On Structured Filtering-Clustering: Global Error Bound and Optimal First-Order Algorithms

Nhat Ho[*,‡]   **Tianyi Lin**[*,⋄]   **Michael I. Jordan**[⋄]

University of California, Berkeley[⋄]
University of Texas, Austin[‡]

]

Abstract

The filtering-clustering models, including trend filtering and convex clustering, have become an important source of ideas and modeling tools in machine learning and related fields. The statistical guarantee of optimal solutions in these models has been extensively studied yet the investigations on the computational aspect have remained limited. In particular, practitioners often employ the first-order algorithms in real-world applications and are impressed by their superior performance regardless of ill-conditioned structures of difference operator matrices, thus leaving open the problem of understanding the convergence property of first-order algorithms. This paper settles this open problem and contributes to the broad interplay between statistics and optimization by identifying a *global error bound* condition, which is satisfied by a large class of dual filtering-clustering problems, and designing a class of *generalized dual gradient ascent* algorithm, which is *optimal* first-order algorithms in deterministic, finite-sum and online settings. Our results are new and help explain why the filtering-clustering models can be efficiently solved by first-order algorithms. We also provide the detailed convergence rate analysis for the proposed algorithms in different settings, shedding light on their potential to solve the filtering-clustering models efficiently. We also conduct experiments on real datasets and the numerical results demonstrate the effectiveness of our algorithms.

1 Introduction

We are interested in the filtering-clustering models which are given by

$$\min_{\beta \in \mathbb{R}^d} \ \Phi(\beta) := f(\beta) + \lambda \left(\sum_{k=1}^{n} \|D_k \beta\|_p \right), \qquad (1.1)$$

where $f : \mathbb{R}^d \mapsto \mathbb{R}$ is a strongly convex loss function and has Lipschitz continuous gradient, $D_k \in \mathbb{R}^{m \times d}$ are called *discrete difference operator* matrices for $1 \leq k \leq n$, and $\lambda > 0$ is a regularization parameter and $p \in \mathbb{N}^+$ is a regularization index. Common applications of the filtering-clustering models in Eq. (1.1), such as ℓ_1-trend filtering and ℓ_2-convex clustering, are associated with squared Euclidean loss function f and poorly conditioned matrices D_k. Such ill-conditioning poses several challenges for first-order algorithms to achieve fast and stable convergence.

The filtering-clustering models, including trend filtering and convex clustering, cover a wide range of application problems arising from machine learning and statistics. Examples of trend filtering applications include nonparametric regression (Kim et al., 2009; Tibshirani, 2014; Lin et al., 2017; Padilla et al., 2018a; Guntuboyina et al., 2020), adaptive estimators in graphs (Wang et al., 2016b; Padilla et al., 2018b), and time series analysis (Leser, 1961). Convex clustering has been proposed as an alternative to traditional clustering approaches, e.g., K-means clustering and hierarchical clustering, and was well known for its appealing robustness and stability properties (Hocking et al., 2011; Zhu et al., 2014; Tan and Witten, 2015; Wu et al., 2016; Radchenko and Mukherjee, 2017).

Recent years have witnessed much progress on the statistical aspects of filtering-clustering models. Indeed, the global solutions of these models have been shown to enjoy the desirable properties (Tibshirani, 2014; Zhu et al., 2014; Tan and Witten, 2015; Wang et al., 2016b; Wu et al., 2016; Radchenko and Mukherjee, 2017; Padilla et al., 2018b; Guntuboyina et al., 2020).

However, investigations on the computational theory for the filtering-clustering models is relatively scarce. Indeed, there have been a flurry of optimization algorithms that were proposed for solving the filtering-clustering models in the literature. For example, the trend filtering algorithms include primal-dual interior-point method (PDIP) (Kim et al., 2009), alternating direction method of multipliers (ADMM) (Ramdas and Tibshirani, 2016) and specialized Newton's method (Wang et al., 2016b). The convex clustering algorithms include ADMM, alternating minimization algorithm (AMA) (Chi and Lange, 2015), projected dual gradient ascent (Wang et al., 2016a) and semismooth Newton's method (Sun et al., 2021). Most of these existing approaches were built on first-order algorithmic frameworks and have been recognized as the benchmark in the literature due to their simplicity and ease-of-the-implementation. Practitioners employ them in real-world applications and often observe superior performance (Chi and Lange, 2015; Ramdas and Tibshirani, 2016; Wang et al., 2016a) regardless of ill-conditioned structures of difference operator matrices. This is somehow surprising since the first-order algorithms are known to suffer from the slow sublinear convergence rate for general convex problems, thus leaving open the problem of understanding the convergence property of the first-order algorithms as applied to the filtering-clustering models. To our knowledge, there is currently a paucity of computational theory that can uncover the mysterious success of these first-order algorithms and further lead to better algorithms.

Contributions. In this paper, we study the computational theory concerning the filtering-clustering models. Our contributions are summarized as follows:

1. We analyze the structure of the filtering-clustering model in Eq. (1.1) and prove that a *global* error bound condition holds for the dual filtering-clustering model when $p = 1$ or $p \in [2, +\infty]$. Our results are nontrivial: first of all, the filtering-clustering model is not amenable to the existing techniques developed in Wang and Lin (2014) since the non-smooth term in the objective function of the dual form does not a polyhedral epigraph when $p \in [2, +\infty)$. Second, the filtering-clustering model in (1.1) can not be formulated as an $\ell_{1,p}$-regularized problem for some $p \geq 1$. As such, our results are not a straightforward consequence of Zhou et al. (2015).

2. We propose a class of first-order primal-dual optimization algorithms for solving the filtering-clustering model with an optimal linear convergence rate. There are two fundamental reasons for the non-triviality of the result: (i) The objective function of the filtering-clustering model is non-strongly convex since $D_k^\top$ are not full column rank; (ii) the gradient of the objective function of the dual filtering-clustering model is inaccessible.

3. In addition to deterministic counterparts, we propose and analyze a class of stochastic first-order primal-dual optimization algorithms for solving the filtering-clustering model in Eq. (1.1). In the finite-sum setting, the proposed stochastic variance reduced first-order algorithms attain an optimal linear convergence rate. In the online setting, we conduct a similar analysis and show that our algorithm achieve the optimal rate up to log factors. Notably, our algorithms are build on celebrated gradient-based algorithms from Allen-Zhu (2017) and Rakhlin et al. (2012).

Notation. We denote $[n]$ as to the set $\{1, 2, \ldots, n\}$. For $p \in [1, +\infty]$, the notion $\|\cdot\|_p$ denotes ℓ_p-norm and $\|\cdot\|$ denotes the Euclidean norm for a vector and the operator norm for a matrix. For all $q \geq 1$, $\mathbb{B}_q = \{\alpha \in \mathbb{R}^m \mid \|\alpha\|_q \leq 1\}$ refers to a ℓ_q-norm unit ball in $\mathbb{R}^m$ and $\bar{D}_q = \max_{\mathbf{x},\mathbf{x}' \in \mathbb{B}_q} \|\mathbf{x} - \mathbf{x}'\|$ refers to a diameter of ℓ_q-norm unit ball in ℓ_2-norm. For simplicity, we let $\mathbb{B}_q^n$ denote the product of n unit balls in ℓ_q-norm. For a convex function f, ∂f refers to the subdifferential of f. If f is differentiable, $\partial f = \{\nabla f\}$ where ∇f is the gradient vector of f. For any closed set $\mathcal{S}$, we let $d(\mathbf{x}, \mathcal{S})$ denote the distance between $\mathbf{x}$ and $\mathcal{S}$. If $\mathcal{S}$ is convex, we let $\mathcal{N}_\mathcal{S}(\mathbf{x})$ denote the normal cone to $\mathcal{S}$ at $\mathbf{x}$. Lastly, given a tolerance $\varepsilon \in (0, 1)$, the notation $n = \Omega(b(\varepsilon))$ and $n = O(b(\varepsilon))$ stand for lower bound $n \geq c_1 \cdot b(\varepsilon)$ and upper bound $n \leq c_2 \cdot b(\varepsilon)$ where $c_1, c_2 > 0$ are independent of $1/\varepsilon$. In addition, $n = \Theta(b(\varepsilon))$ means that upper and lower bounds hold true.

2 Related Work

Regarding the computation theory for deterministic first-order optimization algorithms, the best possible that we can expect is linear rate of convergence (Nesterov, 2018). Such convergence result is commonly established under some additional assumptions on problem structure; e.g., strong convexity, global and local error bound (Pang, 1987, 1997; Zhou and So, 2017; Drusvyatskiy and Lewis, 2018) and restricted strongly convexity (Negahban et al., 2012; Negahban and Wainwright, 2012). The former two assumptions are standard in the optimization literature and the error bound condition can be interpreted as a relaxation of strong convexity. Local error bound provides a guarantee for the *asymptotic* linear convergence of feasible descent algorithms (Luo and Tseng, 1992, 1993; Tseng, 2010), conditional gradient method (Beck and Shtern, 2017), ADMM (Hong and Luo, 2017), and proximal gradient

Nhat Ho[*,‡], Tianyi Lin[*,◊], Michael I. Jordan[◊]

algorithms (Drusvyatskiy and Lewis, 2018). Many application problems were shown to satisfy error bound conditions. For example, Zhou et al. (2015) proved local error bound condition for $\ell_{1,p}$-norm regularized problems and Wang and Lin (2014) derived a clean form of global error bound for a class of nonstrongly convex problems. In addition, Karimi et al. (2016) established the linear convergence of proximal gradient algorithms in various non-strongly convex settings under the Polyak-Lojasiewicz condition (another relaxation of strong convexity). Very recently, Jane et al. (2021) have introduced a new framework based on the error bound, which enables us to provide some sufficient conditions for linear convergence and applicable approaches for calculating linear convergence rates of these first-order algorithms for a class of structured convex problems. On the other hand, the restricted strongly convexity played an important role in statistical learning literature and forms the basis for deriving optimal rate of first-order algorithms as applied to high-dimensional statistical recovery problems with sparsity-induced regularization (Agarwal et al., 2012; Wang et al., 2014; Loh and Wainwright, 2015). However, the aforementioned works do not contain (or imply) *any global error bound analysis* for the filtering-clustering models in Eq. (1.1) due to the regularization term $\sum_{k=1}^{n} \|D_k\beta\|_p$ and thus can not explain the success of first-order algorithms, e.g., projected dual gradient ascent and ADMM, for solving the filtering-clustering models.

Another line of relevant work focuses on first-order primal-dual optimization algorithms for convex-concave saddle-point problems; see, e.g., Chen and Rockafellar (1997); Palaniappan and Bach (2016); Wang and Xiao (2017); Zhang and Xiao (2017) and the references therein. More specifically, some works have derived the linear convergence results if the model is associated with either a strongly convex-concave structure (Chen and Rockafellar, 1997; Palaniappan and Bach, 2016), together with efficient computed proximal mappings for nonsmooth terms. Nevertheless, these assumptions are not satisfied by the filtering-clustering models in Eq. (1.1). An alternative way to establish the linear convergence of first-order primal-dual optimization algorithms is based on the construction of a potential function which decreases at a linear rate (Wang and Xiao, 2017; Zhang and Xiao, 2017); however, it remains open how to construct such function for the filtering-clustering models. In addition, our algorithm is an inexact first-order primal-dual algorithm and thus resembles a generic inexact proximal algorithms (Schmidt et al., 2011). However, their analysis is conducted under strong convexity and can not be extended to the filtering-clustering models.

3 Preliminaries

We flesh out the basic filtering-clustering model in Eq. (1.1) and then turn to specific examples of this model. We also discuss primal and dual forms of the filtering-clustering model and demonstrate that its structure is suitable for the development of first-order primal-dual optimization algorithms.

3.1 Filtering-clustering model

The goal of this paper is to find a way to efficiently compute an optimal solution of the filtering-clustering model in Eq. (1.1). Formally, we have

Definition 3.1 *A point $\beta^\star \in \mathbb{R}^d$ is an optimal solution of the filtering-clustering model in Eq. (1.1) if $\Phi(\beta^\star) \leq \Phi(\beta)$ for all $\beta \in \mathbb{R}^d$.*

Since the convergence of first-order optimization algorithms to an optimal solution will depend on the gradient of the loss function f in its neighborhood, it is necessary to impose Lipschtiz continuity conditions on the gradient ∇f. We also impose the strong convexity on f and remark that this assumption is standard in the filtering-clustering model arising from real-world application problems (Kim et al., 2009; Chi and Lange, 2015; Ramdas and Tibshirani, 2016). However, we notice that $\sum_{k=1}^{n} \|D_k\beta\|_p$ is nonsmooth and does not admit an efficiently computed proximal mapping (even when $p = 1$). Thus, the proximal gradient algorithms can not be directly applied to the filtering-clustering model in Eq. (1.1) with linear convergence.

Definition 3.2 *f is ℓ-gradient Lipschitz if $\forall \beta, \beta' \in \mathbb{R}^d$, $\|\nabla f(\beta) - \nabla f(\beta')\| \leq \ell \|\beta - \beta'\|$.*

Definition 3.3 *f is μ-strongly convex if $\forall \beta, \beta' \in \mathbb{R}^d$, $\|\nabla f(\beta) - \nabla f(\beta')\| \geq \mu \|\beta - \beta'\|$[1].*

In online setting, we impose unbiased and boundedness conditions on the stochastic gradient oracle that has become standard in the literature.

Definition 3.4 *$g(\cdot, \xi)$ is unbiased and bounded if for $\forall \beta \in \mathbb{R}^d$, $\mathbb{E}[g(\beta, \xi)] = \nabla f(\beta)$ and $\mathbb{E}[\|g(\beta, \xi)\|^2] \leq G^2$ for a constant $G > 0$.*

Assumption 3.1 *The loss function f is ℓ-gradient Lipschitz and μ-strongly convex. The stochastic gradient oracle $g(\cdot, \xi)$ is unbiased and bounded. The optimal set is nonempty.*

---

[1]The definition here is a consequence of the standard definition in the textbook (Nesterov, 2018); that is, $(\beta - \beta')^\top (\nabla f(\beta) - \nabla f(\beta')) \geq \mu \|\beta - \beta'\|^2$. Nevertheless, we define it like this for simplicity.

Under Assumption 3.1, the objective function Φ in Eq. (1.1) is *non-smooth but strongly convex* and at least one optimal solution exists. Thus, the filtering-clustering model has a unique optimal solution $\beta^\star$. In general, a first-order optimization algorithm can not return an exact optimal solution in finite time. As such, we define an ε-optimal solution of the filtering-clustering model in Eq. (1.1).

Definition 3.5 $\beta \in \mathbb{R}^d$ *is an ε-optimal solution of the filtering-clustering model in Eq. (1.1) if $\|\beta - \beta^\star\|^2 \leq \varepsilon$ given that $\beta^\star$ is a unique optimal solution.*

With these notions in mind, one opts to develop first-order algorithms in that the required number of gradient evaluations to return an ε-optimal solution has the logarithmic dependence on $1/\varepsilon$ (deterministic or finite-sum setting) and the linear dependence on $1/\varepsilon$ (online setting) under Assumption 3.1.

3.2 Specific instances

We provide some typical examples of the filtering-clustering model in real-world application and give an overview of the existing algorithms.

ℓ_1-trend filtering (Kim et al., 2009; Tibshirani, 2014) has been recently recognized as a benchmark approach to nonparametric regression in the area of statistical learning. The problem setup is given as follows: for $1 \leq i \leq \bar{n}$, we have a few input/response pairs $(\mathbf{x}_i, y_i)$ such that $y_i = f_0(\mathbf{x}_i) + w_i$ where $w_1, \ldots, w_{\bar{n}}$ are independent and identically distributed random variables. Given an integer $k \geq 0$, the k-th order ℓ_1-trend filtering is implemented by solving the ℓ_1-regularized least-squares problem:

$$\min_{\beta \in \mathbb{R}^{\bar{n}}} \ \frac{1}{2}\|y - \beta\|^2 + \lambda\|D^{(k+1)}\beta\|_1, \qquad (3.1)$$

where $\lambda \geq 0$ is a regularization parameter and $D^{(k+1)} \in \mathbb{R}^{(\bar{n}-k-1)\times\bar{n}}$ is a discrete difference operator of order $k + 1$. For example, we have

$$D^{(1)} = \begin{bmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{bmatrix}.$$

In general, the nonzero entries in each row of the matrix $D^{(k+1)}$ are the $(k + 1)$-th row of Pascal's triangle with alternating signs. When $k = 0$, the ℓ_1-trend filtering model in Eq. (3.1) can be interpreted as a special instance of 1-dimensional total variation denoising (Rudin et al., 1992) and fused Lasso (Tibshirani et al., 2005). Two classical algorithms for solving ℓ_1-trend filtering model include primal-dual interior-point

algorithm (PDIP) (Kim et al., 2009) and alternating direction of multiplier method (ADMM) (Ramdas and Tibshirani, 2016). Despite their superior practical performance, these algorithms lack theoretical guarantees. In fact, we are not aware of any provably linearly convergent first-order algorithms that have been proposed for solving ℓ_1-trend filtering model in Eq. (3.1).

