OJMO

Open Journal of Mathematical Optimization

Yifan Sun & Francis Bach

Screening for a Reweighted Penalized Conditional Gradient Method

Volume 3 (2022), article no. 3 (35 pages)

https://doi.org/10.5802/ojmo.14

Article submitted on December 6, 2021, revised on May 20, 2022,
accepted on June 10, 2022.

Open Journal of Mathematical Optimization is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/

CENTRE MERSENNE

Screening for a Reweighted Penalized Conditional Gradient Method

Yifan Sun
Stonybrook University - Department of Computer Science, Stonybrook, New York, USA
yifan.sun@stonybrook.edu

Francis Bach
INRIA - Département d'Informatique de l'Ecole Normale Supérieure PSL Research University Paris, France
francis.bach@inria.fr

Abstract

The conditional gradient method (CGM) is widely used in large-scale sparse convex optimization, having a low per iteration computational cost for structured sparse regularizers and a greedy approach for collecting nonzeros. We explore the sparsity acquiring properties of a general penalized CGM (P-CGM) for convex regularizers and a reweighted penalized CGM (RP-CGM) for nonconvex regularizers, replacing the usual convex constraints with gauge-inspired penalties. This generalization does not increase the per-iteration complexity noticeably. Without assuming bounded iterates or using line search, we show $O(1/t)$ convergence of the gap of each subproblem, which measures distance to a stationary point. We couple this with a screening rule which is safe in the convex case, converging to the true support at a rate $O(1/(\delta^2))$ where $\delta \geq 0$ measures how close the problem is to degeneracy. In the nonconvex case the screening rule converges to the true support in a finite number of iterations, but is not necessarily safe in the intermediate iterates. In our experiments, we verify the consistency of the method and adjust the aggressiveness of the screening rule by tuning the concavity of the regularizer.

Digital Object Identifier 10.5802/ojmo.14

Keywords Dual screening, conditional gradient method, atomic sparsity, reweighted optimization.

Acknowledgments This work was funded in part by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). We also acknowledge support the European Research Council (grant SEQUOIA 724063). The first is also funded in part by AXA pour la recherche and Kamet Ventures, as well as a Google focused award.

1 Introduction

Conditional gradient methods (CGMs) are used in constrained optimization to quickly arrive at sparse solutions of large-scale optimization problems. In this paper, we generalize their applicability to nonconvex penalized (unconstrained) problems and investigate safe screening methods to obtain sparse supports in finite time. We describe these problems as

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \;\; f(x) + \phi(r_{\mathcal{P}}(x)), \tag{1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a convex loss function with an L-Lipschitz continuous gradient, $\phi : \mathbb{R}_+ \to \mathbb{R}$ is a strictly convex monotonically increasing function, and $r_{\mathcal{P}} : \mathbb{R}^d \to \mathbb{R}_+$ a nonconvex variant of a gauge function, defined as the solution to

$$r_{\mathcal{P}}(x) = \min_{c_p \geq 0} \left\{ \sum_{p\in\mathcal{P}_0} \gamma(c_p)p : \sum_{p\in\mathcal{P}_0} c_p p = x \right\} \tag{2}$$

for some concave monotonically increasing function $\gamma : \mathbb{R}_+ \to \mathbb{R}_+$. Here, $\mathcal{P}_0$ is a finite collection of vectors in $\mathbb{R}^d$. In the usual nonzero sparsity case, this penalty reduces to well-studied nonconvex penalties like SCAD, LSP, or p-"norms" for $0 < p < 1$. Problems of this form arise in machine learning, compressed sensing, low-rank matrix factorization, etc., and are often observed in practice to be more effective sparsifiers than their convex relaxations [17].

In particular, we solve (1) using the following iteration scheme

$$s^{(t)} = \operatorname*{argmin}_{s\in\mathbb{R}^d} \nabla f(x^{(t)})^T s + \overline{h}^{(t)}(s), \tag{Min-Maj}$$

$$x^{(t+1)} = (1-\theta^{(t)})x^{(t)} + \theta^{(t)} s^{(t)}, \tag{Merge}$$

where $\overline{h}^{(t)}(s)$ is a local convexification of $\phi(r_{\mathcal{P}}(s))$ at $x^{(t)}$. We call this the reweighted penalized conditional gradient method (RP-CGM), as it resembles both the conditional gradient method (CGM) in sparse convex optimization and reweighting schemes in majorization-minorization methods for nonconvex optimization.

▶ **Example 1.** The ℓ_1 norm is formed by picking $\mathcal{P}_0 = \{\pm e_1, \dots, \pm e_d\}$ the signed unit bases, and $\gamma(\xi) = \xi$. Then the solution to (2) is always unique and can be expressed in closed form as $r_{\mathcal{P}}(x) = \|x\|_1$. Picking instead a concave penalty $\gamma(\xi) = 2\sqrt{\xi}$ leads to the variation $r_{\mathcal{P}}(x) = 2\sum_i \sqrt{|x_i|}$ the "half norm". Similar transformations also lead to the smoothed capped absolute deviation (SCAD) penalty, minimum concave penalty (MCP), etc. (See Table 1.)

By using a generalized convex aggregate penalty ϕ, we can sweep the space between constrained and unconstrained problems, via the penalty's tunable curvature: maximum curvature reduces to the usual constrained problem, and minimum curvature to the usual LASSO penalty problem. The addition of the nonconvex elementwise term γ strengthens the sparsifying behavior. However, because of the sometimes erratic way that the conditional gradient method picks step directions, simple implementations of these features easily lead to divergence. Therefore, a main contribution of this work is to identify carefully the conditions on ϕ and γ such that these two modified CGMs perform optimally.

The other main contribution of this work concerns *safe screening*, in which the variable search space is reduced dynamically by identifying which components will safely not appear in the converged solution. For example, in nonzero sparsity, we identify early on the indices i in which we are guaranteed that $x_i^* = 0$, in hopes of prematurely estimating the solution sparsity pattern. This technique is intended to reduce memory and computational cost.

1.1 Related work

1.1.1 Conditional gradient method

When $h(s) = \iota_{\overline{\mathcal{P}}}(s)$ the indicator for s in $\mathcal{P}$, the proposed method is the conditional gradient method (CGM) [24, 29]. Also called the Frank-Wolfe method, it has been studied since the 50s and was revitalized recently [39] for its success at quickly estimating solutions to sparse optimization problems. Because this foundational method serves as a baseline, we will refer to it as the "vanilla CGM".

This method is particularly useful when the computation of the supporting hyperplane in the (Min-Maj) step is cheap (e.g., when $\mathcal{P}$ is the unit ball of the ℓ_1-norm or a group norm). Much work has come from expanding its use to general (atomic) norms [20, 38, 39, 62] with many variations such as backward steps [42, 59] and fully-corrective steps [65]. Many connections between the CGM and existing methods have also been discovered, such as to mirror descent [2], cutting plane method [72], and greedy coordinate-wise methods [20]. In its simplest version (with no away-steps, line search, or strongly convex assumptions on f or $\mathcal{P}$) the minimum duality gap in CGM converges at rate $O(1/t)$ [24].

1.1.2 Convex gauge function

When $\gamma(c_p) = c_p$, we define $\kappa_{\mathcal{P}}(x) := r_{\mathcal{P}}(x)$, which reduces to the usual convex gauge function for the closed convex set $\mathcal{P}$ [30, 60]. Gauge functions can be seen as generalized versions of the ℓ_1-norm, which is a convex promoter of nonzero vector sparsity, and include penalties like the total variation (TV) norm, nuclear norm, OWL norm [71], OSCAR norm [6], and general conic constraints. Several works have looked at optimization over general gauges [30, 32] and in particular for sparse optimization [15, 39].

1.1.3 Penalized CGM

When $\overline{h}^{(t)}(s)$ is a convex penalty, we refer to the proposed method as the penalized CGM (P-CGM). Compared to CGM, P-CGM has been much less studied [36, 50, 69], and has appeared under different names, like regularized coordinate minimization [23]. An $O(1/t)$ convergence rate has been shown for specific smooth functions [50], with

bounded assumptions on iterates [2], or with improvement steps to ensure boundedness of sublevel sets [36, 69]. When f is quadratic and for a special form of ϕ, the P-CGM can be shown to be equivalent to a form of the iterative shrinkage method, and under proper problem conditioning, has linear convergence [9, 10].

1.1.4 Reweighted methods for nonconvex minimization

Our main algorithmic novelty is to solve a sequence of reweighted penalized CGM (RP-CGM) iterations in order to accommodate nonlinear γ, which appear in nonconvex penalties like SCAD or MCP penalties in difference-of-convex or majorization-minimization methods. This results in a nonconvex penalty $h(x)$, which in practice have been shown to have superior sensing properties [17, 21, 26, 33, 48, 56, 67, 68]. We leverage these observations to improve the screening properties of RP-CGM; by increasing the concavity of γ, we can create an aggressive support recovery method based on an easily computable duality-gap-like residual.

1.1.5 Applications

A main use case of CGMs is in finding generalized sparse solutions to convex losses [15, 39], where the ℓ_1-norm penalty, which promotes element-wise sparsity [13, 14, 22, 63], is generalized to gauge functions that promote sparsity with respect to "atoms", or low dimensional facets of a convex set. This generalizes sparse optimization to applications such as low-rank matrix optimization [31, 69] and grouped feature extraction [6, 64, 71]. Additionally, these atoms may be feasible solutions to combinatorial problems, such as in submodular optimization [1] and object tracking [16]. CGM has also been applied to a variety of machine learning tasks, such as graphical models [41], multitask learning [61], SVMs [43], particle filtering [44], and deep learning [5, 57].

1.1.6 Safe screening

A *screening rule* returns an estimate of the support of x^* given a noisy approximation x. The screening rule is *safe* if there are no false positives (and called *sure* if there are no false negatives). Safe screening rules for LASSO were first proposed by [25], and have since been extended to a number of smooth losses and generalized penalties [7, 28, 46, 49, 51, 66]. An interesting related work is the "stingy coordinate descent" method [40] for LASSO, which optimizes the sparse regularized problem in a CGM-like manner, but uses screening to dynamically skip steps; this kind of method can be extended to P-CGM as well for generalized atoms. In nonconvex optimization, support recovery is discussed by [12] for handling nonlinear constraints which are iteratively linearized, and screening rules by [58] are proposed for a reweighted proximal gradient method.

1.2 Contributions and outline

We analyze the support recovery and convergence properties of P-CGM and RP-CGM on (1). We assume that the loss function f is L-smooth, the function ϕ grows at least asymptotically quadratically, the function γ has slope bounded away from 0 and $+\infty$, and the set $\mathcal{P}_0$ is either finite or a union of a finite set and a nonoverlapping cone. We give three main contributions.

- Under mild assumptions the RP-CGM converges to a stationary point. In particular, *without boundedness assumptions on iterates*, using the deterministic step size schedule of $\theta^{(t)} = 2/(1+t)$, the function value error and gap-like residual of RP-CGM converge as $O(1/t)$.
- We offer an online gap-based screening rule, which at each iteration removes some of the non-support atoms of the true solution x^*. This method is safe for convex penalties and a useful heuristic for nonconvex penalties; for all penalties it converges in finite time to the true support. Having this information can improve caching for improving subproblem efficiency, and can be used in two-stage methods if the method is ended early.
- In general, CGM without line search or away steps does not guarantee finite-time support recovery. We thus give a finite-time support identification rate of $O(1/\delta^2)$ on the post-screened atoms, where δ is a problem-dependent conditioning parameter that measures its distance to degeneracy.

We present the RP-CGM in three stages, with increasing complexity. In Section 2 we consider the nonconvex element-wise penalty, giving the key intuition behind the general method, with simple proofs and analysis. In Section 3 we consider the generalized convex gauge penalized problem, using P-CGM, and show how to handle simple recession cones in $\mathcal{P}$. Finally, in Section 4, we introduce reweighting of the gauge penalties, and give fully general convergence results and screening rules. Experimental results suggest promising method behavior in Section 5.

2 Reweighted Penalized CGM for simple sparse recovery

In this section, we introduce the RP-CGM over problems intending to regularize for nonzero elementwise sparsity. The goal is to present a simple implementation of the full method, to clearly describe the implementation and screening steps, and give intuition to its analysis. Later, we will expand the analysis for more generalized problems.

We begin by considering the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ F(x) := f(x) + \underbrace{\phi(r(x))}_{h(x)}, \qquad r(x) = \sum_{i=1}^{d} \gamma(|x_i|). \tag{3}$$

This is the simplification of (1) with $r := r_{\mathcal{P}}$ and $\mathcal{P}_0 = \{\pm e_1, \dots, \pm e_d\}$ the signed unit basis. The more general case of the $r_{\mathcal{P}}$ gauge-like penalty follows a similar analysis to what is presented in this section, and can be viewed intuitively as sparsity in a preimage space.

2.1 Reweighted penalized CGM

Inspired by methods in majorization-minimization and difference-of-convex literature, we propose the RP-CGM, which at each iteration takes a penalized conditional gradient step over the following convex proxy problem

$$\min_{x\in\mathbb{R}^d}\ \overline{F}(x; x^{(t)}) := f(x) + \phi\left(r(x^{(t)}) - \bar{r}(x^{(t)}; x^{(t)}) + \bar{r}(x; x^{(t)})\right), \tag{4}$$

where $\bar{r}(x;\bar{x}) := \sum_i \gamma'(|\bar{x}_i|)|x_i|$ is the linearized function of r with reference point $\bar{x}$. We summarize the linearized function in terms of a slope and offset

$$w_i = \gamma'(|x_i^{(t)}|), \qquad r_0 = r(x^{(t)}) - \bar{r}(x^{(t)}, x^{(t)}).$$

The RP-CGM on (3) runs by repeatedly iterating

$$s^{(t)} = \underset{s\in\mathbb{R}^d}{\text{argmin}}\ \nabla f(x)^T s + \phi\left(r_0 + \sum_i w_i|s_i|\right), \tag{5}$$

$$x^{(t+1)} = x^{(t)} + \theta^{(t)}(s^{(t)} - x^{(t)}), \tag{6}$$

for some predetermined decaying step size sequence $\theta^{(t)} = O(1/t)$. We decompose step (5) as follows. First, assigning the reweighted variables

$$u_i = s_i w_i, \quad v_i = -\nabla f(x)_i / w_i; \tag{7}$$

then (5) is equivalently expressed as

$$u = \underset{u\in\mathbb{R}^d}{\text{argmax}}\ v^T u - \phi(r_0 + \|u\|_1), \tag{8}$$

which incidentally is also the conjugate function of $g(u) = \phi(r_0 + \|u\|_1)$. Now, we further simplify the task by dividing u into a direction and magnitude

$$\widehat{u} = \frac{1}{\|u\|_1} u, \qquad \xi = \|u\|_1.$$

Then, because $\widehat{u}$ and ξ can be optimized *independently*, (8) can be further simplified to two *separable* problems:

$$\widehat{u} = \underset{u\in\mathbb{R}^d}{\text{argmax}}\,\{v^T u : \|u\|_1 = \xi\}, \qquad \xi = \underset{\xi}{\text{argmax}}\ \|v\|_\infty\, \xi - \phi(r_0 + \xi).$$

Solving for $\widehat{u}$ is exactly the same as the usual LMO for vanilla CGM, and is simply $\widehat{u} = \mathbf{sign}(v_k)e_k$ where $k = \text{argmax}_k\ |v_k|$. Solving for ξ is at worse a 1-D convex optimization problem, which can be solved efficiently via bisection. However, if we pick ϕ cleverly, then recognizing that the convex conjugate $\phi^*(\nu) = \max_\xi \nu\xi - \phi(\xi)$, then the optimal $\xi + r_0 = (\phi^*)'(\nu)$ the derivative of ϕ^*. (To relate to the vanilla CGM, where $\phi(\xi) = \iota_{\cdot\le 1}(\xi)$, the convex conjugate $\phi^*(\nu) = \nu$ and is always optimized at $\xi = 1$.) This leads to the efficient generalization of CGM in Alg. 1.

Algorithm 1 RP-CGM on simple sparse optimization

procedure RP-CGM(f, ϕ, γ, T)
  Initialize with any $x^{(0)} \in \mathbb{R}^d$.
  for $t = 1, \dots, T$ **do**
    Compute negative gradient $z = -\nabla f(x^{(t)})$.
    Compute reweighting terms in three steps. ▷ Reweight
      1. Compute weights $w_i = \gamma'(|x_i^{(t)}|)$ for $i = 1, \dots, d$.
      2. Compute offset $r_0 = r(x^{(t)}) - \bar{r}(x^{(t)}; x^{(t)})$.
      3. Compute reweighted negative gradient $v_i = z_i / w_i$ for $i = 1, \dots, d$.
    Compute next atom $s = \xi \, \mathbf{sign}(v_k) e_k$ in two steps. ▷ Min-maj
      1. Find the maximizing index $k = \operatorname{argmax}_i \ |v_i|$.
      2. Compute the magnitude $\xi = (\phi^*)'(\|v\|_\infty) - r_0$.
    Update $x^{(t+1)} = (1 - \theta^{(t)}) x^{(t)} + \theta^{(t)} s$ where $\theta^{(t)} = 2/(1+t)$. ▷ Merge
  return $x^{(T)}$

2.1.1 The convex penalty function ϕ

The vanilla CGM is written as an optimization function over a bounded set

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \ \{f(x) : x \in \mathcal{P}\}, \tag{9}$$

where $\mathcal{P}$ is some closed compact set. For example, a common choice of $\mathcal{P}$ is a norm ball. By introducing ϕ, we allow the problem statement to generalize not just to convex sets, but convex penalties as well. Specifically, let us first constrain $\gamma(x_i) = |x_i|$. Then if $\phi(\xi) = \iota_{\cdot \le 1}(\xi)$ is an indicator function, then (3) is equivalent to (9) where $\mathcal{P}$ is the ℓ_1-norm ball. On the other extreme, if we allow $\phi(\xi) = \xi$, (3) resembles the usual LASSO penalized problem for sparse optimization. This type of problem poses a big problem in the RP-CGM world, since the conjugate function $\phi^*(\nu) = \iota_{\cdot \le 1}(\nu)$ and the recovered ξ will either be 0 (no step) or $+\infty$ (diverge right away). Therefore, it is clear that some curvature must be imposed upon ϕ for Algorithm 1 to be convergent.

▷ Assumption 1 (Lower quadratic bound). We assume ϕ is lower-bounded by a quadratic function $\phi(\xi) \ge \mu_\phi \xi^2 - \phi_0$, for some $\mu_\phi > 0$ and ϕ_0.

This minimum curvature assumption is also essential for convergence analysis. Under the usual CGM framework, each new iterate $s \in \mathcal{P}$ is by design bounded, so as long as $\theta^{(t)}$ decays, convergence is guaranteed. In the P-CGM and RP-CGM case, Assumption 1 is much weaker than boundedness, and leads to the following growth property.

▶ **Lemma 2.** *If Assumption 1 holds, then ϕ^* is smooth everywhere, and the derivative of ϕ^* is asymptotically nonexpansive; e.g., for some finite-valued ξ_0, $(\phi^*)'(\nu) \le \frac{\nu}{\mu_\phi} + \xi_0$.*

The proof is in Appendix A. Since $\xi = (\phi^*)'(\nu)$ will be the magnitude of each new step, this Lemma says that ξ can grow at most linearly with ν, the magnitude of the gradient. We can interpret this as a relaxation of a boundedness assumption to a controlled growth assumption, which is not fully general, but still much more relaxed.

