Some Primal-Dual Theory for Subgradient Methods for Strongly Convex Optimization

Benjamin Grimmer[*] Danlin Li[†]

Abstract

We consider (stochastic) subgradient methods for strongly convex but potentially nonsmooth non-Lipschitz optimization. We provide new equivalent dual descriptions (in the style of dual averaging) for the classic subgradient method, the proximal subgradient method, and the switching subgradient method. These equivalences enable $O(1/T)$ convergence guarantees in terms of both their classic primal gap and a not previously analyzed dual gap for strongly convex optimization. Consequently, our theory provides these classic methods with simple, optimal stopping criteria and optimality certificates at no added computational cost. Our results apply to a wide range of stepsize selections and of non-Lipschitz ill-conditioned problems where the early iterations of the subgradient method may diverge exponentially quickly (a phenomenon which, to the best of our knowledge, no prior works address). Even in the presence of such undesirable behaviors, our theory still ensures and bounds eventual convergence.

1 Introduction

The study of gradient methods for iteratively solving nonsmooth convex minimization problems dates back to as early as the 60s, see [1]. In recent decades, interest in first-order methods for optimization has resurged in popularity throughout data science and machine learning domains due to their low iteration cost and scalability. This has led to the development of a range of new gradient methods [2–9]. Here, we instead focus on improving performance guarantees for classic subgradient methods, the natural extensions of gradient descent to nonsmooth settings.

We consider general convex minimization problems of the following form

$$p_\star = \begin{cases} \min & f_0(x) + r(x) \\ \text{s.t.} & f_s(x) \leq 0 \qquad \forall s = 1 \dots m \end{cases} \tag{1.1}$$

where the functions $f_s \colon \mathcal{E} \to \mathbb{R}$, $s = 0, \dots, m$, are (strongly) convex but may be nonsmooth and not globally Lipschitz continuous and $r \colon \mathcal{E} \to \mathbb{R} \cup \{\infty\}$ is convex, closed, and simple, all defined over a finite-dimensional Euclidean space $\mathcal{E}$. We will consider iteratively solving problems of this general form via a stochastic switching proximal subgradient method. This general method corresponds to the classic subgradient method when $r = 0$ and $m = 0$, the proximal subgradient method when $m = 0$, and the switching subgradient method of [10] when $r = 0$. Formal assumptions and definitions are deferred to Section 2.

This work provides equivalent dual descriptions and new primal-dual convergence rates for all of these classic subgradient methods. Although our theory will be developed for stochastic,

[*] Johns Hopkins University, Department of Applied Mathematics and Statistics, grimmer@jhu.edu

[†] Johns Hopkins University, Department of Applied Mathematics and Statistics, dli91@alumni.jh.edu

non-Lipschitz problems with $r \neq 0$ and $m \neq 0$, we first briefly present our results without these generalities to showcase and introduce the key ideas.

The Classic Setting: Primal-Dual Theory for the Subgradient Method.
Supposing $r = 0$ and $m = 0$, the problem (1.1) reduces to unconstrained minimization of $f_0 \colon \mathcal{E} \to \mathbb{R}$. We assume access to an oracle capable of producing a subgradient at each iteration $g_0(x) \in \partial f_0(x) := \{g \mid f_0(y) \geq f_0(x) + \langle g, y - x\rangle \quad \forall y \in \mathcal{E}\}$. Note when f_0 is μ-strongly convex (that is, $f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex), each subgradient $g_0(x) \in \partial f_0(x)$, provides a quadratic lower bound

$$f_0(y) \geq f_0(x) + \langle g_0(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2 \quad \forall y \in \mathcal{E} \ . \tag{1.2}$$