Graph ℓ_1-trend filtering (Wang et al., 2016b) can be seen as an extension of ℓ_1-trend filtering to graph problems. The problem setup is given as follows: let $G = (V, E)$ denote a graph consisting of a set of nodes $V = \{1, 2, \ldots, \bar{n}\}$ and undirected edges $E = (e_1, \ldots, e_{\bar{m}})$. Given $k \geq 0$ and a few inputs associated with the nodes, $y = (y_1, \ldots, y_{\bar{n}}) \in \mathbb{R}^{\bar{n}}$, the k-th order graph ℓ_1-trend filtering is implemented by solving the following ℓ_1-regularized least squares problem:

$$\min_{\beta \in \mathbb{R}^{\bar{n}}} \ \frac{1}{2}\|y - \beta\|^2 + \lambda\|\Delta^{(k+1)}\beta\|_1, \qquad (3.2)$$

where $\lambda \geq 0$ is a regularization parameter and $\Delta^{(k+1)} \in \mathbb{R}^{\bar{m}\times\bar{n}}$ is a discrete graph difference operator of order $k + 1$. Concretely, the explicit form of each row of $\Delta^{(1)}$ is given by: (if $e_l = (i, j)$ for $1 \leq l \leq m$)

$$\Delta_l^{(1)} = (0, \ldots, \underbrace{-1}_{i}, \ldots, \underbrace{1}_{j}, \ldots, 0).$$

Recursively, $\Delta^{(k+1)}$ can be represented by

$$\Delta^{(k+1)} = \begin{cases} (\Delta^{(1)})^\top \Delta^{(k)} & \text{if } k \text{ is odd,} \\ \Delta^{(1)}\Delta^{(k)} & \text{if } k \text{ is even.} \end{cases}$$

When $k = 0$, the graph ℓ_1-trend filtering model in Eq. (3.2) can be interpreted as a special instance of graph fused Lasso (Tibshirani and Taylor, 2011). Two popular algorithms for solving graph ℓ_1-trend filtering model include ADMM and specialized Newton's method (Wang et al., 2016b). However, the theoretical convergence guarantee for these algorithms is missing, and it remains open whether a linearly convergent first-order algorithm exists for solving graph ℓ_1-trend filtering model in Eq. (3.2) or not.

ℓ_2-convex clustering has been proposed as an effective alternative to classical clustering approaches in the literature and can be represented as a convex optimization model (Hocking et al., 2011). The problem setup is given as follows: given a number of inputs $\{x_1, x_2, \ldots, x_{\bar{n}}\} \subseteq \mathbb{R}^{\bar{n}}$, the ℓ_2-convex clustering is implemented by solving the following ℓ_2-regularized least square problem:

$$\min_{\{\beta_1,\ldots,\beta_{\bar{n}}\}} \frac{1}{2}\left(\sum_{i=1}^{\bar{n}}\|x_i - \beta_i\|^2\right) + \lambda\left(\sum_{1\leq i<j\leq\bar{n}} w_{ij}\|\beta_i - \beta_j\|_2\right), \qquad (3.3)$$

Nhat Ho[*,‡], Tianyi Lin[*,◇], Michael I. Jordan[◇]

where λ and $\{w_{ij}\}_{1 \leq i < j \leq \bar{n}}$ are both positive parameters. The benchmark first-order algorithms that have been used in practice for solving ℓ_2-convex clustering model in Eq. (3.3) include ADMM and alternating minimization algorithm (AMA) (Chi and Lange, 2015). However, these algorithms are mostly heuristic and lack the solid theoretical guarantee. As an alternative, Sun et al. (2021) have proposed to use the second-order algorithm (e.g., semismooth Newton method) for solving ℓ_2-convex clustering model and proved a linear convergence under certain condition. To the best of our knowledge, there are no provably linearly convergent first-order algorithms that have been proposed for solving ℓ_2-convex clustering model in Eq. (3.3).

3.3 Primal-dual filtering-clustering model

We give an alternative form of the filtering-clustering model in Eq. (1.1) by starting with a convex-concave saddle-point formulation. Suppose $q = \frac{p}{p-1}$, we reformulate the filtering-clustering model equivalently as

$$\min_{\beta \in \mathbb{R}^d} \max_{\alpha \in \mathbb{B}_q^n} \ f(\beta) - \lambda \alpha^\top D \beta, \quad \alpha := \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}, D := \begin{bmatrix} D_1 \\ \vdots \\ D_n \end{bmatrix},$$
(3.4)

where $\mathbb{B}_q^n$ is the product of n unit balls in ℓ_q-norm.

The saddle-point formulation in Eq. (3.4) is different from the existing saddle-point formulation in Lan and Zhou (2017) and Zhang and Xiao (2017). To be more specific, the formulation in the previous works as applied to the filtering-clustering model in Eq. (1.1) is given by

$$\min_{\beta \in \mathbb{R}^d} \max_{\alpha \in \mathbb{R}^d} \ \alpha^\top \beta - f^\star(\alpha) + \lambda \left(\sum_{k=1}^{n} \|D_k \beta\|_p \right).$$

where $f^\star : \mathbb{R}^d \to \mathbb{R}$ is the convex conjugate of the loss function f; see Rockafellar (2015) for the definition. Here the proximal mapping of $\sum_{k=1}^{n} \|D_k \beta\|_p$ can not be efficiently computed in the filtering-clustering model (Parikh et al., 2014). Therefore, the algorithms developed in Lan and Zhou (2017) and Zhang and Xiao (2017) are not suitable for solving the filtering-clustering model in Eq. (1.1). Second, the saddle-point formulation in Eq. (3.4) also differs from the existing linearly constrained formulation in Ramdas and Tibshirani (2016). In particular, the formulation in Ramdas and Tibshirani (2016) as applied to the filtering-clustering model in Eq. (1.1) is given by

$$\min_{\beta \in \mathbb{R}^d, \alpha_k \in \mathbb{R}^m} \ f(\beta) + \lambda \left(\sum_{k=1}^{n} \|\alpha_k\|_p \right), \quad \text{s.t. } \alpha_k = D_k \beta.$$

Based on the convex-concave saddle point formulation in Eq. (3.4), we obtain the dual filtering-clustering

model as follows,

$$\min_{\alpha \in \mathbb{B}_q^n} \ \bar{f}(\alpha) := f^\star(\lambda D^\top \alpha),$$
(3.5)

where $f^\star : \mathbb{R}^d \mapsto \mathbb{R}$ is the convex conjugate of the loss function f. As such, the dual filtering-clustering model in Eq. (3.5) is derived by,

$$\min_{\beta \in \mathbb{R}^d} \max_{\alpha \in \mathbb{B}_q^n} \ f(\beta) - \lambda \alpha^\top D \beta \iff \max_{\alpha \in \mathbb{B}_q^n} \min_{\beta \in \mathbb{R}^d} \ f(\beta) - \lambda \alpha^\top D \beta$$

$$\iff \max_{\alpha \in \mathbb{B}_q^n} \ -f^\star(\lambda \alpha^\top D) \iff \min_{\alpha \in \mathbb{B}_q^n} \ f^\star(\lambda \alpha^\top D).$$

Notably, the above dual filtering-clustering model is amenable to structure analysis and algorithmic design. Indeed, $f^\star$ is a smooth and strongly convex function and $\mathbb{B}_q$ is a structured and bounded convex set with efficient projection when $q = 1, 2, +\infty$. In the sequel, we demonstrate that a global error bound condition is satisfied for the dual filtering-clustering model in Eq. (3.5), making it possible to develop a class of first-order optimization algorithms with desirable convergence guarantee in different settings.

4 Global Error Bound Condition

We prove that the global error bound (GEB) condition holds for the dual filtering-clustering model in Eq. (3.5). Indeed, we investigate the special structure of the dual filtering-clustering model and use them to prove that the GEB condition holds for the dual filtering-clustering model when $q \in [1, 2] \cup \{+\infty\}$ under certain conditions, which correspond to the filtering-clustering model in Eq. (1.1) with $p \in \{1\} \cup [2, +\infty]$. All the proof details are deferred to the appendix.

4.1 Problem structure

We investigate the special structure of the dual filtering-clustering model concerning the objective function and the optimal set. Before our formal analysis, we recall some notations:

$$\begin{aligned} f^\star(z) &= \min_{\beta \in \mathbb{R}^d} \ f(\beta) - z^\top \beta, \\ \beta^\star(\alpha) &= \operatorname*{argmin}_{\beta \in \mathbb{R}^d} \ f(\beta) - \lambda \alpha^\top D \beta, \\ \bar{f}(\alpha) &= f^\star(\lambda D^\top \alpha). \end{aligned}$$

where $f^\star$ is the convex conjugate of the loss function f and $\bar{f}$ is the objective function of the dual filtering-clustering model in Eq. (3.5).

Lemma 4.1 *Under Assumption 3.1, $f^\star$ is $\frac{1}{\mu}$-gradient Lipschitz and $\frac{1}{\ell}$-strongly convex.*

Lemma 4.2 *Under Assumption 3.1, $\beta^\star$ is a Lipschitz function over $\mathbb{B}_q^n$ with parameter $\frac{\lambda \|D\|}{\mu}$.*

Lemma 4.3 *Under Assumption 3.1, $\bar{f}$ is differentiable and $\frac{\lambda^2\|D\|^2}{\mu}$-gradient Lipschitz.*

From the above lemmas, we see that $f^\star$ and $\bar{f}$ have favorable structures from an algorithmic point of view.

4.2 GEB condition and ULC property

We introduce the notion of *upper Lipschitz continuity* (ULC) and use it to prove a sufficient condition for the dual filtering-clustering model in Eq. (3.5) to satisfy the global error bound (GEB) condition. In particular, we first introduce the quantity $d(\alpha, \Omega^\star)$ that measures the distance between a point α and the optimal set $\Omega^\star$. Note that this quantity is generally not accessible since $\Omega^\star$ is unknown. As an alternative, we consider a function $R : \mathbb{R}^{mn} \to \mathbb{R}^{mn}$, which we refer to as *residual function*. Formally, it is given by

$$R(\alpha) := \mathcal{P}_{\mathbb{B}_q^n}(\alpha - \nabla\bar{f}(\alpha)) - \alpha. \qquad (4.1)$$

It is easy to verify that $R(\alpha) = 0$ if and only if $\alpha \in \Omega^\star$. Notably, the quantity $R(\alpha)$ can be computed for any $\alpha \in \mathbb{R}^{mn}$ given the access to $\nabla\bar{f}(\cdot)$. As such, this suggests that $\|R(\alpha)\|$ is a reasonable surrogate for characterizing the optimality of $\alpha \in \mathbb{B}_q^n$. Then, it is natural to ask whether $\|R(\alpha)\|$ is related to $d(\alpha, \Omega^\star)$ or not, further motivating the GEB condition for the dual filtering-clustering model in Eq. (3.5).

Definition 4.1 *The dual filtering-clustering model in Eq. (3.5) satisfies a GEB condition if there exists $\tau > 0$ such that $d(\alpha, \Omega^\star) \leq \tau\|R(\alpha)\|$ for all $\alpha \in \mathbb{B}_q^n$.*

The GEB condition is a relaxed notion of global strong convexity (Pang, 1987). After removing the constraint set $\mathbb{B}_q^n$, we see that $R(\alpha) = -\nabla\bar{f}(\alpha)$ and the GEB condition is satisfied when $\bar{f}$ is strongly convex. The above definition is only given for the sake of completeness and does not give the insights on the range of q in which the GEB condition holds for the dual filtering-clustering model in Eq. (3.5). This is because the residual function R is in the abstract form and it remains elusive which value of q can guarantee that $d(\alpha, \Omega^\star) \leq \tau\|R(\alpha)\|$ holds for some $\tau > 0$. Zhou et al. (2015) have presented an alternative approach based on the notion of ULC property of set-valued mapping and used it to identify the range of p in which the local error bound condition holds for $\ell_{1,p}$-norm regularized problems. Combined with some nontrivial modifications, we show that this approach can be adopted here.

We compute an upper bound for τ in a special example of the dual filtering-clustering model in Eq. (3.5). Since the graph ℓ_1-trend filtering reduces to ℓ_1-trend filtering if we choose a proper graph, we focus on ℓ_1-trend filtering. Our example shows that a dimension-free upper bound for τ in some application problems.

Consider fused Lasso (Tibshirani et al., 2005) which is the ℓ_1-trend filtering with $k = 0$. For simplicity, we assume that $y = 0$ is an input data and $\lambda = 1$. Then, the problem in Eq. (3.1) is given by

$$\min_{\beta \in \mathbb{R}^n} \frac{1}{2}\|\beta\|^2 + \|D\beta\|_1, \text{ s.t. } D = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{bmatrix}.$$
$$(4.2)$$

which implies the dual filtering-clustering model:

$$\min_{\alpha \in \mathbb{R}^{\bar{n}-1}} \frac{1}{2}\|D^\top\alpha\|^2, \quad \text{s.t.} -1 \leq \alpha_i \leq 1, \ \forall 1 \leq i \leq \bar{n}-1.$$

It is easy to verify that

$$DD^\top \succeq \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \dots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \dots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \dots & 1 & 0 \\ 0 & 0 & 0 & \dots & 0 & 1 \end{bmatrix}$$

which implies that $\bar{f}(\alpha) = (1/2)\|D^\top\alpha\|^2$ is 1-strongly convex. Combining it with Drusvyatskiy and Lewis (2018, Corollary 3.6), we have $\tau \leq 4$.

Remark 4.4 *Computing an upper bound of τ is very difficult in general and has been a research topic by itself in the community. Indeed, Drusvyatskiy and Lewis (2018) provided a comprehensive qualitative treatment of error bound with quantitative studies for special cases and Wang and Lin (2014) studied machine learning problems and algorithms using error bound condition. We are also aware of a recent work (Pena et al., 2018) that proposed an algorithm to compute the error bound. Can we compute a bound of τ for special graph ℓ_1-trend filtering? We leave it to future work.*

We present our main results on the GEB condition holds for the dual filtering-clustering model when $q \in [1, 2] \cup \{+\infty\}$, which correspond to the filtering-clustering model in Eq. (1.1) with $p \in \{1\} \cup [2, +\infty]$.

Theorem 4.5 *Under Assumption 3.1 and let $q \in \{1, +\infty\}$, the GEB condition holds for the dual filtering-clustering model in Eq. (3.5).*

Note that $\mathbf{g}^\star = \nabla\bar{f}(\alpha^\star)$ for all optimal solutions $\alpha^\star$ (see Proposition A.1), we let $\mathcal{J} = \{j \in [n] \mid \mathbf{g}_j^\star \neq 0\}$ be the set of indices of nonzero coordinates.

Theorem 4.6 *Under Assumption 3.1 and let $q \in (1, 2]$ and $\mathcal{J} = [n]$, the GEB condition holds for the dual filtering-clustering model in Eq. (3.5).*

Remark 4.7 *$\mathcal{J} = [n]$ is necessary for deriving the GEB condition when $q \in (1, 2]$ (see Appendix A.7 for*

Algorithm 1 Generalized Dual Gradient Ascent

Input: learning rate $\eta > 0$ and tolerance $\varepsilon > 0$.
Initialization: $\beta_0 \in \mathbb{R}^d$, $\alpha_1 \in \mathbb{B}_q^n$ and the other tolerance $\hat{\varepsilon} > 0$.
for $t = 1, 2, \ldots, T$ **do**
   $\beta_t \leftarrow \textsc{InnerLoop}(f, \lambda, D, \alpha_t, \beta_{t-1}, \hat{\varepsilon})$.
   $\alpha_{t+1} \leftarrow \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta\lambda D\beta_t)$.
end for
Return: β_{T+1}.

---

an example) and also assumed in Jane et al. (2021, Theorem 6) for proving the metric subregularity. As argued in Jane et al. (2021, Remark 2), this assumption is mild in terms of applications since an optimal solution often lie on the boundary of a constraint set.

5 Algorithmic Framework

We propose and analyze a *generalized dual gradient ascent* (GDGA) algorithm for solving Eq. (1.1). It can be interpreted as an inexact forward-backward splitting algorithm (Tseng, 1991, 2000) and reduces to the alternating minimization algorithm (AMA) (Chi and Lange, 2015) when specialized to convex clustering.

5.1 Generalized dual gradient ascent

Our GDGA algorithm is a first-order optimization algorithm that only accesses the gradient of f, the matrix D and the regularization parameter λ. Since $f^\star$ does not admit an analytic form in general, we design a subroutine which returns an approximation of $\nabla f^\star$ by minimizing the smooth and strongly convex objective $f(\beta) - \lambda\alpha_t^\top D\beta$ with respect to β; see Algorithm 1.

