □

Now we are in a position to state our main convergence theorem.

Theorem 2 (Convergence)**.** *Let $f \in C^1_L(\mathcal{X})$ such that $\inf_{\mathcal{X}} f > -\infty$ and g be lsc, convex on $\mathcal{X}$. Let $\{x^k\} \subset \mathcal{X}$ be a sequence generated by (6), and let condition (8) hold. Then, there exists a limit point x^* of the sequence $\{x^k\}$, and a constant $K > 0$, such that $\|\rho(x^*)\| \leq K\epsilon(x^*)$. Moreover, if the sequence $\{F(x^k)\}$ converges, then for every limit point x^* of $\{x^k\}$ it holds that $\|r(x^*)\| \leq K\epsilon(x^*)$.*

Proof. Corollary 1 reduces the proof to a case where the arguments of [13] become applicable; thus, we may conclude convergence; we omit details to avoid repetition. □

3 Incremental NOCOPS

We now specialize NOCOPS to the large-scale setting with decomposable $F(x)$, that is, to

$$\text{minimize} \quad \big(F(x) := \sum\nolimits_{t=1}^{T} f_t(x)\big) + g(x), \tag{28}$$

where each $f_t : \mathbb{R}^n \to \mathbb{R}$ is a $C^1_{L_t}(\mathcal{X})$ function (let $L_t \leq L$ for simplicity), and g, $\mathcal{X}$ are as before.

In machine learning and optimization it has long been known that for decomposable objectives it can be advantageous to replace the full gradient $\nabla F(x)$ by an *incremental gradient* $\nabla f_{r(t)}(x)$, where $r(t)$ is some suitable index. Incremental methods have been extensively analyzed in the setting of backpropagation algorithms [11, 13], a setting that corresponds to $g(x) \equiv 0$ in our case. For $g(x) \neq 0$, the stochastic generalized gradient methods of [19] or the perturbed generalized methods of [17] apply. But as previously argued, these methods suffer from problems similar to those faced by ordinary subgradient methods; so we may prefer proximal splitting methods instead.

To specialize NOCOPS for solving (28), we propose the following iteration:

$$\begin{aligned} x^{k+1} &= \mathcal{M}\big(x^k - \eta_k \sum\nolimits_{t=1}^{T} \nabla f_t(z^t)\big) \\ z^1 = x^k, \quad z^{t+1} &= \mathcal{O}(z^t - \eta\nabla f_t(z^t)), \quad t = 1, \ldots, T-1. \end{aligned} \tag{29}$$

Here, $\mathcal{O}$ and $\mathcal{M}$ are appropriate nonexpansive maps, choosing which we get different algorithms. For example, when $\mathcal{X} = \mathbb{R}^n$, $g(x) \equiv 0$, and $\mathcal{M} = \mathcal{O} = \mathrm{Id}$, then (29) reduces to the problem class considered in [27]. If $\mathcal{X}$ is a closed convex set, $g(x) \equiv 0$, $\mathcal{M} = \Pi_{\mathcal{X}}$, and $\mathcal{O} = \mathrm{Id}$, then (29) reduces to a method that is essentially implicit in [27]. Note, however, that in this case, the constraints are enforced *only once* every major iteration; the minor iterates (z^t) may be infeasible.

We introduce below four variants of (29); to our knowledge, *all four* are novel.

1. $\mathcal{X} = \mathbb{R}^n$, $g(x) \not\equiv 0$, $\mathcal{M} = P^g_\eta$, and $\mathcal{O} = \mathrm{Id}$; this is a penalized unconstrained problem, and the penalty is applied *once* every major iteration.
2. $\mathcal{X} = \mathbb{R}^n$, $g(x) \not\equiv 0$, $\mathcal{M} = P^g_\eta$, and $\mathcal{O} = P^g_\eta$; this is a penalized unconstrained problem, but now the penalty is applied at every minor iteration.
3. $\mathcal{X}$ is a closed convex set, $g(x) = \psi(x) + \delta(\cdot|\mathcal{X})$ (where ψ may be zero or nonzero), $\mathcal{M} = P^g_\eta$, and $\mathcal{O} = \mathrm{Id}$; this is a penalized, constrained problem, and the penalty is applied once every major iteration.
4. Same as variant 3, except that $\mathcal{O} = P^g_\eta$; this is a penalized, constrained problem, and the penalty is applied at every minor iteration.