▶ **Example 3** (Monomials)**.** For $1 \le \alpha, \beta \le +\infty$, the following $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ and $\phi^* : \mathbb{R}_+ \to \mathbb{R}_+$ form a conjugate pair:

$$\phi(\xi) = \frac{1}{\alpha} \xi^\alpha, \qquad \phi^*(\nu) = \frac{1}{\beta} \nu^\beta, \qquad \frac{1}{\alpha} + \frac{1}{\beta} = 1.$$

In particular, in the case that $\alpha = 1$, then $\beta \to +\infty$, and the function

$$\phi^*(\nu) = \lim_{\beta \to +\infty} \frac{1}{\beta} \nu^\beta = \begin{cases} 0, & \nu \le 1 \\ +\infty, & \nu > 1. \end{cases}$$

In this case, whenever $\nu > 1$ then $\phi^*(\nu) = +\infty$; we exclude this case as P-CGM will not converge. When $\alpha \ge 2$, ϕ is strongly convex and we can show $O(1/t)$ convergence of P-CGM. When $1 < \alpha < 2$, $\phi^*(\nu)$ is finite and the iterates are well-defined, but the method may converge or diverge.

Table 1 A list of several popular concave penalties, and their slope behavior at extremities. The last entry shows the effect of the piecewise construction, which becomes linear with non-zero slope at large values of ξ.

	$\gamma(c)$	$\lim_{c\to 0}\gamma'(c)$	$\lim_{c\to+\infty}\gamma'(c)$
Fractional fns	$q^{-1}c^q$, $0<q<1$	$+\infty$	0
LSP	$\log(1+\lvert c\rvert/\theta)$ for $\theta>0$	θ^{-1}	0
SCAD	$\begin{cases}\lambda\lvert c\rvert & \lvert c\rvert\le\lambda,\\ \frac{-c^2+2\theta\lambda\lvert c\rvert-\lambda^2}{2(\theta-1)} & \lambda<\lvert c\rvert\le\theta\lambda,\\ (\theta+1)\lambda^2/2 & \lvert c\rvert\ge\theta\lambda,\end{cases}$ for $\theta>2$	λ	0
MCP	$\begin{cases}\lambda\lvert c\rvert-c^2/(2\theta) & \lvert c\rvert\le\theta\lambda,\\ \theta\lambda^2/2 & \lvert c\rvert>\theta\lambda,\end{cases}$ for $\theta>0$	λ	0
Locally convex	(12), given γ_0 and $\bar\xi$	$\lim_{c\to 0}\gamma_0'(c)$	$\gamma_0'(\bar\xi)$

▶ **Example 4** (Barrier functions)**.** Consider

$$\phi(\xi)=-\frac{1}{\beta}\log(C-\xi)-\frac{\xi}{C\beta}+\frac{\log(C)}{\beta}, \tag{10}$$

which is a log-barrier penalization function for $\xi\le C$; as $\beta\to+\infty$, $\phi(\xi)$ approaches the indicator function for this constraint. Its conjugate is

$$\phi^*(\nu)=C\nu-\beta^{-1}\log(C\beta\nu+1),$$

achieved at $\xi=C^2\beta\nu/(C\beta\nu+1)$. For all $C>0,\beta>0$, and $\nu\neq-(C\beta)^{-1}$, both ϕ^* and ξ^* exist and are finite. Note also the implicit constraint, as $\phi(\kappa_{\mathcal{P}}(x))$ is finite only if $x\in C\mathcal{P}$.

2.2 The concave sparsifier γ

The function γ is inspired by concave regularization functions like the LSP or fractional p-norms, that have been shown in practice to more aggressively enforce sparsity. Other popular concave penalties are listed in Table 1; a more complete table is given by [33, 58].

The linearization (4), given γ concave, is a majorant of (3)

$$\sum_{i=1}^d \underbrace{\gamma(|x_i^{(t)}|)}_{r(x^{(t)})}+\underbrace{\gamma'(|x_i^{(t)}|)(|x_i|-|x_i^{(t)}|)}_{r_0:=\bar r(x;x^{(t)})-\bar r(x^{(t)};x^{(t)})}\ge\sum_{i=1}^d\gamma(|x_i|)=r(x) \tag{11}$$

and is exactly equal when $x^{(t)}$ reaches a stationary point. However, actually computing the reweighted LMO can be numerically ill-defined if $w_i=\gamma'(|x_i^{(t)}|)$ is either 0 or $+\infty$, since the reweighted variables (7) will be ill-defined. This leads us to impose Assumption 2 on γ.

▷ Assumption 2 (γ). Assume that $\gamma:\mathbb{R}_+\to\mathbb{R}_+$ is concave, monotonically increasing, and differentiable everywhere on its domain, and the derivative $\gamma'(\xi)$ is lower and upper bounded by

$$0<\gamma_{\min}:=\lim_{\xi\to\infty}\gamma'(\xi)\le\gamma'(\xi)\le\lim_{\xi\to 0^+}\gamma'(\xi)=:\gamma_{\max}<+\infty,\quad\forall\,\xi\ge 0.$$

Additionally, $\gamma(0)=0$.

Note that the standard nonconvex sparsifiers (SCAD, MCP, LSP, p-norm for $p<1$) do not satisfy these assumptions, and when used directly in this reweighting scheme will cause numerical instability. Therefore, we make the following modifications, to ensure stability of RP-CGM.

- In cases where $\gamma'(\xi)\to+\infty$ as $\xi\to 0$, we modify to $\widehat\gamma(\xi)=\gamma(\xi+\xi_0)$ for some hyperparameter $\xi_0>0$.

- In cases where $\gamma'(\xi) \to 0$ as $\xi \to +\infty$, we use a piecewise linear extension given a "boundary point" $\bar{\xi}$:

$$\widehat{\gamma}(\xi) = \begin{cases} \gamma(\xi), & 0 \le \xi \le \bar{\xi}, \\ \gamma'(\bar{\xi})(\xi - \bar{\xi}) + \gamma(\bar{\xi}), & \xi > \bar{\xi}. \end{cases} \tag{12}$$

  See also Figure 1.

It is interesting to note that though we do not use the "full effect" of these canonical sparsifiers, we are able to leverage their aggressive sparsifying effect. When even a very small amount of nonconvex curvature is present, we notice a significant benefit in the numerical experiments in terms of screening and sparsification of the final solution.

2.3 Stationary points and support recovery

We define the *support of* x as the indices of the nonzeros as $\mathbf{supp}(x) = \{i : x_i \neq 0\}$. For a method producing iterates $x^{(1)}, x^{(2)}, \to x^*$, we say that this method has *recovered the support at iteration* $\bar{t}$ if for all $t \ge \bar{t}$, $\mathbf{supp}(x^{(t)}) = \mathbf{supp}(x^*)$.

For a continuous function $h : \mathbb{R}^d \to \mathbb{R}$, the point x^* is a *Clarke stationary point* of (3) if $0 \in \nabla f(x^*) + \partial h(x^*)$ where $\partial h(x) = \mathbf{conv}\{\lim_{x' \to x} \nabla h(x')\}$ is the *Clarke subdifferential* of h at x [18, 19]. Given Assumptions 1 and 2, the Clarke subdifferential for $h(x)$ is [1]

$$(\partial h(x))_i = \begin{cases} \{g_\phi \, \mathbf{sign}(x_i) \, \gamma'(|x_i|) : g_\phi \in \partial \phi(r(x))\} & x_i \neq 0, \\ \phi'(0) \cdot [-\gamma_{\max}, \gamma_{\max}], & x_i = 0, \end{cases}$$

where we use the $\cdot$ notation here for scaling elements in a set ($\alpha \cdot \mathcal{S} = \{\alpha x : x \in \mathcal{S}\}$). In other words, in cases where $\phi'(r(x))$ exists, the optimality conditions can be summarized as follows: x^* is a stationary point of (3) if

$$\begin{aligned} x_i^* \neq 0 \;&\Rightarrow\; -\nabla f(x^*)_i = \phi'(r(x)) \, \gamma'(|x_i|) \\ x_i^* = 0 \;&\Rightarrow\; -\nabla f(x^*)_i \in \phi'(r(x)) \cdot [-\gamma_{\max}, \gamma_{\max}]. \end{aligned}$$

▶ **Example 5.** Suppose that $\gamma(x_i) = |x_i|$ and $\phi(\xi) = \frac{1}{2}\xi^2$. Since $h(x) = \phi(r(x))$ is convex in this example, the Clarke subdifferential reduces to the usual convex subdifferential, and can be expressed element-wise

$$(\partial h(x))_i = \|x\|_1 \cdot \begin{cases} [-1, 1], & x_i = 0, \\ \{1\}, & x_i > 0, \\ \{-1\}, & x_i < 0. \end{cases}$$

The optimality conditions can also be summarized in terms of "wiggle room"; that is, whenever $x_i = 0$, then $\nabla f(x)_i$ lies in an interval. But when $x_i \neq 0$, $\nabla f(x)_i$ must take a specific value. Duality will then allow the element-wise gradient to act as a sparsity indicator. (See also [53, 69].)

▶ **Example 6.** Consider the concave regularizer $h(x) := \left(\sum_i \sqrt{|x_i| + \xi_0}\right)^2$. This construction arises from $\phi(\xi) = \xi^2$ and $r(\xi) = \sqrt{|\xi| + \xi_0}$. Its Clarke-subdifferential can be expressed element-wise

$$(\partial h(x))_i = \left(\sum_j \sqrt{|x_j| + \xi_0}\right) \cdot \begin{cases} [-\xi_0^{-1/2}, \xi_0^{-1/2}], & x_i = 0, \\ \left\{\frac{1}{\sqrt{|x_i| + \xi_0}}\right\}, & x_i > 0, \\ \left\{\frac{-1}{\sqrt{|x_i| + \xi_0}}\right\}, & x_i < 0. \end{cases}$$

Again, note that the duality conditions show "wiggle room" in the values of $\nabla f(x)$ at stationary $x = x^*$, for the indices for which $x_i = 0$. However, in the case of nonconvex functions γ, the gradient at optimality is less informative, since $\gamma'(|x_i|)$ changes with different input values, and moreover is not necessarily maximal when $|x_i| > 0$. For this reason, designing screening rules is nontrivial for nonconvex penalty functions, and fully safe rules may not prove fully efficient.

[1] In general, $\phi(x)$ may not be differentiable for all x. However, since ϕ is convex and only defined on $\mathbb{R}_+$, then $\phi'(0) := \lim_{\xi \to 0^+} \frac{\phi(\xi) - \phi(0)}{\xi}$ must exist.

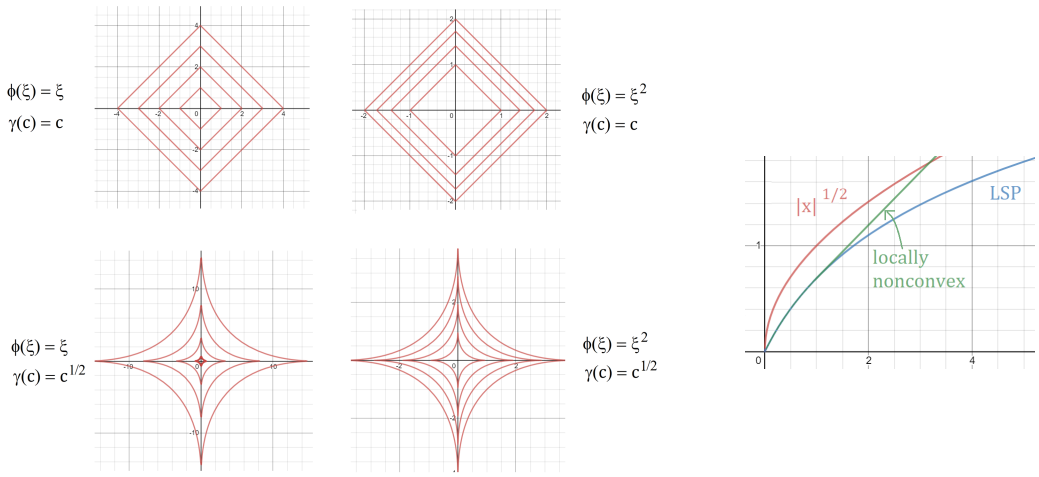

Figure 1 Transformations ϕ and γ. Left: Level sets for the penalty $h(x) = \phi(\sum_i \gamma(|x_i|))$. The concave penalty γ increases the "spike-ness"; the convex penalty ϕ increases the effect of the aggregate value. Right: Three example functions of γ. RP-CGM will behave erratically when $\gamma_{\min} = 0$ (red and blue) and $\gamma_{\max}$ is unbounded (red), so we use a penalty that is bounded on both ends (green = concave + linear).

2.4 Duality

We now give the primal and Fenchel dual formulations of (3) given a reference point $\bar{x}$:

$$\text{(P-simple)} \quad \min_{x\in\mathbb{R}^d} \overline{F}(x;\bar{x}) := f(x) + \underbrace{\phi\left(r_0 + \sum_i w_i|x_i|\right)}_{=:\bar{h}(x;\bar{x})}$$

$$\text{(D-simple)} \quad \max_z \overline{F}_D(z;\bar{x}) := -f^*(-z) \underbrace{-\phi^*\left(\max_i \frac{|z_i|}{w_i}\right) + r_0\left(\max_i \frac{|z_i|}{w_i}\right)}_{\bar{h}^*(z;\bar{x})}.$$

Here, we define $r_0 := r(\bar{x}) - \bar{r}(\bar{x};\bar{x})$ and $w_i = \gamma'(|\bar{x}_i|)$. Given $\bar{x}$, both primal and dual objective functions are convex. In particular, the duality gap of this convexified subproblem, using a primal candidate x and dual candidate $z = -\nabla f(x)$, can be expressed as

$$\mathbf{gap}(x;\bar{x}) = \underbrace{f(x) + f^*(\nabla f(x))}_{=x^T\nabla f(x)} + \bar{h}(x;\bar{x}) + \bar{h}^*(-\nabla f(x);\bar{x})$$

and adds little overhead when used to monitor the progress of Alg. 1. Now, we will show that $\mathbf{gap}(x;\bar{x})$ is an effective residual measurement, and indeed converges to 0 at the usual $O(1/t)$ rate.

2.5 Convergence of RP-CGM

We begin with an unusual twist on a usual assumption.

▷ Assumption 3 (L-smoothness). We assume that f is convex and L-smooth w.r.t. $\|\cdot\|_1$:

$$f(y) - f(x) \le \nabla f(x)^T(y-x) + \frac{L}{2}\|y-x\|_1^2, \quad \forall\, x, y. \tag{13}$$

An important consequence of (13) is that, while the set of minimizers of (P-simple) may not necessarily be unique, their gradient $\nabla f(\bar{x}^*)$ will be unique. Specifically, (13) implies that

$$f(x) - f(y) \ge \nabla f(y)^T(x-y) + \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|_\infty^2, \quad \forall\, x, y \tag{14}$$

and in particular taking $y = \bar{x}^*$ where $-\nabla f(\bar{x}^*) \in \partial \bar{h}(\bar{x}^*; \bar{x})$, we have

$$f(x) + \bar{h}(x; \bar{x}) - f(\bar{x}^*) - \bar{h}(\bar{x}^*; \bar{x}) \geq \frac{1}{2L}\|\nabla f(x) - \nabla f(\bar{x}^*)\|_\infty^2, \quad \forall\, x$$

and thus x is optimal only if $\nabla f(x) = \nabla f(\bar{x}^*)$.

Under Assumption 3, we first show that the duality gap of the original nonconvex problem (3) is (as expected) bounded away from 0, and is thus an inadequate measure of suboptimality.

▶ **Proposition 7** (Duality gap of nonconvex regularizer)**.** *For $h(x) = \phi(r(x))$,*

$$h(x) - h^{**}(x) \geq \phi(r(x)) - \phi(\gamma_{\min}\|x\|_1).$$

Proof. First, given the conjugate function

$$h^*(z) := \sup_x x^T z - \phi(r(x))$$

and picking

$$\begin{cases} x_i \to \mathbf{sign}(z_i) \cdot \alpha & \text{for one } |z_i| = \|z\|_\infty, \\ x_i = 0 & \text{otherwise,} \end{cases}$$

gives

$$h^*(z) \geq \alpha\|z\|_\infty - \phi(\gamma(\alpha)).$$

Since ϕ is monotonically increasing and γ is concave, we have the majorant property of the linearizer, and

$$\phi(\gamma(\alpha)) \leq \phi(\gamma'(\alpha_0) \cdot (\alpha - \alpha_0) + \gamma(\alpha_0)).$$

Therefore

$$h^*(z) \geq \sup_\alpha\ \alpha\|z\|_\infty - \phi(\gamma'(\alpha_0) \cdot (\alpha - \alpha_0) + \gamma(\alpha_0)) \geq \phi^*\left(\frac{\|z\|_\infty}{\gamma'(\alpha_0)}\right) + \underbrace{(\gamma'(\alpha_0)\alpha_0 - \gamma(\alpha_0))}_{\geq 0}\left(\frac{\|z\|_\infty}{\gamma'(\alpha_0)}\right).$$

In particular, since this holds for any α_0, $h^*(z) \geq \phi^*\left(\frac{\|z\|_\infty}{\gamma_{\min}}\right)$. Therefore,

$$h^{**}(x) \leq \sup_z\ x^T z - \phi^*\left(\frac{\|z\|_\infty}{\gamma_{\min}}\right) = \phi(\|x\|_1\gamma_{\min}). \qquad \blacktriangleleft$$

In other words, the duality gap of the original nonconvex problem is somewhat useless for screening, since it does not converge to 0. Instead, we measure convergence via the gap of the linearized problem at $\bar{x} = x$.

▶ **Proposition 8** (Residual)**.** *The duality gap $\mathbf{gap}(x, z; \bar{x})$ between (P-simple) and (D-simple) at primal variable x and dual variable $z = -\nabla f(x)$, with reference point $\bar{x}$, satisfies (at $\bar{x} = x$)*

$$\mathbf{gap}(x, z; x) \geq 0\ \ \forall\, x, \qquad \mathbf{gap}(x, z; x) = 0 \iff \text{x is a stationary point of (3).}$$

Proof. Since $\mathbf{gap}(x, z; x)$ is a duality gap, it is always nonnegative. Explicitly, denote $\nu = \max_i\left(\frac{|\nabla f(x)_i|}{\gamma'(|x_i|)}\right)$. Then, since $f(x) + f^*(\nabla f(x)) = \nabla f(x)^T x$,

$$\begin{aligned}
\mathbf{gap}(x, z; x) &= x^T\nabla f(x) + \phi(r(x)) + \phi^*(\nu) + (\bar{r}(x; x) - r(x)) \cdot \nu \\
&\overset{(a)}{\geq} x^T\nabla f(x) + r(x)\nu + (\bar{r}(x; x) - r(x)) \cdot \nu \\
&= x^T\nabla f(x) + \underbrace{\sum_i w_i|x_i|}_{\|w\odot x\|_1} \cdot \underbrace{\max_j\left(\frac{|\nabla f(x))_j|}{w_j}\right)}_{\|\nabla f(x)\oslash w\|_\infty} \\
&\overset{(b)}{\geq} x^T\nabla f(x) - x^T\nabla f(x) = 0
\end{aligned}$$

where $\odot$ and $\oslash$ represent element-wise multiplication and division, respectively. Tightness of (a) occurs iff Fenchel–Young is satisfied with equality, e.g. $\nu \in \partial\phi(r(x))$. Tightness of (b) occurs iff

$$\max_j \frac{|\nabla f(x)_j|}{\gamma'(|x_j|)} = \frac{-\nabla f(x)_i \cdot \mathbf{sign}(x_i)}{\gamma'(|x_i|)}, \quad \forall\, x_i \neq 0. \tag{15}$$

Combining these two observations, then $\mathbf{gap}(x, -\nabla f(x); x) = 0$ if and only if

$$-\nabla f(x)_i \in \begin{cases} \{-g_\phi\, \gamma'(|x_i|) : g_\phi \in \partial\phi(r(x))\}, & x_i < 0, \\ \{g_\phi\, \gamma'(|x_i|) : g_\phi \in \partial\phi(r(x))\}, & x_i > 0, \\ \phi'(0) \cdot [-\gamma'(0), \gamma'(0)], & x_i = 0, \end{cases}$$

which is the condition for $x = x^*$ a stationary point of (3). ◀

▶ **Theorem 9** (Convergence of RP-CGM, simple case)**.** *Pick any $x^{(0)} \in \mathbb{R}^d$ where $h(x^{(0)})$ is finite. Define the sequence $x^{(t)}$, $t = 1, \ldots$ by the steps dictated in (Min-Maj) and (Merge), using the step size sequence $\theta^{(t)} = 2/(1+t)$. Given Assumptions 1, 2, 3, then*

$$F(x^{(t)}) - F(x^*) = O(1/t), \qquad \min_{t' \leq t} \mathbf{res}(x^{(t')}) = O(1/t).$$

This is a special case of Theorem 33, which is proven in Section 4 and Appendix B. The proof is inductive, and shows that $O(1/t)$ behavior "kicks in" at a large enough t; explicit constants are given in Section 4.