To develop a primal-dual understanding, we consider the following reformulation

$$\begin{aligned} p_\star = \min_{x\in\mathcal{E}} f_0(x) &= \min_{x\in\mathcal{E}} \max_{y\in\mathcal{E}} f_0(y) + \langle g_0(y), x - y\rangle + \frac{\mu}{2}\|x - y\|_2^2 \\ &= \min_{x\in\mathcal{E}} \max_{\substack{y_0,\dots y_T \in \mathcal{E} \\ \lambda_0,\dots\lambda_T \geq 0 \\ \sum_{k=0}^T \lambda_k > 0}} \frac{\sum_{k=0}^T \lambda_k \left(f_0(y_k) + \langle g_0(y_k), x - y_k\rangle + \frac{\mu}{2}\|x - y_k\|_2^2\right)}{\sum_{k=0}^T \lambda_k} \end{aligned}$$

where the first line replaces f_0 by the maximum of its quadratic subgradient lower bounds and the second extends this to the maximum combination of such lower bounds. Any point $x \in \mathcal{E}$ provides a "primal solution" with primal gap $f_0(x) - p_\star$. Any collection of points y_k and weights λ_k provide a "dual solution", which produces a lower bound on $f_0(x)$ of $m^{(T)}(x) = \frac{\sum_{k=0}^T \lambda_k\left(f_0(y_k) + \langle g_0(y_k), x-y\rangle + \frac{\mu}{2}\|x-y_k\|_2^2\right)}{\sum_{k=0}^T \lambda_k}$ and a dual gap $p_\star - \inf m^{(T)}$.

The subgradient method builds a sequence of primal solutions $\{x_k\}$ by repeatedly moving in negative subgradient directions using a sequence of stepsizes $\alpha_k > 0$

$$x_{k+1} = x_k - \alpha_k g_0(x_k) \ . \tag{1.3}$$

The method of dual averaging builds dual solutions yielding models $m^{(k)}$ by repeatedly minimizing this lower bounding model (with an optional β_k regularization term) and then incorporating a new subgradient $g_0(y_{k+1})$ into the next model from the resulting point y_{k+1} with weight $\lambda_{k+1} > 0$

$$\begin{cases} m^{(k)}(y) = \dfrac{\sum_{i=0}^k \lambda_i \left(f_0(y_i) + \langle g_0(y_i), y - y_i, \rangle + \frac{\mu}{2}\|y - y_i\|_2^2\right)}{\sum_{i=0}^k \lambda_i} \\ y_{k+1} = \operatorname{argmin}\left\{ m^{(k)}(y) + \dfrac{\beta_k}{2\sum_{i=0}^k \lambda_i}\|y - y_0\|_2^2 \right\} \ . \end{cases} \tag{1.4}$$

If one sets $\mu = 0$, this corresponds to the dual averaging method of Nesterov [3]. Other variations of dual averaging from the literature are discussed in Section 2.1.

In the not strongly convex setting of $\mu = 0$, Nesterov [3] showed when $\beta_k = \bar{\beta} > 0$ is constant, these two methods are equivalent whenever $\alpha_k = \lambda_k/\bar{\beta}$. That is, they produce the same sequence of iterates $x_k = y_k$. Our Theorem 3.1 extends this equivalence to potentially strongly convex settings, showing these two methods are equivalent whenever $\alpha_k = \lambda_k/(\mu\sum_{i=0}^k \lambda_i + \bar{\beta})$ and $\mu + \bar{\beta} > 0$. Such results allow one to equivalently view the classic subgradient method as either iteratively building a primal solution converging down to optimal or building a lower bound converging up to optimal.