Which of the four variants one prefers depends on the complexity of the constraint set $\mathcal{X}$ and the regularizer $g(x)$. However, the analysis of all four variants is similar, so we present details only for the fourth, as it is the most general.

```
Input: $\{\nabla f_t(X)\}$, $P^g_\eta$: subroutines
Output: Approximate solution to (29)
$t \leftarrow 1$; $k \leftarrow 0$; $x^k \in \mathcal{X}$;
while $\neg$ converged do
    $z^1 \leftarrow x^k$; $\alpha_1 \leftarrow \eta_k$;
    if $t < T$ then
        Get $t$-th gradient $g^t = \nabla f_t(z^t)$;
        Compute $z^{t+1} \leftarrow P^g_{\alpha_t}(z^t - \alpha_t g^t)$;
        Aggregate gradient $F^k \leftarrow F^k + g^t$;
        $t \leftarrow t + 1$;     update stepsize $\alpha_t$;
    else
        $x^{k+1} = P^g_{\eta_k}(x^k - \eta_k F^k)$;
        $t \leftarrow 0$;     $k \leftarrow k + 1$;
        Check for convergence;
    end
end
```

Algorithm 1: Incremental NOCOPS.

3.1 Convergence

We begin by rewriting (29) in a form that matches the main iteration (6):

$$x^{k+1} = \mathcal{M}\big(x^k - \eta_k \sum\nolimits_{t=1}^{T} \nabla f_t(z^t)\big) \quad = \quad \mathcal{M}\left(x^k - \eta_k \nabla F(x^k) + \eta_k \vartheta(x^k)\right),$$

where the error term at a general x is given by $\vartheta(x) := \sum_{t=1}^{T}\big(f_t(x) - f_t(z^t)\big)$. We must ensure that the norm of the error term is bounded. Lemma 5 proves this.

Lemma 5 (Bounded error)**.** *If for all $x \in \mathcal{X}$, $\|\nabla f_t(x)\| \le M$ and $\|\partial g(x)\| \le G$, then $\|\vartheta(x)\| \le K_1, \quad$ for some constant $K_1 > 0$.*

Proof. First, observe that if z^{t+1} is computed by (29), $\mathcal{O} = P^g_\eta$, and $s^t \in \partial g(z^t)$, then

$$\|z^{t+1} - z^t\| \quad \le \quad 2\eta\|\nabla f_t(z^t) + s^t\|. \tag{30}$$

Using (30) we can bound the error incurred upon using z^t instead of x^k. Specifically, if $x \equiv x^k$, and

$$\epsilon_t := \|\nabla f_t(z^t) - \nabla f_t(x)\|, \quad t = 1, \ldots, T, \tag{31}$$

then the following bound holds (for details of the proof, please see appendix):

$$\epsilon_t \le 2\eta L \sum\nolimits_{j=1}^{t-1} (1 + 2\eta L)^{t-1-j} \|\nabla f_j(x) + s^j\|, \quad t = 2, \ldots, T. \tag{32}$$

Now we use (32) and some simplification to finish the proof. Noting that $\epsilon_1 = 0$, we have

$$\begin{aligned} \|\vartheta(x)\| \le \sum\nolimits_{t=2}^{T} \epsilon_t \quad &\overset{(32)}{\le} \quad 2\eta L \sum\nolimits_{t=2}^{T} \sum\nolimits_{j=1}^{t-1} (1 + 2\eta L)^{t-1-j} \beta_j \\ &= 2\eta L \sum\nolimits_{t=1}^{T-1} \beta_t \left(\sum\nolimits_{j=0}^{T-t-1} (1 + 2\eta L)^j\right) \quad = \quad \sum\nolimits_{t=1}^{T-1} \beta_t \big((1 + 2\eta L)^{T-t} - 1\big) \\ &\le \sum\nolimits_{t=1}^{T-1} (1 + 2\eta L)^{T-t} \beta_t \quad \le \quad (1 + 2\eta L)^{T-1} \sum\nolimits_{t=1}^{T-1} \|\nabla f_t(x) + s^t\| \\ &\le C_1(T-1)(M+G) =: K_1. \qquad \square \end{aligned}$$

Remark 2. Lemma 5 implies that if in the error condition (8), we let $\eta \to 0$, then $\eta\|\vartheta(x)\| \to 0$.

