A hybrid proximal generalized conditional gradient method and application to total variation parameter learning

Enis Chenchene\textsuperscript{1} Alireza Hosseini\textsuperscript{2} Kristian Bredies\textsuperscript{1}

Abstract—In this paper, we present a new method for solving optimization problems involving the sum of two proper, convex, lower semicontinuous functions, one of which has Lipschitz continuous gradient. The proposed method has a hybrid nature that combines the usual forward–backward and the generalized conditional gradient method. We establish a convergence rate of $O(k^{-1/2})$ under mild assumptions with a specific step-size rule and show an application to a total variation parameter learning problem, which demonstrates its benefits in the context of nonsmooth convex optimization.

I. INTRODUCTION

Given a Hilbert space $H$, the generalized conditional gradient method is a powerful tool to solve

$$\min_{u \in H} f(u) + g(u),$$

where $f$ and $g$ are suitable convex, proper, lower semicontinuous functions \cite{1}–\cite{5}. The general iteration reads, for $k \in \mathbb{N}$,

$$\begin{cases}
\nu^k \in \arg\min_{v \in H} \langle \nabla f(u^k), v \rangle + g(v), \\
u^{k+1} = u^k + \theta_k(v^k - u^k),
\end{cases}$$

where $\theta_k$ is a step-size that can be obtained, e.g., by line-search, backtracking, or satisfying a certain step-size rule.

The generalized conditional gradient method has an interesting connection with the more popular forward–backward method \cite{15}. Indeed, adding and removing in (1) a quadratic term

$$\min_{u \in H} f(u) - \frac{1}{2} \|u\|^2_H + g(u) + \frac{1}{2} \|u\|^2_P,$$

where $\|u\|^2_P := \langle u, Pu \rangle$ and $P : H \rightarrow H$ is a self-adjoint, positive definite, bounded, linear operator, then applying the generalized conditional gradient with $F(u) := f(u) - \frac{1}{2} \|u\|^2_P$, $G(u) := g(u) + \frac{1}{2} \|u\|^2_P$ and $\theta_k = 1$ for all $k$ gives

$$u^{k+1} = u^k \in \arg\min_{v \in H} \langle \nabla f(u^k) - Pu^k, v \rangle + g(v) + \frac{1}{2} \|v\|^2_P$$

$$= \arg\min_{v \in H} g(v) + \frac{1}{2} \|v - (u^k - P^{-1} \nabla f(u^k))\|^2_P.$$

This coincides with the celebrated (preconditioned) forward–backward method with respect to the metric induced by $P$. Note that, of course, if $P = 0$, one retrieves again the generalized conditional gradient method.

While sharing the same $O(k^{-1})$ rate on the objective function, both methods have distinct desirable properties.

Among others, see, e.g., \cite{16}, one advantage of conditional gradient-type methods over proximal methods is that it allows avoiding the computation of potentially costly projections or proximity operators in favor of cheaper linearized problems \cite{17}. On the other hand, the forward–backward method has a broader range of applicability, as it does not require, for instance, any coercivity assumption on $g$, and enjoys stronger convergence guarantees on the iterates.

Although the two methods have evolved independently over time, resulting in accelerated \cite{3}, \cite{18}, adaptive \cite{19}, \cite{20}, and inexact/stochastic variants \cite{21}, \cite{22}, the in-between scenario, i.e., with $P$ only positive semidefinite, remains unexplored to our knowledge. In this paper, we bridge this gap by introducing a hybrid method that we call Hybrid Proximal Generalized Conditional Gradient (HPGCG). We show that the proposed method can indeed benefit from the advantages of both worlds keeping the low cost-per-iteration of conditional gradient-type methods but allowing a broader range of applicability, and stronger convergence guarantees. Note as well that, in the same spirit of \cite{6}, the proposed method can also be understood as an instance of a degenerate forward–backward method.

The paper is organized as follows. In Section II, we present the algorithm as well as its convergence analysis. In Section III we show that the proposed method is particularly suitable to solve a total variation (TV) parameter learning problem.

II. THE PROPOSED METHOD AND ITS CONVERGENCE ANALYSIS

Throughout, we assume that $f : H \rightarrow \mathbb{R}$ is convex and Fréchet differentiable with Lipschitz continuous gradient, $g : H \rightarrow \mathbb{R} \cup \{\infty\}$ is proper, convex and lower semicontinuous, $f + g$ is coercive. Under these assumptions, problem (1) always admits an optimal solution. Eventually, we assume that $P : H \rightarrow H$ is a bounded, linear, positive semidefinite operator such that $\frac{1}{2} \| \cdot \|^2_P + g$ is strongly coercive.