This new dual understanding enables us to develop primal-dual convergence theory for the subgradient method. We find the dual model built by the subgradient method converges up to optimal at the same rate as the iterates converge down. For example, if $\bar{\beta} = 0$ and $\|g_0(x_k)\|_2 \leq M$ uniformly, a special case of our Theorem 3.2 implies for any selection of α_k and λ_k with $\alpha_k = \lambda_k/(\mu\sum_{i=0}^k \lambda_i)$, the subgradient method (or equivalently dual averaging) has primal gap, dual gap, and distance to optimality all converge with

$$\underbrace{\frac{\sum_{k=0}^T \lambda_k f(x_k)}{\sum_{k=0}^T \lambda_k} - p_\star}_{\texttt{Primal Gap}} + \underbrace{p_\star - \inf m^{(T)}}_{\texttt{Dual Gap}} + \underbrace{\frac{\mu}{2}\|x_{T+1} - x_{\texttt{OPT}}\|_2^2}_{\texttt{Distance To Optimal}} \leq M^2 \frac{\sum_{k=0}^T \lambda_k \alpha_k}{\sum_{k=0}^T \lambda_k} . \tag{1.5}$$

Selecting dual weights $\lambda_k = k+1$, corresponding to primal stepsizes $\alpha_k = 2/\mu(k+2)$ as considered in [11, Section 3.2], this recovers and extends their optimal primal rate $O(M^2/\mu T)$ to have a matching dual term. This dual theory enables a computable optimal stopping criteria (assuming μ is known) as a primal-dual gap $\frac{\sum_{k=0}^T \lambda_k f_0(x_k)}{\sum_{k=0}^T \lambda_k} - \inf m^{(T)}$ less than ϵ occurs within $O(1/\epsilon)$ steps ensuring both primal and dual accuracies of at least ϵ. To the best of our knowledge, no such criterion has been known, even for the classic subgradient method.

Given our primal-dual equivalence, our results can be equally seen as providing primal-dual convergence guarantees for dual averaging. In this sense, we improve the prior best dual averaging theory due to Deng et al. [12, Corollary 8] who showed a primal rate of $O\left(M^2/\mu T\right)$ when $\lambda_k = k+1$ and $\beta_k = \mu(T+2)$ is constant. Note such rates are optimal (by the example presented in [13]), meaning no faster objective gap convergence rate in terms of any of M, μ, T can be achieved.

1.1 Our Contributions

This work develops primal-dual equivalences and convergence theory beyond the above classic subgradient method setting. We consider the general problem (1.1) including an additive composite objective, functional constraints, and stochastic subgradients.

- **Dual Equivalences and Primal-Dual Convergence Theory.** Our theory considers a *Stochastic Switching Proximal Subgradient Method* for the general problem class (1.1), which, as special cases, captures projected, proximal, and switching subgradient methods as well as gradient descent. We introduce a new dual averaging method for this general problem class, *Stochastic Lagrangian Proximal Dual Averaging*, which our Theorem 3.1 shows is equivalent to the stochastic switching proximal subgradient method under proper selection of primal stepsizes α_k and dual weights λ_k and β_k. From this equivalence, our Theorem 3.2 presents new primal-dual convergence rate guarantees for these general methods.
- **Computable Stopping Criteria.** Our theory identifies dual certificates implicitly built by the range of considered subgradient methods. These certificates enable new computable stopping criteria (assuming μ is known), halting once the primal-dual gap is at most a target accuracy ϵ. The associated Lagrange multipliers may further be valuable when the subgradient method is used as a subroutine of a larger computation.
- **New Non-Lipschitz Analysis Bounds for Early Divergence Phenomena.** Often, nonsmooth optimization analysis focuses on Lipschitz continuous functions. Such theory is limited to functions that asymptotically grow at most linearly. Our analysis uses a non-Lipschitz model, allowing up to quadratic growth. By doing so, our theory provides new linear primal-dual convergence guarantees for gradient descent in smooth optimization and new guarantees for minimizing a sum $f_0 = h_1 + h_2$ with smooth h_1 and nonsmooth but Lipschitz h_2, which is overall neither Lipschitz nor smooth. Numerics showcasing highly non-monotone behaviors of subgradient methods on such problems are shown in Section 4, where our theory still provides reasonably accurate predictions.

Outline. Section 2 introduces the considered primal and dual subgradient methods, related literature, and the assumptions needed for our theory. Section 3 states and proves our equivalence between these primal and dual perspectives and our improved primal-dual guarantees. Finally, Section 4 concludes with some numerical validation.