Given the error bounds established by Lemma 5, convergence results for Algorithm 1 follow immediately from the more the general Theorem 2; we omit details to avoid repetition.

4 Application

The main contribution and focus of this paper is the new NOCOPS framework, and studying a specific application is not one of the aims of this paper. Nevertheless, we do illustrate NOCOPS's empirical performance on a challenging nonconvex problem, namely, (penalized) nonnegative matrix factorization:

$$\min_{X,\ A\geq 0} \quad \tfrac{1}{2}\|Y - XA\|_{\mathrm{F}}^2 + \psi_0(X) + \sum\nolimits_{t=1}^{T} \psi_t(a_t), \tag{33}$$

where Y is an $m \times T$ matrix, X is $m \times K$, and A is $K \times T$ with $a_1, \ldots, a_T$ as its columns. Problem (33) extends the famous nonnegative matrix factorization (NMF) problem [6] by allowing Y to be arbitrary (not necessarily nonnegative) and adding nonsmooth regularizers on X and A.

A similar class of problems was recently also studied in [5], but with a crucial difference: the formulation in [5] does not allow nonsmooth regularizers on X (the class of problems studied in [5] is a subset of those our framework allows). On a technical note, [5] consider stochastic optimization methods whose analysis requires perturbations to disappear in the limit; while our method is deterministic and our analysis does not rely on disappearing perturbations.

Following [5] we rewrite (33) in a form more amenable to NOCOPS, that is,

$$\min\nolimits_X \quad \phi(X) := \sum\nolimits_{t=1}^{T} f_t(X) + g(X), \tag{34}$$

where $g(X)$ captures both $\psi_0(X)$ and the constraints on X. Each $f_t(X)$ is defined as

$$f_t(X) := \min\nolimits_a \quad \tfrac{1}{2}\|y_t - Xa\|^2 + g_t(a), \quad 1 \leq t \leq T, \tag{35}$$

where $g_t(a)$ captures both $\psi_t(a)$ and the constraints on a_t. Whenever (35) attains its unique minimum, say a^*, then $f_t(X)$ is differentiable and we have $\nabla_X f_t(X) = (y_t - Xa^*)(a^*)^T$. Thus, we can instantiate Algorithm 1; all we need is a subroutine for solving (35).[1]

In our experiments, we consider the following two instances of (34): (i) $g(X) = \delta(X| \geq 0)$; and (ii) $g(X) = \lambda\|X\|_1 + \delta(X| \geq 0)$. We select the g_t's to be matching, so that $g_t = \delta(a| \geq 0)$, and $g_t(a) = \gamma\|a\|_1 + \delta(a| \geq 0)$ are used. Choice (i) solves standard NMF, while choice (ii) solves a sparse-NMF problem. We remark that in general, when penalizing X one should either constrain a_t or penalize it, otherwise one can get degenerate solutions.

To provide the reader with a baseline, on the basic NMF problem we compare NOCOPS against the well-tuned C++ toolbox SPAMS [5]. Obviously, the comparisons are not fair to NOCOPS, because unlike SPAMS, it is implemented in MATLAB. Fortunately, our MATLAB implementation already runs very competitively, and unlike SPAMS, also allows factorizing sparse matrices. We note that since our subroutines depend heavily on matrix-vector operations, a well-tuned C++ implementation of NOCOPS should run at least 3-4 times (based on initial experiments) faster than our MATLAB version, especially for sparse matrices.

We compute NMF on the following data matrices:

1. CBCL Face Database [28] (dense, size 361×2429); we compute a rank-49 factorization.
2. Yale B Database [29]; (dense, size 32256×2414); we compute a rank-64 factorization.
3. Random matrix (dense, size 4000×4000, entries in $[0, 1]$); we compute a rank-64 factorization, and penalize X by $\lambda\|\cdot\|_1$, and A by $\gamma\|A\|_1$, with $(\lambda, \gamma) = (10^2, 10^{-4})$.
4. Pajek connectivity matrix for Internet routers (sparse, size 124,651 × 124,651, density $1.3 \cdot 10^{-5}$, from the UFL sparse matrix collection, ID: 1505 [30]); we compute a rank-4 factorization; here $(\lambda, \gamma) = (10^{-2}, 10^{-6})$ were used.

