Optimal Convergence Rates for the Proximal Bundle Method

Mateo Díaz[*] Benjamin Grimmer[†]

Abstract

We study convergence rates of the classic proximal bundle method for a variety of nonsmooth convex optimization problems. We show that, without any modification, this algorithm adapts to converge faster in the presence of smoothness or a Hölder growth condition. Our analysis reveals that with a constant stepsize, the bundle method is adaptive, yet it exhibits suboptimal convergence rates. We overcome this shortcoming by proposing nonconstant stepsize schemes with optimal rates. These schemes use function information such as growth constants, which might be prohibitive in practice. We complete the paper with a new parallelizable variant of the bundle method that attains near-optimal rates without prior knowledge of function parameters. These results improve on the limited existing convergence rates and provide a unified analysis approach across problem settings and algorithmic details. Numerical experiments support our findings and illustrate the effectiveness of the parallel bundle method.

1 Introduction

Convex optimization has played a fundamental role in recent developments in high-dimensional statistics, signal processing, and data science. Large-scale applications have motivated researchers to develop first-order methods with computationally simple iterations. Although impressive in scope, these methods often require delicate parameter tunning involving geometrical information about the objective function. Thus, imposing an obstacle for practitioners that rarely have access to such information.

In this work, we develop efficiency guarantees for *proximal bundle methods*, which date back to the 70s, that solve unconstrained convex minimization problems

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\, f(x) \tag{1.1}$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a proper closed convex. Our core finding is that classic bundle methods, without any modification, are adaptive, which means that they speed up in the presence of smoothness or error bounds, with little to no tunning.

Proximal bundle methods were independently proposed in [22] and [38]. They are conceptually similar to model-based methods [4, 9, 31]. That is, methods that update their iterates by applying a proximal step to an approximation of the function, known as the model f_k:

$$x_{k+1} \leftarrow \underset{x}{\operatorname{argmin}}\, f_k(x) + \frac{\rho_k}{2}\|x - x_k\|^2.$$

[*] Center for Applied Mathematics, Cornell University Ithaca, NY 14850, USA; people.cam.cornell.edu/md825/

[†] ORIE, Cornell University Ithaca, NY 14850, USA; people.orie.cornell.edu/bdg79/

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650441.

Unlike these schemes, bundle methods only update their iterates x_k when the decrease in objective value is at least a fraction of the decrease that the model predicted. Moreover, bundle methods incorporate information from past iterations into their models, allowing f_k to capture more than the just objective's geometry near x_k.

This seemingly subtle change has a rather surprising consequence: the iterates generated by a bundle method, with *any* constant parameter configuration, converge to a minimizer of f; see [18, Thm. 4.9], [15, Thm. XV.3.2.4], or [34, Thm. 7.16] for different variations of this result. This stands in harsh contrast to other first-order algorithms; for example, gradient descent and its accelerated variants rely on selecting a stepsize inversely proportional to the level of smoothness. Similarly, subgradient methods rely on carefully controlled decreasing stepsize sequences. These simpler algorithms may fail to converge when the stepsizes are not carefully managed. Thus, providing a compelling reason to consider bundle methods.

Although bundle methods are known to converge under a number of assumptions [2,5,12,17,27–30] and have been successfully used in applications [6, 35, 36], nonasympotic guarantees have remained mostly evasive. The purpose of this paper is to close this gap. We study convergence rates for finding an ε-minimizer, e.g., $f(x) - \inf f \leq \varepsilon$, under a variety of different assumptions on f. We consider settings where the objective function is either M-Lipschitz continuous

$$|f(x) - f(y)| \leq M\|x - y\| \qquad \text{for all } x, y \in \mathbb{R}^d \tag{1.2}$$

or differentiable with an L-Lipschitz gradient, often referred to as L-smoothness,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \qquad \text{for all } x, y \in \mathbb{R}^d \ . \tag{1.3}$$

In either setting, we investigate the method's rate of convergence with and without the presence of Hölder growth

$$f(x) - \inf f \geq \mu \cdot \operatorname{dist}(x, X^*)^p \qquad \text{for all } x \in \mathbb{R}^d \ , \tag{1.4}$$

where $X^* = \{x \mid f(x) = \inf f\}$ is the set of minimizers.[1] Particularly important cases are when $p = 1$ and $p = 2$, which correspond to sharp growth (μ-SG) [3] and quadratic growth (μ-QG), generalizing strong convexity, respectively.

1.1 Contributions

Our **first contribution** is to establish convergence rates under every realizable combination of continuity/smoothness (1.2) or (1.3) and growth assumptions (1.4), see Table 1. Full theorem statements are given in Section 2 and apply for any Hölder growth exponent (rather than just the cases of $p = 1$ and $p = 2$ shown in the table). Our analysis technique is fairly general as we apply it seamlessly to every combination of assumptions as well as different stepsize rules. We show rates for any constant stepsize $\rho_k = \rho$, which tend to be suboptimal. Yet, they improve under amenable geometry. Tuning the constant ρ to depend on a target accuracy ϵ yields faster convergence rates. Further, we propose nonconstant stepsize rules ρ_k with two clear advantages: they yield yet faster convergence and their convergence does not slow down after reaching the target accuracy.

The existing convergence theory for the proximal bundle method applies to settings comparable to the first two rows of our table. Kiwiel [20] derived a $O(\epsilon^{-3})$ convergence rate for Lipschitz problems, which agrees with our theory. Du and Ruszczynski [10] and subsequently Liang and Monteiro [25] showed a $O(\log(1/\epsilon)/\epsilon)$ convergence rate for Lipschitz, strongly convex problems, which we improve on by removing the extra logarithmic term and thus achieve the optimal convergence rate for

[1] Here $\operatorname{dist}(x, S) = \inf_{y \in S} \|x - y\|$.

Assumptions		Rate for generic ρ	Rate for tuned ρ	Rate for adaptive ρ_k
M-Lipschitz	No Growth	$O\left(\frac{M^2\|x_0-x^*\|^4}{\rho\epsilon^3}\right)$	$O\left(\frac{M^2\|x_0-x^*\|^2}{\epsilon^2}\right)$	$O\left(\frac{M^2\|x_0-x^*\|^2}{\epsilon^2}\right)$
	μ-QG	$O\left(\frac{M^2}{\min\{\mu,\rho\}\epsilon}\right)$	$O\left(\frac{M^2}{\mu\epsilon}\right)$	$O\left(\frac{M^2}{\mu\epsilon}\right)$
	μ-SG	$O\left(\frac{M^2}{\rho\epsilon}\right)$	$O\left(\frac{M^2}{\mu^2}\sqrt{\frac{\Delta_f}{\epsilon}}\right)$	$O\left(\frac{M^2}{\mu^2}\log\left(\frac{\Delta_f}{\epsilon}\right)\right)$
L-Smooth	No Growth	$O\left(\frac{L^3\|x_0-x^*\|^2}{\rho^2\epsilon}\right)$	$O\left(\frac{L\|x_0-x^*\|^2}{\epsilon}\right)$	$O\left(\frac{L\|x_0-x^*\|^2}{\epsilon}\right)$
	μ-QG	$O\left(\frac{L^3}{\rho^2\mu}\log\left(\frac{\Delta_f}{\epsilon}\right)\right)$	$O\left(\frac{L}{\mu}\log\left(\frac{\Delta_f}{\epsilon}\right)\right)$	$O\left(\frac{L}{\mu}\log\left(\frac{\Delta_f}{\epsilon}\right)\right)$

Table 1: Convergence rates. We denote $\Delta_f := f(x_0) - \inf f$. The first column applies for any choice of the parameter ρ, showing progressively faster convergence as more structure is introduced. The second column shows the rate after optimizing the choice of ρ. The third column further improves these by allowing nonconstant stepsizes ρ_k.

this setting of $O(1/\epsilon)$. To our knowledge, the rest of our convergence results apply to wholly new settings for the proximal bundle method. In all of the M-Lipschitz settings considered, we show that using a nonconstant stepsize the bundle method attains the optimal nonsmooth convergence rate. In the L-smooth settings considered, the bundle method converges at the same rate as gradient descent. Although, unlike gradient descent, our convergence theory applies to any configuration of its algorithmic parameters.

Our **second contribution** is proposing a parallelizable variant of the bundle method that avoids the reliance on tuning a stepsize or sequence of stepsizes based on potentially unrealistic knowledge of underlying problem constants. This approach too seamlessly falls under the umbrella of our analysis. It attains the optimal nonsmooth convergence rates for Lipschitz problems with any level of Hölder growth, up to the cost of running a logarithmic number of instances of the bundle method in parallel.

1.2 Related work

In 2000, Kiwiel [20] gave the first convergence rate for the proximal bundle method, showing that an ϵ-minimizer x_k is found with $k \leq O\left(\frac{\|x_0-x^*\|^4}{\epsilon^3}\right)$. More recently, Du and Ruszczyński [10] gave the first analysis of bundle methods when applied to problems satisfying a quadratic growth bound. In this case, an ϵ-minimizer is found within $O(\log(1/\epsilon)/\epsilon)$ iterations. Following this, Liang and Monteiro [25] showed a variant of the proximal bundle method with proper stepsize selection attains the optimal convergence rate for convex and strongly convex optimization, up to logarithmic terms.

Despite historically having weaker convergence rate guarantees than simple alternatives like the subgradient method, bundle methods have persisted as a method of choice for nonsmooth convex optimization. See [11, 23] as a survey of much of the bundle method literature. In practice, bundle methods have proven to be efficient methods for solving many nonsmooth problems (see [6, 35, 36] for further discussion). Extensions that apply to nonconvex problems have been considered in in [2, 5, 12, 17, 27–30] and as well as an extension to problems where only an inexact first-order oracle is available in [7, 13, 26].

Stronger convergence rates have been established for related level bundle methods [24], which share many core elements with proximal bundle methods. Variations of level bundle methods were

```
Algorithm 1: Proximal Bundle Method
  Data: $z_0 = x_0 \in \mathbb{R}^n$, $f_0(z) = f(x_0) + \langle g_0, z - x_0 \rangle$
  Step $k$: $(k \geq 0)$
      Compute candidate iterate $z_{k+1} \leftarrow \operatorname*{argmin}_{z \in \mathcal{X}} f_k(z) + \frac{\rho_k}{2}\|z - x_k\|^2$.
      If $\beta(f(x_k) - f_k(z_{k+1})) \leq f(x_k) - f(z_{k+1})$                    (Descent step)
          set $x_{k+1} \leftarrow z_{k+1}$,
      Else                                                                            (Null step)
          set $x_{k+1} \leftarrow x_k$.
      Update $f_{k+1}$ and $\rho_{k+1}$ without violating Assumption A.
```

studied in [19] and [21]. The results of Lan [21] are particularly impressive as their proposed method has optimal convergence rates for both smooth and nonsmooth problems while requiring little input.

Outline Section 2 introduces the Proximal Bundle Method and provides the formal convergence guarantees under different regularity assumptions. This section also introduces simple stepsize rules that guarantee optimal convergence rates for all nonsmooth settings. Practical implementations of these rules require access growth constants of the function. To bypass this issue, in Section 3 we propose an adaptive parallel bundle method that exhibits nearly the same convergence rates without knowledge of such constants. We complement our findings with numerical experiments in Section 4. Finally, Section 5 presents a broadly applicable proof technique to analyze bundle methods and uses it to establish the theoretical results.