2 Preliminaries and Algorithm Definitions

Recall we are interested in the family of problems

$$\begin{cases} \min & f_0(x) + r(x) \\ \text{s.t.} & f_s(x) \le 0 \qquad \forall s = 1 \dots m \ . \end{cases}$$

This extends the previously discussed classic subgradient method setting in three ways.

First, we allow for additive composite objectives $f_0 + r$ for any closed convex $r\colon \mathcal{E} \to \mathbb{R} \cup \{\infty\}$. We address this added term by assuming r is sufficiently simple that its proximal operator $\operatorname{prox}_{\alpha_k, r}(z) := \operatorname{argmin}_x \left\{ r(x) + \frac{1}{2\alpha_k}\|x - z\|_2^2 \right\}$ can be evaluated at each iteration. For example, if r is an indicator function for a closed convex constraint set, its proximal operator corresponds to projection onto that set. Since $r(x) + \frac{1}{2\alpha_k}\|x - z\|_2^2$ is strongly convex, it has a unique minimizer described by the following equivalence

$$z_+ = \operatorname{prox}_{\alpha_k, r}(z) \qquad \Longleftrightarrow \qquad \frac{1}{\alpha_k}(z - z_+) \in \partial r(z_+) \ . \tag{2.1}$$

Second, we allow for m strongly convex functional constraints $f_s(x) \le 0$ for $s = 1, \dots, m$. We address these added terms by "switching": Given a current iterate x, only one function $f_{s(x)}$ will be considered in that iteration. This function will be chosen as $s(x) = 0$ if x is feasible (that is, $f_s(x) \le 0$ for all $s = 1 \dots m$), otherwise $s(x)$ can be chosen generically as any violated constraint $f_{s(x)}(x) > 0$. Third, we allow for stochasticity in our subgradient oracles for each function, denoted by $g_s(x;\xi)$ such that $\mathbb{E}_\xi g_s(x;\xi) \in \partial f_s(x)$. Note this trivially captures deterministic methods by selecting $g_s(x;\xi)$ constant with respect to ξ.

Below, we introduce our considered primal and dual subgradient methods for solving such problems. We introduce these using disjoint notations but will in Theorem 3.1 show these are, in fact, the same algorithm in that their iterate sequences are identical.

A General Primal Subgradient Method. As a primal algorithm, consider the *Stochastic Switching Proximal Subgradient Method* with stepsizes $\alpha_k > 0$ and sequence of iterates $\{x_k\}$ defined by

$$x_{k+1} = \begin{cases} \operatorname{prox}_{\alpha_k, r}(x_k - \alpha_k g_0(x_k;\xi_k)) & \text{if } x_k \text{ is feasible} \\ x_k - \alpha_k g_{s(x_k)}(x_k;\xi_k) & \text{otherwise.} \end{cases} \tag{2.2}$$

for i.i.d. sampled ξ_k. Throughout, we always assume $\alpha_k > 0$, strictly. When $m = 0$, this is the standard (stochastic) proximal subgradient method; when $r = 0$, this is the (stochastic) switching subgradient method. Note when $m > 0$, only limited stochasticity can be allowed since an exact determination of feasibility is required to decide the switching variable $s(x_k)$.