Algorithm 1 is simple, matrix-free, and amenable to distributed implementation. It outperforms the existing approaches (Kim et al., 2009; Wang et al., 2016b; Ramdas and Tibshirani, 2016) which require matrix decomposition and suffer from scalability. Specialized to ℓ_2-convex clustering model, Algorithm 1 becomes inexact AMA (Chi and Lange, 2015) and exhibits fast convergence in practice. It is also worth mentioning that the semismooth Newton method (Sun et al., 2021) can outperform our algorithm by exploiting the special structure of Jacobian matrix in ℓ_2-convex clustering model. However, it appears to be difficult to extend such approach to solve general filtering-clustering model due to the complicated nonsmooth term $\sum_{k=1}^n \|D_k\beta\|_p$. Further, the subroutine can be constructed based on different algorithmic components. Indeed, $\beta_t \leftarrow \textsc{InnerLoop}(f, \lambda, D, \alpha_t, \beta_{t-1}, \hat{\varepsilon})$ is an $\hat{\varepsilon}$-minimizer of $f(\beta) - \lambda\alpha_t^\top D\beta$ satisfying that $\|\beta_t - \beta^\star(\alpha_t)\| \leq \hat{\varepsilon}$ where $\beta^\star(\alpha_t) := \text{argmin}_{\beta \in \mathbb{R}^d} f(\beta) -$

$\lambda\alpha_t^\top D\beta$. The subroutine implements some fast first-order optimization algorithms with an initial point β_{t-1}. For example, we can apply celebrated Nesterov's accelerated gradient descent (Nesterov, 2018) in the deterministic setting and optimal stochastic gradient descent (Rakhlin et al., 2012) in the online setting. If $f(\beta) = (1/2)\|\beta - y\|^2$, this subroutine can be removed since we have the analytic form of $\beta^\star(\alpha)$.

Algorithm 1 enjoys a solid theoretical guarantee. Indeed, we prove that it achieves linear convergence without counting the number of gradient or stochastic gradient oracles used in the subroutine. Although the proof idea is not new but follows the standard strategy (Luo and Tseng, 1992, 1993; Wang and Lin, 2014) given the global error bound condition previously established for the dual filtering-clustering models in Eq. (3.5), we provide the rigorous computational theory for the filtering-clustering model, shedding the light on great potential of AMA to pursue a high-accurate solution in practice.

5.2 Complexity of GDGA algorithmic framework

We establish the linear convergence of the GDGA algorithmic framework without counting the number of gradient or stochastic gradient oracles used in the subroutine. We hope to remark that our results are nontrivial due to the following reasons:

First, Eq. (1.1) is strongly convex but nonsmooth and the term $\sum_{k=1}^n \|D_k\beta\|_p$ is not computationally favorable. Indeed, the proximal mapping of $\sum_{k=1}^n \|D_k\beta\|_p$ does not have the closed form even when $p = 1$ and can not be efficiently computed in general. This makes proximal gradient algorithm (Parikh et al., 2014) not applicable. Second, $f^\star$ is smooth and strongly convex. However, the objective function $\bar{f}$ of the dual filtering-clustering model in Eq. (3.5) is non-strongly convex since a matrix $D^\top$ can be degenerate in real application, e.g., the graph ℓ_1-trend filtering model with $\bar{m} \gg \bar{n}$. In addition, $\nabla\bar{f}$ is not available to the algorithm and necessities an efficient subroutine in both theory and practice. Finally, the objective function of the saddle-point formulation in Eq. (3.4) is strongly convex in β but linear in α with a ball constraint set $\mathbb{B}_q^n$. As such, we can not derive the linear convergence of a few existing primal-dual optimization algorithms as a consequence of the existing results (Palaniappan and Bach, 2016; Wang and Xiao, 2017).

We present our main theorem and refer the readers to the appendix for the full version of Theorem 5.1.

Theorem 5.1 *Under Assumption 3.1 and set* $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1,\lambda^2\}}\})$ *in Algorithm 1 and* $\hat{\varepsilon} > 0$

Algorithm 2 Simplified GDGA

Input: learning rates $\eta > 0$ and tolerance $\varepsilon > 0$.
Initialization: $\beta_0 \in \mathbb{R}^d$ and $\alpha_1 \in \mathbb{B}_q^n$.
for $t = 1, 2, \ldots, T$ **do**
    $\beta_t = \lambda D^\top \alpha_t + y$.
    $\alpha_{t+1} \leftarrow \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \lambda D \beta_t)$.
end for
Return: β_{T+1}.

Algorithm 3 Deterministic GDGA

Input: learning rates $\eta > 0$ and tolerance $\varepsilon > 0$.
Initialization: $\beta_0 \in \mathbb{R}^d$ and $\alpha_1 \in \mathbb{B}_q^n$.
for $t = 1, 2, \ldots, T$ **do**
    $\beta_t \leftarrow \textsc{InnerLoopAgd}(f(\cdot), \lambda, D, \alpha_t, \beta_{t-1}, \hat{\varepsilon})$.
    $\alpha_{t+1} \leftarrow \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \lambda D \beta_t)$.
end for
Return: β_{T+1}.

Algorithm 4 Stochastic Variance Reduced GDGA

Input: learning rates $\eta > 0$ and tolerance $\varepsilon > 0$.
Initialization: $\beta_0 \in \mathbb{R}^d$ and $\alpha_1 \in \mathbb{B}_q^n$.
for $t = 1, 2, \ldots, T$ **do**
    $\beta_t \leftarrow \textsc{InnerLoopKatyusha}(\frac{1}{M}(\sum_{i=1}^M f_i(\cdot)), \lambda, D, \alpha_t, \beta_{t-1}, \hat{\varepsilon})$.
    $\alpha_{t+1} \leftarrow \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \lambda D \beta_t)$.
end for
Return: β_{T+1}.

Algorithm 5 Stochastic GDGA

Input: learning rates $\eta > 0$ and tolerance $\varepsilon > 0$.
Initialization: $\beta_0 \in \mathbb{R}^d$ and $\alpha_1 \in \mathbb{B}_q^n$.
for $t = 1, 2, \ldots, T$ **do**
    $\beta_t \leftarrow \textsc{InnerLoopSgd}(\mathbb{E}_P[F(\cdot, \xi)], \lambda, D, \alpha_t, \beta_{t-1}, \hat{\varepsilon})$.
    $\alpha_{t+1} \leftarrow \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \lambda D \beta_t)$.
end for
Return: β_{T+1}.

---

properly. Then, the number of iterations to return an ε-optimal solution is $T = O(\log(1/\varepsilon))$.

Corollary 5.2 *Under same setting as Theorem 5.1 with stopping criterion in the subroutine as $\mathbb{E}[\delta_t] \leq \hat{\varepsilon}$, the number of iterations required by the stochastic GDGA algorithm to return an ε-optimal solution is bounded by $T = O(\log(1/\varepsilon))$.*

The key ingredient here that enables an improved linear convergence rate is that a global error bound condition holds for the dual filtering-clustering model in Eq. (3.5). If the dual form of any optimization problems admits same structure, we can develop an algorithm with the same convergence guarantee. However, it is nontrivial to prove a global error bound for specific problems and the case of the dual filtering-clustering model is rare. As such, it is difficult to generalize our approach to generic convex optimization problems.

We study the case with $f(\beta) = (1/2)\|\beta - y\|^2$ and demonstrate that Algorithm 2 suffices to solve many widely used filtering-clustering models, e.g., ℓ_1-trend filtering and ℓ_2-convex clustering.

Theorem 5.3 *Under Assumption 3.1 and set $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1, \lambda^2\}}\})$ in Algorithm 2. Then, the number of gradient evaluations to return an ε-optimal solution is $N = O(\log(1/\varepsilon))$.*

For the deterministic setting in which ∇f is accessible by $\textsc{InnerLoop}$, we do not have the closed-form minimizer of $f(\beta) - \alpha_t^\top D\beta$ and obtain an $\hat{\varepsilon}$-minimizer by implementing $\textsc{InnerLoop}$ using AGD.

Theorem 5.4 *Under Assumption 3.1 and set $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1, \lambda^2\}}\})$ and $\hat{\varepsilon} > 0$ in Algorithm 3. The number of gradient evaluations to return an ε-optimal solution is $N = O(\sqrt{\ell/\mu} \log(1/\varepsilon))$.*

For the finite-sum setting in which the loss function f is of the form $\frac{1}{M} \sum_{i=1}^M f_i$, we obtain an $\hat{\varepsilon}$-minimizer by implementing $\textsc{InnerLoop}$ using Katyusha.

Theorem 5.5 *Under Assumption 3.1 and set $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1, \lambda^2\}}\})$ and $\hat{\varepsilon} > 0$ in Algorithm 4. The number of component gradient evaluations to get ε-optimal solution is $N = O((M + \sqrt{\ell M/\mu}) \log(1/\varepsilon))$.*

Theorem 5.5 guarantees the linear convergence rate for Algorithm 4 which outperforms Algorithm 3 in terms of the required number of component gradient evaluations by shaving off $\sqrt{M}$.

For the online setting where $f(\cdot) = \mathbb{E}_P[F(\cdot, \xi)]$ and $\xi \sim P$ is the streaming data that should be processed incrementally without having access to all data, we implement the subroutine using SGD.

Theorem 5.6 *Under Assumption 3.1 and set $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1, \lambda^2\}}\})$ and $\hat{\varepsilon} > 0$ in Algorithm 5. The number of stochastic gradient evaluations to return an ε-optimal solution is $N = O((1/\varepsilon) \log(1/\varepsilon))$.*

Theorem 5.6 guarantees the sublinear convergence rate for Algorithm 5. Our results match the lower bound of any stochastic first-order algorithm for nonsmooth and strongly convex setting up to log factors, showing that Algorithm 5 is the best possible that we can expect.

Remark 5.7 *We apply different subroutines when the problem has different structure so that the best possible theoretical convergence rate can be achieved. However, the choice of subroutine in practice is different and it is possible that specific subroutines are suitable for solving the particular problems. This is beyond the scope of our paper and we leave the answers to future work.*
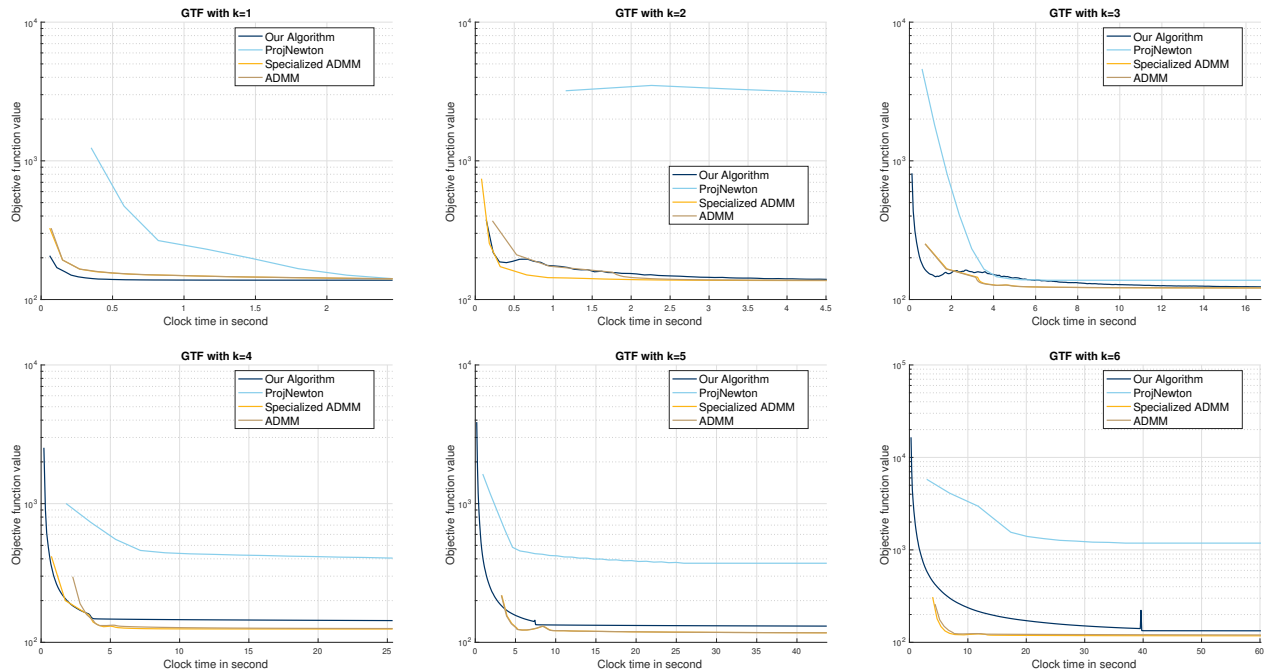
Nhat Ho[*,‡], **Tianyi Lin**[*,⋄], **Michael I. Jordan**[⋄]

Figure 1: Comparison of Algorithm 2, ADMM, specialized ADMM, and projected Newton method on medium image.

6 Experiments

We carefully conduct the experiments on ℓ_1-trend filtering to demonstrate the effectiveness of Algorithm 2. The baseline approaches include ADMM, specialized ADMM (Ramdas and Tibshirani, 2016) and projected Newton method (Wang et al., 2016b) and we consider three real images with various sizes: 128 by 128 pixels (small), 256 by 256 pixels (medium) and 512 by 512 pixels (large)[2]. All the algorithms are evaluated as the order k varies in the discrete difference operator $D^{(k+1)}$ in which the evaluation metric is the objective function value. We choose $\lambda = 0.2$ in our experiment and consider the adaptive step-size $\eta_t > 0$ based on Barzilai-Borwein rule (Fletcher, 2005).

We present the numerical results on different images in Figure 1 (see Appendix for other figures). Algorithm 2 is comparable with ADMM and specialized ADMM, all of which significantly outperform projected Newton method. More specifically, Algorithm 2 is the best when $k = 1$ remain effective as k increases. Note that ADMM and specialized ADMM are robust since they conduct matrix decomposition which significantly alleviates the ill-conditioning of the matrix $D^{(k+1)}$ as k increases. For small datasets, the ADMM-type methods can work quite well since the matrix inversion is not an issue. However, the ADMM-type methods are likely to fail for large datasets since the matrix in-

version becomes a computational bottleneck. In contrast, Algorithm 2 is matrix-free and we find that the Barzilai-Borwein rule can speed up the algorithm by exploring the curvature information and alleviate the ill-conditioning. As such, our algorithms can be good alternatives to the ADMM-type methods in practice.

7 Conclusion

This paper contributes to the landscape of computational aspect of filtering-clustering model in Eq. (1.1) by identifying a global error bound condition, which is satisfied by a large class of dual filtering-clustering problems, and designing a class of generalized dual gradient ascent algorithm, which is *optimal* first-order algorithms in the deterministic, finite-sum and online settings. Our results are new and shed the light on superior performance of several first-order optimization algorithms as applied to solve the filtering-clustering models in practice. We also conduct experiments on real datasets and the numerical results demonstrate the effectiveness of our algorithms.

Acknowledgments

---

[2]These images can be found at: `http://sipi.usc.edu/database/database.php?volume=misc`

References

A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics*, 40(5):2452–2482, 2012.

Z. Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. In *STOC*, pages 1200–1205. ACM, 2017.

H. H. Bauschke and J. M. Borwein. On projection algorithms for solving convex feasibility problems. *SIAM Review*, 38(3):367–426, 1996.

A. Beck and S. Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1-2):1–27, 2017.

G. H. G. Chen and R. T. Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

E. C. Chi and K. Lange. Splitting methods for convex clustering. *Journal of Computational and Graphical Statistics*, 24(4):994–1013, 2015.

D. Drusvyatskiy and A. S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

R. Fletcher. On the Barzilai-Borwein method. In *Optimization and control with applications*, pages 235–256. Springer, 2005.

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.

A. Guntuboyina, D. Lieu, S. Chatterjee, and B. Sen. Adaptive risk bounds in univariate total variation denoising and trend filtering. *Annals of Statistics*, 48(1):205–229, 2020.

T. Hocking, J.-P. Vert, F. Bach, and A. Joulin. Clusterpath: An algorithm for clustering using convex fusion penalties. In *ICML*, 2011.

M. Hong and Z-Q. Luo. On the linear convergence of the alternating direction method of multipliers. *Mathematical Programming*, 162(1-2):165–199, 2017.

J. Y. Jane, X. Yuan, S. Zeng, and J. Zhang. Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems. *Set-Valued and Variational Analysis*, pages 1–35, 2021.

A. Jourani. Hoffman's error bound, local controllability, and sensitivity analysis. *SIAM Journal on Control and Optimization*, 38(3):947–970, 2000.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak-Lojasiewicz condition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 795–811. Springer, 2016.

S-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009.

G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature, 2020.

G. Lan and Y. Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, pages 1–49, 2017.

C. Leser. A simple method of trend construction. *Journal of the Royal Statistical Society: Series B*, 23:91–107, 1961.

K. Lin, J. L. Sharpnack, A. Rinaldo, and R. J. Tibshirani. A sharp error analysis for the fused Lasso, with application to approximate changepoint screening. In *NIPS*, 2017.

P-L. Loh and M. J. Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.

Z-Q. Luo and P. Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2):408–425, 1992.