2.6 Convex support recovery and screening

To understand how gap-based screening works, suppose first that for some x, we magically have a bound on the gradient error over all indices:

$$\epsilon > |(\nabla f(x))_i - (\nabla f(x^*))_i|, \quad \forall\, i.$$

Then the value of the true maximum gradient at the stationary point is at most ϵ away from the maximum value of the current gradient, e.g.

$$\|\nabla f(x^*)\|_\infty \geq \|\nabla f(x)\|_\infty - \epsilon.$$

Moreover, if at any index k,

$$|\nabla f(x)|_k < \|\nabla f(x)\|_\infty - 2\epsilon \leq \|\nabla f(x^*)\|_\infty - \epsilon,$$

this implies that the highest possible value that $|\nabla f(x^*)|_k$ could be is

$$|\nabla f(x^*)|_k \leq |\nabla f(x)|_k + \epsilon < \|\nabla f(x^*)\|_\infty;$$

in other words, index k cannot possibly be maximal. Therefore, it must be that at optimality, $x^*_k = 0$. The last missing detail is the observation that the duality gap gives us this ϵ bound explicitly.

Now we formalize this notion. From optimality conditions, $\bar{x}^*$ minimizes (P-simple) if

$$-\frac{\nabla f(\bar{x}^*)_i}{\alpha w_i} \in \begin{cases} \{\mathbf{sign}(\bar{x}^*_i)\}, & \bar{x}^*_i \neq 0 \\ [-1, 1], & \bar{x}^*_i = 0, \end{cases} \tag{16}$$

for some $\alpha \in \partial_\xi \phi(r_0 + \xi)$ at $\xi = \sum_i w_i \bar{x}^*_i$. In other words, for this convex reweighting problem, the sparsity pattern of $\bar{x}^*$ can be partially ascertained from $\nabla f(\bar{x}^*)$, in that the set of nonzeros of $\bar{x}^*$ must be contained in the set of maximal indices of the reweighted $\nabla f(\bar{x}^*)$. Formally, define

$$\mathbf{dsupp}(x; \bar{x}) := \left\{ i : \frac{|\nabla f(x)_i|}{\gamma'(|\bar{x}_i|)} = \max_j \frac{|\nabla f(x)_j|}{\gamma'(|\bar{x}_j|)} \right\}. \tag{17}$$

Then the optimality condition (16) states that $\mathbf{supp}(\bar{x}^*) \subseteq \mathbf{dsupp}(\bar{x}^*; \bar{x})$, where $\bar{x}^*$ minimizes (P-simple). We are in particular interested in $\bar{x}^* = x^*$ the stationary point of (3). From this observation, we have our first screening property.

▶ **Proposition 10** (Screening for simple sparsity). *If* $\|\nabla f(x) - \nabla f(x^*)\|_\infty \leq \epsilon$, *then*

$$\|\nabla f(x)\|_\infty - |\nabla f(x)_i| > 2\epsilon\gamma_{\max} \;\Rightarrow\; x_i^* = 0. \tag{18}$$

Proof. First, define $v_i = \frac{|\nabla f(x)_i|}{w_i}$ and $\bar{v}_i^* = \frac{|\nabla f(x^*)_i|}{w_i}$. Then

$$\frac{\|\nabla f(x^*) - \nabla f(x)\|_\infty}{\gamma_{\max}} \leq \frac{\|\nabla f(x^*) - \nabla f(x)\|_\infty}{\max_i w_i} \leq \|v - \bar{v}^*\|_\infty.$$

By optimality conditions $\bar{v}_i^* < \|\bar{v}^*\|_\infty \Rightarrow \bar{x}_i^* = 0$. Thus, (18) implies

$$2\epsilon > \frac{\|\nabla f(x)\|_\infty - |\nabla f(x)_i|)}{\gamma_{\max}} \geq \|\bar{v}\|_\infty - \bar{v}_i$$

and therefore $\|\bar{v}^*\|_\infty - \bar{v}_i^* \leq \|\bar{v}\|_\infty - \bar{v}_i + 2\epsilon < 0$. ◀

▶ **Proposition 11** (Residual bound on gradient error). *Define* $D(x) = \sum_{i=1}^d \gamma(|x_i|) - \gamma(|x_i^*|) - \gamma'(|x_i|)(|x_i - x_i^*|)$ *the linearization error at* x. *Denoting* x^* *a stationary point of* (3), *then*

$$\|\nabla f(x^*) - \nabla f(x)\|_\infty \leq \frac{LD(x)}{2\gamma_{\min}} + \sqrt{\frac{L^2 D(x)^2}{4\gamma_{\min}^2} + L\mathbf{res}(x) + LD(x)\frac{\|\nabla f(x)\|_\infty}{\gamma_{\min}}}.$$

Note that if $r(x) = \|x\|_1$ then $D(x) = 0$. Since this proposition is a consequence of Proposition 34 in Section 4, we leave the proof for then.

From these two properties, we immediately get a screening rule for (3):

▶ **Theorem 12** (Screening rule). *For any* x, *define*

$$\mathcal{I}_\epsilon^{(t)} = \left\{ i : \|\nabla f(x)\|_\infty - |\nabla f(x)_i| \geq \epsilon + 2\sqrt{L\mathbf{gap}(x;x) + \epsilon} \right\}. \tag{19}$$

If

$$\epsilon \geq \frac{LD(x)}{\gamma_{\min}} \max\left\{ \frac{1}{2}, \frac{LD(x)}{4\gamma_{\min}} + \|\nabla f(x)\|_\infty \right\},$$

then $x_i^* = 0$ *for all* $i \in \mathcal{I}_\epsilon^{(t)}$, *where* x^* *any minimizer of* (3).

Note that in the convex case ($\gamma(|x_i|) = |x_i|$) then $D(x) = 0$ and $\epsilon = 0$ is a safe choice, for all x. In the general case, since we do not know $D(x)$, we cannot guarantee the safety of an intermediate iterate; however, since $D(x^*) = 0$ by definition of stationary point, then $x^{(t)} \to x^*$ implies $D(x^{(t)}) \to 0$. Picking any decaying sequence $\epsilon^{(t)} \to 0$, therefore, forms a heuristic rule that converges to the true support.

2.6.1 Degeneracy and support recovery guarantee

Following the terminology introduced in [37], we say that x^* is a *degenerate solution* if $\mathbf{supp}(x^*) \neq \mathbf{dsupp}(x^*; x^*)$; that is, there exists i where

$$x_i^* = 0 \quad \text{and} \quad \frac{|\nabla f(x^*)_i|}{\gamma'(|x_i^*|)} = \max_j \frac{|\nabla f(x^*)_j|}{\gamma'(|x_j^*|)}.$$

To characterize nearly degenerate solutions, we define

$$\delta_i(x) = \max_j \frac{|\nabla f(x)_j|}{\gamma'(|x_j|)} - \frac{|\nabla f(x)_i|}{\gamma'(|x_i|)}, \quad \delta_{\min}(x) = \min_{i:x_i=0} \delta_i(x),$$

and the quantity $\delta_{\min}(x^*)$ expresses the distance to degeneracy for this solution. This can be interpreted as a complementary slackness-like condition in duality, where both the primal and dual variables are jointly active. While we may reasonably believe that many real world problems with randomized data do not lead to degenerate solutions, near-degenerate solutions do pose problems for screening and manifold identification [11, 37, 45].

▶ **Corollary 13.** *If* $\delta_{\min} > 0$, *then for a method* $x^{(t)} \to x^*$, *the screening rule* (19) *with* $\epsilon = 0$ *identifies* $\mathbf{supp}(x^*)$ *after a finite number of iterations* $\bar{t}$; *that is, for all* $t \geq \bar{t}$, $\mathcal{I}_0^{(t)} = \mathbf{supp}(x^*)$. *In the convex case* ($\gamma(|x_i|) = |x_i|$), *this occurs when* $\|\nabla f(x^*) - \nabla f(x)\|_\infty \leq \delta_{\min}/3$, *which occurs at* $\bar{t} = O(1/\delta_{\min}^2)$.

3 P-CGM for general convex sparse optimization

Our goal is to now extend the studies of the previous section to solve the generalized sparse optimization problem (1). The key addition is the introduction of the "gauge-like" function $r_{\mathcal{P}}(x)$, but which uses the sparsifying properties of γ. In this section, we will focus on problem (32) when it is convex; namely, we assume that $\gamma(c) = c$. Just as studying the convex LASSO brings to light many of the sparse recovery properties illustrated from the nonconvex problem (3), we will first study the convex penalized version of (32) to gain intuition, and present the full extension in the next section.

3.1 Gauge penalized problems

The penalized CGM (P-CGM) solves problems of the form

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ f(x) + \phi(\kappa_{\mathcal{P}}(x)), \tag{20}$$

where $\kappa_{\mathcal{P}}(x)$ is the *gauge function* [15, 30] defined by a set $\mathcal{P}$ at point x:

$$\kappa_{\mathcal{P}}(x) := \min_{c_p\geq 0}\left\{\sum_{p\in\mathcal{P}_0} c_p : \sum_{p\in\mathcal{P}_0} c_p p = x\right\}. \tag{21}$$

This function generalizes the 1-norm to more size-measuring functions that include norms, semi-norms, and convex cone restrictions. It is useful to compare (21) with the definition usually given in convex analysis literature [8, 60], where the gauge function over a closed convex set $\mathcal{P}$ is defined as

$$\kappa_{\mathcal{P}}(x) := \inf\{\mu \geq 0 : x \in \mu\mathcal{P}\}. \tag{22}$$

In fact, this is equivalent to (21). In particular, when $\mathcal{P}$ is the convex hull of a set of atoms, $\kappa_{\mathcal{P}}(x)$ can be used to promote sparsity with respect to those atoms. The corresponding "dual gauge" is the support function

$$\sigma_{\mathcal{P}}(z) = \max_{s\in\mathcal{P}_0} s^T z$$

which is closely related to the generalized LMO

$$\mathbf{LMO}_{\mathcal{P}}(z) = \underset{s\in\mathcal{P}_0}{\text{argmax}}\ s^T z.$$

If $\kappa_{\mathcal{P}}$ is a norm, then $\sigma_{\mathcal{P}}$ is the usual dual norm [8, 60]. A key feature of the CGM is that this LMO is often cheap to compute in practice, and despite weaker convergence guarantees compared to higher order methods, often converges quickly when x^* is sparse with respect to structured $\mathcal{P}_0$. (See also Table 2.)

▶ **Example 14** (ℓ_1 norm)**.** We start with the usual sparsity case of the ℓ_1 norm. In this case, $\sigma_{\mathcal{P}} = \|\cdot\|_\infty$ is the dual norm of $\|\cdot\|_1$. Then, by setting the optimality condition $0 \in \partial g(x^*)$ and decomposing by index, at optimality

$$\begin{cases} (-\nabla f(x^*))_i = \|x^*\|_1\, \mathbf{sign}(x_i^*) & \text{if } x_i^* \neq 0, \\ (-\nabla f(x^*))_i \in \|x^*\|_1 \cdot [-1, 1] & \text{if } x_i^* = 0. \end{cases}$$

In words, the gradient of f along a coordinate for which the optimal variable is nonsmooth with respect to $\kappa_{\mathcal{P}}$ is allowed "wiggle room"; in contrast, if $g(x)$ is smooth in the direction of x_i then the gradient is fixed. In terms of support recovery, $\max_i |\nabla f(x^*)_i| = \|x^*\|_1$ and additionally, if $|\nabla f(x^*)_i| < \|x^*\|_1$ then it must be that $x_i^* = 0$.

▶ **Example 15** (Weighted ℓ_1 norm)**.** The convex majorant in Section 2 specifically considered $\kappa_{\mathcal{P}}(x) = \sum_i w_i|x_i|$, for weights $w_i > 0$. Here, $\mathcal{P}_0 = \{\pm w_1^{-1}e_1, \ldots, \pm w_d^{-1}e_d\}$, with corresponding "dual gauge" $\sigma_{\mathcal{P}}(z) = \max_i \frac{|z_i|}{|w_i|}$, and the LMO follows exactly the steps for the bounded maximization computation in (5). Note also that the optimality conditions of (20) for this choice of $\kappa_{\mathcal{P}}(x)$ is

$$\begin{cases} \frac{|\nabla f(x^*)_i|}{|w_i|} = \max_j \left(\frac{|\nabla f(x^*)_j|}{|w_j|}\right) \mathbf{sign}(x_i^*) & \text{if } x_i^* \neq 0, \\ \frac{|\nabla f(x^*)_i|}{|w_i|} \in \max_j \left(\frac{|\nabla f(x^*)_j|}{|w_j|}\right) \cdot [-1, 1] & \text{if } x_i^* = 0. \end{cases}$$

It exactly characterizes the optimality conditions for (P-simple). Later, we will generalize this reweighting technique for general atomic sets $\mathcal{P}_0$, to construct the convex majorant of the general nonconvex problem (1).

▶ **Example 16** (Latent group norm)**.** For the task of selecting a sparse collection of overlapping subvectors, such as in gene identification, the latent group norm was proposed in [55]. For $x \in \mathbb{R}^d$, given a collection of overlapping groups $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_K\}$ where $\mathcal{G}_k \subset \{1, \dots, d\}$, this norm a gauge function,

$$\kappa_{\mathcal{P}}(x) = \|x\|_{\mathcal{G}} := \min_{s_k \in \mathbb{R}^d} \left\{ \sum_{k=1}^K \|s_k\|_2 : x = \sum_{k=1}^K s_k,\ (s_k)_i = 0\ \ \forall\, i \notin \mathcal{G}_k \right\}. \tag{23}$$

In particular, (23) is the solution to (21) when

$$\mathcal{P}_0 = \left\{ \frac{1}{\sqrt{|\mathcal{G}_k|}} e_{\mathcal{G}_k},\ k = 1, \dots, K \right\}, \qquad (e_{\mathcal{G}_k})_i = \begin{cases} 1, & i \in \mathcal{G}_k, \\ 0, & \text{else.} \end{cases}$$

Then $\sigma_{\mathcal{P}}(z) = \max_{k=1,\dots,K}\ \|z_{\mathcal{G}_k}\|_2$. Now consider (20) for some smooth ϕ. Then at optimality, decomposing $x^* = \sum_k s_k^*$, for each group k,

$$\begin{cases} \|z_{\mathcal{G}_k}^*\|_2 = \phi'(\kappa_{\mathcal{P}}(x^*)), & \text{if } \|s_k^*\|_2 > 0, \\ \|z_{\mathcal{G}_k}^*\|_2 \le \phi'(\kappa_{\mathcal{P}}(x^*)), & \text{if } \|s_k^*\|_2 = 0. \end{cases}$$

Screening in this case refers to identifying the subvectors where, at optimality, $\|s_k^*\|_2$ might be nonzero; however, just as support identification in the 1-norm case does not imply that the values of x_i^* are known, in a similar vein here it does not imply that the values of s_k^* are known.

Both P-CGM and RP-CGM can be efficiently implemented for the latent group norm. However, a key numerical issue is that computing the group norm $\|x\|_{\mathcal{G}}$ when the groups overlap is computationally burdensome (requires solving complex subproblems) and is needed in the gap computation. Nevertheless, since gap computations are used only infrequently for monitoring progress and for screening, this overhead can be mitigated. (Note that computing the dual norm, and thus the LMO, is comparatively computationally cheap / trivial.)

▶ **Example 17** (Nuclear norm)**.** For a matrix $X \in \mathbb{R}^{m \times n}$, the nuclear norm $\|X\|_*$, defined as the sum of singular values of matrix X, can be expressed as a gauge over the infinite set

$$\mathcal{P}_0 = \{uv^T : u \in \mathbb{R}^m,\ v \in \mathbb{R}^n\}.$$

Because $\mathcal{P}_0$ is not a finite set, screening in this scenario will most likely not be very efficient, or even useful. However, CGM is indeed frequently applied to this version of $\mathcal{P}_0$, in order to promote low-rank matrix solutions, and applying P-CGM to spectral problems is a central application in [69]. In particular, while computing the nuclear norm requires a full spectral calculation, computing the dual norm, the spectral norm, is often much cheaper using fast spectral methods, and can often be compressible [70].

Table 2 summarizes these examples and key properties. Gauges and support functions for convex sets are fundamental objects in convex analysis, and are discussed more by [8, 30, 32, 60].

▶ **Example 18** (Total variation (TV) "norm")**.** We now investigate a case where $\mathcal{P}_0$ contains a direction of recession, which introduces some ambiguity into our construct. Specifically, we investigate the TV norm, which is often used in signal processing as a "smoothing regularizer":

$$\|x\|_{\mathrm{TV}} = \sum_{i=2}^n |x_i - x_{i-1}|.$$

A common way to express this in matrix/vector notation is to introduce a difference matrix

$$D = \begin{bmatrix} I & 0 \end{bmatrix} - \begin{bmatrix} 0 & I \end{bmatrix} \in \mathbb{R}^{d-1,d},$$

and $\|x\|_{\mathrm{TV}} = \|Dx\|_1$. This norm can be viewed as a gauge over atoms $\mathcal{P}_0 = \{b_1, \dots, b_{d-1}\} \cup \{c\mathbf{1} : c \in \mathbb{R}\}$ where for $\mathbf{1}$ the all-ones vector,

$$b_k = \beta_k - \frac{1}{n}\beta_k^T \mathbf{1}, \qquad \beta_k = (\underbrace{1, 1, \dots 1}_{k}, \underbrace{0, 0, ..0}_{n-k}) \in \mathbb{R}^n.$$

Table 2 Common norms, their atoms, support functions, and their LMOs. In particular, computing each LMO is computationally cheap, especially compared to computing the proximal operator of the gauge, or even the gauge itself.