A General Dual Subgradient Method. To give a dual approach to solving (1.1), consider the following equivalent minimax formulation

$$\begin{aligned}
p_\star &= \min_{x\in\mathcal{E}} \max_{u_1,\dots u_s\geq 0} f_0(x) + \sum_{s=1}^{m} u_s f_s(x) + r(x) \\
&= \min_{x\in\mathcal{E}} \max_{\substack{y_0,\dots,y_T\in\mathcal{E} \\ \lambda_0,\dots,\lambda_{T-1}\geq 0 \\ \sum_{k<T:\ s(y_k)=0}\lambda_k>0 \\ n_{k+1}\in\partial r(y_{k+1})}} \mathbb{E}_\xi \left(\sum_{s=0}^{m} \sum_{\substack{k<T \\ s(y_k)=s}} \lambda_k \left(f_{s(y_k)}(y_k) + \langle g_{s(y_k)}(y_k;\xi_k), x-y_k\rangle + \frac{\mu}{2}\|x-y_k\|_2^2 \right) \right. \\
&\qquad \left. + \sum_{\substack{k<T \\ s(y_k)=0}} \lambda_k \left(r(y_{k+1}) + \langle n_{k+1}, x-y_{k+1}\rangle \right) \right) \frac{1}{\sum_{k<T:\ s(y_k)=0}\lambda_k}
\end{aligned}$$

where the first equality is the standard primal Lagrangian formulation and the second equality replaces each function by a combination of its lower bounds. Hence, any selection of values for y_k, λ_k, n_k gives a dual solution and a lower bound on $p_\star$.

As a dual algorithm, consider the following *Stochastic Lagrangian Proximal Dual Averaging* with dual weights $\lambda_k > 0$ and regularization parameters $\beta_k \geq 0$. Throughout, we always assume $\lambda_k > 0$, strictly. Its sequence of iterates $\{y_k\}$ based on i.i.d. sampled ξ_k is defined as follows: At iteration k, construct the following (unnormalized) aggregations for each function based on the previous $k-1$ iterations as

$$\begin{aligned}
F_s^{(k-1)}(y) &:= \sum_{i<k:\ s(y_i)=s} \lambda_i \left(f_{s(y_i)}(y_i) + \langle g_{s(y_i)}(y_i;\xi_i), y-y_i\rangle + \frac{\mu}{2}\|y-y_i\|_2^2 \right) \\
R^{(k-1)}(y) &:= \sum_{i<k:\ s(y_i)=0} \lambda_i \left(r(y_{i+1}) + \langle n_{i+1}, y-y_{i+1}\rangle \right) \\
M^{(k-1)}(y) &:= \sum_{s=0}^{m} F_s^{(k-1)}(y) + R^{(k-1)}(y)
\end{aligned}$$

where $n_{i+1} = \frac{-1}{\lambda_i}(\nabla M^{(i-1)}(y_{i+1}) + \lambda_i(g_0(y_i;\xi_i) + \mu(y_{i+1}-y_i)) + \beta_i(y_{i+1}-y_0)) \in \partial r(y_{i+1})$ (see Lemma 2.1 for verification of this subdifferential containment). At iteration $k=0$, these empty summations are understood to take value zero. Then, based on the switching selection $s(y_k)$, a new weighted model is constructed as

$$U^{(k)}(y) := \begin{cases} \lambda_k \left(f_0(y_k) + \langle g_0(y_k;\xi_k), y-y_k\rangle + \frac{\mu}{2}\|y-y_k\|_2^2 + r(y) \right) & \text{if } y_k \text{ is feasible} \\ \lambda_k \left(f_{s(y_k)}(y_k) + \langle g_{s(y_k)}(y_k;\xi_k), y-y_k\rangle + \frac{\mu}{2}\|y-y_k\|_2^2 \right) & \text{otherwise.} \end{cases}$$

The Lagrangian proximal dual averaging method then iterates by minimizing the aggregation of past models $M^{(k-1)}$ plus the new model $U^{(k)}$ (and an optional extra regularization term)

$$y_{k+1} = \operatorname{argmin} \left\{ M^{(k-1)}(y) + U^{(k)}(y) + \frac{\beta_k}{2}\|y-y_0\|_2^2 \right\} . \tag{2.3}$$

Note our definitions for the updated aggregate model $M^{(k)}$ are chosen such that y_{k+1} will also be the unique minimizer of $M^{(k)}(y) + \frac{\beta_k}{2}\|y-y_0\|_2^2$.