Figure 1 reports summarizes our experimental results. In the first row, in addition to SPAMS, we include running times for Lee and Seung's algorithm, and our implementation of alternating (nonnegatively constrained) least squares. From the graph we see that our MATLAB implementation of NOCOPS runs only slightly slower that the state-of-the-art method in SPAMS. The plots also show a dashed line that hints at what might be achievable with a faster C++ implementation of NOCOPS.

[1] In practice, it is better to use *mini-batches*, and we used them for all the online algorithms compared.

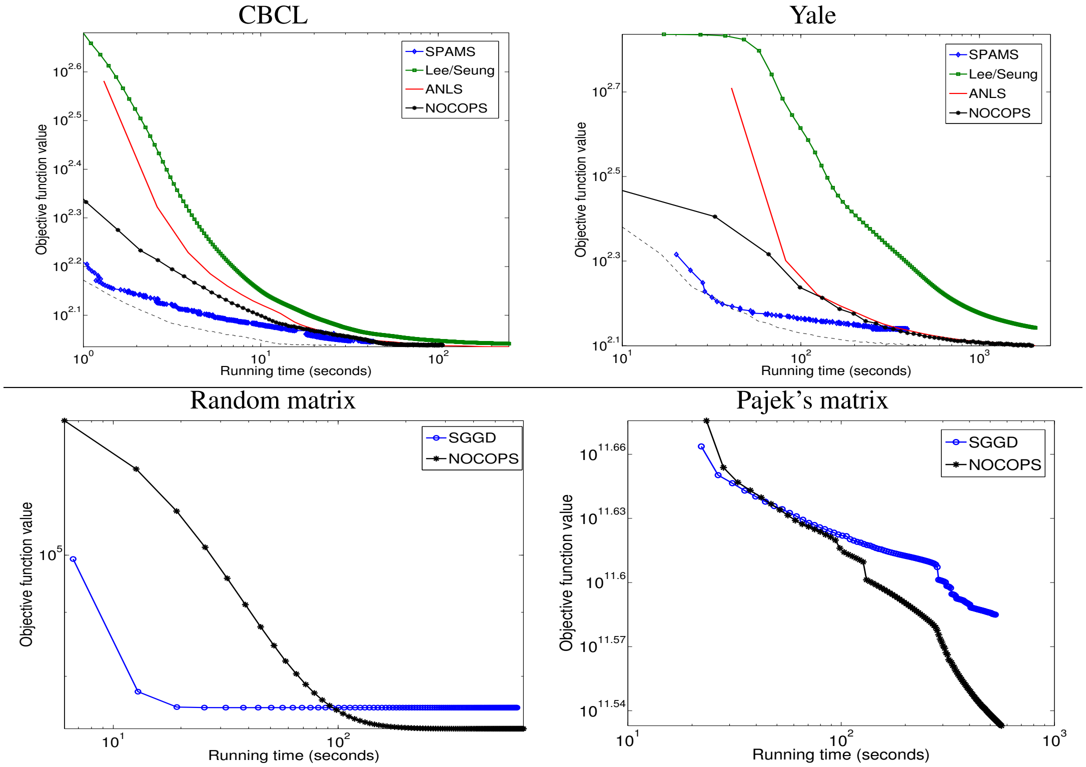

Figure 1: Top row: running times on CBCL and Yale B data. Bottom line, SGGD against NOCOPS. The densities of solutions returned (left to right) were (100%, 58.8%) for SGGD, and (61.1%, 0.1%) for NOCOPS. Initial objective values and very small runtimes have been suppressed for clarity of presentation.

The second row in Figure 1 shows numerical results that compare the stochastic generalized gradient (SGGD) algorithm of [19] against NOCOPS, when started at exactly the same point. As is well-known, SGGD requires careful stepsize tuning to be competitive. Thus, we searched over a range of possible stepsize choices, and have reported the results with the best choices found. NOCOPS also requires some stepsize tuning, but significantly lesser than SGGD. Finally, we note that as predicted, the solutions returned by NOCOPS, often have objective function values better than SGGD, and always achieved greater sparsity.