Gauge $\kappa_{\mathcal{P}}(x)$	Atoms $\mathcal{P}_0$	Support fn $\sigma_{\mathcal{P}}(z)$	LMO(z)
1-norm $\|x\|_1$	$\{\pm e_1, \dots, \pm e_d\}$	$\|z\|_\infty$	$\mathbf{sign}(z_k)e_k$, $k = \operatorname{argmax}_k \ \|z_k\|$
Mapped 1-norm $\|P^{-1}x\|_1$	$\{\pm p_1, \dots, \pm p_d\}$	$\|Pz\|_\infty$	$\mathbf{sign}(p_k^T z)p_k$, $k = \operatorname{argmax}_i \ \|p_i^T z\|$
Group norm $\sum_{i=1}^K \|x_{\mathcal{G}_i}\|_2$,	$\left\{\frac{1}{\sqrt{\|\mathcal{G}_1\|}}e_{\mathcal{G}_1}, \dots, \frac{1}{\sqrt{\|\mathcal{G}_K\|}}e_{\mathcal{G}_K}\right\}$	$\max_k \|z_{\mathcal{G}_k}\|_2$	$\frac{1}{\sqrt{\|\mathcal{G}_k\|}}e_{\mathcal{G}_k}$, $k = \operatorname{argmax}_k \ \|z_{\mathcal{G}_k}\|_2$
TV norm $\|Dx\|_1$	$\{b_k\}_{k=1}^d \cup \{c\mathbf{1} : c \geq 0\}$ $\beta_k = (\mathbf{1}_k, \mathbf{0}_{n-k})$ $b_k = \beta_k - \frac{1}{n}\beta_k^T\mathbf{1}$	$\|D^\dagger z\|_\infty$ if $z \in \mathbf{range}(D^T)$, $+\infty$ else.	b_k, $k = \operatorname{argmax}_i \ \|(D^\dagger z)_i\|$

In particular, for any constant vector x, $\|x\|_{\mathrm{TV}} = 0$. This adds an unbounded direction for the support function; specifically

$$\sigma_{\mathcal{P}}(z) = \begin{cases} \|u\|_\infty & \text{if } z = D^T u, \\ +\infty & \text{else,} \end{cases}$$

and thus the LMO is not always defined. Note here that if $z \in \mathbf{range}(D^T)$, then $u = D(D^T D)^{-1}z$ is uniquely determined; this inspires an "effective band-aid" to deal with directions of recession.

3.1.1 Gauges with directions of recessions

The *recession cone of* $\mathcal{P}$ [8, 60] is defined as

$$\mathbf{rec}(\mathcal{P}) = \{r : cr \in \mathcal{P} \ \ \forall\, c \geq 0\}.$$

Whenever $\mathcal{P}$ has a direction of recession, CGM struggles as the LMO can return an infinite atom. We offer to isolate optimization over this set separately. In particular, suppose

$$\mathcal{P}_0 = \mathcal{P}_0' \cup \mathcal{K}, \qquad \mathcal{P}_0' \cap \mathcal{K} = \emptyset,$$

where $\mathcal{P}_0'$ is a finite set, and thus defining $\mathcal{P}$ as the convex hull of $\mathcal{P}_0'$ ensures that $\mathcal{P}$ is compact. Then we rewrite (20) as

$$\underset{x \in \mathbf{cone}(\mathcal{P}), y \in \mathcal{K}}{\text{minimize}} \ f(x+y) + \phi(\kappa_{\mathcal{P}}(x)) \tag{24}$$

where $\mathbf{cone}(\mathcal{P}) := \{\alpha x : \alpha \in \mathbb{R}_+, x \in \mathcal{P}\}$ is the conic hull of $\mathcal{P}$. At each iteration, x takes a conditional gradient step, and y is updated through a full minimization. (In the case of the TV norm, this simply means that the LMO is applied to a de-meaned $\widehat{x} = x - \frac{1}{d}x^T\mathbf{1}$.) Since the portion of the solution in $\mathcal{K}$ is minimized exactly at each step, from this point on we only consider the support recovery properties for recovering the atoms in $\mathcal{P}_0'$.

▷ Assumption 4 (Atomic set conditions). $\mathcal{P}_0 = \mathcal{P}_0' \cup \mathcal{K}$ where $\mathcal{P}_0'$ is a finite set of atoms and $\mathcal{K}$ is the recession cone; moreover, $\mathcal{P}_0' \cap \mathcal{K} = \emptyset$. We denote $\mathcal{P} = \mathbf{conv}(\mathcal{P}_0')$.

3.2 Generalized smoothness

To ensure the uniqueness of $\mathbf{dsupp}_{\mathcal{P}}(-\nabla f(x^*))$ and to give a useful gap bound, we again need a notion of smoothness on f. We again use our unusual twist on the gauge penalty.

▶ **Definition 19.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ is L-smooth with respect to $\mathcal{P}$ if for all $x, y \in \mathbb{R}^d$:*

$$f(x) - f(y) \leq \nabla f(y)^T (x - y) + \frac{L}{2} \kappa_{\mathcal{P}}(x - y)^2. \tag{25}$$

The purpose of this generalized notion is that sometimes, given the data, tighter bounds can be computed [54]. It is similar in spirit to the notion of *relative smoothness* [3, 47] which facilitates the analysis of generalized proximal gradient methods, where the 2-norm squared proximity measure is replaced by a Bregman divergence. For CGM, it is more computationally efficient to consider generalized gauges as the penalty generalization, which we incorporate to the generalized smoothness definition. Additionally, the subadditivity property of gauges assists with bounding the iterates, a crucial step in the convergence proof.

▷ Assumption 5 (Generalized smoothness). The convex function f is L-smooth w.r.t. $\widetilde{\mathcal{P}} := \mathcal{P} \cup (-\mathcal{P})$.

▶ **Example 20** (Quadratic function)**.** Suppose that $f(x) = \frac{1}{2}\|Ax\|_2^2 + b^T x$. Then

$$L = \begin{cases} L_1 := (\max_i \|A_{:,i}\|_2)^2, & \kappa_{\mathcal{P}} = \|\cdot\|_1, \\ L_2 := \|A\|_2^2, & \kappa_{\mathcal{P}} = \|\cdot\|_2, \\ L_\infty := (\sum_i \|A_{:,i}\|_2)^2, & \kappa_{\mathcal{P}} = \|\cdot\|_\infty. \end{cases}$$

While norm bounds would give $d^2 L_1 \geq d L_2 \geq L_\infty$, the actual values in A might lead to tighter inequalities.

The relationship to usual smoothness is as follows. Suppose that f is L_2-smooth in the usual sense (with respect to $\|\cdot\|_2$). Then since $\mathbf{diam}(\mathcal{P})\kappa_{\mathcal{P}}(x) \geq \|x\|_2$, it follows that $L \leq \mathbf{diam}(\mathcal{P}) L_2$. In this way, we refine the analysis of CGM by absorbing the usual "set size" term into L, which in certain cases may be smaller than $\mathbf{diam}(\mathcal{P})L_2$.

▶ **Proposition 21** (Uniqueness of gradient)**.** *If (25) holds and $0 \in \mathbf{int}\ \mathcal{P}$, then $\nabla f(x^*)$ is unique at the optimum.*

The same logical argument as before applies, as "smoothness" in the primal corresponds to "strong convexity" (w.r.t. $\|\cdot\|_\infty$) in the dual.

3.3 Generalized support recovery

Given a solution to (21), define the *decomposition of x with respect to $\mathcal{P}_0$* as tuples c_p, p, extracted via the mapping $\mathbf{coeff}_{\mathcal{P}}(x, p) = c_p$. The *support of x with respect to $\mathcal{P}_0$* is

$$\mathbf{supp}_{\mathcal{P}}(x) = \{p : \mathbf{coeff}_{\mathcal{P}}(x, p) > 0 \text{ in (21)}\}. \tag{26}$$

For general $\mathcal{P}$, neither the decomposition nor the support of x is unique. As before, we say the support recovery is achieved if one such support $\mathbf{supp}_{\mathcal{P}}(x^*)$ of the limiting point $x^{(t)} \to x^* \in \mathcal{X}^*$ is revealed. The reduction to the support definition in the previous section occurs when $\mathcal{P}_0 = \{\pm e_1, \ldots, \pm e_d\}$ the signed standard basis. Then $\mathbf{supp}_{\mathcal{P}}(x)$ is unique, and explicitly $\mathbf{supp}_{\mathcal{P}}(x) = \{\mathbf{sign}(x_i)\, e_i : x_i \neq 0\}$.

▶ **Proposition 22** (Support optimality condition)**.** *Consider the general convex sparse optimization problem (20) where $\phi : \mathbb{R}_+ \to \mathbb{R}_+$ is a monotonically nondecreasing function. Then for any x^* a minimizer of (20),*

$$-\nabla f(x^*)^T x^* = \kappa_{\mathcal{P}}(x^*) \sigma_{\mathcal{P}}(-\nabla f(x^*)). \tag{27}$$

and

$$p \in \mathbf{supp}(x^*) \Rightarrow -\nabla f(x^*)^T p = \sigma_{\mathcal{P}}(-\nabla f(x^*)). \tag{28}$$

This is the gauge equivalent of "nonzero primal gives maximal dual", and is referred to in [27] as *alignment*. We now generalize the definition of dual support from (17):

$$\mathbf{dsupp}_{\mathcal{P}}(x) := \{p : -\nabla f(x)^T p = \sigma_{\mathcal{P}}(-\nabla f(x))\},$$

and Property 22 says that for any x, $\mathbf{supp}_{\mathcal{P}}(x) \subseteq \mathbf{dsupp}_{\mathcal{P}}(x)$. Finally, as in the previous section, we express this distance as $\delta_{\min}(x^*)$, where

$$\delta_p(x) = \sigma(-\nabla f(x)) + p^T \nabla f(x), \qquad \delta_{\min}(x) = \min_{p \in \mathcal{P}} \{\delta_p(x) : p \notin \mathbf{supp}_{\mathcal{P}}(x)\}$$

for any support of x. In particular, $\delta_{\min}(x^*) = 0$ means the problem is degenerate.

3.4 Duality and gap

For ϕ monotonically nondecreasing, the convex function $h(x) = \phi(\kappa_{\mathcal{P}}(x))$ has conjugate $h^*(z) = \phi^*(\sigma_{\mathcal{P}}(z))$. This gives the primal-dual pair

$$\begin{array}{llll} \text{(P-convex)} & \min_{x,y,w} & f(w) + \phi^*(\kappa_{\mathcal{P}}(x)) & \\ & \text{st} & w = x + y,\ y \in \mathcal{K} & \end{array} \qquad \begin{array}{llll} \text{(D-convex)} & \max_z & -f^*(-z) - \phi^*(\sigma_{\mathcal{P}}(z)) & \\ & \text{st} & z \in \mathcal{K}^\circ & \end{array}$$

where $\mathcal{K}^\circ$ is the polar cone of $\mathcal{K}$. The duality gap between (P-convex) and (D-convex) can be written as

$$\mathbf{gap}(x, y, z) = f(x + y) + h(x) + f^*(-z) - h^*(z) + \iota_{\mathcal{K}^\circ}(z)$$

where $\iota_{\mathcal{K}^\circ}(z) = +\infty$ if z is not dual-feasible, and 0 otherwise.

▶ **Lemma 23** (Feasible gradient). *Take $z := -\nabla_x f(x + y)$. Then $z = -\nabla_y f(x + y)$. Additionally, if $y = \operatorname{argmin}_{y' \in \mathcal{K}} f(x + y')$ then $z \in \mathcal{K}^\circ$.*

Proof. The first part is true from chain rule. Then by optimality condition, z is in the normal cone

$$z^T(y - y') \le 0, \quad \forall\, y' \in \mathcal{K}.$$

Since $0 \in \mathcal{K}$, this implies $z^T y \le 0$, which means $z \in \mathcal{K}^\circ$. ◀

From Lemma 23, the LMO step acquires s where for $z := -\nabla_x f(x + y)$,

$$-z^T s + h(s) = \min_{s'} -z^T s' + h(s) = -h^*(z).$$

Additionally, by Fenchel–Young's inequality, we know that $f(x) + f^*(\nabla f(x)) = \nabla f(x)^T x$, and thus we can simplify the gap to an online-computable quantity

$$\mathbf{gap}(x, y, \nabla_x f(x + y)) = -\nabla_x f(x + y)^T (s - x) + h(x) - h(s).$$

▶ **Proposition 24** (Gap bounds gradient error). *Given a primal feasible x and denote the optimum variable as*

$$x^* = \operatorname*{argmin}_{x'} \min_{y \in \mathcal{K}} f(x + y) + h(x).$$

Furthermore, denote $y = \operatorname{argmin}_{y' \in \mathcal{K}} f(x + y')$ and $y^* = \operatorname{argmin}_{y' \in \mathcal{K}} f(x^* + y')$. Then the duality gap bounds the gradient error

$$\mathbf{gap}_{\mathcal{P}}(x, y, -\nabla f(x + y)) \ge \frac{1}{2L} \sigma_{\widetilde{\mathcal{P}}}(\nabla f(x + y) - \nabla f(\bar{x}^* + \bar{y}^*))^2. \tag{29}$$

Proof. Since the conjugate of $h(x) = \phi(\kappa_{\mathcal{P}}(x))$ is $h^*(z) = \phi^*(\sigma_{\mathcal{P}}(z))$, then

$$\phi^*(\sigma_{\mathcal{P}}(z)) = \sup_x x^T z - \phi(\kappa_{\mathcal{P}}(x)) \ge (x^*)^T z - \phi(\kappa_{\mathcal{P}}(x^*)). \tag{30}$$

Then denoting $z = -\nabla f(x + y)$,

$$\begin{aligned} \mathbf{gap}_{\mathcal{P}}(x, y, z) &= \underbrace{f(x + y) + f^*(-z)}_{\text{Fenchel Young}} + \phi(\kappa_{\mathcal{P}}(x)) + \underbrace{\phi^*(\sigma_{\mathcal{P}}(-\nabla f(x)))}_{(30)}, \\ &\ge (x^* - x)^T z + \underbrace{y^T \nabla f(x)}_{y \in \mathcal{K}, \nabla f(x) \in \mathcal{K}^\circ} + \underbrace{\phi(\kappa_{\mathcal{P}}(x)) - \phi(\kappa_{\mathcal{P}}(x^*))}_{\text{convexity of } h}, \\ &\ge (x - x^*)^T z + \phi(\kappa_{\mathcal{P}}(x)) - \phi(\kappa_{\mathcal{P}}(x^*)) \end{aligned}$$

since $\mathcal{K}$ and $\mathcal{K}^\circ$ are polar cones and thus $y^T\nabla f(x) \leq 0$. Next, recognizing that $h(x) = \phi(\kappa_\mathcal{P}(x))$ is convex, we pick $-\nabla f(x^* + y^*) \in \partial h(x^*)$ and use convexity to further reduce to the result:

$$\mathbf{gap}_\mathcal{P}(x, y, z) \geq (x - x^*)^T(z^* - z) \geq \frac{1}{2L}\sigma_{\widetilde{\mathcal{P}}}(z^* - z)^2. \quad \blacktriangleleft$$

▶ **Theorem 25** (Support identification of screened P-CGM)**.** *Given Assumptions 1, 2, 4, 5, then the screening rule for convex penalties*

$$\mathcal{I}^{(0)} = \mathcal{P}_0, \qquad \mathcal{I}^{(t)} = \mathcal{I}^{(t-1)} \setminus \{p \in \mathcal{P}_0 : p \in \mathcal{I}_0(x) \text{ for } x = x^{(t)}\},$$

is safe and convergent:

$$\mathcal{I}^{(t)} \supseteq \mathbf{supp}_\mathcal{P}(x^*),\ \forall\, t, \quad \text{and} \quad \mathcal{I}^{(t)} = \mathbf{supp}_{\mathcal{P}(x^*)}(x^*),\ t \geq t',$$

where t' is such that

$$\sqrt{L \min_{i \leq t'} \mathbf{gap}(x^{(i)}, -\nabla f(x^{(i)}); x^{(i)})} < \delta_{\min}/3, \tag{31}$$

which happens at a rate $t' = O(1/(\delta_{\min}^2))$.

Proof. This is a direct consequence to Theorems 33 and 35. ◀

Note that Theorem 25 imposes no conditions on the sequence $\theta^{(k)}$, or choice of ϕ, f, etc., except L-smoothness of f. In other words, for any method where the gap is easily computable and its convergence rate known, then a corresponding screening rule and support identification rate automatically follow. Additionally, computing L may be challenging, depending on $\kappa_\mathcal{P}$; as shown previously, at the very least it may require a full pass over the data. However, this is a one-time calculation per dataset, and can be estimated if data are assumed to be drawn from specific distributions (as in sensing applications).

3.5 Invariance

One appealing feature of the CGM is that the iteration scheme and analysis can be done in a way that is invariant to both linear scaling and translation. However when the gauge function is not used as an indicator, this translation invariance vanishes; in general, $\kappa_\mathcal{P}(x) \neq \kappa_{\mathcal{P}+\{b\}}(x + b)$. Therefore the generalized problem formulation (32) is only linear (not translation) invariant.

▶ **Example 26.** Consider $\kappa_\mathcal{P}(x) = \|x\|_1$ for $x \in \mathbb{R}^2$. Take specifically $x = (-1, -1)$ and $b = (1, 1)$. Then $\kappa_\mathcal{P}(x) = 2$, but $\kappa_{\mathcal{P}+\{b\}}(x + b) = \kappa_{\mathcal{P}+\{b\}}(0) = 0 \neq 2$.

▶ **Proposition 27** (Invariance properties)**.** *Define $\mathcal{Q} = A\mathcal{P}$, and $f(x) = g(Ax)$. Define $w = Ax$ where A has full column rank. Then, using (22) and chain rule, the following hold*

- *$f(x) = g(w)$ and $\nabla f(x) = A^T\nabla g(w)$,*
- *$\kappa_\mathcal{P}(x) = \kappa_\mathcal{Q}(w)$ and $\sigma_\mathcal{P}(-\nabla f(x)) = \sigma_\mathcal{Q}(-\nabla g(w))$,*
- *$\mathbf{LMO}_\mathcal{Q}(-\nabla g(w)) = A\,\mathbf{LMO}_\mathcal{P}(-\nabla f(x))$,*
- *$f(x) = g(Ax + b)$ is L-smooth w.r.t. $\mathcal{P}$ iff g is L-smooth w.r.t. $\mathcal{Q}$.*

4 RP-CGM for general nonconvex sparse optimization

Finally, we consider the complete RP-CGM, which expands the method presented in Section 2 to generalized gauge penalties. The fully generalized optimization problem is

$$\min_x\ f(x) + \underbrace{\phi(r_\mathcal{P}(x))}_{h(x)}, \qquad r_\mathcal{P}(x) = \left\{ \min_{c_p \geq 0} \sum_{p \in \mathcal{P}_0} \gamma(c_p) : \sum_{p \in \mathcal{P}_0} c_p p = x \right\}. \tag{32}$$

By imposing the concave transformation on c_p, we effectively gain the same effect as the nonconvex regularizer on the ℓ_1 norm in Section 2. For the most part, much of the analysis will seem very similar to that in Section 2, especially in the proofs of key concepts, which we therefore put in the appendix to avoid repetitiveness. We also use much of the same assumptions (1, 2, 3) and analyses for the scalar functions γ and ϕ.

▶ **Lemma 28** (Smoothness equivalences). *Suppose that f is L-smooth with respect to $\mathcal{P}$. Then the following also holds:*

1. *Expansiveness*

$$(\nabla f(x) - \nabla f(y))^T(x-y) \geq \frac{1}{2L}(\sigma_{\mathcal{P}}(\nabla f(x) - \nabla f(y))^2 + \sigma_{\mathcal{P}}(\nabla f(y) - \nabla f(x))^2), \tag{33}$$

2. *Strongly convex conjugate*

$$f(y) - f(x) \geq \nabla f(x)^T(y-x) + \frac{1}{2L}\sigma_{\mathcal{P}}(\nabla f(y) - \nabla f(x))^2. \tag{34}$$

The proof is in Appendix A.