2 Bundle Methods

In this section, we formally define the family of proximal bundle methods that our theory applies to. We present the convergence rates for the classic method with constant stepsizes. Additionally, we introduce and analyze nonconstant stepsize rules that guarantee faster convergence rates.

Proximal bundle methods work by maintaining a model function $f_k \colon \mathbb{R}^n \to \mathbb{R}$ at each iteration k and a current iterate x_k. The method computes a candidate for the next iterate as

$$z_{k+1} = \operatorname*{argmin}_{z \in \mathcal{X}} f_k(z) + \frac{\rho_k}{2}\|z - x_k\|^2.$$

However, unlike other model-based algorithms, bundle methods do not necessarily move their next iterate to z_{k+1}. Instead, it first checks whether the candidate z_{k+1} has at least $\beta \in (0, 1)$ fraction of the decrease in objective value that our model $f_k(\cdot)$ predicts. If it does, it updates $x_{k+1} = z_{k+1}$ as the next iterate, this is called a *Descent Step*. Otherwise the method keeps the iterate the same $x_{k+1} = x_k$ and updates the model function f_{k+1}, called a *Null Step*.

The proximal bundle method is stated fully in Algorithm 1. Our analysis does not presume a particular parametrization or form of the models. We only assume that the models satisfy mild assumptions, typical of bundle methods in the literature. To state the assumptions, note the first-order optimality conditions define a subgradient

$$s_{k+1} = \rho_k(z_{k+1} - x_k) \in \partial f_k(z_{k+1}) \quad \text{for each } k \geq 0$$

where $\partial f(x) = \{g \mid f(x') \geq f(x) + \langle g, x' - x \rangle \;\; \forall x' \in \mathbb{R}^d\}$ denotes the subdifferential of f at x.

Assumption A. *Let* $\left\{f_k : \mathbb{R}^d \to \mathbb{R}\right\}$ *and* $\{\rho_k\}$ *be the sequence of models and stepsizes used throughout the execution of a bundle method. Assume that for any iteration* $k \geq 0$*, the next model* f_{k+1} *and stepsize* ρ_{k+1} *satisfy the following:*

1. ***Minorant.***
$$f_{k+1}(x) \leq f(x) \qquad \text{for all } x \in \mathbb{R}^d \ . \tag{2.1}$$

2. ***Subgradient lowerbound.*** *There is a subgradient* $g_{k+1} \in \partial f(z_{k+1})$ *such that*
$$f_{k+1}(x) \geq f(z_{k+1}) + \langle g_{k+1}, x - z_{k+1}\rangle \qquad \text{for all } x \in \mathbb{R}^d \ . \tag{2.2}$$

3. ***Model subgradient lowerbound.*** *After a null step* k
$$f_{k+1}(x) \geq f_k(z_{k+1}) + \langle s_{k+1}, x - z_{k+1}\rangle \qquad \text{for all } x \in \mathbb{R}^d \ . \tag{2.3}$$

4. ***Constant stepsize between null steps.*** *After a null step* k
$$\rho_{k+1} = \rho_k \ . \tag{2.4}$$

The first two conditions are natural as they ensure that a new model incorporates first-order information from the objective at z_{k+1}. The third condition is mild and, intuitively, requires the new model to retain some of the approximation accuracy of the previous model. The last assumption is trivial to enforce and guarantees the algorithm only changes its stepsize after it decides to move.

2.1 Bundle Method Model Function Choices

Several methods for constructing model functions f_k that satisfy (2.1)-(2.3) have been considered. In practice, the main consideration lies in weighing the potentially greater per iteration gains from having more complex models against the lower iteration costs from having simpler models.

Full-Memory Proximal Bundle Method. The earliest proposed bundle methods [22, 38] rely on using all of the past subgradient evaluations to construct the models as

$$f_{k+1}(x) = \max_{j=0\ldots k+1} \left\{f(z_j) + \langle g_j, x - z_j\rangle\right\} . \tag{2.5}$$

In this case, solving the quadratically regularized subproblem at each iteration amounts to solving a quadratic programming problem.

Finite Memory Proximal Bundle Method. Alternatively using cut-aggregation [16, 18], the collection of $k+1$ lower bounds used by (2.5) can be simplified down to just two linear lower bounds. The only two necessary lower bounds are exactly those required by (2.2) and (2.3). Namely, one could construct the model functions as

$$f_{k+1}(x) = \max\left\{f_k(z_{k+1}) + \langle \rho_k(z_{k+1} - x_k), x - z_{k+1}\rangle,\ f(z_{k+1}) + \langle g_{k+1}, x - z_{k+1}\rangle\right\} . \tag{2.6}$$

Then the subproblem that needs to be solved at each iteration can be done in closed form, see (5.3). Hence the iteration cost using this model is limited primarily by the cost of one subgradient evaluation.

Spectral Bundle Methods. Both of the above models rely on constructing piecewise linear models of the objective. For more structure problems, richer models can be constructed. For example, in eigenvalue optimization or more broadly semidefinite programming, better spectral lower bounds can be constructed instead of using simple polyhedral bounds [14, 32]. Primal-dual convergence rate guarantees for such spectral bundle methods were recently developed by Ding and Grimmer [8].

2.2 Convergence Rates from Constant Stepsize Choice

We now formalize our convergence theory for the proximal bundle method using any constant choice of the stepsize parameter $\rho_k = \rho$ and any $\beta \in (0, 1)$. These guarantees match those claimed in the first column of Table 1. After each theorem, we remark on the tuned choice of ρ that gives rise to the claimed rate in the second column of Table 1. We start by considering the setting where only Lipschitz continuity is assumed.

Theorem 2.1 (**Lipschitz**). *For any M-Lipschitz convex objective function f, consider applying the bundle method using a constant stepsize $\rho_k = \rho$. Then for any $0 < \epsilon \le f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most*

$$\frac{2\rho D^2}{\beta\epsilon} + \left\lceil \frac{\log\left(\frac{f(x_0)-f(x^*)}{\rho D^2}\right)}{-\log(1-\beta/2)} \right\rceil_+$$

and the number of null steps is at most

$$\frac{12\rho M^2 D^4}{\beta(1-\beta)^2\epsilon^3} + \frac{8M^2}{\beta(1-\beta)^2\rho^2 D^2}$$

where $D^2 = \sup_k \|x_k - x^*\|^2 < \infty$.

Remark 1. *It follows from [34][(7.64)] that $D^2 \le \|x_0 - x^*\|^2 + \frac{2(1-\beta)(f(x_0)-f^*)}{\beta\rho}$. Alternatively, if the level sets of f are bounded, the fact that $f(x_k)$ is non-increasing ensures $D^2 \le \sup\{\|x - x^*\|^2 \mid f(x) \le f(x_0)\}$.*

Remark 2. *Selecting $\rho = \epsilon/D^2$ gives an overall complexity bound of*

$$O\left(\frac{M^2 D^2}{\epsilon^2}\right)$$

and matches the optimal rate for nonsmooth, Lipschitz convex optimization.

If instead of Lipschitz continuity of the objective, we assume the objective has Lipschitz gradient, the bundle method adapts to give the following faster rate.

Theorem 2.2 (**Smooth**). *For any L-smooth convex objective function f, consider applying the bundle method using a constant stepsize $\rho_k = \rho$. Then for any $0 < \epsilon \le f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most*

$$\frac{2\rho D^2}{\beta\epsilon} + \left\lceil \frac{\log\left(\frac{f(x_0)-f(x^*)}{\rho D^2}\right)}{-\log(1-\beta/2)} \right\rceil_+$$

and the number of null steps is at most

$$\frac{4(L+\rho)^3}{(1-\beta)^2\rho^3}\left(\frac{2\rho D^2}{\beta\epsilon}+\left\lceil\frac{\log\left(\frac{f(x_0)-f(x^*)}{\rho D^2}\right)}{\log(1-\beta/2)}\right\rceil_+ +1\right)$$

where $D^2 = \sup_k \|x_k - x^*\|^2 < \infty$.

Remark 3. *Selecting* $\rho = L$ *gives an overall complexity bound of*

$$\frac{16LD^2}{\beta(1-\beta)^2\epsilon}\ .$$

This matches the standard convergence rate for gradient descent.

Next, we reconsider the settings of Lipschitz continuity and smoothness with additional structure in the form of a Hölder growth bound. We find that the convergence guarantees divide into three regions depending on the growth exponent p, whether it is large, equal to, or smaller than 2. Here two is the critical exponent value since the proximal subproblem is adding in quadratic regularization. Regardless, as p decreases, the bundle method converges faster.

Theorem 2.3 (**Lipschitz with Hölder growth**). *For any* M*-Lipschitz objective function* f *satisfying the Hölder growth condition* (1.4)*, consider applying the bundle method using a constant stepsize* $\rho_k = \rho$*. Then for any* $0 < \epsilon \leq f(x_0) - f(x^*)$*, the number of descent steps before an* ϵ*-minimizer is found is at most*

$$\begin{cases} \dfrac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + \left\lceil\dfrac{\log\left(\frac{f(x_0)-f(x^*)}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{-\log(1-\beta/2)}\right\rceil_+ & \text{if } p > 2 \\ \left\lceil\dfrac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta\min\{\mu/2\rho, 1/2\})}\right\rceil & \text{if } p = 2 \\ \left\lceil\dfrac{\log\left(\frac{(\rho/\mu^{2/p})^{1/(1-2/p)}}{\epsilon}\right)}{-\log(1-\beta/2)}\right\rceil_+ + \dfrac{2\rho(f(x_0)-f^*)^{2/p-1}}{(1-2^{1-2/p})\beta\mu^{2/p}} & \text{if } 1 \leq p < 2 \end{cases}$$

and the number of null steps is at most

$$\begin{cases} \dfrac{12\rho M^2}{(1-2/p)\beta(1-\beta)^2\mu^{4/p}\epsilon^{3-4/p}} + \dfrac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} & \text{if } p > 2 \\ \dfrac{2M^2}{\beta(1-\beta)^2\min\{\mu/2\rho, 1/2\}\rho\epsilon} & \text{if } p = 2 \\ \dfrac{4M^2}{\beta(1-\beta)^2\rho\epsilon} + \dfrac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}}C & \text{if } 1 \leq p < 2 \end{cases}$$

with $C = \max\left\{\frac{(f(x_0)-f^*)^{4/p-3}}{(\rho/\mu^{2/p})^{(4/p-3)/(1-2/p)}}, 1\right\}\min\left\{\frac{1}{1-2^{-|4/p-3|}}, \left\lceil\log_2\left(\frac{f(x_0)-f^*}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)\right\rceil\right\}$.