Z-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

S. Negahban and M. J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

Y. Nesterov. *Lectures on Convex Optimization*, volume 137. Springer, 2018.

O. H. M. Padilla, J. Sharpnack, Y. Chen, and D. M. Witten. Adaptive non-parametric regression with the K-NN fused Lasso. *Arxiv Preprint: 1807.11641*, 2018a.

O. H. M. Padilla, J. Sharpnack, J. G. Scott, and R. J. Tibshirani. The DFS fused lasso: Linear-time denoising over general graphs. *Journal of Machine Learning Research*, 18(1):1–36, 2018b.

B. Palaniappan and F. Bach. Stochastic variance reduction methods for saddle-point problems. In *NeurIPS*, pages 1416–1424, 2016.

J-S. Pang. A posteriori error bounds for the linearly-constrained variational inequality problem. *Mathematics of Operations Research*, 12(3):474–484, 1987.

J-S. Pang. Error bounds in mathematical programming. *Mathematical Programming*, 79(1):299–332, 1997.

N. Parikh, S. Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

J. Pena, J. Vera, and L. Zuluaga. An algorithm to compute the hoffman constant of a system of linear constraints. *ArXiv Preprint: 1804.08418*, 2018.

P. Radchenko and G. Mukherjee. Convex clustering via ℓ_1 fusion penalization. *Journal of the Royal Statistical Society: Series B*, 79(5):1527–1546, 2017.

A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, pages 1571–1578, 2012.

A. Ramdas and R. J. Tibshirani. Fast and flexible ADMM algorithms for trend filtering. *Journal of Computational and Graphical Statistics*, 25(3):839–858, 2016.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 2015.

L. I. Rudin, S. Osher, and E. Faterni. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60:259–268, 1992.

M. Schmidt, N. L. Roux, and F. Bach. Convergence rates of inexact proximal-gradient methods for convex optimization. In *NIPS*, pages 1458–1466, 2011.

D. Sun, K-C. Toh, and Y. Yuan. Convex clustering: Model, theoretical guarantee and efficient algorithm. *The Journal of Machine Learning Research*, 22:1–32, 2021.

K. M. Tan and D. Witten. Statistical properties of convex clustering. *Electronic Journal of Statistics*, 9(2):2324–2347, 2015.

R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108, 2005.

R. J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1):285–323, 2014.

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *Annals of Statistics*, 39(3):1335–1371, 2011.

P. Tseng. Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM Journal on Control and Optimization*, 29(1):119–138, 1991.

P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010.

J. Wang and L. Xiao. Exploiting strong convexity from data with primal-dual first-order algorithms. In *ICML*, pages 3694–3702, 2017.

P-W. Wang and C-J. Lin. Iteration complexity of feasible descent methods for convex optimization. *The Journal of Machine Learning Research*, 15(1):1523–1548, 2014.

Q. Wang, P. Gong, S. Chang, T. S. Huang, and J. Zhou. Robust convex clustering analysis. In *ICDM*, pages 1263–1268. IEEE, 2016a.

Y. Wang, J. Sharpnack, A. J. Smola, and R. J. Tibshirani. Trend filtering on graphs. *The Journal of Machine Learning Research*, 17(1):3651–3691, 2016b.

Z. Wang, H. Liu, and T. Zhang. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Annals of Statistics*, 42(6):2164–2201, 2014.

C. Wu, S. Kwon, X. Shen, and W. Pan. A new algorithm and theory for penalized regression-based clustering. *The Journal of Machine Learning Research*, 17:1–25, 2016.

Y. Zhang and L. Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *The Journal of Machine Learning Research*, 18(1):2939–2980, 2017.

Z. Zhou and A. M-C. So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017.

Z. Zhou, Q. Zhang, and A. M-C. So. $\ell_{1,p}$-norm regularization: error bounds and convergence rate analysis of first-order methods. In *ICML*, pages 1501–1510. JMLR. org, 2015.

C. Zhu, H. Xu, C. Leng, and S. Yan. Convex optimization procedure for clustering: Theoretical revisit. In *NIPS*, 2014.

Supplementary Material: On Structured Filtering-Clustering: Global Error Bound and Optimal First-Order Algorithms

A   Postponed Proofs in Section 4

This section first lays out the detailed proofs for Lemma 4.1, Lemma 4.2 and Lemma 4.3. Then, we introduce an *upper Lipschitz continuity* (ULC) property of a set-valued mapping which suffices to guarantee the GEB condition via appeal to the techniques in Zhou et al. (2015). Finally, we provide the detailed proofs for Theorem 4.5 and Theorem 4.6.

A.1   Proof of Lemma 4.1

We first show that $\alpha \in \partial f(\beta) \Leftrightarrow \beta \in \partial f^\star(\alpha)$. Indeed, if $\alpha \in \partial f(\beta)$, we have $f^\star(\alpha) = \alpha^\top \beta - f(\beta)$. This implies that $\beta \in \partial f^\star(\alpha)$ since, for all $\alpha' \in \mathbb{R}^m$, the following inequality holds true,

$$f^\star(\alpha') - f^\star(\alpha) \geq (\alpha')^\top \beta - f(\beta) - (\alpha^\top \beta - f(\beta)) = (\alpha' - \alpha)^\top \beta.$$

Conversely, if $\beta \in \partial f^\star(\alpha)$, the above argument implies that $\alpha \in \partial f^{\star\star}(\beta)$. Note that f is proper and convex. Rockafellar (2015, Theorem 12.2) shows that $f = f^{\star\star}$ and $\alpha \in \partial f(\beta)$.

We are ready to prove that $f^\star$ is $\frac{1}{\mu}$-gradient Lipschitz. Indeed, since f is μ-strongly convex and differentiable, the gradient mapping ∇f is one-to-one and $\partial f(\beta)$ is a singleton. This implies that $\partial f^\star(\alpha) = \{\nabla f^\star(\alpha)\}$ and

$$\|\nabla f^\star(\alpha_1) - \nabla f^\star(\alpha_2)\| = \|\beta_1 - \beta_2\| \leq \frac{\|\nabla f(\beta_1) - \nabla f(\beta_2)\|}{\mu} = \frac{\|\alpha_1 - \alpha_2\|}{\mu}.$$

By similar arguments, we obtain that $f^\star$ is $\frac{1}{\ell}$-strongly convex. Putting these pieces yields the desired result.

A.2   Proof of Lemma 4.2

Since f is μ-strongly convex, $\beta^\star(\alpha)$ is uniquely determined given $\alpha \in \mathbb{B}_q^n$ and hence well-defined. We see from the optimality condition that $\nabla f(\beta^\star(\alpha)) = \lambda D^\top \alpha$. Therefore, the following inequality holds true for any $\alpha', \alpha \in \mathbb{B}_q^n$:

$$\|\beta^\star(\alpha') - \beta^\star(\alpha)\| \leq \frac{\|\nabla f(\beta^\star(\alpha')) - \nabla f(\beta^\star(\alpha))\|}{\mu} \leq \frac{\lambda\|D^\top \alpha' - D^\top \alpha\|}{\mu} \leq \left(\frac{\lambda\|D\|}{\mu}\right)\|\alpha' - \alpha\|.$$

Putting these pieces together yields the desired result.

A.3   Proof of Lemma 4.3

We see from the definition of $\bar{f}$ that $\bar{f}(\alpha) = f^\star(\lambda D^\top \alpha)$. Lemma 4.1 shows that $f^\star$ is smooth and strongly convex under Assumption 3.1. Therefore, we obtain that $\bar{f}$ is differentiable and $\nabla \bar{f}(\alpha) = \lambda D \nabla f^\star(\lambda D^\top \alpha)$ using the chain rule. Since $\nabla f(\beta^\star(\alpha)) = \lambda D^\top \alpha$ and $\alpha \in \partial f(\beta) \Leftrightarrow \beta \in \partial f^\star(\alpha)$ (cf. the proof of Lemma 4.1), we have $\beta^\star(\alpha) = \nabla f^\star(\lambda D^\top \alpha)$. Putting these pieces together yields that $\nabla \bar{f}(\alpha) = \lambda D \beta^\star(\alpha)$. For all $\alpha', \alpha \in \mathbb{B}_q$, we have

$$\|\nabla \bar{f}(\alpha') - \nabla \bar{f}(\alpha)\| \leq \|\lambda D(\nabla f^\star(\lambda D^\top \alpha') - \nabla f^\star(\lambda D^\top \alpha_2))\| \leq \left(\frac{\lambda^2 \|D\|^2}{\mu}\right)\|\alpha_1 - \alpha_2\|.$$

Putting these pieces together yields the desired result.

Nhat Ho[∗,‡], Tianyi Lin[∗,⋄], Michael I. Jordan[⋄]

A.4 ULC property

We start with some mathematical notions. Indeed, let $\mathcal{Y}$ and $\mathcal{Z}$ be Euclidean spaces. A mapping $\Gamma : \mathcal{Y} \mapsto \mathcal{Z}$ is said to be a *set-valued mapping*, or equivalently, a multifunction if for any $y \in \mathcal{Y}$, we have $\Gamma(y) \subseteq \mathcal{Z}$. The graph of Γ is defined by $\{(y, z) \in \mathcal{Y} \times \mathcal{Z} \mid z \in \Gamma(y)\}$. We now define a notion of upper Lipschitz continuity (ULC).

Definition A.1 *A set-valued mapping* $\Gamma : \mathcal{Y} \mapsto \mathcal{Z}$ *has the ULC property at* $y \in \mathcal{Y}$ *if* $\Gamma(y)$ *is nonempty and closed, and there exist constants* $\kappa > 0$ *and* $\delta > 0$ *such that for any* $y \in \mathcal{Y}$ *with* $\|y' - y\| \leq \delta$, *we have* $\Gamma(y') \subseteq \Gamma(y) + \kappa\|y' - y\|\mathbb{B}_2$ *where* $\mathbb{B}_2$ *is the unit* ℓ_2-*norm ball in* $\mathcal{Z}$ *and "+" is the Minkowski sum of two sets.*

Proceeding a further step, we opt to prove a sufficient condition which guarantees that the dual filtering-clustering model in Eq. (3.5) satisfies a GEB condition. Let $\Sigma(\cdot, \cdot)$ be the set-valued mapping given by

$$\Sigma(\mathbf{s}, \mathbf{g}) := \{\alpha \in \mathbb{B}_q^n \mid D^\top \alpha = \mathbf{s}, -\mathbf{g} \in \mathcal{N}_{\mathbb{B}_q^n}(\alpha)\}, \quad \text{for all } (\mathbf{s}, \mathbf{g}) \in \mathbb{R}^d \times \mathbb{R}^{mn}. \tag{A.1}$$

Before proceeding to our main result, we summarize the relationship between the set-valued mapping Σ and the optimal set $\Omega^\star$ in the following proposition.

Proposition A.1 *Under Assumption 3.1, we have* $\Omega^\star = \Sigma(\mathbf{s}^\star, \mathbf{g}^\star)$ *for a pair* $(\mathbf{s}^\star, \mathbf{g}^\star)$ *satisfying that* $\mathbf{s}^\star = D^\top \alpha^\star$ *and* $\mathbf{g}^\star = \nabla \bar{f}(\alpha^\star)$ *for all* $\alpha^\star \in \Omega^\star$.

Proof. Since the dual filtering-clustering model in Eq. (3.5) is a convex optimization problem, the first-order optimality condition is both necessary and sufficient. Therefore, we have

$$\Omega^\star = \{\alpha^\star \in \mathbb{R}^{mn} \mid 0 \in \nabla \bar{f}(\alpha^\star) + \mathcal{N}_{\mathbb{B}_q^n}(\alpha^\star)\}. \tag{A.2}$$

We first show that $\Omega^\star \subseteq \Sigma(\mathbf{s}^\star, \mathbf{g}^\star)$. Let $\alpha^\star \in \Omega^\star$, we notice that $\bar{f}(\alpha) = f^\star(\lambda D^\top \alpha)$ and $f^\star$ is strongly convex (cf. Lemma 4.1). This implies that $\lambda D^\top \alpha^\star$ remains the same for all $\alpha^\star \in \Omega^\star$ and $\mathbf{s}^\star = D^\top \alpha^\star$ is well defined. In addition, $\mathbf{g}^\star = \nabla \bar{f}(\alpha^\star) = \lambda D \nabla f^\star(\lambda D^\top \alpha) = \lambda D \nabla f^\star(\lambda \mathbf{s}^\star)$. This implies that $\mathbf{g}^\star$ is well defined. Since $0 \in \nabla \bar{f}(\alpha^\star) + \mathcal{N}_{\mathbb{B}_q^n}(\alpha^\star)$, we have $-\mathbf{g}^\star \in \mathcal{N}_{\mathbb{B}_q^n}(\alpha^\star)$. Therefore, we have $\alpha^\star \in \Sigma(\mathbf{s}^\star, \mathbf{g}^\star)$.

Conversely, let $\alpha^\star \in \Sigma(\mathbf{s}^\star, \mathbf{g}^\star)$, we have $\mathbf{s}^\star = D^\top \alpha^\star$ and $-\mathbf{g}^\star \in \mathcal{N}_{\mathbb{B}_q^n}(\alpha^\star)$. In addition, $\mathbf{g}^\star = \nabla \bar{f}(\alpha^\star)$. Therefore, we conclude from Eq. (A.2) that $\alpha^\star \in \Omega^\star$. □

Given the result of Proposition A.1, we prove that the ULC property of Σ guarantees that the GEB condition holds for the dual filtering-clustering model in Eq. (3.5).

Theorem A.2 *Under Assumption 3.1, the GEB condition holds for the dual filtering-clustering model in Eq. (3.5) if the mapping* Σ *in Eq. (A.1) has the ULC property at* $(\mathbf{s}^\star, \mathbf{g}^\star)$ *where a pair* $(\mathbf{s}^\star, \mathbf{g}^\star)$ *is given in Proposition A.1.*

Proof. Since Σ has the ULC property at $(\mathbf{s}^\star, \mathbf{g}^\star)$, there exist constants $\kappa > 0$ and $\delta > 0$ such that for all $(\mathbf{s}, \mathbf{g}) \in \mathbb{R}^d \times \mathbb{R}^{mn}$ with $\|(\mathbf{s}, \mathbf{g}) - (\mathbf{s}^\star, \mathbf{g}^\star)\| \leq \delta$, we have

$$\Sigma(\mathbf{s}, \mathbf{g}) \subseteq \Sigma(\mathbf{s}^\star, \mathbf{g}^\star) + \kappa\|(\mathbf{s}, \mathbf{g}) - (\mathbf{s}^\star, \mathbf{g}^\star)\|\mathbb{B}_2. \tag{A.3}$$

We recall the residual function $R(\alpha) = \mathcal{P}_{\mathbb{B}_q^n}(\alpha - \nabla \bar{f}(\alpha)) - \alpha$ in Eq. (4.1) and define two functions $\mathbf{s}^+ : \mathbb{B}_q^n \to \mathbb{R}^d$ and $\mathbf{g}^+ : \mathbb{B}_q^n \to \mathbb{R}^{mn}$ by

$$\mathbf{s}^+(\alpha) := D^\top(\alpha + R(\alpha)), \quad \mathbf{g}^+(\alpha) := \nabla \bar{f}(\alpha) + R(\alpha). \tag{A.4}$$

Since $\mathbb{B}_q^n$ is convex and bounded, $R(\cdot)$ is Lipschitz continuous (Rockafellar, 2015). Combining it with the Lipschitz continuity of $\nabla \bar{f}$ implies that $\mathbf{s}^+(\cdot)$ and $\mathbf{g}^+(\cdot)$ are Lipschitz continuous. This together with Proposition A.1 implies that there exists a constant $\rho > 0$ such that,

$$\|(\mathbf{s}^+(\alpha), \mathbf{g}^+(\alpha)) - (\mathbf{s}^\star, \mathbf{g}^\star)\| \leq \delta, \quad \text{for all } \alpha \in \mathbb{B}_q^n \cap \{d(\alpha, \Omega^\star) \leq \rho\}. \tag{A.5}$$

By the definition of the residual function R, we have

$$\alpha + R(\alpha) = \underset{z \in \mathbb{B}_q^n}{\operatorname{argmin}} \ z^\top \nabla \bar{f}(\alpha) + \frac{1}{2}\|z - \alpha\|^2.$$

By the optimality condition, we have

$$-\nabla \bar{f}(\alpha) - R(\alpha) \in \mathcal{N}_{\mathbb{B}_q^n}(\alpha + R(\alpha)). \tag{A.6}$$

Combining Eq. (A.4) and Eq. (A.6) yields that $\alpha + R(\alpha) \in \Sigma(\mathbf{s}^+(\alpha), \mathbf{g}^+(\alpha))$ for all $\alpha \in \mathbb{B}_q^n$. This together with Eq. (A.3) and Eq. (A.5) yields that

$$d(\alpha + R(\alpha), \Sigma(\mathbf{s}^\star, \mathbf{g}^\star)) \leq \kappa \|(\mathbf{s}^+(\alpha), \mathbf{g}^+(\alpha)) - (\mathbf{s}^\star, \mathbf{g}^\star)\|, \quad \text{for all } \alpha \in \mathbb{B}_q^n \cap \{d(\alpha, \Omega^\star) \leq \rho\}.$$