▶ **Lemma 29** (Uniqueness of gradient). *Suppose Assumption 4 holds. If* (25) *holds, then at the global optimum x^*, $-\nabla f(x^*) = z^* + w^*$ where $z^* \in \mathcal{K}^\circ$ is unique and $z^T w^* \in \mathcal{K}$.*

Proof. Assume that $f(x) = f(x^*)$ for some $x \neq x^*$, x feasible. Then by optimality conditions, $\nabla f(x^*)^T(x^* - x) \leq 0$, and thus

$$\underbrace{f(x) - f(x^*)}_{=0} \geq \underbrace{\nabla f(x^*)^T(x - x^*)}_{\geq 0} + \frac{1}{2L}\sigma_{\mathcal{P}}(\nabla f(x) - \nabla f(x^*))^2,$$

which implies that $\sigma_{\mathcal{P}}(\nabla f(x) - \nabla f(x^*)) = 0$. This means that the vector $w = \nabla f(x) - \nabla f(x^*)$ cannot have any component in $\mathbf{cone}(\mathcal{K}^\circ)$, e.g. it is orthogonal to any $z \in \mathcal{K}$. ◀

4.1 Support recovery

As it was for $\kappa_{\mathcal{P}}$, the domain of $r_{\mathcal{P}}$ is $\mathbf{cone}(\mathcal{P})$. However, the support of $\kappa_{\mathcal{P}}(x)$ and $r_{\mathcal{P}}(x)$ are often not equivalent.

▶ **Example 30** (Different optimal support). Consider $\kappa_{\mathcal{P}}(x) = \|x\|_1$ and $r_{\mathcal{P}}(x) = \frac{1}{\sqrt{2}}\sum_i \sqrt{|x_i|}$. The constrained optimization problem

$$\underset{x}{\text{minimize}} \quad f(x) := -4x_1 - 3x_2 - 4x_3 \qquad \text{subject to} \quad \kappa_{\mathcal{P}}(x) \leq 1$$

has optimal solution $x^* = (1/2, 0, 1/2)$. We verify this from the normal cone condition, where

$$\nabla f(x^*)^T(x - x^*) \geq -\underbrace{\|\nabla f(x^*)\|_\infty \|x\|_1}_{\leq 4} + 4 \geq 0.$$

Note that $r_{\mathcal{P}}(x^*) = 1$ as well. However, taking $\bar{x} = (0, \sqrt{2}, 0)$ also yields $r_{\mathcal{P}}(\bar{x}) = 1$, and has a lower objective value

$$f(\bar{x}) = -3\sqrt{2} \approx -4.24 < -4 = f(x^*).$$

▶ **Example 31** (Different gauge support). The problem can be made even worse, in that the support of x w.r.t. $r_{\mathcal{P}}$ may not even intersect with that w.r.t. $\kappa_{\mathcal{P}}$. Suppose that

$$\mathcal{P}_0 = \left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 3 \\ 0 \end{bmatrix} \right\},$$

and consider $x = (6, 6)$. Then, taking $\gamma(c) = \sqrt{c}$, we have two options

$$x = (0,3) + (3,0), \quad \kappa_{\mathcal{P}}(x) = 2, \quad r_{\mathcal{P}}(x) = 2\sqrt{2} \approx 2.8,$$
$$x = 6 \cdot (1,1), \qquad \kappa_{\mathcal{P}}(x) = 6, \quad r_{\mathcal{P}}(x) = \sqrt{6} \approx 2.4.$$

In other words, the support $\mathbf{supp}_{\mathcal{P}}(x)$ as defined in (26) may not be the support created by the nonconvex gauge $r_{\mathcal{P}}(x)$, which is often sparser. More generally, $r_{\mathcal{P}}(x)$ does not act merely as a concave transformation on the weights c_p in $\kappa_{\mathcal{P}}$, as even the atoms themselves may be selected differently. However, it is worth noting that this scenario does not happen for the ℓ_1 norm or the TV norm, which have unique and consistent supports across choices of monotonically increasing γ.

Overall, the question of nonunique support of a given vector x over atoms $\mathcal{P}_0$ is an interesting one, but not a focus of this paper, which focuses on cases where the support is always unique.

4.2 Stationary points

We can rewrite (32), as the combined optimization problem over c_p, $p \in \mathcal{P}_0$:

$$\underset{c_p \geq 0}{\text{minimize}} \quad f\left(\sum_{p\in\mathcal{P}_0} c_p p\right) + \phi\Bigg(\underbrace{\sum_{p\in\mathcal{P}_0} \gamma(c_p)}_{=:\xi}\Bigg). \tag{35}$$

The stationary points of (35) are x satisfying

$$\forall\, p \in \mathcal{P}_0 :\ 0 \in \nabla f(x)^T p + \phi'(\xi)\, \partial\gamma(c_p), \qquad \text{at } x = \sum_{p\in\mathcal{P}} c_p p. \tag{36}$$

Our goal is to find a support of such a stationary point x^*. Given γ smooth everywhere except at 0, note the close similarity between this and the support optimality conditions for convex gauges:

$$\begin{aligned} c_p > 0 \quad &\Rightarrow \quad -p^T\nabla f(x^*) = \alpha\, \gamma'(c_p) && \text{(no wiggle room)},\\ c_p = 0 \quad &\Rightarrow \quad -p^T\nabla f(x^*) \in \alpha \cdot [-\infty, \gamma_{\max}] && \text{(wiggle room exists)}. \end{aligned}$$

Here, the wiggle room condition looks asymmetric, but note that if p and $-p$ is in $\mathcal{P}_0$, then $c_p = c_{-p} = 0$ implies $-p\nabla f(x^*) \in \alpha \cdot [-\gamma_{\max}, \gamma_{\max}]$, recovering the symmetric condition from Section 2. As before, since γ' is a decreasing function, a nonzero coefficient for x^* does not mean a maximal gradient inner product.

4.3 RP-CGM

In the case that $\mathcal{P}_0$ includes directions of recession, we treat them separately by writing $\mathcal{P}_0 = \mathcal{P}_0' \cup \mathcal{K}$ where $\mathcal{P}_0'$ contains the important (finite-sized) atoms and $\mathcal{K}$ contains directions of recession. We define the *reweighted atomic set* for a given reference point x as

$$\mathcal{P}_0(u) = \left\{ \frac{1}{\gamma'(\mathbf{coeff}_{\mathcal{P}}(u,p))} p : p \in \mathcal{P}_0' \right\}, \quad \mathcal{P}(u) = \mathbf{conv}(\mathcal{P}_0(u)).$$

Then $r_{\mathcal{P}}(s;u) = \kappa_{\mathcal{P}(u)}(s)$, with corresponding reweighted support function

$$\sigma_{\mathcal{P}(u)}(z) = \max_{p\in\mathcal{P}_0} \frac{p^T z}{\gamma'(\mathbf{coeff}_{\mathcal{P}}(u,p))}. \tag{37}$$

At each iteration, we take a penalized conditional gradient step toward solving the reweighted gauge optimization problem with dual

$$\begin{aligned} &\text{(P-general)} && \underset{x,y\in\mathcal{K}}{\text{minimize}} && f(x+y) + \phi(r_0 + \kappa_{\mathcal{P}(\bar{x})}(x)),\\ &\text{(D-general)} && \underset{z\in\mathcal{K}^\circ}{\text{maximize}} && -f^*(-z) - \phi^*(\sigma_{\mathcal{P}(\bar{x})}(z)) + r_0 \cdot \sigma_{\mathcal{P}(\bar{x})}(z). \end{aligned}$$

A description of the most generalized version of the reweighted method is given in Algorithm 2.

4.4 Convergence

▶ **Proposition 32** (Residual). *Denoting* $\mathbf{gap}_{\mathcal{P}}(x;x)$ *the gap at* x *with reference* x*, then*

$$\mathbf{gap}_{\mathcal{P}}(x;x) \geq 0 \ \ \forall\, x, \quad \mathbf{gap}_{\mathcal{P}}(x;x) = 0 \iff x \text{ is a stationary point of } (3).$$

The proof follows closely that of Proposition 8; see Appendix A for full details.

▶ **Theorem 33** (Convergence). *Consider* G *large enough such that for all* $t < 6B$, $\Delta^{(t)} t \leq G$ *and* $G > 24A$. *Given Assumptions 1, 2, 4, 5, with iterates* $x^{(t)} + y^{(t)}$ *from algorithm 2, using* $\theta^{(t)} = 2/(t+1)$*, then*

$$\Delta^{(t)} \leq \frac{G}{t+1} \qquad \text{and} \qquad \min_{i\leq t} \mathbf{res}(x^{(i)}) \leq \frac{3G}{2\log(2)(t+1)}.$$

The details of the proof closely mirror steps in previous works, and thus we give the explicit details in Appendix B.

Let us compare Theorem 33 with the usual rates for CGM. In [39], the primal convergence rate for vanilla CGM (with noiseless gradients) is given as $\Delta^{(t)} \leq \frac{2C_f}{t+2}$ where C_f is a curvature constant that depends on the conditioning of f and the size of P. These players appear here in the form of the conditioning of f (quadratic in L/μ), and implicitly $\sigma_{\widetilde{\mathcal{P}}}$ (which grows proportionally with $\mathbf{diam}\ \mathcal{P}$). The new players ν_0, $\gamma_{\min}$, and $\gamma_{\max}$ account for the penalty and nonconvex generalizations.

Algorithm 2 RP-CGM on general nonconvex sparse optimization

procedure RP-CGM(f, ϕ, γ, $\mathcal{P}_0 = \mathcal{P}_0' \cup \mathcal{K}$, max iter T)
  Initialize with any $x^{(0)} \in \mathbf{cone}(\mathcal{P})$ where $\mathcal{P}$ is the convex hull of $\mathcal{P}_0'$, $y^{(0)} \in \mathcal{K}$.

  for $t = 1, \dots, T$ **do**
    Compute the projected negative gradient $z = -\nabla f(x^{(t)} + y^{(t)})$.
    Compute the reweighted atomic set $\mathcal{P}(x)$.
    Compute next atom $s = \xi p$ in two steps. ▷ Pick next atom
      1. Compute direction $p = \mathbf{LMO}_{\mathcal{P}(x)}(z)$.
      2. Compute magnitude $\xi = (\phi^*)'(\sigma_{\mathcal{P}}(z))$.
    Update $x^{(t+1)} = (1 - \theta^{(t)})x^{(t)} + \theta^{(t)} s$ where $\theta^{(t)} = 2/(1+t)$. ▷ Merge
    Update $y^{(t+1)} = \operatorname{argmin}_{y \in \mathcal{K}} \ f(x^{(t+1)} + y)$. ▷ Recession component
  return $x^{(T)} + y^{(T)}$

4.5 Invariance

Finally, we investigate the linear invariance properties of RP-CGM. Specifically, we consider $\mathcal{Q} = A\mathcal{P}$, $f(x) = g(Ax)$, $w = Ax$, $\overline{w} = A\overline{x}$, where A has full column rank. We will have preserved linear invariance if RP-CGM applied to

$$\min_x \{f(x) : x \in \mathcal{P}\} \quad \text{and} \quad \min_w \{g(w) : w \in \mathcal{Q}\}$$

are equivalent. Assume additionally that both $x, \overline{x} \in \mathbf{cone}(\mathcal{P})$. Then the following hold.

- **Penalty.** $r_{\mathcal{P}}(x) = r_{\mathcal{Q}}(w)$. This follows from noting that

  $$x = \sum_{p \in \mathcal{P}} c_p p \iff w = \sum_{p \in \mathcal{P}} c_p (Ap) = \sum_{q \in \mathcal{Q}} c_q' q$$

  and in fact noting that the coefficients are equal ($c_q' = c_{Ap}$).
- **Stationarity.** We construct P with columns containing the atoms in $\mathcal{P}_0'$, and c such that $x = Pc$, $w = Ax = APc$.

  $$P^T \partial r_{\mathcal{P}}(x) = \partial r_{\mathcal{P}}(c) \overset{r_{\mathcal{P}}(c) = r_{\mathcal{Q}}(c)}{=} \partial r_{\mathcal{Q}}(c) = P^T A^T \partial r_{\mathcal{Q}}(w).$$

  Additionally, for any stationary point x^*, if $\nabla f(x^*) \notin \mathbf{cone}(\mathcal{P})$ then there exists a descent direction that is uneffected by the penalty $r_{\mathcal{P}}(x)$, and thus it must be that $\nabla f(x^*) \in \mathbf{cone}(\mathcal{P})$. By the same token, $A^T \nabla g(w^*) \in \mathbf{cone}(\mathcal{P})$. Therefore, the stationary conditions are equivalent: for $x^* = Aw^*$,

  $$0 \in \nabla f(x^*) + P^T \partial r_{\mathcal{P}}(x^*) \iff 0 \in A^T \nabla g(w^*) + A^T \partial r_{\mathcal{Q}}(w^*).$$

Additionally, it can be shown through the chain rule that $A\mathcal{P}(x) = \mathcal{Q}(w)$ and $\mathbf{res}_{\mathcal{P}}(x) = \mathbf{res}_{\mathcal{Q}}(w)$. Overall, this shows that the steps and analysis of RP-CGM are all invariant to linear transformations on x.

4.6 Screening

We now describe the gradient error measured in terms of this "dual gauge", where the symmetrization $\widetilde{\mathcal{P}} := \mathcal{P} \cup -\mathcal{P}$ ensures that $\sigma_{\widetilde{\mathcal{P}}}(z - z^*) = \sigma_{\widetilde{\mathcal{P}}}(z^* - z)$, bounding errors in both directions.

▶ **Proposition 34** (Gap bound on gradient error). *Denote $D(x) = r_{\mathcal{P}}(x) - r_{\mathcal{P}}(x^*) + \bar{r}_{\mathcal{P}}(x; x) - \bar{r}_{\mathcal{P}}(x^*; x)$ the linearization error at x. Denoting x^* a stationary point of (32) and $y(x) = \operatorname{argmin}_{y' \in \mathcal{K}} \ f(x + y')$, then*

$$\sigma_{\widetilde{\mathcal{P}}}(\nabla f(x + y(x)) - \nabla f(x^* + y(x^*))) \leq \frac{LD(x)}{2\gamma_{\min}} + \sqrt{\frac{L^2 D(x)^2}{4\gamma_{\min}^2} + L\mathbf{res}(x) + LD(x)\frac{\sigma_{\widetilde{\mathcal{P}}}(\nabla f(x + y(x)))}{\gamma_{\min}}}.$$

The linearization error $D(x) = 0$ when the regularizer is convex. The proof is similar to that for Proposition 11, and is detailed in Appendix C.

▶ **Theorem 35** (Dual screening). *For any x and some choice of $\epsilon > 0$, define the screened set as*

$$\mathcal{I}_\epsilon(x) = \{p \in \mathcal{P}_0 : \sigma_{\widetilde{\mathcal{P}}}(\nabla f(x)) + p^T \nabla f(x) > \epsilon + 2\sqrt{L\mathbf{res}(x) + \epsilon}\}. \tag{38}$$

Then given Assumptions 1, 2, and 5, if

$$\epsilon \geq \frac{LD(x)}{\gamma_{\min}} \max\left\{\frac{1}{2}, \frac{LD(x)}{4\gamma_{\min}} + \sigma_{\widetilde{\mathcal{P}}}(\nabla f(x))\right\}$$

then $p \notin \mathbf{supp}_{\mathcal{P}}(x^*)$, where x^* is the optimal variable in (20).

In the convex case, $D(x) = 0$, and thus we pick $\epsilon = 0$ in our screening rule. In this scenario, not only does this screening rule achieve finite-iteration support identification, but the finite time $\bar{t}$ depends directly on $\delta_{\min}$.

5 Experiments

In this section, we explore the convergence behavior and screening ability of P-CGM and RP-CGM on compressed sensing (with ℓ_1, group norm, and TV regularization), and on a sparse logistic regression task on a real world dataset. The code for all the experiments is publicly available. [2]

5.1 Sensing experiment

We first compare the various CGM variants on a simple simulated sparse sensing problem (Figures 2 and 3). We solve a least squares problem

$$\underset{x}{\text{minimize}} \ \tfrac{1}{2m}\|Ax - b\|_2^2 + \phi(r_{\mathcal{P}}(x)), \tag{39}$$

where $A \in \mathbb{R}^{m\times n}$ as $A_{ij} \sim \mathcal{N}(0, 1/n)$ i.i.d. for $i = 1, \ldots, m$, $j = 1, \ldots, n$, and for a given x_0 with 10% nonzero sparsity, $b = Ax_0$. Specifically, we pick $m = n = 100$, where perfect sensing is possible, and either sweep or tune all the hyperparameters to investigate each case.

An important modification needed to improve the stability of P-CGM and RP-CGM is to intensely diminish the step size; in particular, using $\theta^{(t)} = 2/(2+t)$ is too aggressive, so instead we use $\theta^{(t)} = 2/(2+t+t_0)$, where t_0 is another tuned hyperparameter. Note that in performance, this does not slow down the convergence or sensing abilities of the P-CGM and RP-CGM, suggesting that this is a more appropriate step size sequence in these regimes (and is still $O(1/t)$). All hyperparamters (α, ρ, t_0) were tuned to present the best results for each individual method. These two collections of figures are presented to illustrate several points:

- The gaps (left column) in all cases converges to 0 or machine precision at about a $O(1/t)$ rate.
- The screen error (right column), measured as the support difference between $x^{(t)}$ and x^* the final converged point, eventually goes to 0, at a speed somewhat correlated with the "aggressiveness" of the method (where RP-CGM is often more aggressive than P-CGM, but all three variants also depend heavily on choice of hyperparameter). Note that higher ρ, smaller θ, and smaller α all correspond to more aggressive methods.
- In contrast, the support error, measured as the support difference between $x^{(t)}$ and x_0 the ground truth, seems to have better performance when the method is *less* aggressive. It is hard to make sweeping conclusions, but suggests that both metrics are essential to evaluate the success of sparse recovery methods.

5.2 Other gauges

We now pursue the sensing problem for more creative choices of $\mathcal{P}_0$.

First, we consider the group norm in cases where when x_0 has a "pulse-like" structure, in that the signal has blocks of nonzero activity, separated by long spans of zero activity. This can be modeled as $x_0 = \sum_i s_0^i$ where s_0^i is a pulse signal across the ith overlapping window. Figure 4 shows the trajectory for such an experiment, where the complementary characteristics between the primal variable and dual norms is visible, as over time, the nonzero blocks of the primal correspond to the maximal blocks of the dual.

Next, we consider the total variation penalization, where $\kappa_{\mathcal{P}}(x) = \sum_i |x_i - x_{i-1}|$, and what is plotted is the cumulative sum of the demeaned $z^{(t)} = -\nabla f(x^{(t)})$. Note that at optimality, the peaks of this dual atom exactly match the "flip points" of $x^{(t)}$.