Remark 4. *When* $p = 2$*, selecting* $\rho = \mu$ *gives an optimal overall complexity bound of* $O(M^2/\mu\epsilon)$*. Selecting* $\rho = O(\epsilon^{1-2/p})$ *matches the optimal rate for Lipschitz optimization with growth exponent* $p > 2$*. When* $p = 1$*, selecting* $\rho = O(1/\sqrt{\epsilon})$ *minimizes this bound, but the resulting sublinear* $O(1/\sqrt{\epsilon})$ *rate falls short of the best possible rate (linear convergence) for sharp, Lipschitz optimization. In the next section where we consider nonconstant stepsizes, this disconnect will be remedied and a linear convergence guarantee will follow.*

Theorem 2.4 (**Smooth with Hölder growth**)**.** *For any L-smooth objective function f satisfying the Hölder growth condition* (1.4)*, consider applying the bundle method using a constant stepsize $\rho_k = \rho$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most*

$$\begin{cases} \dfrac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + \left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{-\log(1-\beta/2)} \right\rceil_+ & \text{if } p > 2 \\ \left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta\min\{\mu/2\rho, 1/2\})} \right\rceil & \text{if } p = 2 \end{cases}$$

and the number of null steps is at most

$$\begin{cases} \dfrac{4(L+\rho)^3}{(1-\beta)^2\rho^3}\left(\dfrac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + \left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{-\log(1-\beta/2)} \right\rceil_+ + 1\right) & \text{if } p > 2 \\ \dfrac{4(L+\rho)^3}{(1-\beta)^2\rho^3}\left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta\min\{\mu/2\rho, 1/2\})} \right\rceil & \text{if } p = 2\ . \end{cases}$$

Remark 5. *Selecting $\rho = L$ gives an overall complexity bound matching gradient descent.*

2.3 Convergence Rates from Improved Stepsize Choice

Picking ρ_k to vary throughout the execution of the bundle method allows for stronger convergence guarantees. These rates are formalized in the following pair of theorems that consider settings with and without Hölder growth. In the latter case, we find that our stepsize choice removes the need for piecewise guarantees around growth exponent $p = 2$, which notably simplifies the statement of our results.

Intuitively, the stepsize choices are aim to mimic the following idealistic (and impractical) stepsize rule that naturally arises from our theory

$$\rho_k = \frac{f(x_k) - f(x^*)}{\|x_k - x^*\|^2}\ . \tag{2.7}$$

The proof techniques we develop could be extended to study other interesting nonconstant stepsizes. For instance, stepsizes that shrink/grow polynomial with the number of descent steps, mirroring those used for subgradient methods. The analysis of such schemes is beyond the scope of this work.

Theorem 2.5 (**Lipschitz**)**.** *For any M-Lipschitz objective function f, consider applying the bundle method using the stepsize policy*

$$\rho_k = (f(x_k) - f(x^*))/D^2 \tag{2.8}$$

with any choice of $D^2 \geq \sup\{\|x - x^*\|^2 \mid f(x) \leq f(x_0)\}$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta/2)} \right\rceil$$

and the number of null steps is at most

$$\left(\frac{1}{1-(1-\beta/2)^2}\right)\frac{2M^2D^2}{(1-\beta)^2\epsilon^2}\ .$$

Theorem 2.6 (**Lipschitz with Hölder growth**). *For any M-Lipschitz objective function f satisfying the Hölder growth condition* (1.4)*, consider applying the bundle method using the stepsize policy*

$$\rho_k = \mu^{2/p}(f(x_k) - f(x^*))^{1-2/p}. \tag{2.9}$$

Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, the number of descent steps before an ϵ-minimizer is found is at most

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta/2)} \right\rceil$$

and the number of null steps is at most

$$\begin{cases} \left(\dfrac{1}{1-(1-\beta/2)^{2-2/p}}\right)\dfrac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} & \textit{if } p > 1 \\ \dfrac{2M^2}{(1-\beta)^2\mu^2}\left\lceil \dfrac{\log\left(\frac{f(x_0)-f(x^*)}{\epsilon}\right)}{-\log(1-\beta/2)} \right\rceil & \textit{if } p = 1 \ . \end{cases}$$

3 The parallel bundle method

We now give a practical scheme for applying the bundle method that attains the same complexity as our optimally tuned nonconstant stepsizes without any knowledge of the presence of smoothness or growth bounds. We do this by employing a logarithmic number of instances of the bundle method with different constant stepsizes in parallel that continually share their progress with each other. By doing so, we recover our optimal rates, up to the cost of running a logarithmic number of algorithms which can be mitigated through parallelization. This scheme is inspired by the ideas of [33].

The core observation behind our parallel method is that our nonconstant stepsize rules (2.8) and (2.9) before an ϵ-minimizer is found are always in the following interval

$$\rho_k \in \left[O(\epsilon), O(\epsilon^{-1})\right] \ .$$

As input, we only assume the following are given: a lower bound $\bar{\rho}$ and an upper bound $2^J\bar{\rho}$ on the range of stepsizes to consider. Provided our stepsize rules (2.8) and (2.9) lie in this interval,

$$\rho_k \in \left[\bar{\rho}, 2^J\bar{\rho}\right] \ ,$$

we are able to recover our optimal convergence rates. Notice that the interval $[\bar{\rho}, 2^J\bar{\rho}]$ can span the whole range of stepsizes needed for our Hölder growth analysis by setting $\bar{\rho} = O(\epsilon)$ and $J = O(\log(1/\epsilon^2))$. Our resulting convergence guarantees only depends logarithmically on the size of this interval (a cost which can be mitigated through parallelization), so $\bar{\rho}$ and $2^J\bar{\rho}$ can bet set conservatively at little cost.

Description of the algorithm. We propose running J copies of the bundle method in parallel, which share their progress with each other as described below. Each bundle method $j \in \{0, \dots J-1\}$ uses a constant stepsize $\rho^{(j)} = 2^j\bar{\rho}$. Denote the iterates of bundle method j by $x_k^{(j)}$ and its model objectives by $f_k^{(j)}$. Each bundle method j proceeds as normal with the only modification being that after it takes a descent step, the algorithm checks if any other bundle method j' has an iterate with an even lower objective value $f(x_k^{(j')}) < f(x_{k+1}^{(j)})$. If such an improvement exists, the bundle method instead descends to the best such iterate, setting

$$\begin{cases} x_{k+1}^{(j)} & \leftarrow x_k^{(j')} \\ f_{k+1}^{(j)}(z) & \leftarrow f(x_k^{(j')}) + \langle g_k^{(j')}, z - x_k^{(j')} \rangle \end{cases}$$

and then proceeds.

For analysis sake, we will assume that each parallel instance of the bundle method operates synchronously, with every instance completing one iteration before any instance completes a second iteration. This process can be implemented sequentially by cycling through the bundle method instances computing one iteration for each before repeating. An asynchronous variant of this procedure could be analyzed as well, using similar techniques as those in [33]. However, this is beyond the focus of this work. Note the choice to use powers of two here is arbitrary. In the following numerical section, we use powers of 10 and 100 demonstrating the effectiveness of this scheme even when using a sparse selection of sample stepsizes.

3.1 Convergence Rates for the Parallel Bundle Method

First, we remark that all of our previous convergence theory for constant stepsizes (Theorems 2.1, 2.2, 2.3, and 2.4) immediately apply to the Parallel Bundle Method fixing $\rho = 2^j\bar{\rho}$ for any $j \in \{0, \dots J-1\}$. This follows as our convergence theory on relies on a lemma ensuring sufficient decrease at each descent step (Lemma 5.1) and the new case of a bundle method restarting at another method's lower objective value iterate can only further improve on this decrease. Hence any individual instance of the bundle method with $\rho^{(j)} = 2^j\bar{\rho}$ in our parallel scheme will converge at least as fast as Theorems 2.1, 2.2, 2.3, and 2.4 guarantee it would converge on its own.

Further and more importantly, when our nonconstant stepsize rules (2.8) and (2.9) lie in the interval $[\bar{\rho}, 2^J\bar{\rho}]$, we find that their convergence theory (Theorems 2.5 and 2.6) also extends to our parallel algorithm. This is formalized as follows.

Theorem 3.1. *For any M-Lipschitz objective function f that satisfies the Hölder growth condition* (1.4)*, consider applying the Parallel Bundle Method with stepsizes $\rho = 2^j\bar{\rho}$ for $j \in \{0, \dots, J-1\}$. Then for any $0 < \epsilon \leq f(x_0) - f(x^*)$, if*

$$\bar{\rho} \leq \frac{1}{4}\mu^{2/p}\min\{\epsilon^{1-2/p}, (f(x_0) - f(x^*))^{1-2/p}\}$$

and

$$J \geq \log_2\left(\frac{\mu^{2/p}(\max\{\epsilon^{1-2/p}, (f(x_0) - f(x^*))^{1-2/p}\}}{4\bar{\rho}}\right) ,$$

then one of our J bundle methods will find an ϵ-minimizer within its first

$$\begin{cases} \left(\frac{2}{1-(1-\beta/2)^{2-2/p}}\right)\frac{16M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} + 2\left\lceil\frac{\log(\frac{f(x_0)-f^*}{\epsilon})}{-\log(1-\beta/2)}\right\rceil & \text{if } p > 1 \\ 2\left(\frac{16M^2}{(1-\beta)^2\mu^2} + 1\right)\left\lceil\frac{\log(\frac{f(x_0)-f^*}{\epsilon})}{-\log(1-\beta/2)}\right\rceil & \text{if } p = 1 \end{cases}$$

iterations.

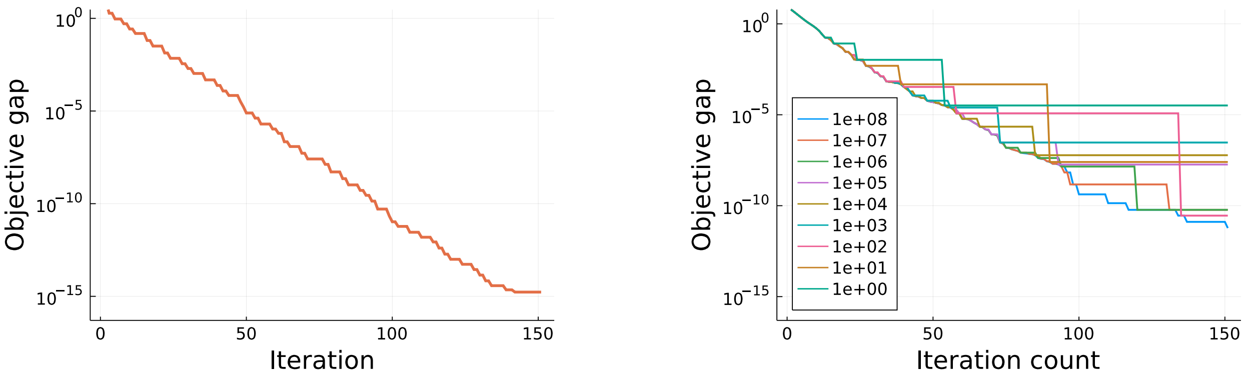

Figure 1: Objective gap against iteration count: using ideal stepsize (2.7) (left) and using the parallel bundle method, plotting each instance deployed with stepsizes from $10^0, \ldots, 10^8$ (right).