Recalling that $\nabla \bar{f}(\alpha) = \lambda D \nabla f^\star(\lambda D^\top \alpha)$ and $\mathbf{g}^\star = \lambda D \nabla f^\star(\lambda \mathbf{s}^\star)$, we obtain from the definition of $\mathbf{s}^+(\cdot)$ and $\mathbf{g}^+(\cdot)$ in Eq. (A.4) that

$$\begin{aligned}
\|\mathbf{s}^+(\alpha) - \mathbf{s}^\star\| &\leq & \|D^\top \alpha - \mathbf{s}^\star\| + \|D\|\|R(\alpha)\|, \\
\|\mathbf{g}^+(\alpha) - \mathbf{g}^\star\| &\leq & \left(\frac{\lambda^2\|D\|}{\mu}\right)\|D^\top \alpha - \mathbf{s}^\star\| + \|R(\alpha)\|.
\end{aligned}$$

In view of Proposition A.1, we have $d(\alpha, \Omega^\star) \leq d(\alpha + R(\alpha), \Sigma(\mathbf{s}^\star, \mathbf{g}^\star)) + \|R(\alpha)\|$. Putting these pieces together yields that, for all $\alpha \in \mathbb{B}_q^n \cap \{d(\alpha, \Omega^\star) \leq \rho\}$, we have

$$d(\alpha, \Omega^\star) \leq \left(\kappa + \frac{\lambda^2\|D\|\kappa}{\mu}\right)\|D^\top \alpha - \mathbf{s}^\star\| + (\kappa\|D\| + \kappa + 1)\|R(\alpha)\|.$$

Letting $\kappa_0 = \max\{\kappa + \frac{\lambda^2\|D\|\kappa}{\mu}, \kappa\|D\| + \kappa + 1\}$ and using the inequality $(a+b)^2 \leq 2(a^2 + b^2)$ yields that, for all $\alpha \in \mathbb{B}_q^n \cap \{d(\alpha, \Omega^\star) \leq \rho\}$, we have

$$d^2(\alpha, \Omega^\star) \leq 2\kappa_0^2(\|D^\top \alpha - \mathbf{s}^\star\|^2 + \|R(\alpha)\|^2). \tag{A.7}$$

Since $f^\star$ is $\frac{1}{\ell}$-strongly convex, we have

$$\|D^\top \alpha - \mathbf{s}^\star\|^2 = \frac{\|\lambda D^\top \alpha - \lambda \mathbf{s}^\star\|^2}{\lambda^2} \leq \frac{\ell}{\lambda^2}(\lambda D^\top \alpha - \lambda \mathbf{s}^\star)^\top(\nabla f^\star(\lambda D^\top \alpha) - \nabla f^\star(\lambda \mathbf{s}^\star)).$$

Let $\alpha^\star$ be the projection of α onto $\Omega^\star$, we have $\mathbf{s}^\star = D^\top \alpha^\star$. It also follows from the definition that $\mathbf{g}^\star = \nabla \bar{f}(\alpha^\star) = \lambda D \nabla f^\star(\lambda \mathbf{s}^\star)$. Then, we have

$$\|D^\top \alpha - \mathbf{s}^\star\|^2 \leq \frac{\ell}{\lambda^2}(\alpha - \alpha^\star)^\top(\nabla \bar{f}(\alpha) - \mathbf{g}^\star). \tag{A.8}$$

For all $u \in \mathcal{N}_{\mathbb{B}_q^n}(\alpha + R(\alpha))$ and $v \in \mathcal{N}_{\mathbb{B}_q^n}(\alpha^\star)$, we obtain from the definition of the normal cone that $(u - v)^\top(\alpha + R(\alpha) - \alpha^\star) \geq 0$. Plugging $u = -\nabla \bar{f}(\alpha) - R(\alpha)$ and $v = -\mathbf{g}^\star$ into the above inequality yields that

$$(\alpha - \alpha^\star)^\top(\nabla \bar{f}(\alpha) - \mathbf{g}^\star) + \|R(\alpha)\|^2 \leq (\mathbf{g}^\star - \nabla \bar{f}(\alpha) + \alpha^\star - \alpha)^\top R(\alpha).$$

Since $\mathbf{g}^\star = \nabla \bar{f}(\alpha^\star)$ and Lemma 4.3 implies that $\nabla \bar{f}$ is Lipschitz continuous, there exists a constant $\kappa_1 > 0$ such that

$$(\mathbf{g}^\star - \nabla \bar{f}(\alpha) + \alpha^\star - \alpha)^\top R(\alpha) \leq \kappa_1 \|\alpha - \alpha^\star\|\|R(\alpha)\|.$$

Since $\|R(\alpha)\|^2 \geq 0$, we have

$$(\alpha - \alpha^\star)^\top(\nabla \bar{f}(\alpha) - \mathbf{g}^\star) \leq \kappa_1 \|\alpha - \alpha^\star\|\|R(\alpha)\|. \tag{A.9}$$

Combining Eq. (A.7), Eq. (A.8) and Eq. (A.9) yields that there exists a constant $\kappa_2 > 0$ such that the following inequality holds true,

$$d^2(\alpha, \Omega^\star) \leq \kappa_2(\|\alpha - \alpha^\star\|\|R(\alpha)\| + \|R(\alpha)\|^2), \quad \text{for all } \alpha \in \mathbb{B}_q^n \cap \{d(\alpha, \Omega^\star) \leq \rho\}.$$

Notice that $d(\alpha, \Omega^\star) = \|\alpha - \alpha^\star\|$. Thus, by solving the above quadratic inequality, we obtain that there exists a constant $\kappa_3 > 0$ such that

$$d(\alpha, \Omega^\star) \leq \kappa_3 \|R(\alpha)\|, \quad \text{for all } \alpha \in \mathbb{B}_q^n \cap \{d(\alpha, \Omega^\star) \leq \rho\}.$$

Since $R(\alpha) = 0$ if and only if $d(\alpha, \Omega^\star) = 0$, the function $h(\alpha) = \frac{d(\alpha, \Omega^\star)}{\|R(\alpha)\|}$ is well defined and continuous over the domain $\mathbb{B}_q^n \cap \{d(\alpha, \Omega^\star) > \rho\}$. Since $\mathbb{B}_q^n$ is convex and bounded, there exists a constant $\kappa_4 > 0$ such that $h(\alpha) \leq \kappa_4$ for all $\alpha \in \mathbb{B}_q^n \cap \{d(\alpha, \Omega^\star) > \rho\}$. Equivalently, we have

$$d(\alpha, \Omega^\star) \leq \kappa_4 \|R(\alpha)\|, \quad \text{for all } \alpha \in \mathbb{B}_q^n \cap \{d(\alpha, \Omega^\star) > \rho\}.$$

Setting $\tau = \max\{\kappa_3, \kappa_4\} > 0$, we obtain that $d(\alpha, \Omega^\star) \leq \tau \|R(\alpha)\|$ for all $\alpha \in \mathbb{B}_q^n$. Therefore, the GEB condition holds for the dual filtering-clustering model in Eq. (3.5). □

Equipped with the result of Theorem A.2, we will prove that the GEB condition holds for the dual filtering-clustering model in Eq. (3.5) when $q \in [1, 2] \cup \{+\infty\}$.

A.5 Proof of Theorem 4.5

We show that Σ has the ULC property when $q = 1$ or $q = +\infty$.

Lemma A.3 *Under Assumption 3.1 and let $q \in \{1, +\infty\}$, the set-valued mapping Σ is a polyhedral multifunction.*

Proof. Since $q \in \{1, +\infty\}$, we have $\mathbb{B}_q^n$ is a polyhedron which implies that its indicator function has a polyhedral epigraph. By the definition of the normal cone, we have $\mathcal{N}_{\mathbb{B}_q^n}$ is the subdifferential of an indicator function of $\mathbb{B}_q^n$. Putting these pieces together with Zhou et al. (2015, Lemma 2) yields that Σ is a polyhedral multifunction. □

With Theorem A.2 and Lemma A.3 at hand, we can prove that the GEB condition holds for the dual filtering-clustering model in Eq. (3.5). Indeed, by Assumption 3.1, we have $\Omega^\star$ is an nonempty set. Proposition A.1 guarantees that there exists a pair $(\mathbf{s}^\star, \mathbf{g}^\star) \in \mathbb{R}^d \times \mathbb{R}^{mn}$ such that $\Omega^\star = \Sigma(\mathbf{s}^\star, \mathbf{g}^\star)$ where the set-valued mapping Σ is given by Eq. (A.1). By Lemma A.3, the set-valued mapping Σ is a polyhedral multifunction. Putting these pieces together with Zhou et al. (2015, Lemma 1) implies that Σ has the ULC property at $(\mathbf{s}^\star, \mathbf{g}^\star)$. Therefore, we conclude from Theorem A.2 that the GEB condition holds for the dual filtering-clustering model in Eq. (3.5).

A.6 Proof of Theorem 4.6

It suffices to show that Σ has the ULC property when $q \in (1, 2]$ under certain conditions. Our first lemma provides the linear regularity of a collection of polyhedral sets. It has been stated as Bauschke and Borwein (1996, Corollary 5.26) and we omit the proof here.

Lemma A.4 *Suppose that $\mathcal{S}_1, \ldots, \mathcal{S}_M$ are a collection of polyhedra in $\mathbb{R}^{mn}$. Then, there exists a constant $\bar{\kappa} > 0$ such that $d(\alpha, \cap_{i=1}^M \mathcal{S}_i) \leq \bar{\kappa} \sum_{i=1}^M d(\alpha, \mathcal{S}_i)$ for all $\alpha \in \mathbb{R}^{mn}$.*

The next proposition provides a concrete representation of $\Omega^\star = \Sigma(\mathbf{s}^\star, \mathbf{g}^\star)$. In particular, we let $\mathcal{J} = \{j \in [n] \mid \mathbf{g}_j^\star \neq 0\}$ be the set of indices of nonzero coordinates.

Proposition A.5 *Under Assumption 3.1 and let the mapping Σ be defined in Eq. (A.1) and $(\mathbf{s}^\star, \mathbf{g}^\star) \in \mathbb{R}^d \times \mathbb{R}^{mn}$ be given in Proposition A.1. Then, we have*

$$\Sigma(\mathbf{s}^\star, \mathbf{g}^\star) = \left\{ \alpha^\star = \begin{bmatrix} \alpha_1^\star \\ \vdots \\ \alpha_n^\star \end{bmatrix} \in \mathbb{R}^{mn} \; \middle| \; \begin{array}{l} D^\top \alpha^\star = \mathbf{s}^\star, \\ \alpha_j^\star = -\frac{v(\mathbf{g}_j^\star)}{\|v(\mathbf{g}_j^\star)\|_q}, \text{ for all } j \in \mathcal{J}, \\ \alpha_j^\star \in \text{int}(\mathbb{B}_q^n), \text{ for all } j \notin \mathcal{J}. \end{array} \right\},$$

where the function $v : \mathbb{R}^m \mapsto \mathbb{R}^m$ is defined by $v(\mathbf{g}) = (\text{sign}(g_1)|g_1|^{\frac{p}{q}}, \ldots, \text{sign}(g_m)|g_m|^{\frac{p}{q}})$. Fixing any $q \in (1, 2]$, there exists some constants $\delta > 0$ and $\nu > 0$ such that, for all $\mathbf{g} \in \mathbb{R}^m$ satisfying that $\|\mathbf{g} - \mathbf{g}^\star\| \leq \delta$, we have $\|v(\mathbf{g}) - v(\mathbf{g}^\star)\| \leq \nu \|\mathbf{g} - \mathbf{g}^\star\|$.

Proof. By the definition of $(\mathbf{s}^\star, \mathbf{g}^\star) \in \mathbb{R}^d \times \mathbb{R}^{mn}$ in Proposition A.1, we have $D^\top \alpha^\star = \mathbf{s}^\star$, $\nabla \bar{f}(\alpha^\star) = \mathbf{g}^\star$ and $\alpha^\star \in \mathbb{B}_q^n$. Then, we consider two different cases: $\mathbf{g}_j^\star = 0$ and $\mathbf{g}_j^\star \neq 0$.

If $\mathbf{g}_j^\star = 0$, the j-th block of $\alpha^\star$ is in the interior of $\mathbb{B}_q^n$ and $\alpha_j^\star \in \text{int}(\mathbb{B}_q^n)$ for all $j \notin \mathcal{J}$.

If $\mathbf{g}_j^\star \neq \mathbf{0}$, the j-th block of $\alpha^\star$ is in the interior of $\mathbb{B}_q^n$ and $\|\alpha_j^\star\|_q = 1$ for all $j \in \mathcal{J}$. By the KKT condition in constrained optimization, $\alpha^\star \in \Omega^\star$ if and only if there exists a multiplier $\mu \geq 0$ such that the following statement holds true,

$$
\begin{aligned}
1 - \|\alpha_j^\star\|_q &\geq 0, \\
\mathbf{s}^\star - D^\top \alpha^\star &= 0, \\
\mathbf{g}_j^\star + \mu \cdot \frac{(\text{sign}((\alpha_j^\star)_1)|(\alpha_j^\star)_1|^{q-1}, \ldots, \text{sign}((\alpha_j^\star)_m)|(\alpha_j^\star)_m|^{q-1})}{\|\alpha_j^\star\|_q^{q/p}} &= 0, \\
\mu(1 - \|\alpha_j^\star\|_q) &= 0.
\end{aligned}
$$

Since $\mathbf{g}_j^* \neq \mathbf{0}$, we have $\mu > 0$ and again $\|\alpha_j^\star\|_q = 1$. Thus, we can solve $\alpha_j^\star$ in terms of $\mathbf{g}_j^\star$ by using the above equations and obtain that

$$
\mathbf{s}^\star - D^\top \alpha^\star = 0, \quad \alpha_j^\star + \frac{v(\mathbf{g}_j^\star)}{\|v(\mathbf{g}_j^\star)\|_q} = 0.
$$

Putting these pieces together yields that

$$
\Sigma(\mathbf{s}^\star, \mathbf{g}^\star) = \left\{ \alpha^\star = \begin{bmatrix} \alpha_1^\star \\ \vdots \\ \alpha_n^\star \end{bmatrix} \in \mathbb{R}^{mn} \; \middle| \; \begin{array}{l} D^\top \alpha^\star = \mathbf{s}^\star, \\ \alpha_j^\star = -\frac{v(\mathbf{g}_j^\star)}{\|v(\mathbf{g}_j^\star)\|_q}, \text{ for all } j \in \mathcal{J}, \\ \alpha_j^\star \in \text{int}(\mathbb{B}_q^n), \text{ for all } j \notin \mathcal{J}. \end{array} \right\}.
$$

Fixing any $q \in (1, 2]$, we have $\frac{p}{q} \geq 1$. Then, the function $s \mapsto \text{sign}(s)|s|^{\frac{p}{q}}$ is continuously differentiable and hence locally Lipschitz (Rockafellar, 2015). So there exist some constants $\delta > 0$ and $\nu > 0$ such that, for all $s', s \in \mathbb{R}$ satisfying that $|s' - s| \leq \delta$, we have

$$
|\text{sign}(s')|s'|^{\frac{p}{q}} - \text{sign}(s)|s|^{\frac{p}{q}}| \leq \nu|s_1 - s_2|.
$$

By the definition of v, we have $\|v(\mathbf{g}) - v(\mathbf{g}^\star)\| \leq \nu\|\mathbf{g} - \mathbf{g}^\star\|$ for all $g \in \mathbb{R}^m$. As a consequence, we reach the conclusion of the proposition. □

Proposition A.5 shows that $\Sigma(\mathbf{s}^\star, \mathbf{g}^\star)$ is closed. Since $\Sigma(\mathbf{s}^\star, \mathbf{g}^\star) \subseteq \mathbb{B}_q^n$, we have $\Sigma(\mathbf{s}^\star, \mathbf{g}^\star)$ is bounded where $(\mathbf{s}^\star, \mathbf{g}^\star) \in \mathbb{R}^d \times \mathbb{R}^{mn}$ is given in Proposition A.1. Now we are ready to prove that Σ has the ULC property when $q \in (1, 2]$ and $\mathcal{J} = [n]$.

By Assumption 3.1, we have $\Omega^\star$ is an nonempty set. Proposition A.1 guarantees that there exists a pair $(\mathbf{s}^\star, \mathbf{g}^\star) \in \mathbb{R}^d \times \mathbb{R}^{mn}$ such that $\Omega^\star = \Sigma(\mathbf{s}^\star, \mathbf{g}^\star)$ where the set-valued mapping Σ is given by Eq. (A.1).