[2] Code link: `https://github.com/yifan0sun/rpcgm`

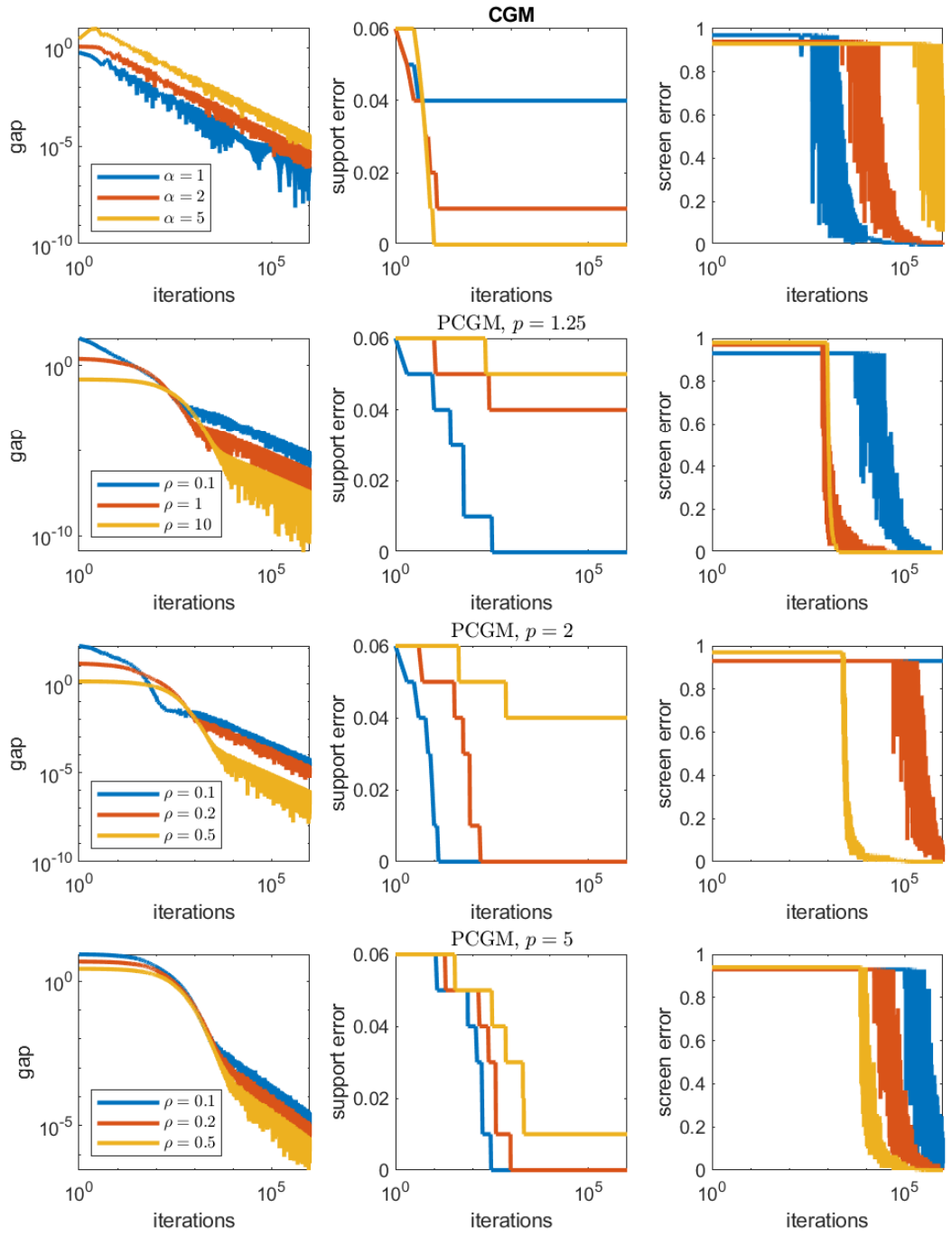

Figure 2 Small sensing experiment, CGM and P-CGM. The first row shows CGM, with $\phi(s) = \iota_{s<\alpha}$. The next three rows show P-CGM where $\phi(s) = \frac{\rho}{p}s^p$, for various values of p and corresponding optimal ρ. For the P-CGM experiments, we also used an iteration offset of $t_0 = 1000$. Offset was not needed for CGM ($t_0 = 0$).

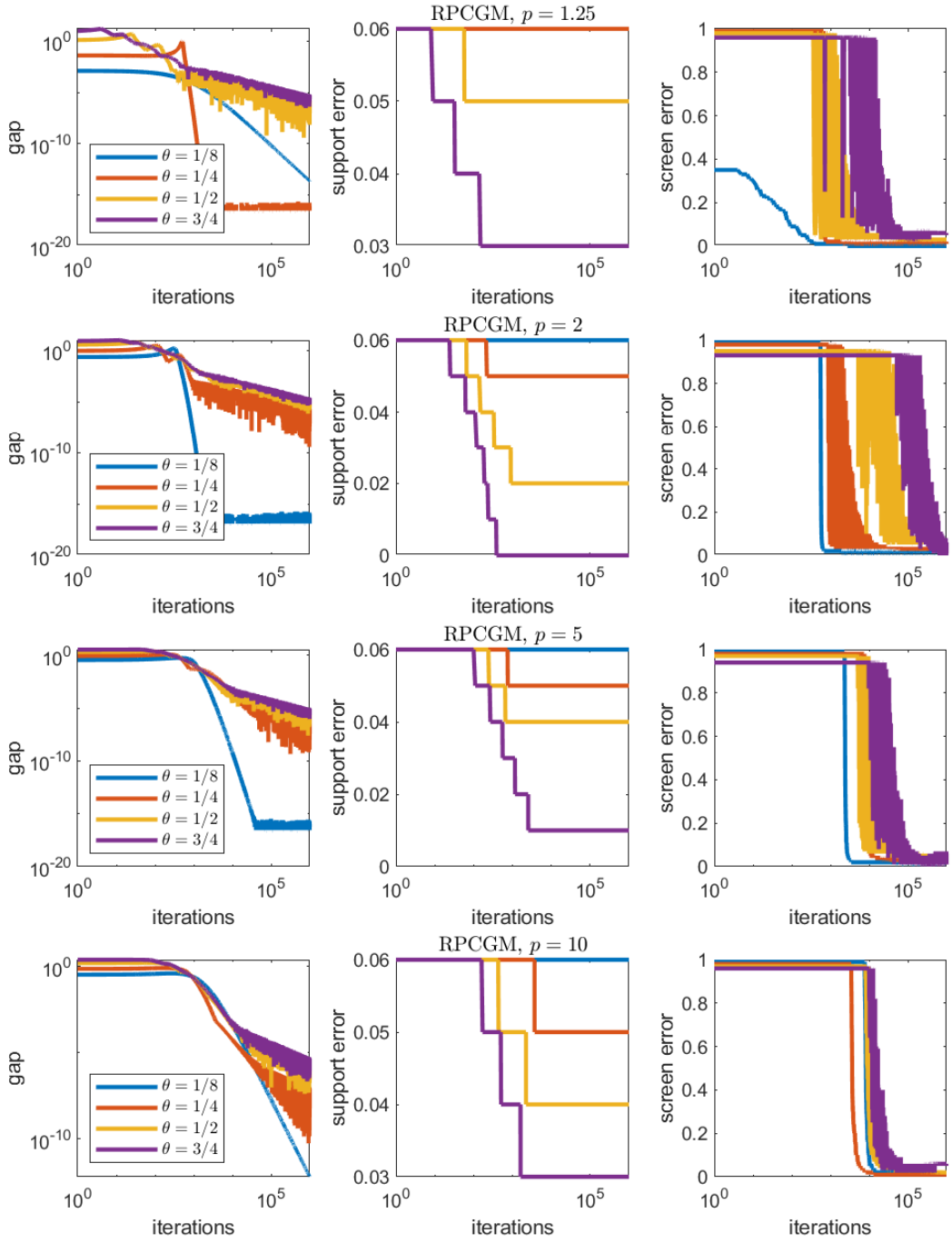

Figure 3 Small sensing experiment, RP-CGM. We again use $\phi(s) = \frac{\rho}{p} s^p$, and use a piecewise LSP function for RP-CGM ($\gamma(w) = \log(1 + |w|/\theta)$ if $|w| \leq c$, and $\gamma(w) = \gamma'(\bar{c})(w - \bar{c})$ for $|w| > \bar{c}$). In addition to what is labeled in the figure, we use $\rho = 0.5, 0.1, 0.01, 0.001$ for $p = 1.25, 2, 5, 10$ respectively (all tuned for best performance). Additionally, we have $t_0 = 1000$.

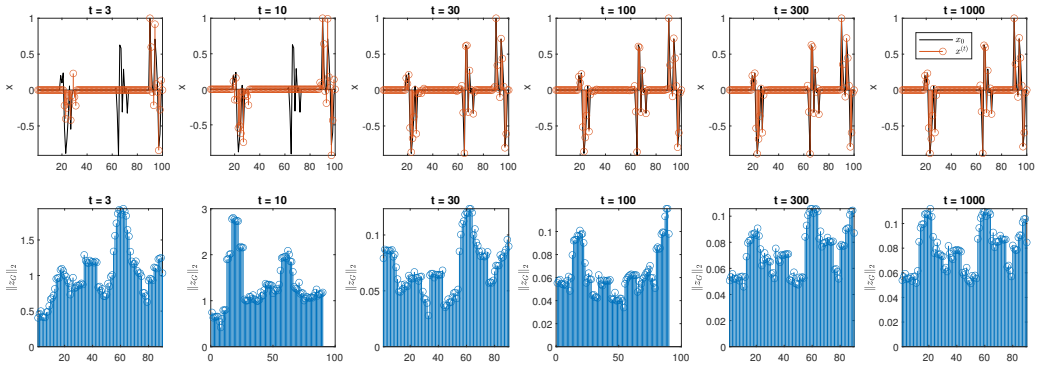

Figure 4 Trajectory of primal variable and dual group norms. Here we investigate RP-CGM where $\theta = 1/2$, $\rho = 0.01$, and $p = 2$. For stability, $t_0 = 100$. The ground truth x_0 contains 3 "pulses", e.g. areas where it is nonzero, and the goal is to fit $x^{(t)}$ to x_0, using this group structure prior.

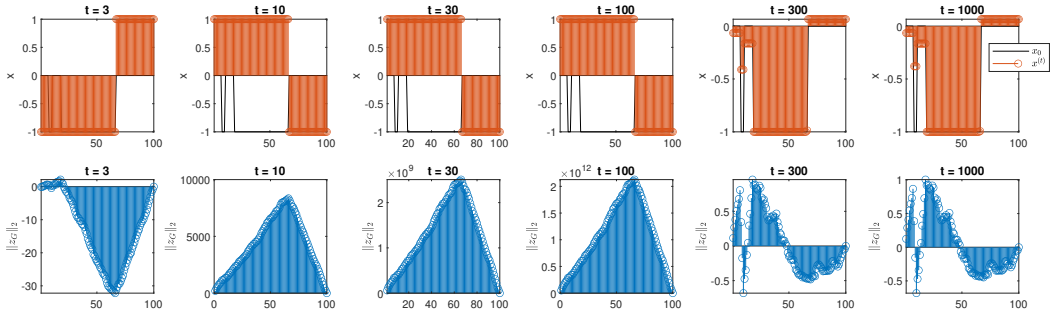

Figure 5 Trajectory of primal variable and dual variables for TV penalty. Here we investigate RP-CGM where $\theta = 3/4$, $\rho = 0.25$, and $p = 2$. For stability, $t_0 = 100$. The ground truth x_0 contains 3 flips, and is otherwise smooth. As before, the goal is to fit $x^{(t)}$ to x_0, using this group structure prior.

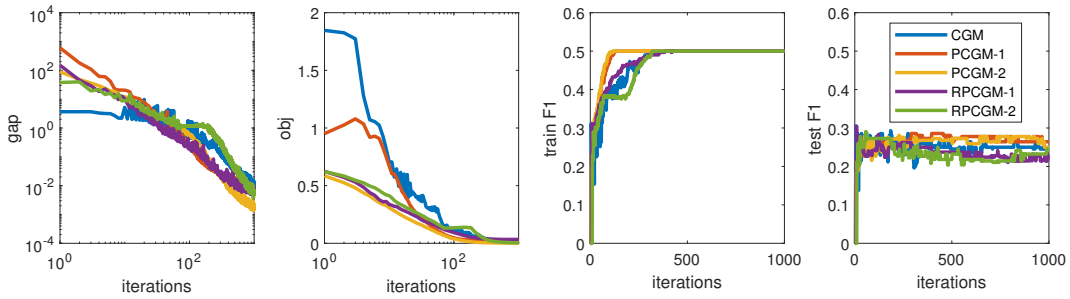

Figure 6 Dorothea classification experiment For CGM, $\alpha = 100$. PCGM-1 and RPCGM-1 uses $p = 2$, and PCGM-2 and RPCGM-2 uses $p = 5$. Additionally, RPCGM-1 and RPCGM-2 both use $\theta = \bar{c} = 1/2$.

5.3 Dorothea experiment

Finally, we consider a "real world" experiment, in which we use these methods to classify the Dorothea dataset [35]. Sparse optimization is essential in this application, which has only 1950 samples but 100000 attributes. Additionally, the dataset is heavily imbalanced, with very few positive labels. We run sparse logistic regression over this dataset, and illustrate the performance of the different methods in Figure 6. Note that the best implementation reaches an F1 score of about 0.3; without regularization, logistic regression achieves a test F1 score of about 0.16, highlighting the importance of sparse regularization.[3]

[3] Our F1 scores are not comparable to SOTA on this task, as we use a weak classifier (better models, like boosted decision trees, do not have differentiable f) and do not account for label imbalance. It is possible that the score may be improved with more involved data science techniques, which is not the focus of this work.

6 Discussion

This work considers two variations of the conditional gradient method (CGM): the P-CGM, which accommodates gauge-based penalties in place of constraints, and the RP-CGM, which allows concave transformations of the gauges. The gauges may be induced by compact sets, but also accomodate "simple" directions of recession. We give a convergence rate to a stationary point, and propose a gradient screening rule and support recovery guarantee. Compared with proximal methods, these CGM-based methods often have a much cheaper per-iteration cost; e.g. in the group norm, computing the LMO (without reweighting) is trivial compared to even computing the gauge function itself. Additionally, the almost-for-free computation of the gap and residual quantity makes screening a very small computational addition.

The key challenge in showing the convergence of these methods is controlling the size of each $s^{(t)}$. This was trivial in the CGM case when $s^{(t)}$ was constrained in a compact set; when transformed to a penalty, we require a minimum amount of curvature of ϕ at $\xi \to +\infty$, and we restrict γ to only having strict concavity over a finite support. However, as shown in the numerical results, these restrictions do not greatly inhibit the sparsifying effects of the penalty functions.

After determining convergence behavior, we then implement gap-based screening, which allows for knowledge of the true solution's sparsity pattern without completing optimization. This is a deliberate tool to reduce computational cost, and can be used in a number of ways. For nonzero sparsity or group sparsity, we can simply avoid computation over the "determined zeros". For problems where the solution is significantly sparse, a 2-stage solving technique can be used, where after enough zero components have been screened away, the problem can be solved over the reduced support using a more powerful (e.g. 2nd order) method. And, for problems with a very large number of atoms that need to be explicitly queried at each iteration (e.g. in submodular optimization) we can significantly reduce the search space. Therefore we believe these techniques have many practical benefits in a number of applications.

Finally, we do not incorporate away step [34, 42]. In implementation, they are somewhat orthogonal to the extensions provided in this work; an away-step implementation of P-CGM and RP-CGM can be directly implemented, and its analysis is a subject for future work.

A General lemmas

Proof of Lemma 2

Proof. Assume that ϕ_0 is as large as possible; e.g., there exists some finite ξ_0 where $\phi(\xi_0) = \mu\xi_0^2 - \phi_0$. By convexity, for all $\nu \in \partial\phi(\xi)$,

$$\phi(\xi) - \phi(\xi_0) \leq \nu(\xi - \xi_0).$$

Additionally, by the assumption,

$$\mu\xi^2 - \phi_0 \leq \phi(\xi), \quad \forall\, \xi.$$

Therefore,

$$\mu(\xi^2 - \xi_0^2) \leq \phi(\xi) - \phi(\xi_0) \leq \nu(\xi - \xi_0),$$

and therefore, for $\xi \geq \xi_0$,

$$\nu \geq \mu\frac{(\xi + \xi_0)(\xi - \xi_0)}{\xi - \xi_0} = \mu\xi + \mu\xi_0 \iff \xi \leq \mu^{-1}\nu - \xi_0.$$

Thus, for any ξ, $\nu \in \partial\phi(\xi)$ must satisfy $\xi \leq \max\{\xi_0, \mu^{-1}\nu - \xi_0\} \leq \mu^{-1}\nu + \xi_0$. By Fenchel Young, this must apply to all $\xi \in \partial\phi^*(\nu)$. ◀

Proof of Lemma 28

Proof. The proof largely follows from [52], mildly adapted.

- First prove (25) $\Rightarrow$ (33). Construct $g(x) = f(x) - x^T \nabla f(y)$, which is convex, also L-smooth, and has minimum at $x = y$. Then, for any w,

  $$g(y) \leq g(x+w) \overset{(a)}{\leq} g(x) + \nabla g(x)^T w + \frac{L}{2}\kappa_{\mathcal{P}}(w)^2,$$

  where (a) is since g is L smooth and convex.

  Now pick

  $$w \in \frac{1}{L}\sigma_{\mathcal{P}}(-\nabla g(x))\partial\sigma_{\mathcal{P}}(-\nabla g(x)),$$

  which implies

  $$\frac{L}{\sigma_{\mathcal{P}}(-\nabla g(x))} w \in \underset{\kappa_{\mathcal{P}}(u)\leq 1}{\operatorname{argmax}} \langle u, -\nabla g(x)\rangle = \partial\sigma_{\mathcal{P}}(-\nabla g(x)),$$

  and thus

  $$\kappa_{\mathcal{P}}(w) = \frac{\sigma_{\mathcal{P}}(-\nabla g(x))}{L},$$

  and

  $$\langle w, -\nabla g(x)\rangle = \frac{1}{L}\sigma_{\mathcal{P}}(-\nabla g(x))^2.$$

  Then

  $$\frac{L}{2}\kappa_{\mathcal{P}}(w)^2 = \frac{1}{2L}\sigma_{\mathcal{P}}(-\nabla g(x))^2,$$

  and plugging in the construction for g gives

  $$g(y) - g(x) \leq \underbrace{\nabla g(x)^T w + \frac{L}{2}\kappa_{\mathcal{P}}(w)^2}_{-\frac{1}{2L}\sigma_{\mathcal{P}}(-\nabla g(x))^2}$$

  $$\iff f(y) - f(x) \leq (y-x)^T \nabla f(y) - \frac{1}{2L}\sigma_{\mathcal{P}}(\nabla f(y) - \nabla f(x))^2.$$

  Applying the last inequality twice gives

  $$(y-x)^T(\nabla f(y) - \nabla f(x)) \leq \frac{1}{2L}((\sigma_{\mathcal{P}}(\nabla f(x) - \nabla f(y))^2 + (\sigma_{\mathcal{P}}(\nabla f(y) - \nabla f(x))^2).$$

- Now prove (25) $\Rightarrow$ (34). Using the same g as before, consider

  $$\min_z \; g(x) + \langle \nabla g(x), z - x\rangle + \frac{L}{2}\kappa_{\mathcal{P}}(x-z)^2 = \min_w \langle \nabla g(x), w\rangle + \frac{L}{2}\kappa_{\mathcal{P}}(w)^2.$$

  Using optimality conditions, picking $w = z - y$, we have

  $$0 \in \nabla g(x) + L\kappa_{\mathcal{P}}(w)\partial\kappa_{\mathcal{P}}(w) \iff -\frac{1}{L\kappa_{\mathcal{P}}(w)}\nabla g(x) = \underset{\sigma_{\mathcal{P}}(u)\leq 1}{\operatorname{argmax}} \langle u, w\rangle,$$

  which implies

  $$\sigma_{\mathcal{P}}(-\nabla g(x)) = L\kappa_{\mathcal{P}}(w), \qquad -\frac{1}{L\kappa_{\mathcal{P}}(w)}\langle w, \nabla g(x)\rangle = \kappa_{\mathcal{P}}(w).$$

  so

  $$\langle w, -\nabla g(x)\rangle = L\kappa_{\mathcal{P}}(w)^2 = \frac{1}{L}\sigma_{\mathcal{P}}(-\nabla g(x))^2,$$

  and overall

  $$g(y) \geq \min_z \; g(x) + \langle \nabla g(x), z - x\rangle + \frac{L}{2}\kappa_{\mathcal{P}}(x-z)^2 = g(x) - \frac{1}{2L}\sigma_{\mathcal{P}}(-\nabla g(x))^2.$$

  Plugging in f gives

  $$f(y) - f(x) \geq (y-x)^T \nabla f(y) - \frac{1}{2L}\sigma_{\mathcal{P}}(\nabla f(y) - \nabla f(x))^2.$$ ◀

Proof of Proposition 22

Proof. Without loss of generality, we assume $0 \in \mathcal{P}$, since $\kappa_{\mathcal{P}} = \kappa_{\mathcal{P}\cup\{0\}}$. Denote $z^* = -\nabla f(x^*)$. Then the optimality condition for (20) is

$$z^* \in \partial h(x^*) \overset{(\star)}{=} \alpha \partial \kappa_{\mathcal{P}}(x^*), \quad h(x) := \phi(\kappa_{\mathcal{P}}(x)) \tag{40}$$

for some $\alpha \in \partial\phi(\xi)$ at $\xi = \kappa_{\mathcal{P}}(x^*)$. Here, $(\star)$ is a result from [4, Corollary 16.72].