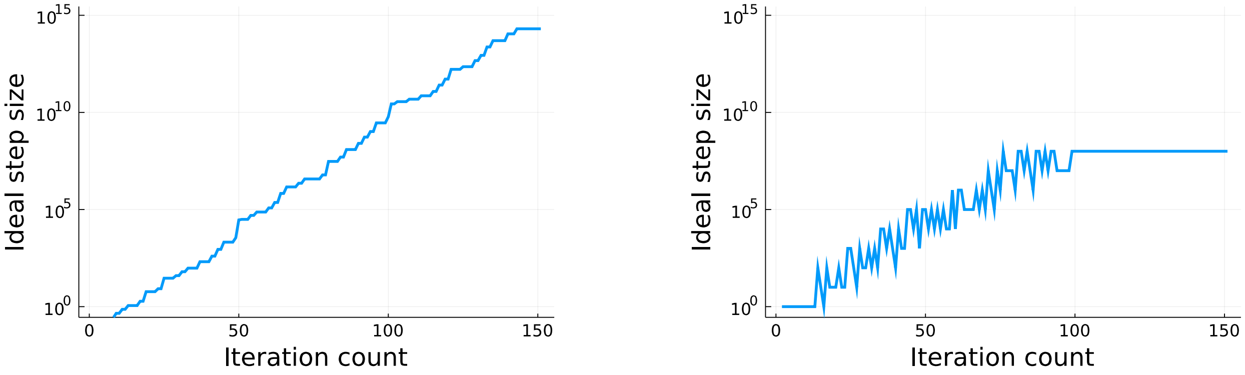

Figure 2: Stepsize against iteration count: using ideal stepsize (2.7) (left) and using the parallel bundle method (right).

4 Numerical experiments

In this section, we present two examples that illustrate numerically the theory for the bundle method. These experiments were implemented in Julia, see `https://github.com/mateodd25/proximal-bundle-method`.

4.1 Sharp linear regression

The first experiment aims to exemplify the fast convergence of the bundle method under sharp growth. We consider a simple linear regression problem of the form

$$\min_{x \in \mathbb{R}^d} f(x) := \|Ax - b\|$$

where $A \in \mathbb{R}^{n \times d}$ is a matrix and $b = Ax^\star$ for a fixed $x^\star$. This problem is equivalent to the classic least-squares problem after taking squares. Yet, without the square it is well known that for Gaussian matrices, $(A)_{ij} \sim N(0, \frac{1}{\sqrt{n}})$, this function is sharp and Lipchitz continuous provided n is large enough.

We generate a random Gaussian matrix $A \in \mathbb{R}^{100 \times 50}$ and random solution $x^\star \sim N(0, I_d)$. We run two algorithms: the proximal bundle method with the "ideal" stepsize (2.7) and the parallel bundle method described in Section 3. The ideal stepsize is impractical since it requires knowing the optimal solution. However, the theoretical analysis shows that it gives optimal convergence rates. In fact, the stepsizes proposed in our results (2.8) and (2.9) try to mimic its behavior. Thus, the method with ideal stepsize serves as a point of comparison. The parallel bundle method uses 9 parallel instances with stepsizes in $\rho \in \{1, 10, \dots, 10^8\}$. We let both methods run for 150 iterations.

Figure 1 displays the objective gap $f - \min f$ against the iteration count for both methods. On the other hand, Figure 2 shows the stepsize used at each iteration. For the parallel bundle method, we display the stepsize used by the last instance to reduce the best objective value seen.

As the theory predicts the convergence of both methods is linear. The bundle method with ideal stepsize exhibits steady progress and reaches an objective gap of $1.70 \cdot 10^{-15}$, while the parallel version slows down around 100 iterations and only achieves $5.87 \cdot 10^{-12}$. This behavior is explained by the stepsize plots. Figure 2 plots how the parallel algorithm roughly emulates the ideal stepsize until it exceeds the largest instance's stepsize 10^8. After which, the instance with stepsize 10^8 consistently leads the method's progress, albeit sublinearly.

4.2 Support Vector Machine

To illustrate the adaptive features of the parallel bundle method we consider the standard Support Vector Machine (SVM) formulation: we are giving datapoints $(x_1, y_1), \dots (x_n, y_n)$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{\pm 1\}$ and our goal is to solve

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum \max\{0, 1 - y_i \langle w, x_i \rangle\} + \frac{\lambda}{2} \|w\|^2 \tag{4.1}$$

where $\lambda \in \mathbb{R}$ is a fixed constant. This problem is not smooth due to the first term. For this experiment we compare against a subgradient method based on Pegasos [37], a state-of-the art solver for SVM. Our vanilla implementation of the parallel bundle method is not tuned for efficiency and does not aim to be competitive with commercial solvers. Instead, we aim to show that an out-of-the-box implementation is immediately comparable to a specialized first-order method for this problem.

We generate SVM problems using three datasets from the LIBSVM Binary Classification Database [1]. In particular, we use `colon-cancer`, `duke`, and `leu`.[2] We preprocess the data by deleting empty features, normalizing the features, and adding an extracomponent $x_k = (x_k, 1)$ to allow for affine functions.

The implementation of the subgradient algorithm updates

$$w_{k+1} \leftarrow (1 - \eta_k \lambda) w_k + \eta_k \sum_{i=1}^{n} \mathbf{1}\{1 \leq y_i \langle w_k, x_i \rangle\} y_i x_i$$

where $\eta_k = \frac{1}{\lambda k}$ and $\mathbf{1}\{\cdot\}$ is one if $\cdot$ holds true and zero otherwise. This is analogous to Pegasos with the exception that it does full instead of stochastic subgradient evaluations. Of course, knowledge of λ is necessary for the implementation of this method.

For the parallel bundle method, we use stepsizes 11 instances with constant stepsizes

$$\rho \in \{10^{-15} \cdot 100^j \mid j = 0, \dots, 10\}.$$

We run both methods for 2000 iterations and measure the objective gap $f - \min f$. To compute the minimum we use Gurobi with accuracy set to 10^{-10}. Figure 3 plots the gap against while varying the regularizer coefficient within $\lambda \in \{0.001, 0.01, 0.1, 0.5, 1.5, 2.0\}$.

[2] We refer the reader to LIBSVM for the origin of each of these datasets.

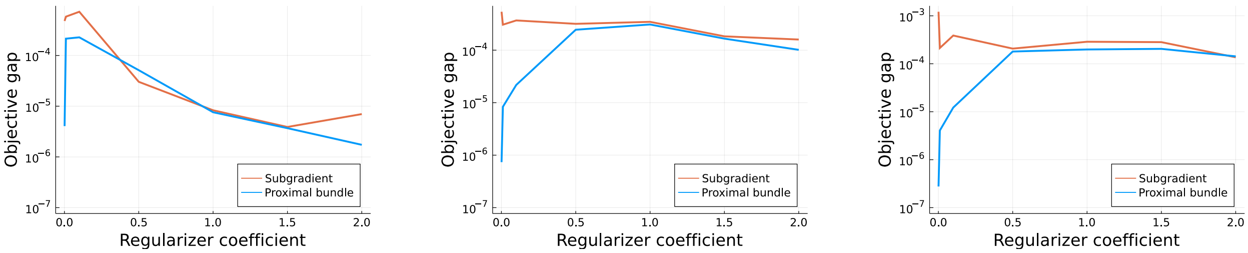

Figure 3: Objective gap against coefficient λ for the three problems, solved by a subgradient method and by the parallel bundle method: `colon-cancer` (left), `duke` (center), and `leu` (right).

In this simple setting, the parallel bundle method out of the box performs similarly to the tuned subgradient method while only requiring a constant amount of extra work (that can be parallelized). We see that the parallel method with the same parameter configuration can handle a wide range of parameters λ. While for small λ the performance of the subgradient method tends to deteriorate, the performance of the bundle method improves (outperforming the subgradient method by several orders of magnitude).

5 Analysis

In this section, we develop the proofs of the convergence rates. We start by introducing the general strategy that we use to establish all of our results and then specialize it to each scenario.

5.1 Analysis Overview and Proof Sketch

Each iteration of the bundle method can be viewed as an attempt to mimic the proximal point method, using the model f_k instead of the true objective function f. At each iteration k, we denote the objective gap of the proximal subproblem, called the *proximal gap*, by

$$\Delta_k := f(x_k) - \left(f(\bar{x}_{k+1}) + \frac{\rho_k}{2}\|\bar{x}_{k+1} - x_k\|^2\right)$$

where $\bar{x}_{k+1} = \operatorname{argmin}_{x\in\mathbb{R}^d} \{f(x) + \frac{\rho_k}{2}\|x - x_k\|^2\}$.

Regardless of which continuity, smoothness and growth assumptions are made, our analysis works by relating the proximal steps computed by the bundle method on the models f_k to proximal steps on f. The following pair of observations show that the behavior on both descent steps and null steps is controlled by the proximal gap Δ_k.

(i) **Descent steps attain decrease proportional to the proximal gap.**

Lemma 5.1. *A descent step, at iteration k, has*

$$f(x_{k+1}) \leq f(x_k) - \beta\Delta_k.$$

(ii) **The number of consecutive null steps is bounded by the proximal gap.**

Lemma 5.2. *A descent step, at iteration k, followed by T consecutive null steps has at most*

$$T \leq \frac{2G_{k+1}^2}{(1-\beta)^2\rho_{k+1}\Delta_{k+1}}$$

where $G_{k+1} = \sup\{\|g_{t+1}\| \mid k \leq t \leq k+T\}$. *This simplifies to*

$$T \leq \begin{cases} \dfrac{2M^2}{(1-\beta)^2\rho_{k+1}\Delta_{k+1}} & \textit{if } f \textit{ is } M\textit{-Lipschitz, or} \\ \dfrac{4(L+\rho_k)^3}{(1-\beta)^2\rho_{k+1}^3} & \textit{if } f \textit{ is } L\textit{-smooth .} \end{cases}$$

With these two observations in hand, convergence guarantees for the bundle method follow from specifying any choice of the parameter ρ_k. Given a choice of ρ_k, bounding the proximal gap is a classic, well-understand problem, independent from the details of the bundle method being used. Standard analysis [34] of the proximal gap shows the following bound for any minimizer x^*.

Lemma 5.3. *For any* $x_k \in \mathbb{R}^n$, *the proximal gap is lower bounded by*

$$\Delta_k \geq \begin{cases} \dfrac{1}{2\rho_k}\left(\dfrac{f(x_k)-f(x^*)}{\|x_k-x^*\|}\right)^2 & \textit{if } f(x_k)-f(x^*) \leq \rho_k\|x_k-x^*\|^2 \\ f(x_k)-f(x^*)-\dfrac{\rho_k}{2}\|x_k-x^*\|^2 & \textit{otherwise.} \end{cases} \tag{5.1}$$

Our ideal stepsize (2.7) is chosen to balance the two cases of this classic bound.

All of our analysis follows directly from applying these core lemmas. We bound the number of descent steps by combining Lemmas 5.1 and 5.3 to give a recurrence relation describing the decrease in the objective gap. Then Lemmas 5.2 and 5.3 together allow us to bound the number of consecutive null steps between each of these descent steps, which can then be summed up to bound the total number of iterations required.