5 Discussion

We presented a new framework called NOCOPS, which solves a broad class of nonconvex composite objective problems. NOCOPS builds on the general analysis of [13], and extends it to admit problems that are *strictly* more general. NOCOPS permits nonvanishing perturbations, which is a useful practical feature. We exploited the perturbation analysis to derive both batch and incremental versions of NOCOPS. Finally, experiments with medium to large matrices showed that NOCOPS is competitive with state-of-the-art methods; NOCOPS was also seen to outperform the stochastic generalized gradient method.

We conclude by mentioning NOCOPS includes numerous algorithms and problem settings as special cases. Example are: forward-backward splitting with convex costs, incremental forward-backward splitting (convex), gradient projection (both convex and nonconvex), the proximal-point algorithm, and so on. Thus, it will be valuable to investigate if some of the theoretical results for these methods can be carried over to NOCOPS. Theoretically, the most important open problem that we would like to analyze is to permit even the regularizer in (1) to be nonconvex—but this might require significantly different convergence analysis.

References

[1] J. Duchi and Y. Singer. Online and Batch Learning using Forward-Backward Splitting. *J. Mach. Learning Res. (JMLR)*, Sep. 2009.

[2] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. In S. Sra, S. Nowozin, and S. J. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.

[3] M. Schmidt, E. van den Berg, M. Friedlander, and K. Murphy. Optimizing Costly Functions with Simple Constraints: A Limited-Memory Projected Quasi-Newton Algorithm. In *AISTATS*, 2009.

[4] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems. *SIAM J. Imgaging Sciences*, 2(1):183–202, 2009.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online Learning for Matrix Factorization and Sparse Coding. *JMLR*, 11:10–60, 2010.

[6] D. D. Lee and H. S. Seung. Algorithms for nonnegative matrix factorization. In *NIPS*, pages 556–562, 2000.

[7] D. Kundur and D. Hatzinakos. Blind image deconvolution. *IEEE Signal Processing Magazine*, 13(3):43 –64, may 1996.

[8] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computation*, 15:349–396, 2003.

[9] O. L. Mangasarian. Mathematical Programming in Neural Networks. *Informs J. Computing*, 5(4):349–360, 1993.

[10] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall PTR, 1st edition, 1994.

[11] D. P. Bertsekas. Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey. Technical Report LIDS-P-2848, MIT, August 2010.

[12] A. A. Gaivoronski. Convergence properties of backpropagation for neural nets via theory of stochastic gradient methods. Part 1. *Optimization methods and Software*, 4(2):117–134, 1994.

[13] M. V. Solodov. Convergence analysis of perturbed feasible descent methods. *J. Optimization Theory and Applications*, 93(2):337–353, 1997.

[14] M. Fukushima and H. Mine. A generalized proximal point algorithm for certain non-convex minimization problems. *Int. J. Systems Science*, 12(8):989–1000, 1981.

[15] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 2007/76, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), September 2007.

[16] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. Optimization Online, Dec. 2010.

[17] M. V. Solodov and S. K. Zavriev. Error stability properties of generalized gradient-type algorithms. *J. Optimization Theory and Applications*, 98(3):663–680, 1998.

[18] Y. M. Ermoliev and V. I. Norkin. Stochastic generalized gradient method with application to insurance risk management. Technical Report IR-97-021, IIASA, Austria, April 1997.

[19] Y. M. Ermoliev and V. I. Norkin. Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization. *Cybernetics and Systems Analysis*, 34:196–215, 1998.

[20] P. L. Combettes and J.-C. Pesquet. Proximal Splitting Methods in Signal Processing. *arXiv:0912.3522v4*, May 2010.

[21] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*. Springer, 1998.

[22] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulation*, 4(4):1168–1200, 2005.

[23] F. H. Clarke. *Optimization and nonsmooth analysis*. John Wiley & Sons, Inc., 1983.

[24] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2004.

[25] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787.