A. HPGCG algorithm

In this section, we present the proposed HPGCG method along with our step-size choice. To do so, we first need to introduce some notation. First, we define $v(u) \in H$ as any minimizer of

$$\min_{v \in H} \langle \nabla f(u), v \rangle - \langle u, v \rangle_p + \frac{1}{2} \|v\|^2_p + g(v),$$

which always admits an optimal solution due to the strong coercivity of $\frac{1}{2} \| \cdot \|^2_p + g$. Then, we set

$$H_u(v) := \langle \nabla f(u), v \rangle - \langle u, v \rangle_p + \frac{1}{2} \|v\|^2_p + g(v),$$

1Institute of Mathematics and Scientific Computing, University of Graz, Graz, Austria. kristian.bredies@uni-graz.at, enis.chenchene@uni-graz.at
2School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Tehran, Iran. hosseini.alireza@ut.ac.ir
Initialize: $u_0 \in \text{Dom}(G)$

for $k = 0, 1, \ldots$ do

$\nu^k = v(u^k)$ according to (3)

Update the step-size as

$$\theta_k = \min \left( 1, \frac{1}{2D_f(u^k, \nu^k)} \right) \quad (4)$$

$u^{k+1} = u^k + \theta_k (\nu^k - u^k)$

end

Algorithm 1: HPGCG algorithm for solving (1).

and define $D(u) := H_u(u) - \inf H_u$. Note that

$$D(u) = \langle \nabla f(u), u - v(u) \rangle + g(u) - g(v(u)) - \frac{1}{2} \|u - v(u)\|_p^2.$$  

We further fix a function $D_f : H \times H \to \mathbb{R}$ with the following properties:

- $D_f(u, v) \geq f(v) - f(u) - \langle \nabla f(u), v-u \rangle$,
- $D_f(u, u + \theta (v - u)) \leq \theta^2 f(u, v)$, for $\theta \in [0, 1]$,
- $D_f$ is bounded on bounded sets.

A simple choice for $D_f$ is $D_f(u, v) = \frac{1}{2} \|u - v\|_2^2$, where $L$ is a Lipschitz constant of $\nabla f$. However, if $f(u) = \frac{1}{2} \|Ru - q\|_2^2$, where $K : H \to K$ is a linear and bounded operator, $K$ a Hilbert space and $q \in K$, we can also pick $D_f(u, v) = \frac{1}{2} \|R(u - v)\|_2^2$, which does not require the knowledge of $L$.

Inspired from [5], we can now present our step-size rule, which is shown in Algorithm 1. It is worth noting that the suggested step-size selection solves

$$\min_{\theta \in [0,1]} f(u) + \theta \langle \nabla f(u), v(u) - u \rangle + \theta^2 D_f(u, v(u)) + (1 - \theta) g(u) + \theta g(v(u)),$$

a problem that majorizes the usual linesearch problem.

B. Convergence Analysis

First, for every $u \in H$, let

$$r(u) := (f + g)(u) - \inf (f + g).$$

We have the following result.

**Lemma 1:** Let $u \in H$, $v = v(u)$ according to (3), and $\theta \in [0, 1]$. Then, for every optimal solution $u^*$ of (1) we have

1) $D(u) + \frac{1}{2} \|u - v\|_p^2 \geq r(u) - \|u - v\|_p \|u^* - v\|_p$,

2) the estimate

$$\theta \left( D(u) + \frac{1}{2} \|u - v\|_p^2 \right) \geq r(u) - \frac{1}{2} \|u - u^*\|_p^2$$

$$+ \frac{1}{2} \|u + \theta (v - u) - u^*\|_p^2 + \frac{\theta}{2} (r - \theta) \|u - v\|_p^2,$$

3) $D(u) \geq 0$ for every $u \in H$ and $D(u) = 0$ if and only if $u$ is an optimal solution to (1).

**Proof:** From the optimality conditions for (3) and the subgradient inequality, we have

$$g(v) + \langle u - v, u^* - v \rangle_p - \langle \nabla f(u), u^* - v \rangle \leq g(u^*). \quad (5)$$

Consequently, (5) and the convexity of $f$ yield

$$D(u) \geq \langle u - v, u^* - v \rangle_p + \langle \nabla f(u), u^* - u \rangle$$

$$+ g(u) - g(u^*) - \frac{1}{2} \|u - v\|_p^2$$

$$\geq r(u) + \langle u - v, u^* - v \rangle_p - \frac{1}{2} \|u - v\|_p^2,$$

which implies 1) via Cauchy–Schwarz. From (6) and the polarization identity, we get

$$\theta (D(u) + \frac{1}{2} \|u - v\|_p^2) \geq \theta r(u) + \langle \theta (u - v), u^* - v \rangle_p$$

$$= \theta r(u) + \langle \theta (u - v), u^* - u \rangle_p + \theta \|u - v\|_p^2$$

$$= \theta r(u) + \frac{1}{2} \|u + \theta (v - u) - u^*\|_p^2 - \frac{\theta^2}{2} \|u - v\|_p^2$$

$$- \frac{1}{2} \|u - u^*\|_p^2 + \theta \|u - v\|_p^2,$$

which proves part 2). By definition, $D \geq 0$. As $v$ is a minimizer of $H_u$, $D(u) = 0$ if and only if

$$u \in \text{arg min}_{v \in H} \langle \nabla f(u), v \rangle - \langle u, v \rangle_p + \frac{1}{2} \|v\|_p^2 + g(v),$$

which, from optimality conditions, is equivalent to $u$ being a minimizer of $f + g$. This completes the proof of part 3).