5.2 Proof of the Descent Step Lemma 5.1

Let $\bar{x}_{k+1} = \operatorname{argmin}\{f(\cdot) + \frac{\rho_k}{2}\|\cdot - x_k\|^2\}$. From (2.1), we have

$$\begin{aligned} f_k(x_{k+1}) &\leq f_k(x_{k+1}) + \frac{\rho_k}{2}\|x_{k+1}-x_k\|^2 \\ &\leq f_k(\bar{x}_{k+1}) + \frac{\rho_k}{2}\|\bar{x}_{k+1}-x_k\|^2 \\ &\leq f(\bar{x}_{k+1}) + \frac{\rho_k}{2}\|\bar{x}_{k+1}-x_k\|^2 \ . \end{aligned}$$

Hence $f(x_k) - f_k(x_{k+1}) \geq \Delta_k$. Since we have assumed that iteration k was a descent step, this implies $(f(x_k) - f(x_{k+1}))/\beta \geq \Delta_k$. Concluding the proof.

5.3 Proof of the Null Step Lemma 5.2

Consider some descent step, at iteration k, followed by T consecutive null steps. Denote the proximal subproblem gap at iteration $k < t \leq k+T$ on the model f_t by

$$\widetilde{\Delta}_t := f(x_{k+1}) - \left(f_t(z_{t+1}) + \frac{\rho_{k+1}}{2}\|z_{t+1}-x_{k+1}\|^2\right).$$

Note every such null step t has the same stepsize $\rho_t = \rho_{k+1}$ and the same proximal center $x_t = x_{k+1}$. The core of this null step bound relies on the following recurrence showing $\widetilde{\Delta}_t$ decreases at each step

$$\widetilde{\Delta}_{t+1} \leq \widetilde{\Delta}_t - \frac{(1-\beta)^2\rho_{k+1}\widetilde{\Delta}_t^2}{2G_{k+1}^2} \ . \tag{5.2}$$

Before deriving this inequality, we show how it completes the proof of this lemma. After T consecutive null steps, the fact that $f_{k+T} \leq f$ ensures $\widetilde{\Delta}_{k+T} \geq \Delta_{k+T} = \Delta_{k+1}$. Thus, to bound T it suffices to bound the minimum iteration at which the reversed inequality hold. By solving the recurrence, see Lemma A.1 in the appendix with $\epsilon = \Delta_{k+1}$, we conclude the number of consecutive null steps is at most

$$T \leq \frac{2G_{k+1}^2}{(1-\beta)^2 \rho_{k+1} \Delta_{k+1}} \ .$$

Now all that remains is to derive the recurrence (5.2). Consider some null step $k < t \leq k+T$ in the sequence of consecutive null steps. We will use the following claim mulitiple times in the proof.

Claim 1. The following inequalities hold true $\|s_{t+1}\|^2 \leq 2\rho_{k+1}\tilde{\Delta}_t \leq G_{k+1}^2$.

Proof of the Claim. Due to the ρ_{k+1}-strongly convexity of the proximal subproblem $f_t(z) + \frac{\rho_{k+1}}{2}\|z - x_{k+1}\|^2$ and the fact that z_{t+1} is its unique minimizer, we derive

$$\begin{aligned} \frac{\rho_{k+1}}{2}\|z_{t+1} - x_{k+1}\|^2 &\leq f_t(x_{k+1}) - \left(f_t(z_{t+1}) + \frac{\rho_{k+1}}{2}\|z_{t+1} - x_{k+1}\|^2\right) \\ &\leq \widetilde{\Delta}_t \\ &\leq \widetilde{\Delta}_{k+1} \leq \frac{1}{2\rho_k}\|g_{k+1}\|^2. \end{aligned}$$

The last inequality follows by (2.2) since

$$\begin{aligned} f_{k+1}(z_{k+2}) &\geq f(x_{k+1}) + \langle g_{k+1}, z_{k+2} - x_{k+1}\rangle \\ &\geq f(x_{k+1}) - \frac{1}{2}\left(\frac{\|g_{k+1}\|^2}{\rho_{k+1}} + \rho_{k+1}\|z_{k+2} - x_{k+1}\|^2\right). \end{aligned}$$

□

Define the necessary lower bound on f_{t+1} given by (2.2) and (2.3) as

$$\tilde{f}_{t+1}(\cdot) := \max\left\{f_t(z_{t+1}) + \langle s_{t+1}, \cdot - z_{t+1}\rangle,\ f(z_{t+1}) + \langle g_{t+1}, \cdot - z_{t+1}\rangle\right\} \leq f_{t+1}(\cdot) \ .$$

Denote the result of a proximal step on $\tilde{f}_{t+1}$ by

$$y_{t+2} = \operatorname{argmin}\left\{\tilde{f}_{t+1}(\cdot) + \frac{\rho_{k+1}}{2}\|\cdot - x_{k+1}\|^2\right\} \ .$$

A simple computation gives an explicit form for the minimizer of this problem

$$\begin{aligned} \theta_{t+1} &= \min\left\{1, \frac{\rho_{k+1}\left(f(z_{t+1}) - f_t(z_{t+1})\right)}{\|g_{t+1} - s_{t+1}\|^2}\right\} \\ y_{t+2} &= x_{k+1} - \frac{1}{\rho_{k+1}}\left(\theta_{t+1} g_{t+1} + (1-\theta_{t+1}) s_{t+1}\right). \end{aligned} \tag{5.3}$$

Hence the objective of the proximal subproblem at iteration $t+1$ is lower bounded by

$$\begin{aligned}
&f_{t+1}(z_{t+2}) + \frac{\rho_{k+1}}{2}\|z_{t+2} - x_{k+1}\|^2 \\
&\geq \tilde{f}_{t+1}(y_{t+2}) + \frac{\rho_{k+1}}{2}\|y_{t+2} - x_{k+1}\|^2 \\
&\geq \theta_{t+1}\left(f(z_{t+1}) + \langle g_{t+1}, y_{t+2} - z_{t+1}\rangle\right) \\
&\qquad + (1-\theta_{t+1})\left(f_t(z_{t+1}) + \langle s_{t+1}, y_{t+2} - z_{t+1}\rangle\right) + \frac{\rho_{k+1}}{2}\|y_{t+2} - x_{k+1}\|^2 \\
&= f_t(z_{t+1}) + \theta_{t+1}\Big(f(z_{t+1}) - f^t(z_{t+1})\Big) \\
&\qquad + \langle \theta_{t+1}g_{t+1} + (1-\theta_{t+1})s_{t+1}, y_{t+2} - z_{t+1}\rangle + \frac{\rho_{k+1}}{2}\|y_{t+2} - x_{k+1}\|^2 \\
&= f_t(z_{t+1}) + \theta_{t+1}\Big(f(z_{t+1}) - f^t(z_{t+1})\Big) \\
&\qquad + \theta_{t+1}^2\|g_{t+1} - s_{t+1}\|^2/\rho_{k+1} + \frac{\rho_{k+1}}{2}\|z_{t+1} - x_{k+1}\|^2 \ ,
\end{aligned}$$

where the first inequality uses that $f_{t+1} \geq \tilde{f}_{t+1}$, the second inequality takes a convex combination of the two affine functions defining $\tilde{f}_{t+1}$, and the second equality uses the definition of y_{t+2}. Thus we have

$$\widetilde{\Delta}_{t+1} \leq \widetilde{\Delta}_t - \theta_{t+1}\left(f(z_{t+1}) - f_t(z_{t+1})\right) + \theta_{t+1}^2\|g_{t+1} - s_{t+1}\|^2/\rho_{k+1} \ .$$

The amount of decrease guaranteed above can be lower bounded as follows

$$\begin{aligned}
&\theta_{t+1}\left(f(z_{t+1}) - f_t(z_{t+1})\right) + \theta_{t+1}^2\|g_{t+1} - s_{t+1}\|^2/\rho_{k+1} \\
&\geq \min\left\{f(z_{t+1}) - f_t(z_{t+1}),\ \frac{2\rho_{k+1}(f(z_{t+1}) - f_t(z_{t+1}))^2}{\|g_{t+1} - s_{t+1}\|^2}\right\} \\
&\geq \min\left\{(1-\beta)\widetilde{\Delta}_t,\ \frac{2\rho_{k+1}(1-\beta)^2\widetilde{\Delta}_t^2}{\|g_{t+1} - s_{t+1}\|^2}\right\} \\
&\geq \min\left\{(1-\beta)\widetilde{\Delta}_t,\ \frac{\rho_{k+1}(1-\beta)^2\widetilde{\Delta}_t^2}{\|g_{t+1}\|^2 + \|s_{t+1}\|^2}\right\} \\
&\geq \min\left\{2\frac{\rho_{k+1}(1-\beta)\widetilde{\Delta}_t^2}{G_{k+1}^2},\ \frac{\rho_{k+1}(1-\beta)^2\widetilde{\Delta}_t^2}{2G_{k+1}^2}\right\} \\
&\geq \frac{\rho_{k+1}(1-\beta)^2\widetilde{\Delta}_t^2}{2G_{k+1}^2}
\end{aligned}$$

where the first inequality uses the definition of θ_{t+1} and drops a norm squared term, the second inequality uses the definition of a null step, and the fourth inequality uses Claim 1 and $\|g_{t+1}\|^2 \leq G_{k+1}^2$. This verifies (5.2) and completes the proof of our general bound.

For any M-Lipschitz objective, our specialized result follows from observing that $G_k \leq M$ as subgradients everywhere are uniformly bounded in norm by the Lipschitz constant. For any L-smooth objective, the following three inequalities hold for any null step t in the sequence of consecutive null steps following a descent step $k < t$:

$$\|g_{t+1}\| \leq \|g_{k+1}\| + L\|z_{t+1} - x_{k+1}\| \tag{5.4}$$

$$\|z_{t+1} - x_{k+1}\| \leq \|g_{k+1}\|/\rho_{k+1} \tag{5.5}$$

$$\|g_{k+1}\| \leq \sqrt{2(L+\rho_{k+1})\Delta_{k+1}} \ . \tag{5.6}$$

Before proving these three inequalities, we note that combined they give the claimed bound as

$$\begin{aligned} G_{k+1} = \sup_t\{\|g_{t+1}\|\} &\leq \sup_t\{\|g_{k+1}\| + L\|z_{t+1} - x_{k+1}\|\} \\ &\leq (1 + L/\rho_{k+1})\|g_{k+1}\| \\ &\leq (1 + L/\rho_{k+1})\sqrt{2(L + \rho_{k+1})\Delta_{k+1}} \end{aligned}$$

and thus $G_{k+1}^2 \leq 2(L + \rho_{k+1})^3\Delta_{k+1}/\rho_{k+1}^2$. First (5.4) follows directly from the gradient being L-Lipschitz continuous. Second (5.5) follows from Claim 1. Third (5.6) follows from the L-smoothness of f and considering the full proximal subproblem $f(z) + \frac{\rho_{k+1}}{2}\|z - x_{k+1}\|^2$ since

$$\begin{aligned} \Delta_{k+1} &= f(x_{k+1}) - \min_z\left\{f(z) + \frac{\rho_{k+1}}{2}\|z - x_{k+1}\|^2\right\} \\ &\geq f(x_{k+1}) - \min_z\left\{f(x_{k+1}) + \langle g_{k+1}, z - x_{k+1}\rangle + \frac{L + \rho_{k+1}}{2}\|z - x_{k+1}\|^2\right\} \\ &= \frac{\|g_{k+1}\|^2}{2(L + \rho_{k+1})} \ . \end{aligned}$$

5.4 Proof of Theorem 2.1

For a constant stepsize $\rho_k = \rho$, we can simplify the lower bound (5.1) to only depend on x_k through a simple threshold on $f(x_k) - f^*$ as

$$\Delta_k \geq \begin{cases} \frac{1}{2\rho}\left(\frac{f(x_k) - f^*}{D}\right)^2 & \text{if } f(x_k) - f^* \leq \rho D^2 \\ \frac{1}{2}\left(f(x_k) - f^*\right) & \text{otherwise.} \end{cases} \tag{5.7}$$

Combining this with Lemma 5.1 gives a recurrence relation describing the decrease in the objective gap $\delta_k = f(x_k) - f^*$ on any descent step k of

$$\delta_{k+1} \leq \begin{cases} \delta_k - \frac{\beta\delta_k^2}{2\rho D^2} & \text{if } \delta_k \leq \rho D^2 \\ (1 - \beta/2)\delta_k & \text{if } \delta_k > \rho D^2 \ . \end{cases}$$

Our analysis of the bundle method then proceeds by considering these two cases separately. In each case, solving the given recurrence relation bounds the number of descent steps and applying Lemma 5.2 bounds the number of null steps.