Define the sets $\mathcal{S}$ and $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$ by

$$
\mathcal{S} := \{\alpha^\star \in \mathbb{R}^{mn} \mid D^\top \alpha^\star = \mathbf{s}^\star\}, \quad \mathcal{S}_j := \left\{ \alpha^\star = \begin{bmatrix} \alpha_1^\star \\ \vdots \\ \alpha_n^\star \end{bmatrix} \in \mathbb{R}^{mn} \; \middle| \; \alpha_j^\star = -\frac{v(\mathbf{g}_j^\star)}{\|v(\mathbf{g}_j^\star)\|_q} \right\}.
$$

By Proposition A.5 and using the assumption that $\mathcal{J} = [n]$, we have

$$
\Sigma(\mathbf{s}^\star, \mathbf{g}^\star) = \mathcal{S} \cap (\cap_{j=1}^n \mathcal{S}_j).
$$

Moreover, $\mathcal{S}, \mathcal{S}_1, \ldots, \mathcal{S}_n$ are all polyhedral subsets of $\mathbb{R}^{mn}$. By Lemma A.4, there exists a constant $\bar{\kappa} > 0$ such that, for any $\alpha \in \mathbb{R}^{mn}$, we have

$$
d(\alpha, \Sigma(\mathbf{s}^\star, \mathbf{g}^\star)) \leq \bar{\kappa} \left(d(\alpha, \mathcal{S}) + \sum_{j=1}^n d(\alpha, \mathcal{S}_j) \right). \tag{A.10}
$$

Thus, it suffices to bound the right-hand side of the above inequality for all $\alpha \in \Sigma(\mathbf{s}, \mathbf{g})$ in which $(\mathbf{s}, \mathbf{g})$ lies in the neighborhood of $(\mathbf{s}^\star, \mathbf{g}^\star) \in \mathbb{R}^d \times \mathbb{R}^{mn}$ and $\Sigma(\mathbf{s}, \mathbf{g})$ is nonempty. Towards that end, we discuss the bound on $d(\alpha, \mathcal{S})$ and $d(\alpha, \mathcal{S}_j)$ separately.

Nhat Ho[*,‡], **Tianyi Lin**[*,◇], **Michael I. Jordan**[◇]

The bound on $d(\alpha, \mathcal{S})$ follows from the well-known Hoffman bound (see Jourani (2000) for example). In particular, there exists a constant $\theta > 0$ such that $d(\alpha, \mathcal{S}) \leq \theta \| D^\top \alpha - \mathbf{s}^\star \|$ for all $\alpha \in \mathbb{R}^{mn}$. In addition, for all $\alpha \in \Sigma(\mathbf{s}, \mathbf{g})$ with $\mathbf{g} \neq 0$, we have $D^\top \alpha = \mathbf{s}$. Putting these pieces together yields that

$$d(\alpha, \mathcal{S}) \leq \theta \| \mathbf{s} - \mathbf{s}^\star \|, \quad \text{for all } \alpha \in \Sigma(\mathbf{s}, \mathbf{g}). \tag{A.11}$$

Proceeding a further step, we bound $d(\alpha, \mathcal{S}_j)$ via appeal to Hölder inequality. In particular, we notice that $\mathbf{g}_j^\star \neq 0$ for all $j \in [n]$. Therefore, there exists a constant $\delta_j > 0$ such that $\|(\mathbf{s}, \mathbf{g}) - (\mathbf{s}^\star, \mathbf{g}^\star)\| \leq \delta_j$ and $\mathbf{g}_j \neq 0$ for all $j \in [n]$. For any $\alpha \in \Sigma(\mathbf{s}, \mathbf{g})$ satisfying that $\|(\mathbf{s}, \mathbf{g}) - (\mathbf{s}^\star, \mathbf{g}^\star)\| \leq \delta_j$, we have

$$\mathbf{s} - D^\top \alpha = 0, \quad \alpha_j + \frac{v(\mathbf{g}_j)}{\|v(\mathbf{g}_j)\|_q} = 0.$$

Since $\mathbf{g}_j, \mathbf{g}_j^\star \neq 0$, we have $\|v(\mathbf{g}_j)\|_q > 0$ and $\|v(\mathbf{g}_j^\star)\|_q > 0$. Putting these pieces together with Proposition A.5, we have

$$d(\alpha, \mathcal{S}_j) \leq \left\| \frac{v(\mathbf{g}_j)}{\|v(\mathbf{g}_j)\|_q} - \frac{v(\mathbf{g}_j^\star)}{\|v(\mathbf{g}_j^\star)\|_q} \right\|,$$

which implies that

$$\left\| \frac{v(\mathbf{g}_j)}{\|v(\mathbf{g}_j)\|_q} - \frac{v(\mathbf{g}_j^\star)}{\|v(\mathbf{g}_j^\star)\|_q} \right\| = \frac{\|v(\mathbf{g}_j) - v(\mathbf{g}_j^\star)\|_q \|v(\mathbf{g}_j)\| + \|v(\mathbf{g}_j) - v(\mathbf{g}_j^\star)\| \|v(\mathbf{g}_j)\|_q}{\|v(\mathbf{g}_j)\|_q \|v(\mathbf{g}_j^\star)\|_q}.$$

Since $\|(\mathbf{s}, \mathbf{g}) - (\mathbf{s}^\star, \mathbf{g}^\star)\| \leq \delta_j$, we have $\|v(\mathbf{g}_j)\|$ and $\|\|v(\mathbf{g}_j)\|_q$ are bounded. Since $q \in (1, 2]$, the Hölder inequality implies that $\|v(\mathbf{g}_j) - v(\mathbf{g}_j^\star)\|_q \leq m^{\frac{2-q}{2q}} \|v(\mathbf{g}_j) - v(\mathbf{g}_j^\star)\|$. Putting these pieces together yields that

$$d(\alpha, \mathcal{S}_j) \leq C \left(m^{\frac{2-q}{2q}} + 1 \right) \|v(\mathbf{g}_j) - v(\mathbf{g}_j^\star)\| \overset{\text{Proposition A.5}}{\leq} C\nu_j \left(m^{\frac{2-q}{2q}} + 1 \right) \|\mathbf{g}_j - \mathbf{g}_j^\star\|. \tag{A.12}$$

Plugging Eq. (A.11) and Eq. (A.12) into Eq. (A.10) yields that

$$d(\alpha, \Sigma(\mathbf{s}^\star, \mathbf{g}^\star)) \leq \left(\bar{\kappa} \max \left\{ \theta, C \left(m^{\frac{2-q}{2q}} + 1 \right) \left(\sum_{j \in \mathcal{J}} \nu_j \right) \right\} \right) \|(\mathbf{s}, \mathbf{g}) - (\mathbf{s}^\star, \mathbf{g}^\star)\|,$$

for any $\alpha \in \Sigma(\mathbf{s}, \mathbf{g})$ with $\|(\mathbf{s}, \mathbf{g}) - (\mathbf{s}^\star, \mathbf{g}^\star)\| \leq \min_{j=1}^n \delta_j$. This implies that Σ has the ULC property at $(\mathbf{s}^\star, \mathbf{g}^\star) \in \mathbb{R}^d \times \mathbb{R}^{mn}$. Therefore, we conclude from Theorem A.2 that the GEB condition holds for the dual filtering-clustering model in Eq. (3.5).

A.7 Counterexample to the case without $\mathcal{J} = [n]$

The following example, which is presented in Jane et al. (2021, Remark 2), shows that the GEB condition fails in general when $q \in (1, 2]$. In particular, we consider

$$\min_{\mathbf{x} \in \mathbb{B}_2} \bar{f}(\alpha) = \frac{1}{2}(\alpha_2 - 1)^2. \tag{A.13}$$

It is clear that $(0, 1)$ is the only point in $\Omega^\star$ and $\mathbf{g}^\star = 0$. Define the sequence

$$\alpha_t = (\cos \theta_t, \sin \theta_t), \quad \theta_t \in (0, \tfrac{\pi}{2}), \quad \theta_t \to \tfrac{\pi}{2},$$

we have $\alpha_t \in \mathbb{B}_2$ and $\alpha_t \to (0, 1)$. Furthermore,

$$R(\alpha_t) = \frac{((1 - \sqrt{1 + \cos^2 \theta_t}) \cos \theta_t, 1 - \sin \theta_t \sqrt{1 + \cos^2 \theta_t})}{\sqrt{1 + \cos^2 \theta_t}}.$$

Let $z_t = \sin \theta_t$, we have

$$\|R(\alpha_t)\| = \frac{\sqrt{2 + 2\cos^2 \theta_t - 2\sqrt{1 + \cos^2 \theta_t}(\cos^2 \theta_t + \sin \theta_t)}}{\sqrt{1 + \cos^2 \theta_t}} = \sqrt{2 - \frac{2 + 2z_t - 2z_t^2}{\sqrt{2 - z_t^2}}}.$$

Therefore, we conclude that

$$\frac{d(\alpha_t, \Omega^\star)}{\|R(\alpha_t)\|} = \sqrt{\frac{(1 - z_t)(2 - z_t^2)}{2 - z_t^2 - (1 + z_t - z_t^2)\sqrt{2 - z_t^2}}}.$$

Letting $t \to +\infty$ and using the L'Hospital's rule, we have

$$\lim_{t \to +\infty} \left(\frac{d(\alpha_t, \Omega^\star)}{\|R(\alpha_t)\|}\right)^2 = \lim_{z \to 1} \frac{3z^2 - 2z - 2}{-2z - (1 - 2z)\sqrt{2 - z^2} + z(1 + z - z^2)(2 - z^2)^{-1/2}} = +\infty.$$

This implies that the GEB condition does not hold true for the model in Eq. (A.13).

B   Postponed Proofs in Section 5

This section lays out the detailed proofs for Theorem 5.1, Corollary 5.2, Theorem 5.4, Theorem 5.5 as well as Theorem 5.6.

B.1   Technical lemmas

We prove several useful technical lemmas. The key quantity here is the distance between β_t and $\beta^\star(\alpha_t) = \operatorname{argmin}_{\beta \in \mathbb{R}^d} f(\beta) - \lambda \alpha_t^\top D\beta$ given by

$$\delta_t := \|\beta_t - \beta^\star(\alpha_t)\|.$$

We also denote $\bar{\alpha}_t$ as the projection of α_t onto the optimal set $\Omega^\star$ of the dual filtering-clustering model in Eq. (3.5), and define the objective gap between $\bar{f}(\alpha_t)$ and $\bar{f}(\bar{\alpha}_t)$ by

$$\Delta_t := \bar{f}(\alpha_t) - \bar{f}(\bar{\alpha}_t).$$

Our first lemma provides a key lower bound for the iterative objective gap $\Delta_t - \Delta_{t+1}$.

Lemma B.1 *Under Assumption 3.1 and let $(\alpha_t, \beta_t)_{t \geq 0}$ be generated by Algorithm 1 with the stepsize $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1, \lambda^2\}}\})$. Then, for any $t \geq 0$, we have*

$$\Delta_t - \Delta_{t+1} \geq \left(\frac{\lambda^2}{4\eta}\right)\|\alpha_t - \alpha_{t+1}\|^2 - \left(\frac{\eta\|D\|^2}{2}\right)\delta_t^2.$$

Proof. Since $\alpha_{t+1} = \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta\lambda D\beta_t)$, we have

$$(\alpha - \alpha_{t+1})^\top(\alpha_{t+1} - \alpha_t + \eta\lambda D\beta_t) \geq 0, \quad \text{for all } \alpha \in \mathbb{B}_q^n.$$

Plugging $\alpha = \alpha_t$ into the above inequality and rearranging yields that

$$\lambda(\alpha_t - \alpha_{t+1})^\top D\beta_t \geq \frac{\|\alpha_t - \alpha_{t+1}\|^2}{\eta}. \tag{B.1}$$

Since $\bar{f}$ is $\frac{\lambda^2\|D\|^2}{\mu}$-gradient Lipschiz (cf. Lemma 4.3), we have

$$(\alpha_t - \alpha_{t+1})^\top \lambda D\beta_t = (\alpha_t - \alpha_{t+1})^\top \nabla\bar{f}(\alpha_t) + (\alpha_t - \alpha_{t+1})^\top(\lambda D\beta_t - \nabla\bar{f}(\alpha_t)) \tag{B.2}$$
$$\leq \bar{f}(\alpha_t) - \bar{f}(\alpha_{t+1}) + \left(\frac{\lambda^2\|D\|^2}{\mu}\right)\|\alpha_t - \alpha_{t+1}\|^2 + (\alpha_t - \alpha_{t+1})^\top(\lambda D\beta_t - \nabla\bar{f}(\alpha_t)).$$

Using $\nabla\bar{f}(\alpha_t) = \lambda D\beta^\star(\alpha_t)$ and the Young's inequality, we have

$$(\alpha_t - \alpha_{t+1})^\top(\lambda D\beta_t - \nabla\bar{f}(\alpha_t)) = (\alpha_t - \alpha_{t+1})^\top(\lambda D\beta_t - \lambda D\beta^\star(\alpha_t)) \tag{B.3}$$
$$\leq \left(\frac{\lambda^2}{2\eta}\right)\|\alpha_t - \alpha_{t+1}\|^2 + \left(\frac{\eta}{2}\right)\|D\beta_t - D\beta^\star(\alpha_t)\|^2 \leq \left(\frac{\lambda^2}{2\eta}\right)\|\alpha_t - \alpha_{t+1}\|^2 + \left(\frac{\eta\|D\|^2}{2}\right)\delta_t^2.$$

Plugging Eq. (B.2) and Eq. (B.3) into Eq. (B.1) yields that

$$
\begin{aligned}
\bar{f}(\alpha_t) - \bar{f}(\alpha_{t+1}) &\geq \left(\frac{\lambda^2}{2\eta} - \frac{\lambda^2 \|D\|^2}{\mu} \right) \|\alpha_t - \alpha_{t+1}\|^2 - \left(\frac{\eta \|D\|^2}{2} \right) \delta_t^2 \\
&\geq \left(\frac{\lambda^2}{4\eta} \right) \|\alpha_t - \alpha_{t+1}\|^2 - \left(\frac{\eta \|D\|^2}{2} \right) \delta_t^2.
\end{aligned}
$$

This completes the proof. □

Our second lemma presents an upper bound for Δ_{t+1} based on $\|\alpha_t - \alpha_{t+1}\|^2$ using a global error bound for the dual filtering-clustering model in Theorem 4.5 and 4.6.

Lemma B.2 *Under Assumption 3.1 and let $(\alpha_t, \beta_t)_{t \geq 0}$ be generated by Algorithm 1 with the stepsize $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1, \lambda^2\}}\})$. Then, for any $t \geq 0$, we have*

$$
\|\alpha_{t+1} - \alpha_t\|^2 \geq \left(\frac{4\tau \eta^2}{17\tau^2 + 14\tau + 1} \right) \Delta_{t+1} - \left(\frac{8\tau^2 \lambda^2 \|D\|^2 \eta^2}{17\tau^2 + 14\tau + 1} \right) \delta_t^2.
$$

Proof. By the definition of $\nabla \bar{f}$ and using the triangle inequality, we have

$$
\begin{aligned}
\|\alpha_t - \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \nabla \bar{f}(\alpha_t))\| &= \|\alpha_t - \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \lambda D \beta^\star(\alpha_t))\| \\
&\leq \|\alpha_t - \alpha_{t+1}\| + \|\alpha_{t+1} - \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \lambda D \beta^\star(\alpha_t))\|.
\end{aligned}
$$

Since $\alpha_{t+1} = \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \lambda D \beta_t)$ and the projection operator is nonexpansive, we have

$$
\|\alpha_{t+1} - \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \lambda D \beta^\star(\alpha_t))\| \leq \eta \lambda \|D\| \delta_t.
$$

We see from Gafni and Bertsekas (1984, Lemma 1) that $(1/\eta)\|\alpha_t - \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \nabla \bar{f}(\alpha_t))\|$ is nonincreasing as a function of $\eta > 0$. Since $\eta \in (0, 1)$, we have

$$
\eta \|\alpha_t - \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \nabla \bar{f}(\alpha_t))\| \leq \|\alpha_t - \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \nabla \bar{f}(\alpha_t))\|.
$$

Putting these pieces together yields that

$$
\|\alpha_t - \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \nabla \bar{f}(\alpha_t))\| \leq \frac{\|\alpha_t - \alpha_{t+1}\|}{\eta} + \lambda \|D\| \delta_t. \tag{B.4}
$$

Since $\bar{\alpha}_t$ is the projection of α_t onto $\Omega^\star$, we derive from Theorem 4.5 and 4.6 that

$$
\|\alpha_t - \bar{\alpha}_t\| \leq \tau \|\alpha - \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \nabla \bar{f}(\alpha_t))\|.
$$

Plugging Eq. (B.4) into the above inequality yields that

$$
\|\alpha_t - \bar{\alpha}_t\| \leq \left(\frac{\tau}{\eta} \right) \|\alpha_t - \alpha_{t+1}\| + \tau \lambda \|D\| \delta_t. \tag{B.5}
$$