**Lemma 2:** Let $\{u^k\}_k$ be a sequence generated by Algorithm 1, then for $k \geq 0$, we get

$$r(u^{k+1}) - r(u^k) \leq -\frac{\theta_k}{2} \langle D(u^k) + \frac{1}{2} \|u^k - v^k\|_p^2 \rangle.$$  

**Proof:** By the definition of $r$, the first property of $D_f$ and convexity of $g$, we get

$$r(u^{k+1}) - r(u^k)$$

$$= f(u^k + \theta_k (v^k - u^k)) - f(u^k) - \theta_k \langle \nabla f(u^k), v^k - u^k \rangle$$

$$+ g(u^k + \theta_k (v^k - u^k)) - g(u^k) + \theta_k \langle \nabla f(u^k), v^k - u^k \rangle$$

$$\leq D_f(u^k, u^k + \theta_k (v^k - u^k)) - \frac{\theta_k}{2} \|u^k - v^k\|_p^2$$

$$+ \theta_k (g(v^k) - g(u^k) - \langle \nabla f(u^k), u^k - v^k \rangle + \frac{1}{2} \|u^k - v^k\|_p^2).$$

Consequently, using the definition of $D(u^k)$ and the second property of $D_f$ we obtain

$$r(u^{k+1}) - r(u^k)$$

$$\leq -\theta_k \langle D(u^k) + \frac{1}{2} \|u^k - v^k\|_p^2 \rangle + \theta_k^2 D_f(u^k, v^k). \quad (9)$$

Now, if $D(u^k) + \frac{1}{2} \|u^k - v^k\|_p^2 \geq 2D_f(u^k, v^k)$, then $\theta_k = 1$ and (9) turns easily into (8). Otherwise, $\theta_k = \frac{D(u^k) + \frac{1}{2} \|u^k - v^k\|_p^2}{2D_f(u^k, v^k)}$, which again leads from (9) to (8).

**Lemma 3:** The sequences $\{u^k\}_k$ and $\{v^k\}_k$ produced by Algorithm 1 are bounded.

**Proof:** From Lemma 2, and 3) of Lemma 1, $\{r(u^k)\}_k$ is non-increasing, thus bounded. As $f + g$ is coercive and $\nabla f$ is Lipschitz continuous, $\{u^k\}_k$ and $\{v^k\}_k$ are bounded as well. As $v^k$ minimizes $H_v$, optimality conditions yield

$$v^k \in (P + \partial g)^{-1}(Pu - \nabla f(u)).$$

Since $P + \partial g$ is strongly coercive, $(P + \partial g)^{-1}$ maps bounded sets to bounded sets, see, e.g., [13, Theorem 3.3]. Thus, $\{v^k\}_k$ is bounded.

**Proposition 1:** Let $\{u^k\}_k$ and $\{v^k\}_k$ be generated by Algorithm 1, then we have

$$\lim_{k \to \infty} D(u^k) + \frac{1}{2} \|u^k - v^k\|_p^2 = 0.$$  

**Proof:** For the sake of notation, we set $\xi_k := D(u^k) + \frac{1}{2} \|u^k - v^k\|_p^2$ for all $k \in \mathbb{N}$. Let $k \in \mathbb{N}$, and note that if $\theta_k = 1$, then, from (8), we get $r(u^{k+1}) - r(u^k) \leq -\frac{1}{2} \xi_k$. If $\theta_k < 1$, since $\{u^k\}_k$ and $\{v^k\}_k$ are bounded, $\{D_f(u^k, v^k)\}_k$ is bounded.
as well (Lemma 3 and third property of $D_f$), and thus there exists a $C > 0$, such that $\theta_k \geq 2C_\infty \xi_k$. Consequently, using again (8), we get $r(u^{k+1}) - r(u^k) \leq -C_\infty^2 |u|$. In both cases, we obtain

$$r(u^{k+1}) - r(u^k) \leq -\min \left\{ \frac{1}{2} \xi_k, C_\infty^2 \right\}. \quad (11)$$

Thus, the right-hand-side of (11) is summable and, in particular, (10) holds.

**Theorem 1:** Let $\{u^k\}_k$ be generated by Algorithm 1. Then, $\{r(u^k)\}_k$ converges monotonically to zero with rate $o(k^{-1/3})$.

**Proof:** From Lemma 2, part 2) of Lemma 1 and the definition of $|u|^k$, we have

$$r(u^k) - r(u^{k+1}) \geq \frac{\theta_k}{2} (D(u^k) + \frac{1}{2} |u^k - v^k|^2)$$

$$\geq \frac{\theta_k}{2} r(u^k) + \frac{1}{4} |u^{k+1} - u^*|^2 - \frac{1}{4} |u^k - u^*|^2.$$ 