5.4.1 Bounding steps with $\delta_k > \rho D^2$. First we show that the number of descent steps with $\delta_k > \rho D^2$ is bounded by

$$\left\lceil \frac{\log\left(\frac{f(x_0) - f^*}{\rho D^2}\right)}{-\log(1 - \beta/2)} \right\rceil_+ \tag{5.8}$$

and the number of null steps with $\delta_k > \rho D^2$ is at most

$$\frac{8M^2}{\beta(1 - \beta)^2\rho^2 D^2} \ . \tag{5.9}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1 - \beta/2)\delta_k$. This immediately bounds the number of descent steps by (5.8). Index the descent steps before a ρD^2-minimizer is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than ρD^2. Define $k_0 = -1$. Then for each $i = 0 \dots n-1$, $f(x_{k_i+1}) - f^* \geq (1-\beta/2)^{i-(n-1)}\rho D^2$. It follows from (5.1) that $\Delta_{k_i+1} \geq (f(x_{k_i+1})-f^*)/2 \geq (1-\beta/2)^{i-(n-1)}\rho D^2/2$. Plugging this into Lemma 5.2 upper bounds the number of consecutive null steps after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1-\beta/2)^{(n-1)-i} \frac{4M^2}{(1-\beta)^2\rho^2 D^2} \ .$$

Summing this over $i = 0 \dots n-1$ bounds the total number of null steps before a ρD^2-minimizer is found by (5.9) as

$$\sum_{i=0}^{n-1} (1-\beta/2)^{(n-1)-i} \frac{4M^2}{(1-\beta)^2\rho^2 D^2} \leq \frac{8M^2}{\beta(1-\beta)^2\rho^2 D^2} \ .$$

5.4.2 Bounding steps with $\rho D^2 \geq \delta_k > \epsilon$.

Now we complete our proof of Theorem 2.1 by bounding the number of descent steps with $\rho D^2 \geq \delta_k > \epsilon$ by

$$\frac{2\rho D^2}{\beta\epsilon} \tag{5.10}$$

and the number of null steps with $\rho D^2 \geq \delta_k > \epsilon$ by

$$\frac{12\rho D^4 M^2}{(1-\beta)^2\epsilon^3} \ . \tag{5.11}$$

After the bundle method has passed objective value ρD^2, the recurrence relation becomes

$$\delta_{k+1} \leq \delta_k - \frac{\beta\delta_k^2}{2\rho D^2}.$$

Solving this recurrence with Lemma A.1 implies $\delta_k > \epsilon$ holds for at most (5.10) descent stwhereeps. Then we can bound the number of null steps between these descent steps by noting (5.7) implies $\Delta_k \geq (f(x_k) - f^*)^2/2\rho D^2 \geq \epsilon^2/2\rho D^2$. Then Lemma 5.2 upper bounds the number of consecutive null steps by $4D^2M^2/(1-\beta)^2\epsilon^2$. Then multiplying this by our bound on the number of descent steps gives (5.11) as

$$\left(\frac{2\rho D^2}{\beta\epsilon} + 1\right) \frac{4D^2M^2}{(1-\beta)^2\epsilon^2} \leq \frac{12\rho D^4 M^2}{\beta(1-\beta)^2\epsilon^3} \ .$$

5.5 Proof of Theorem 2.2

Our bound on the number of descent steps comes directly from Theorem 2.1. Our claimed bound on the total number of null steps follows by multiplying this by the constant bound on the number of consecutive null steps from Lemma 5.2.

5.6 Proof of Theorem 2.3

Assuming Hölder growth (1.4) holds and fixing $\rho_k = \rho$, the lower bound (5.1) simplifies to only depend on a simple threshold with $f(x_k) - f^*$ as

$$\Delta_k \geq \begin{cases} \dfrac{\mu^{2/p}(f(x_k) - f^*)^{2-2/p}}{2\rho} & \text{if } (f(x_k) - f^*)^{1-2/p} \leq \rho/\mu^{2/p} \\ \dfrac{1}{2}\left(f(x_k) - f^*\right) & \text{otherwise .} \end{cases} \tag{5.12}$$

From this, we arrive at a recurrence relation on the objective gap $\delta_k = f(x_k) - f^*$ decrease at each descent step k by plugging this lower bound into Lemma 5.1 of

$$\delta_{k+1} \leq \begin{cases} \delta_k - \dfrac{\beta\mu^{2/p}\delta_k^{2-2/p}}{2\rho} & \text{if } \delta_k^{1-2/p} \leq \rho/\mu^{2/p} \\ (1-\beta/2)\delta_k & \text{if } \delta_k^{1-2/p} > \rho/\mu^{2/p} \ . \end{cases}$$

Our analysis proceeds by considering the two cases of this recurrence and the three cases of $p > 2$, $p = 2$, and $1 \leq p < 2$ separately. In each case, solving the given recurrence relation bounds the number of descent steps and applying Lemma 5.2 bounds the number of null steps.

5.6.1 Given $p > 2$, bounding steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$.

First we show that the number of descent steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is bounded by

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f^*}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)}{-\log(1-\beta/2)} \right\rceil_+ \tag{5.13}$$

and the number of null steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is at most

$$\frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \ . \tag{5.14}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1-\beta/2)\delta_k$. This immediately bounds the number of descent steps by (5.13). Index the descent steps before a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than $(\rho/\mu^{2/p})^{1/(1-2/p)}$. Define $k_0 = -1$. Then for each $i = 0 \ldots n-1$, $f(x_{k_i+1}) - f^* \geq (1-\beta/2)^{i-(n-1)}(\rho/\mu^{2/p})^{1/(1-2/p)}$. It follows from (5.1) that

$$\Delta_{k_i+1} \geq (f(x_{k_i+1}) - f^*)/2 \geq (1-\beta/2)^{i-(n-1)}\left(\rho/\mu^{2/p}\right)^{1/(1-2/p)}/2.$$

Plugging this into Lemma 5.2 upper bounds the number of consecutive null steps after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1-\beta/2)^{(n-1)-i}\frac{4M^2}{(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \ .$$

Summing this over $i = 0 \ldots n-1$ bounds the total number of null steps before a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer is found by (5.14) as

$$\sum_{i=0}^{n-1}(1-\beta/2)^{(n-1)-i}\frac{4M^2}{(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \leq \frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \ .$$

5.6.2 Given $p > 2$, bounding steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$.

Next we show that the total number of descent steps with$(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ is bounded by

$$\frac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} \tag{5.15}$$

and the number of null steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ is at most

$$\frac{12\rho M^2}{(1-2/p)\beta(1-\beta)^2\mu^{4/p}\epsilon^{3-4/p}} \ . \tag{5.16}$$

In this case, the recurrence relation on objective value decrease becomes

$$\delta_{k+1} \leq \delta_k - \frac{\beta\mu^{2/p}\delta_k^{2-2/p}}{2\rho}.$$

Applying Lemma A.1 gives our bound on the number of descent steps with $\delta_k > \epsilon$ in (5.15). Plugging the lower bound $\Delta_k \geq \mu^{2/p}(f(x_k) - f^*)^{2-2/p}/2\rho \geq \mu^{2/p}\epsilon^{2-2/p}/2\rho$ into Lemma 5.2, the number of consecutive null steps after a descent step is at most

$$\frac{4M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} .$$

Then multiplying our limit on consecutive null steps by the number of descent steps between finding a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer and finding an ϵ-minimizer gives the bound (5.16) as

$$\left(\frac{2\rho}{(1-2/p)\beta\mu^{2/p}\epsilon^{1-2/p}} + 1\right)\frac{4M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} \leq \frac{12\rho M^2}{(1-2/p)\beta(1-\beta)^2\mu^{4/p}\epsilon^{3-4/p}} .$$

5.6.3 Given $p = 2$, bounding steps with $\delta_k > \epsilon$. Here both cases of our recurrence relation have a similar form, and so we directly bound the total number of descent steps with $\delta_k > \epsilon$ by

$$\left\lceil \frac{\log\left(\frac{f(x_0)-f^*}{\epsilon}\right)}{-\log(1-\beta\min\{\mu/2\rho, 1/2\})} \right\rceil \tag{5.17}$$

and the number of null steps with $\delta_k > \epsilon$ by

$$\frac{2M^2}{\beta(1-\beta)^2\min\{\mu/2\rho, 1/2\}\rho\epsilon} . \tag{5.18}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1-\beta\min\{\mu/2\rho, 1/2\})\delta_k$. This immediately bounds the number of descent steps by (5.17). Index the descent steps before an ϵ-minimizer is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than ϵ. Define $k_0 = -1$. Then for each $i = 0 \ldots n-1$,

$$f(x_{k_i+1}) - f^* \geq (1-\beta\min\{\mu/2\rho, 1/2\})^{i-(n-1)}\epsilon .$$

It follows from (5.1) that $\Delta_{k_i+1} \geq (1-\beta\min\{\mu/2\rho, 1/2\})^{i-(n-1)}\epsilon/2$. Plugging this into Lemma 5.2 upper bounds the number of consecutive null steps after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1-\beta\min\{\mu/2\rho, 1/2\})^{(n-1)-i}\frac{2M^2}{(1-\beta)^2\rho\epsilon} .$$

Summing this over $i = 0 \ldots n-1$ bounds the total number of null steps before an ϵ-minimizer is found by

$$\sum_{i=0}^{n-1}(1-\beta\min\{\mu/2\rho, 1/2\})^{(n-1)-i}\frac{2M^2}{(1-\beta)^2\rho\epsilon} \leq \frac{2M^2}{\min\{\mu/2\rho, 1/2\}\beta(1-\beta)^2\rho\epsilon} .$$