It suffices to lower bound $\|\alpha_t - \bar{\alpha}_t\|$ using Δ_{t+1}. Since $\alpha_{t+1} = \mathcal{P}_{\mathbb{B}_q^n}(\alpha_t - \eta \lambda D \beta_t)$, we have

$$
(\alpha - \alpha_{t+1})^\top (\alpha_{t+1} - \alpha_t + \eta \lambda D \beta_t) \geq 0, \quad \text{for all } \alpha \in \mathbb{B}_q^n.
$$

Letting $\alpha = \bar{\alpha}_t$ in the above inequality and rearranging the inequality yields that

$$
\lambda(\alpha_{t+1} - \bar{\alpha}_t)^\top D \beta_t \leq \left(\frac{1}{\eta} \right) (\alpha_{t+1} - \bar{\alpha}_t)^\top (\alpha_t - \alpha_{t+1}). \tag{B.6}
$$

Since $\bar{\alpha}_t, \bar{\alpha}_{t+1} \in \Omega^\star$, we have $\bar{f}(\bar{\alpha}_t) = \bar{f}(\bar{\alpha}_{t+1})$. By the convexity of $\bar{f}$, we have

$$
\Delta_{t+1} = \bar{f}(\alpha_{t+1}) - \bar{f}(\bar{\alpha}_{t+1}) = \bar{f}(\alpha_{t+1}) - \bar{f}(\bar{\alpha}_t) \leq (\alpha_{t+1} - \bar{\alpha}_t)^\top \nabla \bar{f}(\alpha_{t+1}). \tag{B.7}
$$

Since $\nabla \bar{f}(\alpha_t) = \lambda D\beta^\star(\alpha_t)$, we have

$$(\alpha_{t+1} - \bar{\alpha}_t)^\top \nabla \bar{f}(\alpha_{t+1}) \leq \underbrace{(\alpha_{t+1} - \bar{\alpha}_t)^\top (\nabla \bar{f}(\alpha_{t+1}) - \nabla \bar{f}(\alpha_t))}_{\mathbf{I}}$$
$$+ \underbrace{(\alpha_{t+1} - \bar{\alpha}_t)^\top (\lambda D\beta^\star(\alpha_t) - \lambda D\beta_t)}_{\mathbf{II}} + \underbrace{\lambda(\alpha_{t+1} - \bar{\alpha}_t)^\top D\beta_t}_{\mathbf{III}}.$$

Since $\bar{f}$ is $\frac{\lambda^2\|D\|^2}{\mu}$-gradient Lipschiz (cf. Lemma 4.3), we have

$$\mathbf{I} \leq \left(\frac{\lambda^2\|D\|^2}{\mu}\right)\|\alpha_{t+1} - \alpha_t\|\|\alpha_{t+1} - \bar{\alpha}_t\|.$$

By the definition of δ_t, we have

$$\mathbf{II} \leq \lambda\delta_t\|D\|\|\alpha_{t+1} - \bar{\alpha}_t\|.$$

In addition, Eq. (B.6) implies that

$$\mathbf{III} \leq \left(\frac{1}{\eta}\right)\|\alpha_t - \alpha_{t+1}\|\|\alpha_{t+1} - \bar{\alpha}_t\|.$$

Therefore, we conclude that

$$(\alpha_{t+1} - \bar{\alpha}_t)^\top \nabla \bar{f}(\alpha_{t+1}) \leq \left(\lambda\delta_t\|D\| + \left(\frac{\lambda^2\|D\|^2}{\mu} + \frac{1}{\eta}\right)\|\alpha_{t+1} - \alpha_t\|\right)\|\alpha_{t+1} - \bar{\alpha}_t\|.$$

By the triangle inequality, we have

$$\|\alpha_{t+1} - \bar{\alpha}_t\| \leq \|\alpha_{t+1} - \alpha_t\| + \|\alpha_t - \bar{\alpha}_t\| \overset{\text{Eq. (B.5)}}{\leq} \left(1 + \frac{\tau}{\eta}\right)\|\alpha_t - \alpha_{t+1}\| + \tau\lambda\|D\|\delta_t.$$

Putting these pieces together yields that

$$(\alpha_{t+1} - \bar{\alpha}_t)^\top \nabla \bar{f}(\alpha_{t+1}) \leq \left(\lambda\delta_t\|D\| + \left(\frac{\lambda^2\|D\|^2}{\mu} + \frac{1}{\eta}\right)\|\alpha_{t+1} - \alpha_t\|\right)\left(\left(1 + \frac{\tau}{\eta}\right)\|\alpha_t - \alpha_{t+1}\| + \tau\lambda\|D\|\delta_t\right).$$

By the definition of η, we have $\eta \leq 1$ and $\frac{\lambda^2\|D\|^2}{\mu} < \frac{1}{\eta}$. Combining these two inequalities with the above inequality yields that

$$(\alpha_{t+1} - \bar{\alpha}_t)^\top \nabla \bar{f}(\alpha_{t+1}) \leq \left(\lambda\delta_t\|D\| + \left(\frac{2}{\eta}\right)\|\alpha_{t+1} - \alpha_t\|\right)\left(\left(\frac{\tau+1}{\eta}\right)\|\alpha_t - \alpha_{t+1}\| + \tau\lambda\|D\|\delta_t\right).$$

which implies that

$$(\alpha_{t+1} - \bar{\alpha}_t)^\top \nabla \bar{f}(\alpha_{t+1}) \leq \left(\frac{2\tau+2}{\eta^2}\right)\|\alpha_{t+1} - \alpha_t\|^2 + \left(\frac{(3\tau+1)\lambda\|D\|}{\eta}\right)\delta_t\|\alpha_{t+1} - \alpha_t\| + \tau\lambda^2\|D\|^2\delta_t^2.$$

By the Young's inequality, we have

$$\delta_t\|\alpha_{t+1} - \alpha_t\| \leq \left(\frac{\tau\eta\lambda\|D\|}{3\tau+1}\right)\delta_t^2 + \left(\frac{3\tau+1}{4\tau\eta\lambda\|D\|}\right)\|\alpha_{t+1} - \alpha_t\|^2.$$

Combining the above two inequalities, we have

$$(\alpha_{t+1} - \bar{\alpha}_t)^\top \nabla \bar{f}(\alpha_{t+1}) \leq \left(\frac{17\tau^2 + 14\tau + 1}{4\tau\eta^2}\right)\|\alpha_{t+1} - \alpha_t\|^2 + 2\tau\lambda^2\|D\|^2\delta_t^2.$$

This together with Eq. (B.7) yields that

$$\|\alpha_{t+1} - \alpha_t\|^2 \geq \frac{4\tau\eta^2\Delta_{t+1}}{17\tau^2 + 14\tau + 1} - \left(\frac{8\tau^2\lambda^2\|D\|^2\eta^2}{17\tau^2 + 14\tau + 1}\right)\delta_t^2.$$

Nhat Ho[*,‡], Tianyi Lin[*,◇], Michael I. Jordan[◇]

This completes the proof. □

Equipped with the bounds of iterative objective gap $\Delta_t - \Delta_{t+1}$ and objective gap Δ_{t+1} in Lemma B.1 and B.2, we prove the main lemma on the number of iterations required by Algorithm 1 to reach a certain threshold with objective gap Δ_t. Before stating that result, we assume the key technical assumption with an error $\hat{\varepsilon}$ for each inner.

$$\hat{\varepsilon} = \frac{\sqrt{\varepsilon}}{2} \cdot \min \left\{ 1, \frac{\lambda\mu}{2\sqrt{\ell}} \sqrt{\frac{\tau}{C(17\tau^2 + (14 + \eta\lambda^2)\tau + 1)}} \right\}, \tag{B.8}$$

where $C > 0$ is defined as

$$C := \frac{2\tau^2\lambda^4\|D\|^2}{17\tau^2 + 14\tau + 1} + \frac{\|D\|^2}{2}.$$

Lemma B.3 *Under Assumption 3.1 and let $(\alpha_t, \beta_t)_{t \geq 0}$ be generated by Algorithm 1 with the stepsize $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1, \lambda^2\}}\})$ and $\hat{\varepsilon} > 0$ satisfies Eq. (B.8). Then, the number of iterations $T > 0$ required to reach $\Delta_T \leq (\frac{\mu^2}{8\ell})\varepsilon$ satisfies that*

$$T \leq 1 + \left(\frac{17\tau^2 + 14\tau + 1}{\tau\lambda^2\eta} \right) \log \left(\frac{16\ell\Delta_1}{\mu^2\varepsilon} \right).$$

Proof. Invoking the results from Lemma B.1 and Lemma B.2, we have

$$\begin{aligned}
\Delta_t - \Delta_{t+1} &\geq \frac{\lambda^2}{4\eta} \left(\left(\frac{4\tau\eta^2}{17\tau^2 + 14\tau + 1} \right) \Delta_{t+1} - \left(\frac{8\tau^2\lambda^2\|D\|^2\eta^2}{17\tau^2 + 14\tau + 1} \right) \delta_t^2 \right) - \left(\frac{\eta\|D\|^2}{2} \right) \delta_t^2 \\
&= \left(\frac{\tau\lambda^2\eta}{17\tau^2 + 14\tau + 1} \right) \Delta_{t+1} - \left(\frac{2\tau^2\lambda^4\|D\|^2}{17\tau^2 + 14\tau + 1} + \frac{\|D\|^2}{2} \right) \eta\delta_t^2.
\end{aligned}$$

Since $\beta_t \leftarrow \text{INNERLOOP}(f, \lambda, D, \alpha_t, \beta_{t-1}, \hat{\varepsilon})$, we have $\delta_t \leq \hat{\varepsilon}$. In addition, we define $\rho > 0$ by

$$\rho := \left(1 + \frac{\tau\lambda^2\eta}{17\tau^2 + 14\tau + 1} \right)^{-1}.$$

Then, for any $t \geq 0$, we find that

$$\Delta_t - \Delta_{t+1} \geq \left(\frac{1}{\rho} - 1 \right) \Delta_{t+1} - C\eta\hat{\varepsilon}^2,$$

which implies that

$$\Delta_{t+1} \leq \rho\Delta_t + \rho C\eta\hat{\varepsilon}^2.$$

Recursively performing the above inequality yields that

$$\Delta_t \leq \rho^{t-1}\Delta_1 + \left(\sum_{j=1}^{t-1} \rho^{t-j} \right) \cdot C\eta\hat{\varepsilon}^2 \leq \rho^{t-1}\Delta_1 + \left(\frac{C\eta}{1 - \rho} \right) \hat{\varepsilon}^2.$$

By the definition of ρ, we have

$$\frac{C\eta}{1 - \rho} = C\eta + \frac{C(17\tau^2 + 14\tau + 1)}{\tau\lambda^2}.$$

By the definition of $\hat{\varepsilon}$ in Eq. (B.8), we have $\Delta_t \leq \rho^{t-1}\Delta_1 + (\frac{\mu^2}{16\ell})\varepsilon$. Therefore, the number of iterations $T > 0$ required to reach $\Delta_T \leq (\frac{\mu^2}{8\ell})\varepsilon$ is

$$T \leq 1 + \left(\frac{17\tau^2 + 14\tau + 1}{\tau\lambda^2\eta} \right) \log \left(\frac{16\ell\Delta_1}{\mu^2\varepsilon} \right).$$

This completes the proof. □

Finally, we present the lemmas which will be used for the finite-sum setting and the online setting. Indeed, we consider the stochastic variant of the GDGA algorithm in these two settings which can be much more efficient than the deterministic counterpart in real applications (Lan, 2020). The stochastic GDGA algorithms are associated with the subroutines which are implemented using stochastic first-order optimization algorithms, e.g., Katyusha (Allen-Zhu, 2017) and optimal SGD (Rakhlin et al., 2012), and the stopping criterion $\mathbb{E}[\delta_t] \leq \hat{\varepsilon}$. We omit the proofs since they resemble the proofs for Lemma B.1, B.2 and B.3 in the deterministic setting.

Lemma B.4 *Under Assumption 3.1 and let $(\alpha_t, \beta_t)_{t \geq 0}$ be generated by Algorithm 1 with the stepsize $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1, \lambda^2\}}\})$. Then, for any $t \geq 0$, we have*

$$\mathbb{E}[\Delta_t] - \mathbb{E}[\Delta_{t+1}] \geq \left(\frac{\lambda^2}{4\eta}\right) \mathbb{E}[\|\alpha_t - \alpha_{t+1}\|^2] - \left(\frac{\eta\|D\|^2}{2}\right) \mathbb{E}[\delta_t^2].$$

Lemma B.5 *Under Assumption 3.1 and let $(\alpha_t, \beta_t)_{t \geq 0}$ be generated by Algorithm 1 with the stepsize $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1, \lambda^2\}}\})$. Then, for any $t \geq 0$, we have*

$$\mathbb{E}[\|\alpha_{t+1} - \alpha_t\|^2] \geq \left(\frac{4\tau\eta^2}{17\tau^2 + 14\tau + 1}\right) \mathbb{E}[\Delta_{t+1}] - \left(\frac{8\tau^2\lambda^2\|D\|^2\eta^2}{17\tau^2 + 14\tau + 1}\right) \mathbb{E}[\delta_t^2].$$

Lemma B.6 *Under Assumption 3.1 and let $(\alpha_t, \beta_t)_{t \geq 0}$ be generated by Algorithm 1 with the stepsize $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1, \lambda^2\}}\})$ and $\hat{\varepsilon} > 0$ satisfies Eq. (B.8). Then, the number of iterations $T > 0$ required to reach $\mathbb{E}[\Delta_T] \leq (\frac{\mu^2}{8\ell})\varepsilon$ satisfies that*

$$T \leq 1 + \left(\frac{17\tau^2 + 14\tau + 1}{\tau\lambda^2\eta}\right) \log\left(\frac{16\ell\Delta_1}{\mu^2\varepsilon}\right).$$

B.2 Proof of Theorem 5.1

We recall that $\bar{\alpha}_t$ is the projection of α_t onto the optimal set of the dual filtering-clustering model in Eq. (3.5) and $\Delta_t = \bar{f}(\alpha_t) - \bar{f}(\bar{\alpha}_t)$ is the objective gap. The following is the full version of Theorem 5.1.

Theorem B.7 *Under Assumption 3.1 and set $\eta \in (0, \min\{1, \frac{\mu}{4\|D\|^2 \max\{1, \lambda^2\}}\})$ in Algorithm 1 and $\tau > 0$ be defined in Theorem 4.5 and 4.6. Given a tolerance $\varepsilon > 0$, we let $\hat{\varepsilon} > 0$ be*

$$\hat{\varepsilon} = \frac{\sqrt{\varepsilon}}{2} \cdot \min\left\{1, \frac{\lambda\mu}{2\sqrt{\ell}}\sqrt{\frac{\tau}{C(17\tau^2 + (14 + \eta\lambda^2)\tau + 1)}}\right\},$$

where $C > 0$ is defined as

$$C := \frac{2\tau^2\lambda^4\|D\|^2}{17\tau^2 + 14\tau + 1} + \frac{\|D\|^2}{2}.$$

Then, the number of iterations required by Algorithm 1 to return an ε-optimal solution is bounded by

$$T \leq 1 + \left(\frac{17\tau^2 + 14\tau + 1}{\tau\lambda^2\eta}\right) \log\left(\frac{16\ell\Delta_1}{\mu^2\varepsilon}\right).$$

Proof. Suppose that $\beta^\star \in \mathbb{R}^d$ is an optimal solution of the filtering-clustering model in Eq. (1.1), we have

$$\|\beta_t - \beta^\star\|^2 \leq 2(\underbrace{\|\beta_t - \beta^\star(\alpha_t)\|^2}_{\mathbf{I}} + \underbrace{\|\beta^\star(\alpha_t) - \beta^\star\|^2}_{\mathbf{II}}).$$

Since $\beta_t \leftarrow \textsc{InnerLoop}(f, \lambda, D, \alpha_t, \beta_{t-1}, \hat{\varepsilon})$ and $\hat{\varepsilon} \leq \frac{\sqrt{\varepsilon}}{2}$, we have $\mathbf{I} \leq \hat{\varepsilon}^2 \leq \frac{\varepsilon^2}{4}$.

Recall that $\bar{\alpha}_t$ is the projection of α_t onto the optimal set of the dual filtering-clustering model in Eq. (3.5). Since the objective of the filtering-clustering model is strongly convex, its optimal solution is unique. Then, by the definition of $\beta^\star(\cdot)$, we have $\beta^\star = \beta^\star(\bar{\alpha}_t)$. Therefore, we conclude that

$$\mathbf{II} \overset{\text{Lemma 4.3}}{\leq} \frac{\|\lambda D^\top \alpha_t - \lambda D^\top \bar{\alpha}_t\|^2}{\mu^2} \overset{\text{Lemma 4.1}}{\leq} \frac{2\ell(f^\star(\lambda D^\top \alpha_t) - f^\star(\lambda D^\top \bar{\alpha}_t))}{\mu^2} = \frac{2\ell\Delta_t}{\mu^2}.$$

If the number of iterations $T > 0$ satisfies that

$$T > 1 + \left(\frac{17\tau^2 + 14\tau + 1}{\tau\lambda^2\eta} \right) \log \left(\frac{16\ell\Delta_1}{\mu^2\varepsilon} \right),$$

we obtain from Lemma B.3 that $\Delta_t \leq (\frac{\mu^2}{8\ell})\varepsilon$. This implies that

$$\|\beta^\star(\alpha_t) - \beta^\star\|^2 \leq \frac{\varepsilon}{4}.$$

Putting these pieces together yields that $\|\beta_t - \beta^\star\|^2 \leq \varepsilon$. □

B.3 Proof of Corollary 5.2

The proof is the same as that in Theorem 5.1. In particular, it suffices to bound $\mathbb{E}[\|\beta^\star(\alpha_t) - \beta^\star\|^2]$ via appeal to Lemma B.6 instead of Lemma B.3.