Since ϕ is monotonically nondecreasing over $\mathbb{R}^+$, $\alpha \geq 0$. If $\alpha = 0$, then $\nabla f(x^*) = 0$ and both results are trivially true. Now consider $\alpha > 0$. Noting that $\kappa_{\mathcal{P}} = \sigma_{\mathcal{P}^\circ}$ where $\mathcal{P}^\circ$ is the polar set of $\mathcal{P}$,

$$\alpha^{-1} z^* = \operatorname*{argmax}_{z \in \mathcal{P}^\circ} (x^*)^T z \iff (z^*)^T x^* = \kappa_{\mathcal{P}^\circ}(z^*)\sigma_{\mathcal{P}^\circ}(x^*) = \kappa_{\mathcal{P}}(x^*)\sigma_{\mathcal{P}}(z^*)$$

which proves (27). Now take the conic decomposition $x^* = \sum_{p\in\mathcal{P}_0} c_p p$ where $c_p \geq 0$, and

$$(x^*)^T z^* = \sum_{p\in\mathcal{P}_0} c_p p^T z^* \leq \underbrace{\left(\sum_{p\in\mathcal{P}_0} c_p\right)}_{=\kappa_{\mathcal{P}}(x^*)} \underbrace{(p^T z^*)}_{\leq \sigma_{\mathcal{P}}(z^*)},$$

which is with equality if and only if $p^T z^* = \sigma_{\mathcal{P}}(z^*)$ whenever $c_p > 0$, proving (28). ◀

Proof of Proposition 32

Proof. Denote $y = \operatorname{argmin}_y \ f(x+y)$, and $z = -\nabla f(x+y)$, and plug in $\kappa_{\mathcal{P}(x)}(x) = \bar{r}_{\mathcal{P}}(x;x)$. Then

$$\begin{aligned}
\mathbf{res}_{\mathcal{P}}(x) &= f(x+y) + f^*(-z) + \phi(r_{\mathcal{P}}(x)) + \phi^*(\sigma_{\mathcal{P}(x)}(z)) + (\kappa_{\mathcal{P}(x)}(x) - r_{\mathcal{P}}(x)) \cdot \sigma_{\mathcal{P}(x)}(z) \\
&\overset{(a)}{=} x^T \nabla f(x+y) + \phi(r_{\mathcal{P}}(x)) + \phi^*(\sigma_{\mathcal{P}(x)}(z)) + (\kappa_{\mathcal{P}(x)}(x) - r_{\mathcal{P}}(x)) \cdot \sigma_{\mathcal{P}(x)}(z) \\
&\overset{(b)}{\geq} x^T \nabla f(x+y) + \underbrace{y^T \nabla f(x+y)}_{\geq 0} + r_{\mathcal{P}}(x)\sigma_{\mathcal{P}(x)}(z) + (\kappa_{\mathcal{P}(x)}(x) - r_{\mathcal{P}}(x)) \cdot \sigma_{\mathcal{P}(x)}(z) \\
&\overset{(b)}{\geq} x^T \nabla f(x+y) + \kappa_{\mathcal{P}(x)}(x) \cdot \sigma_{\mathcal{P}(x)}(z) \\
&\overset{(c)}{\geq} x^T \nabla f(x+y) - x^T \nabla f(x+y) = 0
\end{aligned}$$

where

(a) uses the Fenchel–Young inequality on f and f^*,
(b) uses the Fenchel–Young inequality on ϕ and ϕ^*,
(c) follows since $-\nabla f(x+y) \in \mathcal{K}^\circ$ and $y \in \mathcal{K}$, and thus $y^T z \geq 0$, and
(d) follows from the definition of $\sigma_{\mathcal{P}(x)}$.

Tightness of (b) occurs iff Fenchel–Young is satisfied with equality, e.g.

$$\sigma_{\mathcal{P}(x)}(z) \in \partial\phi(r_{\mathcal{P}}(x)) \tag{41}$$

Tightness of (c) occurs iff

$$\sigma_{\mathcal{P}(x)}(z) = \frac{-\nabla f(x)^T p}{\gamma'(\mathbf{coeff}_{\mathcal{P}}(x;p))}, \quad \forall\ p,\ c_p \neq 0. \tag{42}$$

The "element-wise" optimality conditions for (32) are, for all $p \in \mathcal{P}_0$,

$$\begin{aligned}
\frac{-\nabla f(x)^T p}{\gamma'(\mathbf{coeff}_{\mathcal{P}}(x;p))} &\in \gamma'(c_p) \cdot \partial\phi(r_{\mathcal{P}}(x)) \qquad \text{if } c_p \neq 0 \\
\frac{-\nabla f(x)^T p}{\gamma'(\mathbf{coeff}_{\mathcal{P}}(x;p))} &\leq \gamma'(c_p) \max_{g_\phi \in \partial\phi(r_{\mathcal{P}}(x))} g_\phi \quad \text{if } c_p = 0
\end{aligned}$$

which is true iff (41), (42) hold. ◀

B Convergence results from Section 4

▶ **Lemma 36** (Iterate gauge control). *Given Assumptions 1, 2,5, suppose additionally* $\theta^{(t)} = 2/(t+1)$. *Then*

$$\kappa_{\mathcal{P}}(s^{(t)} - x^{(t)}) \leq \frac{\gamma_{\max}}{\gamma_{\min}\mu}\left(2\sigma_{\widetilde{\mathcal{P}}}(\nabla f(x^* + y^*)) + \sqrt{2L\Delta^{(t)}} + \frac{2}{t(t-1)}\sum_{u=1}^{t-1}\sqrt{2L\Delta^{(u)}}\right) + 2\nu_0\gamma_{\max} + \kappa_{\mathcal{P}}(x^{(0)}).$$

Proof.

$$\begin{aligned}
\kappa_{\mathcal{P}}(s^{(t)} - x^{(t)}) &\overset{\text{subadditive gauge}}{\leq} \kappa_{\mathcal{P}}(s^{(t)}) + \kappa_{\mathcal{P}}(x^{(t)}) \\
&\overset{\text{convexity}}{\leq} \kappa_{\mathcal{P}}(s^{(t)}) + \theta^{(t-1)}\kappa_{\mathcal{P}}(s^{(t-1)}) + (1-\theta^{(t-1)})\kappa_{\mathcal{P}}(x^{(t-1)}) \\
&\overset{\text{recursion}}{\leq} \kappa_{\mathcal{P}}(s^{(t)}) + \kappa_{\mathcal{P}}(x^{(0)}) + \sum_{u=1}^{t-1}\theta^{(u)}\underbrace{\prod_{u'=u+1}^{t-1}(1-\theta^{(u')})}_{=\frac{(u+1)u}{t(t-1)}}\kappa_{\mathcal{P}}(s^{(u)}) \\
&\leq \kappa_{\mathcal{P}}(s^{(t)}) + \kappa_{\mathcal{P}}(x^{(0)}) + \frac{2}{t(t-1)}\sum_{u=1}^{t-1}u\kappa_{\mathcal{P}}(s^{(u)}).
\end{aligned}$$

In general, for any x, z, $\bar{x}$,

$$\kappa_{\mathcal{P}}(x) \leq \gamma_{\max}\kappa_{\mathcal{P}(\bar{x})}(x), \qquad \sigma_{\mathcal{P}}(z) \geq \frac{1}{\gamma_{\min}}\sigma_{\mathcal{P}(\bar{x})}(z).$$

Taking $y^{(u)} = \operatorname{argmin}_{y\in\mathcal{K}}\ f(x^{(u)} + y)$, $z^{(u)} = -\nabla f(x^{(u)} + y^{(u)})$, $z^* = -\nabla f(x^* + y^*)$:

$$\begin{aligned}
\kappa_{\mathcal{P}(\bar{x})}(s^{(u)}) = (\phi^*)'\left(\sigma_{\mathcal{P}(\bar{x})}(z^{(u)})\right) &\overset{\text{Asspt. 1}}{\leq} \mu^{-1}\cdot\sigma_{\mathcal{P}(\bar{x})}(z^{(u)}) + \nu_0 \\
&\overset{\text{Bound on }\gamma'}{\leq} \frac{1}{\mu r_{\min}}\sigma_{\mathcal{P}}(z^{(u)}) + \nu_0 \\
&\overset{\substack{\Delta\text{-ineq +}\\ \text{Prop. 34}}}{\leq} \frac{1}{\mu r_{\min}}\left(\sigma_{\widetilde{\mathcal{P}}}(z^*) + \sqrt{2L\Delta^{(u)}}\right) + \nu_0.
\end{aligned}$$

Putting it all together gives the desired result. ◀

From Lemmas 37 and 36, we arrive at

$$\Delta^{(t+1)} - \Delta^{(t)} \leq -\theta^{(t)}\mathbf{res}_{\mathcal{P}}(x^{(t)}) + (\theta^{(t)})^2\left(B\Delta^{(t)} + B\overline{\Delta}^{(t-1)} + A\right)$$

for constants

$$A = \left(\frac{6L\gamma_{\max}^2}{\mu^2\gamma_{\min}^2}\sigma_{\widetilde{\mathcal{P}}}(-\nabla f(x^* + y^*) + 6\gamma_{\max}\nu_0 + 3\kappa_{\mathcal{P}}(x^{(0)})\right)^2, \qquad B = \frac{3L^2\gamma_{\max}^2}{\mu^2\gamma_{\min}^2}$$

and where $\overline{\Delta}^{(t)}$ is defined as an averaging over square roots, e.g.

$$\sqrt{\overline{\Delta}^{(t)}} = \frac{2}{t(t+1)}\sum_{u=1}^{t}u\sqrt{\Delta^{(u)}}.$$

▶ **Lemma 37** (One step descent). *Suppose* f *is* L*-smooth w.r.t.* $\mathcal{P}$ *(unweighted). Take*

$$x^+ = (1-\theta)x + \theta s, \qquad s = \operatorname*{argmin}_{\tilde{s}}\ \nabla f(x+y)^T\tilde{s} + \bar{h}(s;x)$$

for some $\theta \in (0,1)$. *Define* $y = \operatorname{argmin}_{y\in\mathcal{K}}\ f(x+y)$, $y^+ = \operatorname{argmin}_{y\in\mathcal{K}}\ f(x+y)$. *Then*

$$f(x^+ + y^+) + h(x^+) - f(x+y) - h(x) \leq -\theta\mathbf{res}(x) + \frac{L\theta^2}{2}\kappa_{\mathcal{P}}(s-x)^2.$$

Proof. From L-smoothness we have

$$\begin{aligned} f(x^+ + y^+) - f(x+y) &\le f(x^+ + y) - f(x+y) \\ &\le \nabla f(x+y)^T(x^+ - x) + \frac{L}{2}\kappa_{\mathcal{P}}(x^+ - x) \\ &= \theta \nabla f(x+y)^T(s-x) + \frac{L\theta^2}{2}\kappa_{\mathcal{P}}(s-x)^2 \end{aligned} \tag{43}$$

Denote $\nu = \sigma_{\widetilde{\mathcal{P}}(x)}(-\nabla f(x+y))$. Since $s = \xi\phi'(\nu)$, then

$$\begin{aligned} \nabla f(x+y)^T s + \phi(r_0 + \bar{r}_{\mathcal{P}}(s;x)) &= \min_{\widetilde{s}} \underbrace{\nabla f(x+y)^T \widetilde{s}}_{=-\xi\cdot\nu} + \phi\Big(r_0 + \underbrace{\bar{r}_{\mathcal{P}}(s;x)}_{=\xi}\Big) \\ &= \nu r_0 - \phi^*(\nu). \end{aligned} \tag{44}$$

Also, by definition of residual,

$$\begin{aligned} \mathbf{res}_{\mathcal{P}}(x) &= f(x+y) + f^*(\nabla f(x+y)) + \phi(r_{\mathcal{P}}(x+y)) + \phi^*(\nu) - r_0\cdot\nu \\ &= \underbrace{\nabla f(x+y)^T(x+y)}_{\nabla f(x+y)^T y \ge 0} + \phi(r(x)) + \phi^*(\nu) - r_0\cdot\nu \\ &\ge \nabla f(x+y)^T x + \phi(r(x)) + \phi^*(\nu) - r_0\cdot\nu. \end{aligned} \tag{45}$$

Therefore taking $F(x+y) = f(x+y) + \phi(r(x))$ and combining (43), (44), and (45),

$$\begin{aligned} F(x^+ + y^+) - F(x+y) = &-\theta\mathbf{res}(x) + \theta\left(\phi(r_{\mathcal{P}}(x)) - \phi(r_0 + \bar{r}_{\mathcal{P}}(s;x))\right) \\ &+ \frac{L\theta^2}{2}\kappa_{\mathcal{P}}(s-x)^2 + \phi(r_{\mathcal{P}}(x^+)) - \phi(r_{\mathcal{P}}(x)) \end{aligned}$$

Next, by convexity of ϕ,

$$\begin{aligned} (1-\theta)\phi(r_{\mathcal{P}}(x)) + \theta\phi(r_0 + \bar{r}_{\mathcal{P}}(s;x)) &\ge \phi(r_{\mathcal{P}}(x) + \bar{r}_{\mathcal{P}}(x^+;x) - \bar{r}_{\mathcal{P}}(x;x)) \\ &\overset{\text{majorant}}{\ge} \phi(r_{\mathcal{P}}(x^+;x)) \end{aligned}$$

which leaves the desired result. ◀

▶ **Proposition 38** (Linearized objective value bound)**.** *Given Assumptions 1, 2, 4, 5, then the objective error of each linearized problem decreases as*

$$\Delta^{(t)} = O(1/t).$$

Proof. Define

$$A = \left(\frac{6L\gamma_{\max}^2}{\mu^2\gamma_{\min}^2}\sigma_{\widetilde{\mathcal{P}}}(-\nabla f(x^*+y^*) + 6\gamma_{\max}\nu_0 + 3\kappa_{\mathcal{P}}(x^{(0)})\right)^2, \qquad B = \frac{3L^2\gamma_{\max}^2}{\mu^2\gamma_{\min}^2}.$$

Then putting together lemmas 37, 36 and using the relation $(a+b)^2 \le 2a^2 + 2b^2$ gives

$$\Delta^{(t+1)} - \Delta^{(t)} \le -\theta^{(t)}\mathbf{res}_{\mathcal{P}}(x^{(t)}) + (\theta^{(t)})^2\left(B\Delta^{(t)} + B\overline{\Delta}^{(t-1)} + A\right).$$

where $\overline{\Delta}^{(t)}$ is defined as an averaging over square roots, e.g.

$$\sqrt{\overline{\Delta}^{(t)}} = \frac{2}{t(t+1)}\sum_{u=1}^{t} u\sqrt{\Delta^{(u)}}.$$

Then picking $\bar{t} > 6B$, we get that for all $t \ge \bar{t}$, $B(\theta^{(t)})^2 \le \theta^{(t)}/3$, and therefore

$$\begin{aligned} \Delta^{(t+1)} - \Delta^{(t)} &\le -\theta^{(t)}\underbrace{\mathbf{res}_{\mathcal{P}}(x^{(t)})}_{\ge\Delta^{(t)}} + (\theta^{(t)})^2\left(B\Delta^{(t)} + B\overline{\Delta}^{(t-1)} + A\right) \\ &\le -\theta^{(t)}\Delta^{(t)} + (\theta^{(t)})^2\left(B\Delta^{(t)} + B\overline{\Delta}^{(t-1)} + A\right) \\ &\le -\frac{2\theta^{(t)}\Delta^{(t)}}{3} + (\theta^{(t)})^2\left(B\overline{\Delta}^{(t-1)} + A\right). \end{aligned}$$

We now pick G large enough such that for all $t \leq \bar{t}$, $\Delta^{(t)} \leq G/t$, and $G > 24A$. Since $\Delta^{(t)}$ is always a bounded quantity ($x^{(t)}$ is always feasible), this is always possible. Then, for all $t < \bar{t}$,

$$\sqrt{\overline{\Delta}^{(t)}} \leq \frac{\sqrt{G}}{t(t+1)} \sum_{t'=1}^{t} \sqrt{t'} \overset{(a)}{\leq} \frac{2\sqrt{G}}{3t(t+1)} t^{3/2},$$

where (a) is by integral rule, and so

$$\overline{\Delta}^{(t)} \leq \frac{4Gt}{9(t+1)^2} \leq \frac{G}{2t}.$$

Now we make an inductive step. Suppose that for some t, $\Delta^{(t')} < G/t'$ for all $t' \leq t$. Then

$$\begin{aligned}
\Delta^{(t+1)} &\leq \Delta^{(t)} - \frac{2}{3}\theta^{(t)}\Delta^{(t)} + (\theta^{(t)})^2(A + B\overline{\Delta}^{(t)}) \\
&\leq \frac{G}{t} - \frac{2}{3}\frac{2G}{t+1}\frac{1}{t} + \frac{4}{(t+1)^2}\left(A + \frac{GB}{2t}\right) \\
&= \frac{G}{t+1}\left(\frac{t+1}{t} - \frac{4}{3t} + \frac{4A}{(t+1)G} + \frac{2B}{t(t+1)}\right) \\
&\leq \frac{G}{t+1}\left(1 - \frac{1}{3t} + \frac{4A}{tG} + \frac{2B}{t^2}\right) \\
&= \frac{G}{t+1}\left(1 + \frac{1}{t}\left(-\frac{1}{3} + \underbrace{\frac{4A}{G}}_{<1/6} + \underbrace{\frac{2B}{t}}_{<1/6}\right)\right) \leq \frac{G}{t+1},
\end{aligned}$$

which satisfies the inductive step. ◀

The following is a generalized and modified version of a proof segment from [39], which will be used for proving $O(1/t)$ gap convergence.