In particular, $\{r(u^k)\}_k$ is monotonically non-increasing. Therefore, for each $n \in \mathbb{N}$,

$$\sum_{k=0}^{n-1} \theta_k r(u^k) \leq 2r(u^0) + \frac{1}{2} |u^0 - u^*|^2 \leq C,$$

for some $C > 0$. Hence, $\{\theta_k r(u^k)\}_k$ is summable. Now, if $\theta_k < 1$ then, using part 1) of Lemma 1 and boundedness of $\{D_f(u^k, v^k)\}_k$ (cf., proof of Proposition 1), we get

$$\theta_k \geq C(D(u^k) + \frac{1}{2} |u^k - v^k|^2)$$

$$\geq C(r(u^k) - |u^k - v^k| |u^k - v^k|), \quad (12)$$

for some $C > 0$. From the first inequality in (12), since $D(u^k) \geq 0$ by part 3) of Lemma 2, we get

$$\|u^k - v^k\| \leq \frac{\sqrt{2C}}{\theta_k}, \quad (13)$$

and from the second, we get

$$r(u^k) \leq C^{-1} \theta_k + \|u^k - v^k\| |u^k - v^k| \quad (14)$$

Thus, using (13) and boundedness of $\{v^k\}_k$ (cf., Lemma 3), we get $r(u^k) \leq C(\theta_k + \sqrt{\theta_k})$ for some $C > 0$, and since $\theta_k \leq 1$, $r(u^k) \leq 2C^{-1} \sqrt{\theta_k}$. Therefore, there exists a constant $C \geq 0$ such that for every $k \in \mathbb{N}$ with $\theta_k < 1$, $\theta_k r(u^k) \geq C r(u^k)$. By monotonicity of $\{r(u^k)\}_k$ there exists $C' > 0$ such that $C r(u^k) \leq 1$ for all $k \in \mathbb{N}$. In particular, for all $k \in \mathbb{N}$ with $\theta_k = 1$, $\theta_k r(u^k) \geq C' r(u^k)^3$. Therefore, for all $k \in \mathbb{N}$, $\theta_k r(u^k) \geq \min\{C, C'\} r(u^k)^3$. Thus, $\{r(u^k)\}_k$ is summable and monotonically non-increasing, hence $r(u^k) = o(k^{-1/3})$, see [14, Theorem 3.1.1].

**Remark 1:** If $\theta_k$ is bounded away from zero, then the rate for $r(u^k)$ immediately improves to $o(k^{-1})$. This is the case if $P$ is positive definite. In case $P = 0$, the rate $O(k^{-1})$ can be obtained using, for instance, the arguments in [5]. In the hybrid case, neither of these techniques can be applied. The proof of Theorem 1 accounts for this, yielding $o(k^{-1/3})$, which might still be improved in a future work.

The hybrid nature of HPGCC allows us to state a partial convergence result for the iterates.

**Theorem 2:** Let $\{u^k\}_k$ be generated by Algorithm 1 and $P^{1/2}$ be the square root of $P$. Then, $\{P^{1/2}u^k\}_k$ converges weakly to some $p^* = P^{1/2}u^*$, $u^*$ being a minimizer of (1).

**Proof:** From part 2) of Lemma 1 and Lemma 2 it follows that for all $k \in \mathbb{N}$ and every minimizer $u^*$,

$$\frac{1}{4} \|u^{k+1} - u^*\|_p^2 \leq \frac{1}{4} \|u^k - u^*\|_p^2 + r(u^k) - r(u^{k+1}).$$

Thus, for every minimizer $u^*$ of (1), the sequence $\{\|u^k - u^*\|_p^2\}_k$ converges, cf., [7, Lemma 2, Section 2.2.1]. Now, as $\{P^{1/2}u^k\}_k$ is bounded, it admits weak cluster points. Assume that $p^*$ and $p^{**}$ are two such elements with $P^{1/2}u^k \to p^*$ and $P^{1/2}u^k \to p^{**}$. Then, using that $\{u^k\}_k$ is bounded and all its weak cluster points are minimizers of (1) as a consequence of Theorem 1, it is easy to show that $p^* = P^{1/2}u^*$ and $p^{**} = P^{1/2}u^*$ for two minimizers $u^*$, $u^{**} \in H$. Now, since

$$\langle Pu^k, u^* - u^{**} \rangle = \frac{1}{2} \|u^k - u^{**}\|_p^2 - \frac{1}{2} \|u^k - u^*\|_p^2$$

$$- \frac{1}{2} u^{**}^2 + \frac{1}{2} \|u^k\|_p^2,$$

$\{\langle Pu^k, u^* - u^{**} \rangle\}_k$ converges to some $\lambda \in \mathbb{R}$. Thus, $\{\langle Pu^k, u^* - u^{**} \rangle\}_k$ and $\{\langle Pu^k, u^* - u^{**} \rangle\}_k$ converge to $\lambda$ too. Hence, taking the difference and passing to the limit gives $\|u^* - u^{**}\|_p^2 = 0$, and, thus, $p^* = P^{1/2}u^* = P^{1/2}u^* = p^{**}$. 

**III. TV PARAMETER LEARNING**

The automatic tuning of the regularization parameter for inverse problems is an ongoing challenge that recently featured several new data-driven approaches, see, e.g., [8]. Here, we propose a new learning model and show that the proposed HPGCC method allows us to solve it efficiently.