B.4 Proof of Theorem 5.4

We first present the complexity bound of the AGD-based subroutine in the following proposition.

Proposition B.8 *Under Assumption 3.1, the required number of gradient evaluations in the implementation of* $\beta_t \leftarrow \text{INNERLOOPAGD}(f(\cdot), \lambda, D, \alpha_t, \beta_{t-1}, \hat{\varepsilon})$ *is bounded by*

$$N_t \leq \sqrt{\frac{\ell}{\mu}} \cdot \begin{cases} \log\left(\frac{\|\beta_0 - \beta^\star(\alpha_1)\|}{\hat{\varepsilon}} \right), & t = 1, \\ \log\left(1 + \frac{\lambda\|D\|\|\alpha_t - \alpha_{t-1}\|}{\mu\hat{\varepsilon}} \right), & t \geq 2. \end{cases}$$

Proof. Since f is μ-strongly convex and ℓ-gradient Lipschitz, we have $f(\beta) - \lambda\alpha_t^\top D\beta$ is also μ-strongly convex and ℓ-gradient Lipschiz. For $t = 1$, the initial distance is $\|\beta_0 - \beta^\star(\alpha_1)\|$. By the convergence theory for AGD in Nesterov (2018), we have

$$N_1 \leq \sqrt{\frac{\ell}{\mu}} \log \left(\frac{\|\beta_0 - \beta^\star(\alpha_1)\|}{\hat{\varepsilon}} \right).$$

For $t \geq 2$, the initial distance is $\|\beta_{t-1} - \beta^\star(\alpha_t)\|$. Therefore, we have

$$N_t \leq \sqrt{\frac{\ell}{\mu}} \log \left(\frac{\|\beta_{t-1} - \beta^\star(\alpha_t)\|}{\hat{\varepsilon}} \right).$$

By using the triangle inequality, we have

$$\|\beta_{t-1} - \beta^\star(\alpha_t)\| \leq \|\beta_{t-1} - \beta^\star(\alpha_{t-1})\| + \|\beta^\star(\alpha_{t-1}) - \beta^\star(\alpha_t)\| \leq \hat{\varepsilon} + \|\beta^\star(\alpha_{t-1}) - \beta^\star(\alpha_t)\|.$$

Since $\beta^\star(\alpha)$ is $\frac{\lambda\|D\|}{\mu}$-Lipschitz over $\mathbb{B}_q^n$ (cf. Lemma 4.3), we have

$$\|\beta^\star(\alpha_{t-1}) - \beta^\star(\alpha_t)\| \leq \frac{\lambda\|D\|\|\alpha_t - \alpha_{t-1}\|}{\mu}.$$

Therefore, we have

$$N_t \leq \sqrt{\frac{\ell}{\mu}} \log \left(1 + \frac{\lambda\|D\|\|\alpha_t - \alpha_{t-1}\|}{\mu\hat{\varepsilon}} \right).$$

This completes the proof. □

Equipped with the result in Proposition B.8, we are ready to derive the complexity bound of deterministic GDGA in terms of the number of gradient evaluations.

By the definition, we have $N = N_1 + \sum_{t=2}^{T} N_t$. Then, Proposition B.8 implies that

$$N \leq \sqrt{\frac{\ell}{\mu}} \left((T-1) \log \left(1 + \frac{\lambda \|D\| \bar{D}_q}{\mu \hat{\varepsilon}} \right) + \log \left(\frac{\|\beta_0 - \beta^\star(\alpha_1)\|}{\hat{\varepsilon}} \right) \right).$$

By Theorem 5.1, we have

$$T - 1 \leq \left(\frac{17\tau^2 + 14\tau + 1}{\tau \lambda^2 \eta} \right) \log \left(\frac{16\ell \Delta_1}{\mu^2 \varepsilon} \right).$$

Putting these pieces together yields the desired result.

B.5 Proof of Theorem 5.5

We first present the complexity bound of the Katyusha-based subroutine in the following proposition.

Proposition B.9 *Under Assumption 3.1, the required number of gradient evaluations in the implementation of* $\beta_t \leftarrow \textsc{InnerLoopKatyusha}(f(\cdot) = \frac{1}{M}(\sum_{i=1}^{M} f_i(\cdot)), \lambda, D, \alpha_t, \beta_{t-1}, \hat{\varepsilon})$ *is bounded by*

$$N_t \leq C_{\text{Katyusha}} \cdot \begin{cases} M + \sqrt{\frac{\ell M}{\mu}} \log \left(\frac{\ell \|\beta_0 - \beta^\star(\alpha_1)\|}{\mu \hat{\varepsilon}} \right), & t = 1, \\ M + \sqrt{\frac{\ell M}{\mu}} \log \left(\frac{\ell}{\mu} + \frac{\lambda \ell \|D\| \|\alpha_t - \alpha_{t-1}\|}{\mu^2 \hat{\varepsilon}} \right), & t \geq 2, \end{cases}$$

where $C_{\text{Katyusha}} > 0$ *is a universal constant defined in Allen-Zhu (2017, Theorem 2.1) and independent of* ℓ, μ, $\|D\|$ *and* $1/\varepsilon$.

Proof. Since f is μ-strongly convex and ℓ-gradient Lipschitz, we have $f(\beta) - \lambda \alpha_t^\top D \beta$ is μ-strongly convex and ℓ-gradient Lipschiz. For $t = 1$, the initial distance is $\|\beta_0 - \beta^\star(\alpha_1)\|$. By the convergence theory for Katyusha (Allen-Zhu, 2017, Theorem 2.1) with the suggested stepsize rule $\max\{\frac{2}{3\ell}, \frac{1}{\sqrt{3M\mu\ell}}\}$, we have

$$N_1 \leq C_{\text{Katyusha}} \cdot \left(M + \sqrt{\frac{\ell M}{\mu}} \log \left(\frac{\ell \|\beta_0 - \beta^\star(\alpha_1)\|}{\mu \hat{\varepsilon}} \right) \right).$$

For $t \geq 2$, the initial distance is $\|\beta_{t-1} - \beta^\star(\alpha_t)\|$. Therefore, we have

$$N_t \leq C_{\text{Katyusha}} \cdot \left(M + \sqrt{\frac{\ell M}{\mu}} \log \left(\frac{\ell \|\beta_{t-1} - \beta^\star(\alpha_t)\|}{\mu \hat{\varepsilon}} \right) \right).$$

Applying the similar argument as that in the proof of Proposition B.8, we have

$$N_t \leq C_{\text{Katyusha}} \cdot \left(M + \sqrt{\frac{\ell M}{\mu}} \log \left(\frac{\ell}{\mu} + \frac{\lambda \ell \|D\| \|\alpha_t - \alpha_{t-1}\|}{\mu^2 \hat{\varepsilon}} \right) \right).$$

This completes the proof. □

Equipped with the result in Proposition B.9, we are ready to derive the complexity bound of stochastic variance reduced GDGA in terms of the number of component gradient evaluations.

By the definition, we have $N = N_1 + \sum_{2=1}^{T} N_t$. Then, Proposition B.9 implies that

$$N \leq C_{\text{Katyusha}} \cdot \left(TM + (T-1)\sqrt{\frac{\ell M}{\mu}} \log \left(\frac{\ell}{\mu} + \frac{\lambda \ell \|D\| \bar{D}_q}{\mu^2 \hat{\varepsilon}} \right) + \sqrt{\frac{\ell M}{\mu}} \log \left(\frac{\ell \|\beta_0 - \beta^\star(\alpha_1)\|}{\mu \hat{\varepsilon}} \right) \right).$$

By Corollary 5.2, we have

$$T - 1 \leq \left(\frac{17\tau^2 + 14\tau + 1}{\tau \lambda^2 \eta} \right) \log \left(\frac{16\ell \Delta_1}{\mu^2 \varepsilon} \right).$$

Putting these pieces together yields the desired result.

B.6   Proof of Theorem 5.6

We first present the complexity bound of the SGD-based subroutine in the following proposition.

Proposition B.10 *Under Assumption 3.1, the required number of gradient evaluations in the implementation of $\beta_t \leftarrow \textsc{InnerLoopSgd}(f(\cdot) = \mathbb{E}_P[F(\cdot, \xi)], \lambda, D, \alpha_t, \beta_{t-1}, \hat{\varepsilon})$ is bounded by*

$$N_t \leq \frac{4G^2}{\mu^2 \hat{\varepsilon}^2}, \quad \text{for all } t \geq 0.$$

where $G > 0$ refers to the bound for stochastic gradient oracle in Assumption 3.1.

Proof. Since f is μ-strongly convex and ℓ-gradient Lipschitz, we have $f(\beta) - \lambda \alpha_t^\top D\beta$ is μ-strongly convex and ℓ-gradient Lipschiz. Under Assumption 3.1, we have the stochastic gradient oracle $g(\cdot, \xi)$ is unbiased and bounded: $\mathbb{E}[\|g(\beta, \xi)\|^2] \leq G^2$ for some $G > 0$. By the convergence theory for optimal SGD (Rakhlin et al., 2012, Lemma 1) with the suggested stepsize rule $\frac{1}{\mu k}$ (k refers to SGD iteration count), we have

$$N_t \leq \frac{4G}{\mu^2 \hat{\varepsilon}^2}, \quad \text{for all } t \geq 1.$$

This completes the proof.   □

Equipped with the result in Proposition B.10, we are ready to derive the complexity bound of stochastic GDGA in terms of the number of stochastic gradient evaluations.

By the definition, we have $N = \sum_{t=1}^T N_t$. Then, Proposition B.10 implies that

$$N \leq \left(\frac{4G^2}{\mu^2 \hat{\varepsilon}^2} \right) T.$$

By Corollary 5.2, we have

$$T \leq 1 + \left(\frac{17\tau^2 + 14\tau + 1}{\tau \lambda^2 \eta} \right) \log \left(\frac{16\ell \Delta_1}{\mu^2 \varepsilon} \right).$$

Putting these pieces together yields the desired result.

C   Additional Experimental Results

Setup.   We include ADMM, specialized ADMM (Ramdas and Tibshirani, 2016) and projected Newton method (Wang et al., 2016b) as the baseline approaches and consider three real images with various sizes: 128 by 128 pixels (small image), 256 by 256 pixels (medium image) and 512 by 512 pixels (large image)[3]. All the algorithms are evaluated as the order k varies in the discrete difference operator $D^{(k+1)}$ in which the evaluation metric is the objective function value. We choose $\lambda = 0.2$ in our experiment and consider the adaptive step-size $\eta_t > 0$ based on Barzilai-Borwein rule which is omitted due to space limit.

---

[3]These images can be found at: `http://sipi.usc.edu/database/database.php?volume=misc`
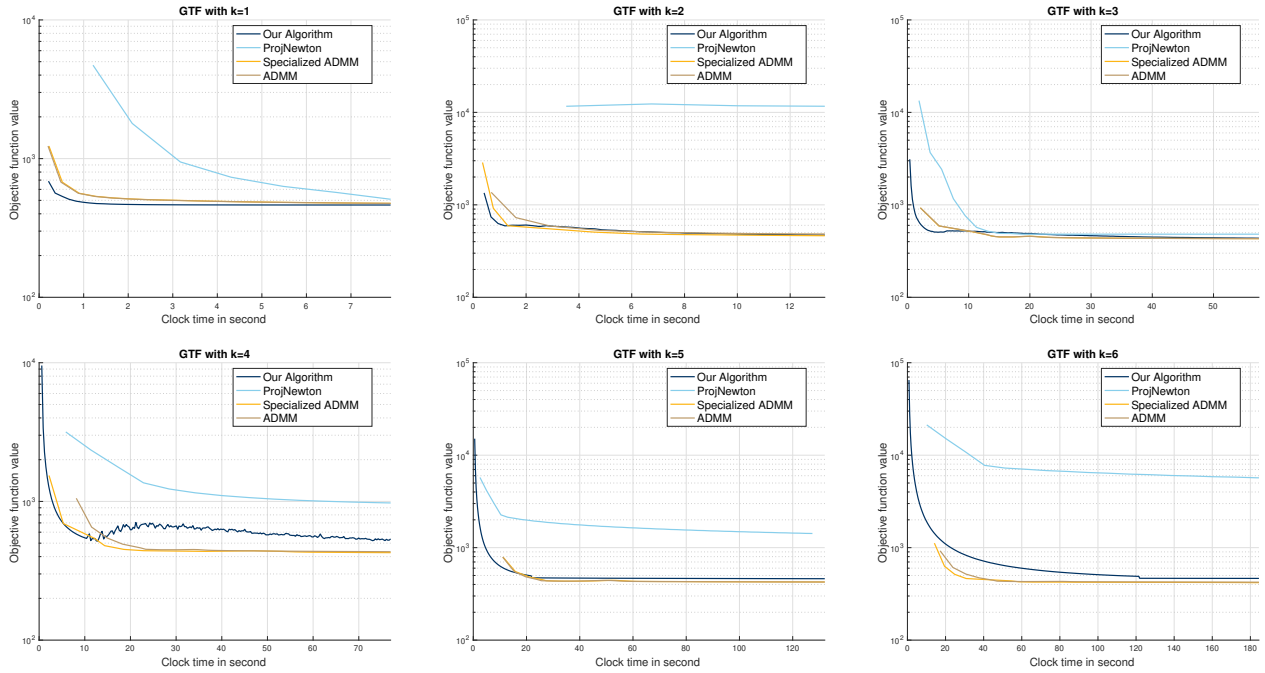
Figure 2: Comparison of Algorithm 2, ADMM, specialized ADMM, and projected Newton method on medium image.
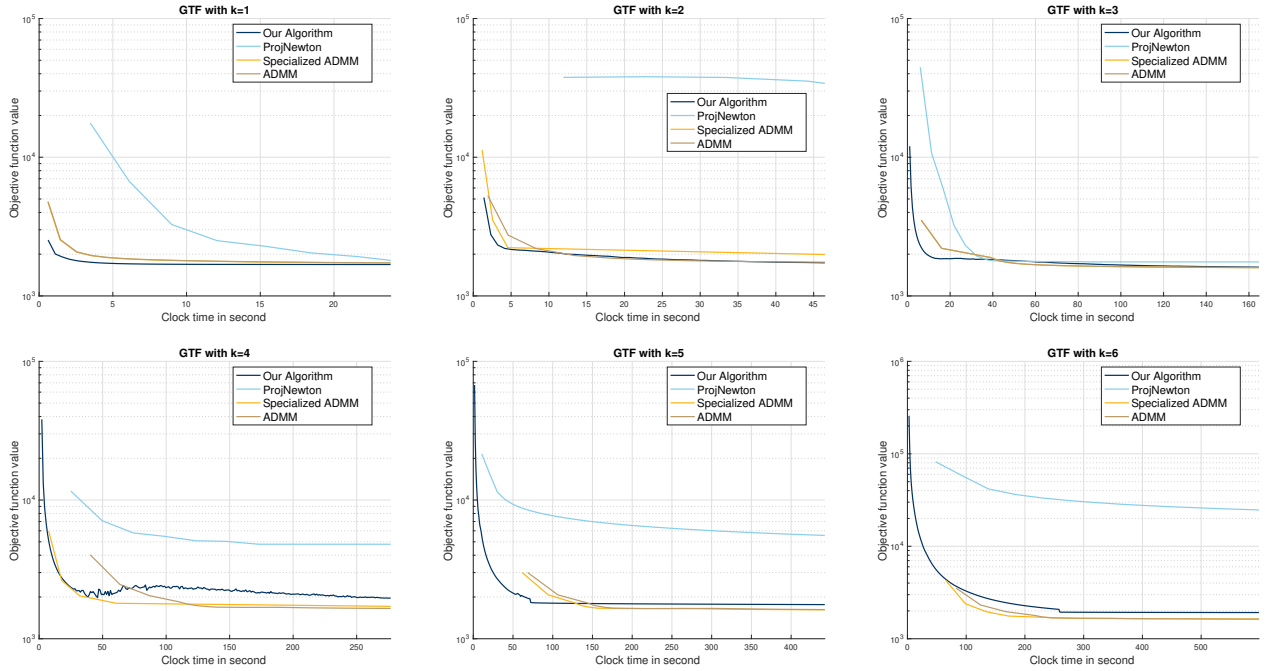

Figure 3: Comparison of Algorithm 2, ADMM, specialized ADMM, and projected Newton method on large image.