[26] E. M. Gafni and D. P. Bertsekas. Two-metric projection methods for constrained optimization. *SIAM Journal on Control and Optimization*, 22(6):936–964, 1984.

[27] M. V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11:23–35, 1998.

[28] K.-K. Sung. *Learning and Example Selection for Object and Pattern Recognition*. PhD thesis, MIT, Artificial Intelligence Laboratory and Center for Biological and Computational Learning, Cambridge, MA, 1996.

[29] K.C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.

[30] T. A. Davis and Y. Hu. The University of Florida Sparse Matrix Collection. *ACM Transactions on Mathematical Software*, 2011. To appear.

Nonconvex proximal splitting: supplement

A. Anonymous anon@anon.org

June 3, 2011

Note: References of the type (M-X) refer to equation (X), or Lemma X, Corollary X, etc., in the main paper.

1 Proof of Lemma M-3

Proof. First consider the "deflected" model-function (the objective function corresponding to P_η)

$$m_g(x, \eta; y, z) := \langle z,\, x - y\rangle + \frac{1}{2\eta}\|x - y\|^2 + g(x), \tag{1}$$

for which we define the (deflected) *Moreau-envelope*:

$$E_g(\eta) := \inf_{x \in \mathcal{X}}\; m_g(x, \eta; y, z). \tag{2}$$

Function μ_g is easily seen to be a convex (see e.g., [3; Theorem 2.26]). Rearranging m_g, we see that the infinum in (2) is attained at $P_\eta^g(y - \eta z)$, which is unique. Thus, $E_g(\eta)$ is differentiable, and

$$E'_g(\eta) = -\frac{1}{2\eta^2}\|P_\eta^g(y - \eta z) - y\|^2 = -\tfrac{1}{2}p(\eta)^2.$$

Since $E_g(\eta)$ is convex, E'_g is an increasing function; equivalently $p(\eta)$ is decreasing.

To prove (M-16), observe that $\hat{E}_g(\gamma) := E_g(1/\gamma)$ is a concave function of γ, because it is a pointwise (indexed by x) infimum of functions linear in γ [1; §3.2.3]. Thus, its derivative

$$\hat{E}'_g(\gamma) = \tfrac{1}{2}\|P_{1/\gamma}^g(x - \gamma^{-1}y) - x\|^2 = q(1/\gamma),$$

is a decreasing function of γ; writing $\eta = 1/\gamma$ completes the argument. □

2 Proof of Theorem M-1

Let $z = \nabla F(x) - \vartheta(x)$, so that (M-20) becomes

$$u = P_\eta(y - \eta z).$$

The directional derivative $\mathrm{d}m_g$ (of m_g, with respect to x, and in the direction w) satisfies at $x = u$ the optimality condition

$$\mathrm{d}m_g(u, \eta; y, z)(w) = \langle z + \eta^{-1}(u - y) + s,\, w\rangle \geq 0, \quad s \in \partial g(u). \tag{3}$$

Upon setting $y = x$, and $w = x - u$ in (3), we obtain the inequality

$$\langle \nabla F(x) - \vartheta(x) + \eta^{-1}(u - x) + s,\, x - u\rangle \geq 0, \tag{4}$$

Proof. By Lemma M-1 we have

$$\phi(u) \leq F(x) + \langle \nabla F(x),\, u - x\rangle + \tfrac{L}{2}\|u - x\|^2 + g(u). \tag{5}$$

Rearranging inequality (4), we obtain

$$\langle \nabla F(x) - \vartheta(x),\, u - x\rangle \quad \leq \quad \langle \eta^{-1}(u - x) + s,\, x - u\rangle. \tag{6}$$