Given $p \in \mathbb{N}$ and a $n := p \times p$ grid, to denoise a degraded image $\xi \in \mathbb{R}^n$ we consider the classical ROF model [9]

$$\min_{u \in \mathbb{R}^n} \frac{1}{2} \|u - \xi\|_2^2 + \alpha \text{TV}(u), \quad (15)$$

where $\alpha$ is a positive parameter, $|\cdot|_2$ is the $l^2$ norm, and TV is the discrete total variation functional, namely $\text{TV}(u) := \|\nabla u\|_1$, where $\nabla$ is the discrete gradient operator defined via standard forward differences and $\|\cdot\|_1$ is defined by $\|v\|_1 = \sum_{i=1}^n |v_i|_2$ for every discrete vector field $v \in \mathbb{R}^{n \times 2}$. For the sake of notation, from now on we often denote

$$f(u) := \frac{1}{2} \|u - \xi\|_2^2, \quad \text{and } g_\alpha(v) := \alpha \|v\|_1. \quad (16)$$

From standard duality theory, see, e.g., [11, Section 19.2] and [12, Section 6.2.1], problem (15) is equivalent to

$$\min_{v \in \mathbb{R}^{n \times 2}} \frac{1}{2} \|\nabla v + \xi\|_2^2 \quad \text{s.t.: } \|v\|_{\infty, 2} \leq \alpha, \quad (17)$$

where $\nabla = -\nabla^*$ is the discrete divergence operator, and $\|\cdot\|_{\infty, 2}$ is the dual norm of $\|\cdot\|_{1, 2}$, which is defined for all $v \in \mathbb{R}^{n \times 2}$ by $\|v\|_{\infty, 2} = \max_{i \in \{1, \ldots, n\}} \|v_i\|_2$. Optimal solutions $u^\alpha$ and $v^\alpha$ to (15) and (17) respectively are often called primal-dual pairs and together can be characterized as solutions of the Fenchel–Rockafellar primal-dual optimality system

$$\nabla \cdot v^\alpha = u^\alpha - \xi, \quad \nabla u^\alpha = \partial g^\alpha_\alpha(v^\alpha), \quad (18)$$

as well as the roots of the primal-dual gap, which is the non-negative function $g_\alpha : \mathbb{R}^n \times \mathbb{R}^{n \times 2} \rightarrow \mathbb{R}_+$ defined by

$$g_\alpha(u, v) := f(u) + g_\alpha(\nabla u) + f^*(\nabla v) + g^\alpha_*(v), \quad (19)$$

where $f^*$ and $g^\alpha_*$ are the Legendre–Fenchel conjugates of $f$ and $g_\alpha$ respectively, cf., [11, Definition 13.1]. In our case,
\[ f^*(u) = \frac{1}{2} \| u + \xi \|^2 - \frac{1}{2} \| \xi \|^2 \text{ for all } u \in \mathbb{R}^n, \text{ and } g_\alpha^* = I_{B_{\alpha}}(\alpha), \]

setting \( B_{\alpha}(\alpha) := \{ v \in \mathbb{R}^{n \times 2} \mid \| v \|_{\infty,2} \leq \alpha \} \) and \( I_{B_{\alpha}}(\alpha) \) the indicator function of \( B_{\alpha}(\alpha) \), i.e., for all \( v \in \mathbb{R}^{n \times 2} \),

\[ I_{B_{\alpha}}(\alpha)(v) = 0 \text{ if } v \in B_{\alpha}(\alpha) \text{ and } I_{B_{\alpha}}(\alpha)(v) = +\infty \text{ else.} \]

**A. Learning problem**

Our objective is to learn a function \( f : \mathbb{R}^n \rightarrow \mathbb{R}_+ \) that given a degraded image \( \bar{x} \in \mathbb{R}^n \) yields a parameter \( \alpha(\bar{x}) \) such that the solution \( u^{\alpha(\bar{x})}(\bar{x}) \) to (15) is as close as possible to the ground-truth \( u^* \). The problem can be formulated from a standard machine-learning perspective as follows.

Given a dataset \( \mathcal{D} := \{ (u^i, \bar{x}^i) \}_{i=1}^N \), where \( u^i \in \mathbb{R}^n \) is a noise-free image (often referred to as ground-truth) and \( \bar{x}^i \in \mathbb{R}^n \) is its degraded, or noisy, version, and a suitable space of functions \( \mathcal{F} \subset \{ \alpha : \mathbb{R}^n \rightarrow \mathbb{R}_+ \} \), we seek a minimizer of

\[
\min_{\alpha, \mathcal{F}} \frac{1}{N} \sum_{i=1}^N \| u_i^\dagger - u_i^{\alpha(\bar{x})} \|^2, \tag{20}
\]

where \( u_i^{\alpha(\bar{x})} \) is the optimal solution to (15) with data \( \bar{x}^i \) as \( \xi \) and regularization parameter \( \alpha(\bar{x}) \) as \( \alpha \).