5.6.4 Given $1 \leq p < 2$, bounding steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$.

Then we show that the number of descent steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is bounded by

$$\frac{2\rho(f(x_0) - f^*)^{2/p-1}}{(1 - 2^{1-2/p})\beta\mu^{2/p}} \tag{5.19}$$

and the number of null steps with $\delta_k > (\rho/\mu^{2/p})^{1/(1-2/p)}$ is at most

$$\frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}}C \tag{5.20}$$

with $C = \max\left\{\frac{(f(x_0)-f^*)^{4/p-3}}{(\rho/\mu^{2/p})^{(4/p-3)/(1-2/p)}}, 1\right\} \min\left\{\frac{1}{1-2^{-|4/p-3|}}, \left\lceil \log_2\left(\frac{f(x_0)-f^*}{(\rho/\mu^{2/p})^{1/(1-2/p)}}\right)\right\rceil\right\}$. Notice that since $p < 2$, the power $1 - 2/p$ of δ_k in the threshold condition of our recurrence is negative. In this case, the recurrence relation on objective value decrease becomes

$$\delta_{k+1} \leq \delta_k - \frac{\beta\mu^{2/p}\delta_k^{2-2/p}}{2\rho} \ .$$

As an intermediate step, for any $i \geq 0$, we first bound the number of descent and null steps with

$$2^{i+1}(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > 2^i(\rho/\mu^{2/p})^{1/(1-2/p)} \ .$$

Since descent steps decreases the objective gap by at least $\beta\mu^{2/p}\delta_k^{2-2/p}/2\rho$, there are at most

$$\frac{2\rho(2^i(\rho/\mu^{2/p})^{1/(1-2/p)})^{2/p-1}}{\beta\mu^{2/p}} = \frac{2^{(2/p-1)i+1}}{\beta}$$

descent steps in this interval. Further, noting that in this interval

$$\Delta_k \geq \frac{\mu^{2/p}(2^i(\rho/\mu^{2/p})^{1/(1-2/p)})^{2-2/p}}{2\rho} = 2^{(2-2/p)i-1}(\rho/\mu^{2/p})^{1/(1-2/p)} \ ,$$

we can bound the number of consecutive null steps following any of these descent steps via Lemma 5.2. Hence there are at most

$$\frac{2^{(4/p-3)i+3}M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}}$$

null steps in this interval.

The bundle method halves its objective value at most $N = \lceil \log_2((f(x_0) - f^*)/(\rho/\mu^{2/p})^{1/(1-2/p)})\rceil$ times before an $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer is found. Then summing up these bounds on the descent and null steps in each interval limits the number of descent steps needed to find a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer by (5.19) as

$$\sum_{i=0}^{N-1} \frac{2^{(2/p-1)i+1}}{\beta} \leq \frac{2}{\beta}\sum_{i=0}^{N-1} 2^{(2/p-1)i} \leq \frac{2^{(2/p-1)(N-1)+1}}{(1-2^{1-2/p})\beta} \leq \frac{2\rho(f(x_0) - f^*)^{2/p-1}}{(1 - 2^{1-2/p})\beta\mu^{2/p}}$$

and similarly, the number of null steps needed by (5.20) as

$$\begin{aligned}
&\sum_{i=0}^{N-1} \frac{2^{(4/p-3)i+3}M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \\
&\leq \frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \sum_{i=0}^{N-1} 2^{(4/p-3)i} \\
&\leq \frac{8M^2}{\beta(1-\beta)^2\rho(\rho/\mu^{2/p})^{1/(1-2/p)}} \\
&\qquad\qquad \max\left\{\frac{(f(x_0)-f^*)^{4/p-3}}{(\rho/\mu^{2/p})^{(4/p-3)/(1-2/p)}}, 1\right\} \min\left\{\frac{1}{1-2^{-|4/p-3|}}, N\right\}
\end{aligned}$$

where the last inequality bounds the geometric sum regardless of the sign of the exponent $4/p-3$.

5.6.5 Given $1 \leq p < 2$, bounding steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$. Finally, we show that the number of descent steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ is bounded by

$$\left\lceil \frac{\log\left(\frac{(\rho/\mu^{2/p})^{1/(1-2/p)}}{\epsilon}\right)}{-\log(1-\beta/2)} \right\rceil \tag{5.21}$$

and the number of null steps with $(\rho/\mu^{2/p})^{1/(1-2/p)} \geq \delta_k > \epsilon$ is at most

$$\frac{4M^2}{\beta(1-\beta)^2\rho\epsilon} \ . \tag{5.22}$$

In this case, our recurrence relation simplifies to have geometric decrease at each descent step $\delta_{k+1} \leq (1-\beta/2)\delta_k$. This immediately bounds the number of descent steps by (5.21). Index the descent steps after a $(\rho/\mu^{2/p})^{1/(1-2/p)}$-minimizer but before an ϵ-minimizer is found by $k_1 < \cdots < k_n$ such that x_{k_n+1} is the first iterate with objective value less than ϵ. Then for each $i = 0 \ldots n-1$, $f(x_{k_i+1}) - f^* \geq (1-\beta/2)^{i-(n-1)}\epsilon$. It follows from (5.1) that $\Delta_{k_i+1} \geq (f(x_{k_i+1}) - f^*)/2 \geq (1-\beta/2)^{i-(n-1)}\,\epsilon/2$. Plugging this into Lemma 5.2 upper bounds the number of consecutive null steps after the descent step k_i by

$$k_{i+1} - k_i - 1 \leq (1-\beta/2)^{(n-1)-i}\,\frac{2M^2}{(1-\beta)^2\rho\epsilon} \ .$$

Summing this over $i = 0 \ldots n-1$ bounds the additional number of null steps before an ϵ-minimizer is found by (5.22) as

$$\sum_{i=0}^{n-1} (1-\beta/2)^{(n-1)-i}\,\frac{2M^2}{(1-\beta)^2\rho\epsilon} \leq \frac{4M^2}{\beta(1-\beta)^2\rho\epsilon} \ .$$

5.7 Proof of Theorem 2.4

Our bound on the number of descent steps comes directly from Theorem 2.3. Our claimed bound on the total number of null steps follows by multiplying this by the constant bound on the number of consecutive null steps from Lemma 5.2.

5.8 Proof of Theorem 2.5

Combining the lower bound $\Delta_k \geq \frac{1}{2}(f(x_k) - f^*)$ with Lemma 5.1 shows linear decrease in the objective every descent step

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\beta}{2}\right)(f(x_k - f^*).$$

Our bound on the number of descent steps follows immediately from this. Combining the lower bound $\Delta_k \geq \frac{1}{2}(f(x_k) - f^*)$ with Lemma 5.2 shows that at most

$$\frac{2M^2D^2}{(1-\beta)^2(f(x_{k+1}) - f^*)^2}$$

null steps occur between each descent step. Denote the sequence of descent steps taken by the bundle method by $k_1, k_2, k_3 \ldots$ and as a base case define $k_0 = -1$. Let k_n be the first descent step finding an ϵ-minimizer, which must have $n \leq \lceil \log_{(1-\beta/2)}(\frac{\epsilon}{f(x_0)-f^*})\rceil_+$. From our linear decrease condition, we know for any $i = 0, 1, 2, 3, \ldots n-1$

$$f(x_{k_i+1}) - f^* \geq (1-\beta/2)^{i-(n-1)}\,\epsilon$$

and from our null step bound, we know for any $i = 0, 1, 2, \ldots n-1$

$$k_{i+1} - k_i - 1 \leq \frac{2M^2D^2}{(1-\beta)^2(f(x_{k_i+1}) - f^*)^2} \leq (1-\beta/2)^{2(i-(n-1))}\frac{2M^2D^2}{(1-\beta)^2\epsilon^2}\ .$$

Then summing up our null step bounds ensures

$$k_n - n \leq \sum_{i=1}^{n}(1-\beta/2)^{2(i-1-(n-1))}\frac{2M^2D^2}{(1-\beta)^2\epsilon^2}.$$

Bounding this geometric series shows us that the bundle method finds an ϵ-minimizer with the number of null steps bounded by

$$\left(\frac{1}{1-(1-\beta/2)^2}\right)\frac{2M^2D^2}{(1-\beta)^2\epsilon^2}\ .$$

5.9 Proof of Theorem 2.6

Our bound on the number of descent steps follows from Theorem 2.5. Our proof of the null step bound follows the same approach as Theorem 2.5 with only minor differences. Applying Lemma 5.2 with our stepsize choice (2.9) bounds the number of consecutive null steps after some descent step k by

$$\frac{2M^2}{(1-\beta)^2\mu^{2/p}(f(x_{k+1}) - f^*)^{2-2/p}}\ .$$

Denote the descent steps $-1 = k_0 < k_1 < k_2 < \ldots$ and suppose the x_{k_n+1} is the first ϵ-minimizer. Then

$$k_{i+1} - k_i - 1 \leq (1-\beta/2)^{(2-2/p)(i-(n-1))}\frac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}}$$

since $f(x_{k_i+1}) - f^* \geq \left(1 - \frac{\beta}{2}\right)^{i-(n-1)}\epsilon$. Summing this up gives

$$k_n - n \leq \sum_{i=1}^{n}(1-\beta/2)^{(2-2/p)(i-1-(n-1))}\frac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}}\ .$$

When $p > 1$, this geometric series shows us that the bundle method finds an ϵ-minimizer with the number of null steps bounded by

$$\left(\frac{1}{1-(1-\beta/2)^{2-2/p}}\right)\frac{2M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}}.$$

When $p = 1$, we have a constant upper bound on the number of null steps following a descent step. Hence the number of null steps is bounded by

$$\frac{2M^2}{(1-\beta)^2\mu^2}\left\lceil\frac{\log\left(\frac{f(x_0)-f^*}{\epsilon}\right)}{-\log(1-\beta/2)}\right\rceil .$$

5.10 Proof of Theorem 3.1

Let $\delta_k = \min_{j\in\{0,\dots,J-1\}}\{f(x_k^{(j)})-f^*\}$ denote the lowest objective gap among all of our J instances of the bundle method after they have taken k synchronous steps. Then the core of our convergence proof is bounding the number of iterations where this lowest objective gap is in the interval

$$(1-\beta/2)^{-n}\epsilon \le \delta_k \le (1-\beta/2)^{-(n+1)}\epsilon .$$

for any integer $0 \le n < N := \left\lceil\frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)}\right\rceil$. Within this interval, we focus on the instance

$$j = \left\lceil\log_2\left(\frac{\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{1-2/p}}{4\bar{\rho}}\right)\right\rceil .$$

This instance of the bundle method's constant stepsize $\rho^{(j)} = 2^j\bar{\rho}$ approximates the stepsize (2.9) as

$$\frac{1}{4}\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{1-2/p} \le \rho^{(j)} \le \frac{1}{2}\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{1-2/p} .$$

Then (5.12) bounds this method's proximal gap before an $(1-\beta/2)^{-n}\epsilon$-minimizer is found by

$$\Delta_k^{(j)} \ge \frac{1}{2}(f(x_k^{(j)})-f^*) \ge (1-\beta/2)^{-n}\epsilon/2 .$$

Letting $\delta_k^{(j)} = f(x_k^{(j)})-f^*$, each descent step k improves method j's objective gap according to the recurrence $\delta_{k+1}^{(j)} \le \min\{(1-\beta/2)\delta_k^{(j)}, \delta_k\}$ where the first term in the minimum comes from Lemma 5.1 and the second term comes from method j taking any further improvement from the other bundle methods. By assumption, we have $\delta_k \le (1-\beta/2)^{-(n+1)}\epsilon$, and so after one descent step $k' > k$ we must have $\delta_{k'+1}^{(j)} \le (1-\beta/2)^{-(n+1)}\epsilon$. Thus after a second descent step $k'' > k'$, our intermediate target accuracy is met as $\delta_{k''+1} \le \delta_{k''+1}^{(j)} \le (1-\beta/2)^{-n}\epsilon$.