Adding and subtracting $\vartheta(x)$ in (5), and then combining with (6) we further obtain

$$\begin{aligned}
\phi(u) &\leq F(x) + \langle \nabla F(x) - \vartheta(x),\, u - x\rangle + \tfrac{L}{2}\|u - x\|^2 + g(u) + \langle \vartheta(x),\, u - x\rangle \\
&\leq F(x) + \langle \eta^{-1}(u - x) + s,\, x - u\rangle + \tfrac{L}{2}\|u - x\|^2 + g(u) + \langle \vartheta(x),\, u - x\rangle \\
&= F(x) + g(u) + \langle s,\, x - u\rangle + \left(\tfrac{L}{2} - \tfrac{1}{\eta}\right)\|u - x\|^2 + \langle \vartheta(x),\, u - x\rangle \\
&\leq F(x) + g(x) - \tfrac{2 - L\eta}{2\eta}\|u - x\|^2 + \langle \vartheta(x),\, u - x\rangle \\
&\leq \phi(x) - \tfrac{2 - L\eta}{2\eta}\|u - x\|^2 + \|\vartheta(x)\|\|u - x\| \\
&\leq \phi(x) - \tfrac{2 - L\eta}{2\eta}\|u - x\|^2 + \eta^{-1}\epsilon(x)\|u - x\|,
\end{aligned}$$

where the third inequality follows from convexity of g, the fourth one from Cauchy-Schwarz, and the last one from the definition of $\epsilon(x)$. Flipping signs, we immediately obtain (3). □

Moreover, we can also prove the following bound

$$\langle \phi'(u),\, x - u\rangle \quad \geq \quad \tfrac{1 - L\eta}{\eta}\|u - x\|^2 - \eta^{-1}\epsilon(x)\|u - x\| \tag{7}$$

Proof. Consider the directional derivative $\mathrm{d}\phi(u; x - u)$, for which using $\phi'(u) = \nabla F(u) + s$, we have

$$\begin{aligned}
\langle \nabla F(u) + s,\, x - u\rangle &= \langle \nabla F(x) - \vartheta(x) + s,\, x - u\rangle - \langle \nabla F(u) - \nabla F(x) + \vartheta(x),\, u - x\rangle \\
&\geq \langle \eta^{-1}(x - u),\, x - u\rangle - \langle \nabla F(u) - \nabla F(x),\, u - x\rangle - \langle \vartheta(x),\, u - x\rangle \\
&\geq (\eta^{-1} - L)\|u - x\|^2 - \langle \vartheta(x),\, u - x\rangle \\
&\geq \tfrac{1 - L\eta}{\eta}\|u - x\|^2 - \|\vartheta(x)\|\|u - x\| \\
&\geq \tfrac{1 - L\eta}{\eta}\|u - x\|^2 - \eta^{-1}\epsilon(x)\|u - x\|.
\end{aligned}$$

□

3 Proof of (M-32)

To that end, we first bound $\|z^{t+1} - z^t\|$ in Lemma 1 below.

Lemma 1 (Bounded increment)**.** *Let z^{t+1} be computed by M-29. Then, we have*

$$\text{if} \quad \mathcal{O} = P^g_\eta \quad \text{and} \quad s^t \in \partial g(z^t), \;\; \text{then} \quad \|z^{t+1} - z^t\| \leq 2\eta\|\nabla f_t(z^t) + s^t\|. \tag{8}$$

Proof. For proving (8), notice that definition (M-4) implies the inequality

$$\begin{aligned}
&\tfrac{1}{2}\|z^{t+1} - z^t + \eta\nabla f_t(z^t)\|^2 + \eta g(z^{t+1}) \leq \tfrac{1}{2}\|\eta\nabla f_t(z^t)\|^2 + \eta g(z^t), \\
&\tfrac{1}{2}\|z^{t+1} - z^t\|^2 \leq \eta\langle \nabla f_t(z^t),\, z^t - z^{t+1}\rangle + \eta(g(z^t) - g(z^{t+1})).
\end{aligned}$$

But since ψ is convex, we know that

$$g(z^{t+1}) \geq g(z^t) + \langle s_t,\, z^{t+1} - z^t\rangle, \quad s_t \in \partial g(z^t).$$

Since g is convex it further follows that

$$\begin{aligned}
\tfrac{1}{2}\|z^{t+1} - z^t\|^2 &\leq \eta\langle s^t,\, z^t - z^{t+1}\rangle + \langle \nabla f_t(z^t),\, z^t - z^{t+1}\rangle \\
&\leq \eta\|s_t + \nabla f_t(z^t)\|\|z^t - z^{t+1}\| \\
\implies \quad &\|z^{t+1} - z^t\| \leq 2\eta\|\nabla f_t(z^t) + s^t\|.
\end{aligned}$$