Problem (20) has a clear bilevel structure that is not amenable to computation. However, we will see that (20) has an elegant connection with the primal-dual gap (19), which can ultimately be used to design a monopole proxy for (20). In the following result, we show that \( g_\alpha \) in (19) can be equivalently expressed as a sum of Bregman divergences. Recall that a Bregman divergence relative to a proper, convex, lower semicontinuous function \( F \) on a Hilbert space \( H \) is defined for all \( u, u' \in H \) and \( p \in \partial F(u') \) by

\[
\mathcal{D}_p^F(u, u') := F(u) - F(u') - \langle p, u - u' \rangle.
\]

**Theorem 3:** Let \( \alpha > 0 \), \( n \in \mathbb{N} \), let \( \nabla : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times 2} \) be the discrete gradient operator, and \( f, \mathcal{G}_\alpha \) be defined as in (16), then \( \mathcal{G}_\alpha \) according to (19) admits the following expression

\[
\mathcal{G}_\alpha(u, v) = \mathcal{D}_{\nabla v, \alpha}^f(u, u^\alpha) + \mathcal{D}_{\alpha}^f(\nabla \cdot v, \nabla \cdot v^\alpha)
\]

\[
+ \mathcal{G}_{\alpha}^v(v, v^\alpha) + \mathcal{G}_{\alpha}^u(\nabla u, \nabla u^\alpha),
\]

for every \( (u, v) \in \mathbb{R}^n \times \mathbb{R}^{n \times 2} \) and every pair of primal-dual solutions \( (u^\alpha, v^\alpha) \).

**Proof:** The proof follows from straightforward computations, recalling that \( \mathcal{G}_\alpha \) vanishes on any primal-dual solution pair.

In our case, we can see that for all \( \alpha > 0 \),

\[
\mathcal{D}_{\nabla v, \alpha}^f(u, u^\alpha) = \frac{1}{2} \| u - u^\alpha \|^2,
\]

where \( u^\alpha \) is the optimal solution to (15) with parameter \( \alpha \) and data \( \xi \). Indeed, we have

\[
\mathcal{D}_{\nabla v, \alpha}^f(u, u^\alpha) = \frac{1}{2} \| u - \xi \|^2 - \frac{1}{2} \| u^\alpha - \xi \|^2 - \langle \nabla \cdot v^\alpha, u - u^\alpha \rangle
\]

\[
= \frac{1}{2} \| u - \nabla \cdot v^\alpha + \xi \|^2 - \frac{1}{2} \| u - u^\alpha \|^2,
\]

where we repeatedly used that \( \nabla \cdot v^\alpha + \xi = u^\alpha \), see (18).

Therefore, from Theorem 3 we have that, for every noise-free image \( u^\dagger \in \mathbb{R}^n \),

\[
\min_{v \in \mathbb{R}^{n \times 2}} \mathcal{G}_\alpha(u^\dagger, v) = \mathcal{G}_\alpha(u^\dagger, v^\alpha)
\]

\[
= \frac{1}{2} \| u^\dagger - u^\alpha \|^2 + \mathcal{D}_{\alpha}^u(\nabla u^\dagger, \nabla u^\alpha) \geq \frac{1}{2} \| u^\dagger - u^\alpha \|^2.
\]

In particular, for a data point \( (u_i^\dagger, \bar{x}^i) \in \mathcal{D} \), the function

\[
\alpha \mapsto \min_{v \in \mathbb{R}^{n \times 2}} \mathcal{G}_\alpha(u_i^\dagger, v), \tag{21}
\]

majorizes the quadratic distance between the reconstructed image \( u_i^{\alpha(\bar{x})}(\bar{x}) \) and the ground-truth \( u_i^\dagger \) and can be used to turn (20) into the following optimization problem

\[
\alpha \in \mathcal{F}, v_1, \ldots, v_N \in \mathbb{R}^{n \times 2} \quad \text{such that} \quad \| v_i \|_{\infty,2} \leq \alpha(\bar{x}) \quad \text{for all} \quad i \in \{1, \ldots, N\}.
\]

**B. Model selection**

It remains to fix \( \mathcal{F} \). Here, different choices can be made. In this paper, we investigate the performance of quadratic models, i.e., with \( \alpha(\bar{x}) = \frac{\bar{x}}{\alpha} \), where \( \bar{x} = [\bar{x}, 1]^\top \) and \( \alpha \) is a symmetric positive semidefinite matrix of size \( (n + 1)^2 \).

With this choice, problem (22) turns into the following convex problem

\[
\min_{(\alpha, v) \in \mathcal{C}_q} \frac{1}{N} \sum_{i=1}^N \| \nabla \cdot v_i + \xi_i \|^2 + \frac{1}{N} \sum_{i=1}^N \bar{x}_i A_i \bar{x}_i \text{TV}(u_i^\dagger), \tag{23}
\]

where \( \mathcal{C}_q \) is the subset of \( \mathcal{F} \times \mathbb{R}^{n \times 2} \) of all positive semidefinite matrices \( A \in \mathbb{R}^{(n+1) \times (n+1)} \) and all \( v = (v_1, \ldots, v_N) \in \mathbb{R}^{n \times 2} \) such that \( \| v_i \|_{\infty,2} \leq \bar{x}_i \text{TV}(u_i^\dagger) \) for all \( i \).