Applying Lemma 5.2 bounds the number of null steps between descent steps by

$$\frac{2M^2}{(1-\beta)^2\rho^{(j)}\Delta_{k+1}^{(j)}} \le \frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}} .$$

Hence the total number of steps before $\delta_k^{(j)} < 2^n\epsilon$ (and consequently $\delta_k < 2^n\epsilon$) is at most

$$2\left(\frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}}+1\right) .$$

Summing over this bound completes our proof. When $p > 1$, this gives

$$\begin{aligned}
&\sum_{n=0}^{N-1} 2\left(\frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}}+1\right)\\
&= 2\sum_{n=0}^{N-1} \frac{16M^2}{(1-\beta)^2\mu^{2/p}((1-\beta/2)^{-n}\epsilon)^{2-2/p}} + 2\left\lceil \frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)}\right\rceil\\
&\leq \left(\frac{2}{1-(1-\beta/2)^{2-2/p}}\right)\frac{16M^2}{(1-\beta)^2\mu^{2/p}\epsilon^{2-2/p}} + 2\left\lceil \frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)}\right\rceil \ .
\end{aligned}$$

When $p = 1$, the number of steps in each of our intervals is constant. Consequently, the total number of iterations before an ϵ minimizer is found is at most

$$\sum_{n=0}^{N-1} 2\left(\frac{16M^2}{(1-\beta)^2\mu^2}+1\right) = 2\left(\frac{16M^2}{(1-\beta)^2\mu^2}+1\right)\left\lceil \frac{\log((f(x_0)-f^*)/\epsilon)}{-\log(1-\beta/2)}\right\rceil \ .$$

References

[1] *Libsvm data: Classification (binary class).* `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html`. Accessed: 2021-05-12.

[2] P. Apkarian, D. Noll, and O. Prot, *A Proximity Control Algorithm to Minimize Nonsmooth and Nonconvex Semi-infinite Maximum Eigenvalue Functions*, J. Convex Anal., 16 (2009), pp. 641–666.

[3] J. V. Burke and M. C. Ferris, *Weak sharp minima in mathematical programming*, SIAM Journal on Control and Optimization, 31 (1993), pp. 1340–1359.

[4] D. Davis and D. Drusvyatskiy, *Stochastic model-based minimization of weakly convex functions*, SIAM Journal on Optimization, 29 (2019), p. 207–239.

[5] W. de Oliveira, *Proximal bundle methods for nonsmooth DC programming*, J. Glob. Optim., 75 (2019), pp. 523–563.

[6] W. de Oliveira, C. A. Sagastizábal, and C. Lemaréchal, *Convex proximal bundle methods in depth: a unified analysis for inexact oracles*, Math. Program., 148 (2014), pp. 241–277.

[7] W. de Oliveira and M. Solodov, *Bundle Methods for Inexact Data*, Springer International Publishing, Cham, 2020, pp. 417–459.

[8] L. Ding and B. Grimmer, *Revisit of spectral bundle methods: Primal-dual (sub)linear convergence rates*, 2020.

[9] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, *Nonsmooth optimization using taylor-like models: error bounds, convergence, and termination criteria*, Math. Program., 185 (2021), pp. 357–383.

[10] Y. Du and A. Ruszczyński, *Rate of Convergence of the Bundle Method*, J. Optim. Theory Appl., 173 (2017), pp. 908–922.

[11] A. Frangioni, *Standard Bundle Methods: Untrusted Models and Duality*, Springer International Publishing, Cham, 2020, pp. 61–116.

[12] W. Hare and C. Sagastizábal, *A Redistributed Proximal Bundle Method for Nonconvex Optimization*, SIAM J. Optim., 20 (2010), pp. 2442–2473.

[13] W. Hare, C. Sagastizábal, and M. Solodov, *A Proximal Bundle Method for Nonsmooth Nonconvex Functions with Inexact Information*, Computational Optimization and Applications, 63 (2016), pp. 1–28.

[14] C. Helmberg and F. Rendl, *A spectral bundle method for semidefinite programming*, SIAM Journal on Optimization, 10 (2000), pp. 673–696.

[15] J.-B. Hiriart-Urruty and C. Lemaréchal, *Acceleration of the Cutting-Plane Algorithm: Primal Forms of Bundle Methods*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1993, pp. 275–330.

[16] K. C. Kiwiel, *An Aggregate Subgradient Method for Nonsmooth Convex Minimization*, Math. Program., 27 (1983), pp. 320–341.

[17] ——, *A Linearization Algorithm for Nonsmooth Minimization*, Mathematics of Operations Research, 10 (1985), pp. 185–194.

[18] K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization.*, Springer, Berlin, 1985.

[19] K. C. Kiwiel, *Proximal Level Bundle Methods for Convex Nondifferentiable Optimization, Saddle-point Problems and Variational Inequalities*, Math. Program., 69 (1995), pp. 89–109.

[20] ——, *Efficiency of Proximal Bundle Methods*, Journal of Optimization Theory and Applications, 104 (2000), pp. 589–603.

[21] G. Lan, *Bundle-Level Type Methods Uniformly Optimal For Smooth And Nonsmooth Convex Optimization*, Mathematical Programming, 149 (2015), pp. 1–45.

[22] C. Lemarechal, *An Extension of Davidon Methods to Nondifferentiable Problems*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1975, pp. 95–109.

[23] C. Lemaréchal, *Lagrangian Relaxation*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001, pp. 112–156.

[24] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov, *New Variants of Bundle Methods*, Math. Program., 69 (1995), pp. 111–147.

[25] J. Liang and R. D. C. Monteiro, *A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes*, 2021.

[26] J. Lv, L. Pang, and F. Meng, *A proximal bundle method for constrained nonsmooth nonconvex optimization with inexact information*, J. Glob. Optim., 70 (2018), pp. 517–549.

[27] J. Lv, L. Pang, N. Xu, and Z. Xiao, *An infeasible bundle method for nonconvex constrained optimization with application to semi-infinite programming problems*, Numer. Algorithms, 80 (2019), pp. 397–427.

[28] R. Mifflin, *A Modification and an Extension of Lemarechal's Algorithm for Nonsmooth Minimization*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1982, pp. 77–90.

[29] N. H. Monjezi and S. Nobakhtian, *A new infeasible proximal bundle algorithm for nonsmooth nonconvex constrained optimization*, Comput. Optim. Appl., 74 (2019), pp. 443–480.

[30] ——, *A filter proximal bundle method for nonsmooth nonconvex constrained optimization*, J. Glob. Optim., 79 (2021), pp. 1–37.

[31] P. Ochs, J. Fadili, and T. Brox, *Non-smooth non-convex bregman minimization: Unification and new algorithms*, J. Optim. Theory Appl., 181 (2019), pp. 244–278.

[32] F. Oustry, *A second-order bundle method to minimize the maximum eigenvalue function*, Math. Program., 89 (2000), pp. 1–33.

[33] J. Renegar and B. Grimmer, *A Simple Nearly-Optimal Restart Scheme For Speeding-Up First Order Methods*, Foundations of Computational Mathematics, (2021).

[34] A. Ruszczynski, *Nonlinear Optimization*, Princeton University Press, Princeton, NJ, USA, 2006.

[35] C. Sagastizábal, *Divide to Conquer: Decomposition Methods for Energy Optimization*, Mathematical Programming, 134 (2012), pp. 187–222.

[36] C. Sagastizábal and M. Solodov, *An Infeasible Bundle Method for Nonsmooth Convex Constrained Optimization without a Penalty Function or a Filter*, SIAM Journal on Optimization, 16 (2005), pp. 146–169.

[37] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, *Pegasos: Primal estimated sub-gradient solver for svm*, Mathematical programming, 127 (2011), pp. 3–30.

[38] P. Wolfe, *A Method of Conjugate Subgradients for Minimizing Nondifferentiable Functions*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1975, pp. 145–173.

A Solutions to Recurrence Relations

Throughout our analysis, we frequently encounter recurrence relations of the form $\delta_{k+1} \leq \delta_k - \alpha\delta_k^q$ for some $\alpha > 0$ and $q > 1$. The follow lemma bounds the number of steps of such a recurrence to reach a desired level of accuracy $\delta_k \leq \epsilon$.

Lemma A.1. *For any $\epsilon > 0$, the recurrence $\delta_{k+1} \leq \delta_k - \alpha\delta_k^q$ has $\delta_k \leq \epsilon$ satisfied by some*

$$k \leq \left\lceil \frac{1}{(q-1)\alpha\epsilon^{q-1}} \right\rceil .$$

Proof. It suffices to show the following upper bound on δ_k as a function of k

$$\delta_k \leq \left(\frac{1}{(q-1)\alpha k}\right)^{1/(q-1)} .$$

First we show this bound holds with $k = 1$. This follows as

$$\delta_1 \leq \delta_0 - \alpha\delta_0^q \leq \max_{\delta\in\mathbb{R}}\{\delta - \alpha\delta^q\} \leq \left(\frac{1}{q\alpha}\right)^{1/(q-1)} .$$

Then we complete our proof by induction using the following *weighted arithmetic-geometric mean (AM-GM) inequality*, which ensures for any $a, \alpha, b, \beta > 0$ we have $a^\alpha b^\beta \leq \left(\frac{\alpha a+\beta b}{\alpha+\beta}\right)^{\alpha+\beta}$. This implies that for any $k \geq 1$, $(k-(q-1)^{-1})(k+1)^{1/(q-1)} \leq k^{q/(q-1)}$ by taking $a = k-(q-1)^{-1}$, $\alpha = 1$, $b = k+1$, $\beta = 1/(q-1)$. By expanding the recurrence at $k+1$ and applying this inequality we get

$$\begin{aligned}
\delta_{k+1} \leq \delta_k - \alpha\delta_k^q &\leq \left(\frac{1}{(q-1)\alpha k}\right)^{1/(q-1)} - \alpha\left(\frac{1}{(q-1)\alpha k}\right)^{q/(q-1)} \\
&= \left(\frac{1}{(q-1)\alpha}\right)^{1/(q-1)} \left(\frac{k}{k^{q/(q-1)}} - \frac{1}{(q-1)k^{q/(q-1)}}\right) \\
&= \left(\frac{1}{(q-1)\alpha}\right)^{1/(q-1)} \frac{k-(q-1)^{-1}}{k^{q/(q-1)}} \\
&\leq \left(\frac{1}{(q-1)\alpha(k+1)}\right)^{1/(q-1)} .
\end{aligned}$$

Proving the result. □