□

Lemma 2 (Incrementality error)**.** *Let $x \equiv x^k$, and define*

$$\epsilon_t := \|\nabla f_t(z^t) - \nabla f_t(x)\|, \quad t = 1, \ldots, T. \tag{9}$$

Then, for each $t \geq 2$, the following bound on the error holds:

$$\epsilon_t \leq 2\eta L \sum\nolimits_{j=1}^{t-1} (1 + 2\eta L)^{t-1-j} \|\nabla f_j(x) + s^j\|, \quad t = 2, \ldots, T. \tag{10}$$

Proof. The proof extends the unconstrained, unpenalized setting of [4] to our setting. We proceed by induction. The base case is $t = 2$, for which we have

$$\epsilon_2 = \|\nabla f_2(z^2) - \nabla f_2(x)\| \leq L\|z^2 - x\| = L\|z^2 - z^1\| \overset{(8)}{\leq} 2\eta L\|\nabla f_1(x) + s^1\|.$$

Assume inductively that (M-32) holds for $t \leq r < T$, and consider $t = r + 1$. In this case we have

$$\begin{aligned} \epsilon_{r+1} &= \|\nabla f_{r+1}(z^{r+1}) - \nabla f_{r+1}(x)\| && \leq L\|z^{r+1} - x\| \\ &= L\left\|\sum\nolimits_{j=1}^{r} (z^{j+1} - z^j)\right\| && \leq L\sum\nolimits_{j=1}^{r} \|z^{j+1} - z^j\| \\ &\overset{\text{Lemma 1}}{\leq} 2\eta L \sum\nolimits_{j=1}^{r} \|\nabla f_j(z^j) + s^j\|. \end{aligned} \tag{11}$$

To complete the induction, first observe that $\|\nabla f_t(z^t) + s^t\| \leq \|\nabla f_t(x) + s^t\| + \epsilon_t$. Thus, invoking the induction hypothesis, we obtain

$$\|\nabla f_t(z^t)\| \leq \|\nabla f_t(x)\| + 2\eta L \sum\nolimits_{j=1}^{t-1} (1 + 2\eta L)^{t-1-j} \|\nabla f_j(x) + s^j\|, \quad t = 2, \ldots, r. \tag{12}$$

Combining inequality (12) with (11) we further obtain

$$\epsilon_{r+1} \leq 2\eta L \sum\nolimits_{j=1}^{r} \left(\|\nabla f_j(x) + s^j\| + 2\eta L \sum\nolimits_{l=1}^{j-1} (1 + L\eta)^{j-1-l} \|\nabla f_l(x) + s^l\| \right).$$

Introducing the shorthand $\beta_j \equiv \|\nabla f_j(x) + s^j\|$, simple manipulation of the above inequality yields

$$\begin{aligned} \epsilon_{r+1} \leq \quad & 2\eta L\beta_r + \sum\nolimits_{l=1}^{r-1} \left(2\eta L + 4\eta^2 L^2 \sum\nolimits_{j=l+1}^{r} (1 + 2\eta L)^{j-l-1} \right) \beta_l \\ = \quad & 2\eta L\beta_r + \sum\nolimits_{l=1}^{r-1} \left(2\eta L + 4\eta^2 L^2 \sum\nolimits_{j=0}^{r-l-1} (1 + 2\eta L)^j) \right) \beta_l \\ = \quad & 2\eta L\beta_r + \sum\nolimits_{l=1}^{r-1} 2\eta L(1 + 2\eta L)^{r-l} \beta_l \quad = \quad 2\eta L \sum\nolimits_{l=1}^{r} (1 + 2\eta L)^{r-l} \beta_l, \end{aligned}$$

which completes the proof. □

References

[1] S. Boyd and L. Vandenberghe. *Convex Optimization.* Cambridge University Press, March 2004. ISBN 0521833787.

[2] P. L. Combettes and J.-C. Pesquet. Proximal Splitting Methods in Signal Processing. *arXiv:0912.3522v4*, May 2010.

[3] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis.* Springer, 1998.

[4] M. V. Solodov. Incremental gradient algorithms with stepsizes bounded away from zero. *Computational Optimization and Applications*, 11:23–35, 1998.