**C. Optimization procedure**

To solve (23) we employ the HPGCC method with

\[
f(v, A) := \frac{1}{2N} \sum_{i=1}^N \| \nabla \cdot v_i + \xi_i \|^2 + \frac{1}{N} \sum_{i=1}^N \bar{x}_i A_i \bar{x}_i \text{TV}(u_i^\dagger),
\]

\[
g(v, A) := \I_{\mathcal{C}_q}(v, A).
\]

Further, we pick \( \| (v, A) \|_F^2 = \lambda \| A \|_F^2 \) with \( \lambda > 0 \), where \( \| \cdot \|_F \) is the Frobenius norm, which from now on we will simply write as \( \| \cdot \| \). Note that, here, we cannot employ the generalized conditional gradient method due to the lack of coercivity of \( g \). Also, the projection onto the constraint set \( \mathcal{C}_q \) does not admit an explicit expression and can only be computed approximately with possibly time-consuming inner procedures. Such a bottleneck can undermine the convergence performance of standard proximal methods that require the computation of the projection onto \( \mathcal{C}_q \). The HPGCC method allows us to circumvent this issue and leads to a low-complexity iterative method, as no projections onto \( \mathcal{C}_q \) would be required.

To see this, recall that at each iteration, given \( v \) and \( A \), we need to solve

\[
\min_{v, A} \langle \nabla_r f(v, A), \hat{v} \rangle + \langle \nabla_A f(v, A) - \lambda A, \hat{A} \rangle + \frac{\lambda}{2} \| \hat{A} \|^2 + g(\hat{v}, \hat{A}),
\]
Data: \( \{ (u_i^+, \xi_i) \}_{i=1}^N \) and \( \lambda > 0 \)
Return: \( A^\infty = \lim_{k \to \infty} A^k \)
Initialize: \( v_0, \ldots, v_0 \in \mathbb{R}^{n \times 2}, A_0 \in \mathbb{R}^{(n+1)^2} \) with \( A_0 \) an Intel Core i7-9700 CPU@3.00GHz.

Data and code can be found at https://github.com/TraDE-OPT/TV-parameter-learning.

Eventually, the proposed HPGCG method when applied to problem (23) turns into the iterative method illustrated in Algorithm 2. Note, in particular, that in Algorithm 2, by Theorem 2, we can expect convergence of \( \{ A^k \}_k \).

D. Numerical experiments

In this section, we present our numerical experiments\(^1\).

We train our model considering a dataset of \( N = 101440 \) patches of size \( n = p \times p \) with \( p = 16 \) extracted from 1109 cartoon images. Working with patches allows us to consider significantly more data points than degrees of freedom, which are of order \( O(n^2) \). This would allow us to avoid overfitting phenomena. To each patch, we apply a Gaussian noise of variance 0.05. We set \( \lambda = 50 \) and run Algorithm 2 choosing \( D_f(u, v) = L/2 \| u - v \|_2^2 \) with \( L = 8/N \).

Remark 2: Note that in Algorithm 2 at iteration \( k \), for (15), the term \( \nabla \cdot v_i^k + \xi_i \) is an approximation of the TV-denoised \( p^k \) patch, which can be used to monitor the reconstruction quality online, see Figure 3.

At every iteration, in order to update the step-size we should compute the residual \( D(A^k, v^k) \), which can also be employed for a stopping rule. Specifically, we stop the iteration as soon as \( D(A^k, v^k) < 10^{-4} \). The residual as a function of the iteration number is shown in Figure 1.

**Experiment 1.** We employ the trained model to denoise a new test image, which we split into 16 × 16 patches. For every single patch \( \xi_i \) we compute the TV-parameter by \( \xi_i^k, \xi_i \) solves (25) with our trained model or given by one of the constants \( N = 1 \) and \( \mathcal{F} \) composed of only non-negative constants, for 100000 iterations (up to a residual of \( \sim 10^{-5} \)). Then, we measure the Mean Squared Error, i.e.,

\[
\text{MSE}_a := \frac{1}{N_i} \sum_{i=1}^{N_i} (\alpha_i^* - \alpha(\xi_i))^2.
\]

We also compute the MSE with respect to eight constant choices spaced evenly from \( 10^{-4} \) to \( 10^{-1} \), i.e., (26) replacing \( \alpha(\xi_i) \) with these constant values. Further, we consider a constant model trained on 1000 patches extracted from the same training set via HPGCG, with \( \lambda = 50 \) and stopped as soon as the residual drops below \( 10^{-5} \) (reached in 15479 iterations). The constant model yielded a value of \( \alpha = 2.713 \times 10^{-2} \). Eventually, for each parameter choice (computed with our trained model or given by one of the constants above), we also measure the Mean Squared Error relative to the reconstructed images, namely

\[
\text{MSE}_u := \frac{1}{N_i} \sum_{i=1}^{N_i} \| u_i^k - u^*(\xi_i) \|_2^2.
\]

\(^1\)All computations were carried out in Python on a PC with 62 GB RAM and an Intel Core i7-9700 CPU@3.00GHz. Data and code can be found at https://github.com/TraDE-OPT/TV-parameter-learning.
or (27) replacing $\alpha(\xi)$ with the above constant values. The five best results are contained in Table I.

| Models | Quadratic | Constant $\alpha = \eta \cdot 10^{-3}$ |
|--------|-----------|--------------------------------------|
|        |           | $\eta = 13.9$ $\eta = 27.13$ $\eta = 37.3$ $\eta = 100$ |
| MSE$_p$ | 3.39 $10^{-4}$ | 9.79 $10^{-4}$ $4.95$ $10^{-4}$ $3.62$ $10^{-4}$ $4.08$ $10^{-4}$ |
| MSE$_u$ | 0.1529 | 0.2917 0.1833 0.1777 0.4764 |

**TABLE I**

**Results.** From Figure 1 we can see that HPGCG, before entering into a sub-linear regime, is able to quickly reach high precision within about a few hundred iterations. From Table I we can also see that the proposed model yields very accurate parameter choices and performs better than constant models in terms of MSE both relative to the parameter choice and the image reconstruction.