▶ **Lemma 39.** *Pick some $0 < T_2 < T_1$ and pick*

$$\bar{k} = \lceil D(k+D)/(D+T_1) \rceil - D \;\Rightarrow\; \frac{D}{D+T_1} \leq \frac{\bar{k}+D}{k+D} \leq \frac{D}{D+T_2}.$$

Then if

$$\frac{C_1(D+T_1)}{D} \leq C_3 \cdot \log\left(\frac{D+T_2}{D}\right),$$

then for all $k > T_1$,

$$\left(\frac{C_1}{D+\bar{k}} + \sum_{i=\bar{k}}^{k} \frac{C_2}{(D+i)^2} - \frac{C_3}{D+i} \cdot \frac{1}{D+k}\right) < 0.$$

Proof. Using integral rule, we see that

$$\sum_{i=\bar{k}}^{k} \frac{1}{(D+i)^2} \leq \int_{z=\bar{k}-1}^{k-1} \frac{1}{(D+i)^2} = \frac{1}{D-1+k} - \frac{1}{D-1+\bar{k}}$$

$$\sum_{i=\bar{k}}^{k} \frac{1}{D+i} \geq \int_{z=\bar{k}}^{k} \frac{1}{D+i} = \log(D+k) - \log(D+\bar{k}).$$

This yields

$$\begin{aligned}
c(k) &:= \frac{C_1}{D+\bar{k}} + \sum_{i=\bar{k}}^{k} \frac{C_2}{(D+i)^2} - \frac{1}{D+i}\cdot\frac{C_3}{D+k} \\
&\leq \frac{C_1}{D+\bar{k}} + \frac{C_2}{D-1+k} - \frac{C_2}{D-1+\bar{k}} + \frac{C_3}{D+k}\cdot(\log(D+\bar{k}) - \log(D+k)) \\
&\leq \frac{C_1(D+T_1)}{D(D+k)} + \underbrace{\frac{C_2}{D-1+k} - \frac{C_2}{D-1+\bar{k}}}_{<0} + \frac{C_3}{D+k}\cdot\log\left(\frac{D}{D+T_2}\right) \\
&\leq \frac{C_1(D+T_1)}{D(D+k)} + \frac{C_3}{D+k}\cdot\log\left(\frac{D}{D+T_2}\right) < 0.
\end{aligned}$$

◀

▶ **Lemma 40** (Generalized non-monotonic gap bound)**.** *Given*

- $\Delta^{(t)} \leq \frac{G_1}{t+D}$ *for some* G_1,
- $\theta^{(t)} = \frac{G_2}{t+D}$ *for some* G_2 *and* D, *and*
- $\Delta^{(t+1)} - \Delta^{(t)}(1+\alpha\theta^{(t)}) \leq -\theta^{(t)}\mathbf{res}(x^{(t)}) + (\theta^{(t)})^2 G_3$ *for some* G_3,

then for

$$G_4 \geq \frac{G_1}{G_2}\frac{(D+2)}{D(\log\left(\frac{D+1}{D}\right))},$$

we have

$$\min_{i\leq t}\mathbf{res}(x^{(i)}) \leq \frac{G_4}{t+D}.$$

Proof. We have

$$\Delta^{(t+1)} - \Delta^{(t)} \leq \alpha\theta^{(t)}\Delta^{(t)} - \theta^{(t)}\mathbf{gap}^{(t)} + G_3(\theta^{(t)})^2.$$

Now assume that for all $i \leq t$, $\mathbf{gap}^{(i)} > \frac{G_4}{t+D}$. Then, telescoping from $\bar{t}$ to t gives

$$\begin{aligned}
\Delta^{(t+1)} &\leq \Delta^{(\bar{t})} + \sum_{i=\bar{t}}^{t}\left(\alpha\theta^{(i)}\Delta^{(i)} - \theta^{(i)}\mathbf{gap}^{(i)} + G_3(\theta^{(i)})^2\right) \\
&< \frac{G_1}{\bar{t}+D} + \sum_{i=\bar{t}}^{t}\left(\alpha\frac{G_1G_2}{(i+D)^2} - \frac{G_2}{i+D}\frac{G_4}{t+D} + \frac{G_3G_2^2}{(i+D)^2}\right).
\end{aligned}$$

Picking $C_1 = G_1$, $C_2 = \alpha G_1G_2 + G_3G_2^2$, $C_3 = G_2G_4$, and invoking Lemma 39, this yields that $\Delta^{(t+1)} < 0$, which is impossible. Therefore, the assumption must not be true. ◀

Piecing everything in this section together gives Theorem 33 (main convergence theorem.)

C Screening proofs from Section 4

Proof of Proposition 34

Proof. First, note that

$$\begin{aligned}
\phi^*(\sigma_{\mathcal{P}(x)}(z)) + r_0\cdot(\sigma_{\mathcal{P}(x)}(z)) &= \sup_y\ y^Tz - \phi(r_0 + \kappa_{\mathcal{P}(x)}(y)) \\
&\geq z^Tx^* - \phi(r_0 + \kappa_{\mathcal{P}(x)}(x^*)). \qquad (46)
\end{aligned}$$

Define $\mathbf{res}(x) = (\overline{F}(x;x) - \overline{F}_D(-\nabla f(x);x)$. Taking $(x, -\nabla f(x))$ as a feasible primal-dual pair and reference point $\bar{x} = x$, and denoting $\epsilon(x) = \phi(r_0 + \kappa_{\mathcal{P}(x)}(x^*)) - \phi(x^*)$, $z = -\nabla f(x+y(x))$, and $z^* = -\nabla f(x^* + y(x^*))$,

then

$$\begin{aligned}
\mathbf{res}(x) = & \underbrace{f(x) + f^*(z)}_{\text{use Fenchel–Young}} + \phi(r(x)) \\
& + \underbrace{\phi^*(\sigma_{\mathcal{P}(x)}(z) - r_0 \cdot (\sigma_{\mathcal{P}(x)}(z))}_{\text{use (46)}} \\
& \geq -z^T(x - x^*) + \phi(r_{\mathcal{P}}(x)) - \phi\left(r_0 + \kappa_{\mathcal{P}(x)}(x^*)\right) \\
& \overset{+\epsilon(x)-\epsilon(x)}{\geq} -z^T(x - x^*) + \underbrace{\phi(r_{\mathcal{P}}(x)) - \phi(r_{\mathcal{P}}(x^*))}_{\text{convex in } x} - \epsilon(x) \\
& \overset{g \in \partial h(x^*)}{\geq} -z^T(x - x^*) + g^T(x - x^*) - \epsilon(x).
\end{aligned}$$

Picking in particular $g = -\nabla f(x^* + y(x^*))$,

$$\mathbf{res}(x) + \epsilon(x) \geq (x - x^*)^T(z^* - z) \overset{(\star)}{\geq} \frac{1}{L}\sigma_{\widetilde{\mathcal{P}}}(z - z^*)^2$$

where $(\star)$ follows from Assumption 5.

Next, note that

$$\begin{aligned}
\epsilon(x) &= \phi(r_{\mathcal{P}}(x) - r_{\mathcal{P}}(x; x) + r_{\mathcal{P}}(x^*; x)) - \phi(r_{\mathcal{P}}(x^*)) \\
&\overset{\text{convex } \phi}{\leq} g_\phi \underbrace{(r_{\mathcal{P}}(x) - r_{\mathcal{P}}(x; x) + r_{\mathcal{P}}(x^*; x) - r_{\mathcal{P}}(x^*))}_{=: D(x)}
\end{aligned}$$

for all $g_\phi \in \partial\phi(r_{\mathcal{P}}(x^*))$, where in general, $D(x) \leq (\gamma_{\max} - \gamma_{\min})\kappa_{\widetilde{\mathcal{P}}}(x - x^*)$ and $D(x) = 0$ if $\gamma(\xi) = \xi$ (convex case). Noting that, at optimality,

$$\partial\phi(r_{\mathcal{P}}(x^*)) \ni \sigma_{\mathcal{P}(x^*)}(z^*) \leq \frac{\sigma_{\widetilde{\mathcal{P}}}(z^*)}{\gamma_{\min}},$$

then

$$\gamma_{\min}\phi'(r(x^*)) \leq \sigma_{\widetilde{\mathcal{P}}}(z^*) \leq \sigma_{\widetilde{\mathcal{P}}}(z) + \sigma_{\widetilde{\mathcal{P}}}(z - z^*)$$

and overall,

$$\begin{aligned}
\sigma_{\widetilde{\mathcal{P}}}(z^* - z)^2 &\leq L\mathbf{res}(x) + L\epsilon(x) \\
&\leq L\mathbf{res}(x) + LD(x)\frac{\sigma_{\widetilde{\mathcal{P}}}(z) + \sigma_{\widetilde{\mathcal{P}}}(z^* - z)}{\gamma_{\min}}.
\end{aligned}$$

This inequality is quadratic in $\sigma_{\widetilde{\mathcal{P}}}(z^* - z)$, which leads to the bound

$$\sigma_{\widetilde{\mathcal{P}}}(z^* - z) \leq \frac{LD(x)}{2\gamma_{\min}} + \sqrt{\frac{L^2D(x)^2}{4\gamma_{\min}^2} + L\mathbf{res}(x) + LD(x)\frac{\sigma_{\widetilde{\mathcal{P}}}(z)}{\gamma_{\min}}}.$$ ◀

References

1 Francis Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, pages 118–126, 2010.

2 Francis Bach. Duality between subgradient and conditional gradient methods. *SIAM J. Optim.*, 25(1):115–129, 2015.

3 Heinz H. Bauschke, Jérôme Bolte, and Marc Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res.*, 42(2):330–348, 2017.

4 Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, volume 408. Springer, 2 edition, 2011.

5 Leonard Berrada, Andrew Zisserman, and M. Pawan Kumar. Deep Frank-Wolfe for neural network optimization. `https://arxiv.org/abs/1811.07591`, 2018.

6 Howard D. Bondell and Brian J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123, 2008.

7 Antoine Bonnefoy, Emiya Valentin, Ralaivola Liva, and Remi Gribonval. Dynamic screening: Accelerating first-order algorithms for the LASSO and group-LASSO. *IEEE Trans. Signal Process.*, 63(19):5121–5132, 2015.

8 Jonathan Borwein and Adrian S. Lewis. *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer, 2010.

9 Kristian Bredies and Dirk A. Lorenz. Iterated hard shrinkage for minimization problems with sparsity constraints. *SIAM J. Sci. Comput.*, 30(2):657–683, 2008.

10 Kristian Bredies, Dirk A. Lorenz, and Peter Maass. A generalized conditional gradient method and its connection to an iterative shrinkage method. *Comput. Optim. Appl.*, 42(2):173–193, 2009.

11 James V. Burke and Jorge J. Moré. On the identification of active constraints. *SIAM J. Numer. Anal.*, 25(5):1197–1211, 1988.

12 Jim Burke. On the identification of active constraints II: The nonconvex case. *SIAM J. Numer. Anal.*, 27(4):1081–1102, 1990.

13 Emmanuel Candès and Justin Romberg. Robust signal recovery from incomplete observations. In *2006 International Conference on Image Processing*, pages 1281–1284. IEEE, 2006.

14 Emmanuel Candès and Terence Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, 2005.

15 Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Found. Comput. Math.*, 12(6):805–849, 2012.

16 Visesh Chari, Simon Lacoste-Julien, Ivan Laptev, and Josef Sivic. On pairwise costs for network flow multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5537–5545, 2015.

17 Xiaojun Chen and Weijun Zhou. *Convergence of reweighted ℓ_1 minimization algorithms and unique solution of truncated ℓ_p minimization*. Department of Applied Mathematics, The Hong Kong Polytechnic University, 2010.

18 Frank H. Clarke. Generalized gradients and applications. *Trans. Am. Math. Soc.*, 205:247–262, 1975.

19 Frank H. Clarke. Nonsmooth analysis and optimization. In *Proceedings of the international congress of mathematicians*, volume 5, pages 847–853. Citeseer, 1983.

20 Kenneth L. Clarkson. Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. *ACM Trans. Algorithms*, 6(4):63, 2010.

21 Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. *Commun. Pure Appl. Math.*, 63(1):1–38, 2010.

22 David L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

23 Miroslav Dudik, Zaid Harchaoui, and Jérôme Malick. Lifted coordinate descent for learning with trace-norm regularization. In *Artificial Intelligence and Statistics*, pages 327–336, 2012.

24 Joseph C. Dunn and S. Harshbarger. Conditional gradient algorithms with open loop step size rules. *J. Math. Anal. Appl.*, 62(2):432–444, 1978.

25 Laurent El Ghaoui, Vivian Viallon, and Tarek Rabbani. Safe feature elimination for the Lasso and sparse supervised learning problems. *Pac. J. Optim.*, 8(4):667–698, 2012.

26 Alina Ene and Adrian Vladu. Improved Convergence for ℓ_1 and ℓ_∞ Regression via Iteratively Reweighted Least Squares. In *International Conference on Machine Learning*, pages 1794–1801, 2019.

27 Zhenan Fan, Halyun Jeong, Yifan Sun, Michael Friedlander, et al. Atomic Decomposition via Polar Alignment: The Geometry of Structured Optimization. *Found. Trends Optim.*, 3(4):280–366, 2020.

28 Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Mind the duality gap: safer rules for the lasso. In *International Conference on Machine Learning*, pages 333–342. PMLR, 2015.

29 Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Nav. Res. Logist. Q.*, 3(1-2):95–110, 1956.

30 Robert M. Freund. Dual gauge programs, with applications to quadratic programming and the minimum-norm problem. *Math. Program.*, 38(1):47–67, 1987.

31 Robert M. Freund, Paul Grigas, and Rahul Mazumder. An Extended Frank–Wolfe Method with "In-Face" Directions, and Its Application to Low-Rank Matrix Completion. *SIAM J. Optim.*, 27(1):319–346, 2017.

32 Michael Friedlander, Ives Macedo, and Ting Kei Pong. Gauge optimization and duality. *SIAM J. Optim.*, 24(4):1999–2022, 2014.

33 Pinghua Gong, Changshui Zhang, Zhaosong Lu, Jianhua Huang, and Jieping Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *International Conference on Machine Learning*, pages 37–45, 2013.

34 Jacques Guélat and Patrice Marcotte. Some comments on Wolfe's 'away step'. *Math. Program.*, 35(1):110–119, 1986.

35 Isabelle Guyon, Steve Gunn, Asa Ben-Hur, and Gideon Dror. Result analysis of the NIPS 2003 feature selection challenge. *Adv. Neural Inf. Process. Syst.*, 17, 2004.

36 Zaid Harchaoui, Anatoli Juditsky, and Arkadi Nemirovski. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Math. Program.*, 152(1-2):75–112, 2015.

37 Warren Hare. Identifying active manifolds in regularization problems. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 261–271. Springer, 2011.

38 Elad Hazan. Sparse approximate solutions to semidefinite programs. In *Latin American Symposium on Theoretical Informatics*, pages 306–316. Springer, 2008.

39 Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, pages 427–435, 2013.

40 Tyler B. Johnson and Carlos Guestrin. Stingy CD: safely avoiding wasteful updates in coordinate descent. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1752–1760, 2017.

41 Rahul G. Krishnan, Simon Lacoste-Julien, and David Sontag. Barrier Frank-Wolfe for marginal inference. In *Advances in Neural Information Processing Systems*, pages 532–540, 2015.

42 Simon Lacoste-Julien and Martin Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In *Advances in Neural Information Processing Systems*, pages 496–504. 2015.

43 Simon Lacoste-Julien, Martin Jaggi, Mark Schmidt, and Patrick Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *International Conference on Machine Learning*, pages 53–61. PMLR, 2013.

44 Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential kernel herding: Frank-Wolfe optimization for particle filtering. In *Artificial Intelligence and Statistics*, pages 544–552. PMLR, 2015.

45 Adrian S. Lewis and Stephen J. Wright. Identifying activity. *SIAM J. Optim.*, 21(2):597–614, 2011.

46 Jun Liu, Zheng Zhao, Jie Wang, and Jieping Ye. Safe screening with variational inequalities and its application to lasso. In *International Conference on Machine Learning*, pages 289–297. PMLR, 2014.

47 Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim.*, 28(1):333–354, 2018.

48 Julien Mairal, Francis Bach, and Jean Ponce. Sparse modeling for image and vision processing. `https://arxiv.org/abs/1411.3230`, 2014.

49 Abed Malti and Cédric Herzet. Safe screening tests for Lasso based on firmly non-expansiveness. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4732–4736. IEEE, 2016.

50 Cun Mu, Yuqian Zhang, John Wright, and Donald Goldfarb. Scalable robust matrix recovery: Frank–Wolfe meets proximal methods. *SIAM J. Sci. Comput.*, 38(5):A3291–A3317, 2016.

51 Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparse multi-task and multi-class models. In *Advances in Neural Information Processing Systems*, pages 811–819, 2015.

52 Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, 2013.

53 Julie Nutini, Mark Schmidt, and Warren Hare. "Active-set complexity" of proximal gradient: How long does it take to find the sparsity pattern? *Optim. Lett.*, 13(4):645–655, 2019.

54 Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke. Coordinate descent converges faster with the Gauss-Southwell rule than random selection. In *International Conference on Machine Learning*, pages 1632–1641, 2015.

55 Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. `https://arxiv.org/abs/1110.0413`, 2011.

56 Peter Ochs, Alexey Dosovitskiy, Thomas Brox, and Thomas Pock. On iteratively reweighted algorithms for nonsmooth nonconvex optimization in computer vision. *SIAM J. Imaging Sci.*, 8(1):331–372, 2015.

57 Wei Ping, Qiang Liu, and Alexander T. Ihler. Learning infinite RBMs with Frank-Wolfe. In *Advances in Neural Information Processing Systems*, pages 3063–3071, 2016.

58 Alain Rakotomamonjy, Gilles Gasso, and Joseph Salmon. Screening rules for lasso with non-convex sparse regularizers. In *International Conference on Machine Learning*, pages 5341–5350. PMLR, 2019.

59 Nikhil Rao, Parikshit Shah, and Stephen J. Wright. Forward-backward greedy algorithms for atomic norm regularization. *IEEE Trans. Signal Process.*, 63(21):5798–5811, 2015.

60 R. Tyrrell Rockafellar. *Convex Analysis*, volume 28. Princeton University Press, 1970.

61 Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, pages 527–538, 2018.

62 Ambuj Tewari, Pradeep K. Ravikumar, and Inderjit S. Dhillon. Greedy algorithms for structurally constrained high dimensional problems. In *Advances in Neural Information Processing Systems*, pages 882–890, 2011.

63 Robert Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc., Ser. B, Stat. Methodol.*, 58(1):267–288, 1996.

64 Marina Vinyes and Guillaume Obozinski. Fast column generation for atomic norm regularization. In *Proceedings of International Conference on Artificial Intelligence and Statistics*, 2017.

65 Balder Von Hohenbalken. Simplicial decomposition in nonlinear programming algorithms. *Math. Program.*, 13(1):49–68, 1977.

66 Jie Wang, Jiayu Zhou, Peter Wonka, and Jieping Ye. LASSO screening rules via dual polytope projection. *Adv. Neural Inf. Process. Syst.*, pages 1070–1078, 2013.

67 R. Wolke and H. Schwetlick. Iteratively reweighted least squares: algorithms, convergence analysis, and numerical comparisons. *SIAM J. Sci. Stat. Comput.*, 9(5):907–921, 1988.

68 Stephen J. Wright, Robert D. Nowak, and Mário A. T. Figueiredo. Sparse reconstruction by separable approximation. *IEEE Trans. Signal Process.*, 57(7):2479–2493, 2009.

69 Yaoliang Yu, Xinhua Zhang, and Dale Schuurmans. Generalized conditional gradient for sparse estimation. *J. Mach. Learn. Theory*, 18(1):5279–5324, 2017.

70 Alp Yurtsever, Madeleine Udell, Joel Tropp, and Volkan Cevher. Sketchy decisions: Convex low-rank matrix optimization with optimal storage. In *Artificial intelligence and statistics*, pages 1188–1196. PMLR, 2017.

71 Xiangrong Zeng and Mário A. T. Figueiredo. The Ordered Weighted ℓ_1 Norm: Atomic Formulation, Projections, and Algorithms. `https://arxiv.org/abs/1409.4271`, 2014.

72 Song Zhou, Swati Gupta, and Madeleine Udell. Limited memory Kelley's method converges for composite convex and submodular objectives. In *Advances in Neural Information Processing Systems*, pages 4414–4424. 2018.