**CONCLUSION**

Our research paper presents a new hybrid method that combines the desirable properties of classical forward–backward and generalized conditional gradient methods. In fact, the proposed method turns into a powerful scheme to solve a new TV parameter learning model without requiring the costly projection onto the convex set $\mathcal{C}_p$, while keeping convergence guarantees for $\{A^k\}_k$. Future research avenues under discussion include devising new stochastic and inexact variants for HPGCG, inspecting different models in the TV parameter learning problem, and further investigating the proposed learning framework for a broader range of applications.

**ACKNOWLEDGMENTS**

K.B. and E.C. have received funding from the European Union’s Framework Programme for Research and Innovation Horizon 2020 (2014–2020) under the Marie Sklodowska-Curie Grant Agreement No. 861137. The Institute of Mathematics and Scientific Computing, to which K.B. and E.C. are affiliated, is a member of NAWI Graz (https://www.nawigraz.at/).

**REFERENCES**

[1] Bredies, K., Lorenz, D. & Maass, P. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Computational Optimization and Applications*. 42, 173–193 (2009).

[2] Bredies, K., Carioni, M., Fanzon, S. & Romero, F. A generalized conditional gradient method for dynamic inverse problems with optimal transport regularization. *Foundations of Computational Mathematics*. (2022).

[3] Bredies, K., Carioni, M., Fanzon, S. & Walter, D. Asymptotic linear convergence of fully-corrective generalized conditional gradient methods. arXiv:2110.06756. (2021).

[4] Bredies, K. & Pikkarainen, H. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*. 19, 190–218 (2013).

[5] Bredies, K. & Lorenz, D. Iterated hard shrinkage for minimization problems with sparsity constraints. *SIAM Journal on Scientific Computing*. 30, 657–683 (2008).

[6] Bredies, K., Chenchene, E., Lorenz, D. & Naldi, E. Degenerate preconditioned proximal point algorithms. *SIAM Journal on Optimization*. 32, 2376–2401 (2022).

[7] Polyak, B. T. Introduction to optimization. *Optimization Software Inc.*, New York. (1987).

[8] De los Reyes, J. C., Schönlieb, C. B. & Valkonen, T. Bilevel parameter learning for higher-order total variation regularisation models. *Journal of Mathematical Imaging and Vision*. 57, 1–25, (2017).

[9] Rudin, L., Osher, S. & Fatemi, E. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*. 60, 259–268 (1992).

[10] Boyd, S. & Vandenberghe, L. Convex optimization. *Cambridge University Press*. (2004).

[11] Bauschke, H. & Combettes, P. Convex analysis and monotone operator theory in Hilbert spaces. *Springer New York*. (2011).

[12] Chambolle, A. & Pock, T. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*. 40, 120–145 (2011).

[13] Bauschke, H., Borwein, J. & Combettes, P. Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces. *Communications in Contemporary Mathematics*. 3, 615–647 (2001).

[14] Knopp, K. Theory and application of infinite series. *Dover Publications*. (1990).

[15] Lions, P. & Mercier, B. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*. 16, 964–979 (1979).

[16] Bonne, I., Rinaldi, F. & Zeffiro, D. Frank–Wolfe and friends: a journey into projection-free first-order optimization methods. *4OR*. 19, 313–345 (2021).

[17] Combettes, P. & Pokutta, S. Complexity of linear minimization and projection on some sets. *Operations Research Letters*. 49, 565–571 (2021).

[18] Beck, A. & Teboulle, M. A Fast Iterative Shrinkage-Thresholding Algorithm for linear inverse problems. *SIAM Journal On Imaging Sciences*. 2, 183–202 (2009).

[19] Wright, S., Nowak, R. & Figueiredo, M. Sparse reconstruction by separable approximation. *IEEE Transactions On Signal Processing*. 57, 2479–2493 (2009).

[20] Gabidullina, Z. Adaptive conditional gradient method. *Journal Of Optimization Theory And Applications*. 183, 1077–1098 (2019).

[21] Silvetti-Falls, A., Molinari, C. & Fadili, J. Inexact and stochastic generalized conditional gradient with augmented lagrangian and proximal step. *Journal Of Nonsmooth Analysis And Optimization*. 2, (2021).

[22] Rosasco, L., Villa, S. & Vü, B. Stochastic forward-backward splitting for monotone inclusions. *Journal Of Optimization Theory And Applications*. 169, 388–406 (2016).