Lemma 2.1. *y_{k+1} is the unique minimizer of $M^{(k)}(y) + \frac{\beta_k}{2}\|y-y_0\|_2^2$.*

Proof. First note that when y_k is infeasible (and so $s(y_k) \neq 0$), $M^{(k)} = M^{(k-1)} + U^{(k)}$. From this, the result is immediate. Now consider when y_k is feasible (and so $s(y_k) = 0$). Since y_{k+1} is the unique minimizer of $M^{(k-1)}(y) + U^{(k)}(y) + \frac{\beta_k}{2}\|y - y_0\|_2^2$, the necessary and sufficient optimality condition ensures y_{k+1} is the unique solution to

$$0 \in \nabla M^{(k-1)}(y_{k+1}) + \lambda_k(g_0(y_k;\xi_k) + \mu(y_{k+1} - y_k) + \partial r(y_{k+1})) + \beta_k(y_{k+1} - y_0) \ .$$

Rewriting this as $\nabla M^{(k-1)}(y_{k+1}) + \lambda_k(g_0(y_k;\xi_k) + \mu(y_{k+1} - y_k)) + \beta_k(y_{k+1} - y_0) \in -\lambda_k \partial r(y_{k+1})$, we see that n_{k+1} is the element of $\partial r(y_{k+1})$ certifying the minimization's optimality. Therefore

$$\begin{aligned} 0 &= \nabla M^{(k-1)}(y_{k+1}) + \lambda_k(g_0(y_k;\xi_k) + \mu(y_{k+1} - y_k) + n_{k+1}) + \beta_k(y_{k+1} - y_0) \\ &= \nabla M^{(k)}(y_{k+1}) + \beta_k(y_{k+1} - y_0) \end{aligned}$$

and so y_{k+1} is a the unique minimizer of $M^{(k)}(y) + \frac{\beta_k}{2}\|y - y_0\|_2^2$. □

Note whenever $s(y_k) = 0$, the step (2.3) corresponds to minimizing r plus a simple quadratic. This amounts to computing a proximal operator for r and so is within the assumed computational oracle model. Whenever $s(y_k) \neq 0$, the step (2.3) minimizes a simple quadratic and can be done in closed form. The following easily verifiable lemmas provide a simple way to maintain the minimum of the aggregate quadratic model.

Lemma 2.2. *Any quadratic function of the form* $Q(z) = c + \langle d, z\rangle + \frac{b}{2}\|z - \hat{z}\|_2^2$ *with* $b > 0$ *is equal to* $Q(z) = \min Q + \frac{b}{2}\|z - \operatorname{argmin} Q\|_2^2$.

Lemma 2.3. *The sum of quadratics* $Q_i(z) = a_i + \frac{b_i}{2}\|z - z_i\|_2^2$ *for* $i = 1, 2$ *with* $b_i > 0$ *equals*

$$(Q_1 + Q_2)(z) = a_{1+2} + \frac{b_{1+2}}{2}\|z - z_{1+2}\|_2^2$$

where $a_{1+2} = a_1 + a_2 + \frac{b_1 b_2}{2(b_1+b_2)}\|z_1 - z_2\|_2^2$, $b_{1+2} = b_1 + b_2$, *and* $z_{1+2} = \frac{b_1}{b_1+b_2}z_1 + \frac{b_2}{b_1+b_2}z_2$.

2.1 Related Work

More General Distance Terms in Dual Minimization. Nesterov's original development [3] and most of the subsequent literature have considered a slightly different model for dual averaging than discussed here. To the best of our knowledge, no previous dual averaging methods handled functional constraints. Instead, they fix $m = 0$. Prior works have primarily fixed $\mu = 0$ (not utilizing the quadratic improvement in lower bound quality from strong convexity) but allowed a more generic distance function in the second term of dual averaging's objective to be minimized at each step. The "standard" dual averaging iteration for unconstrained minimization of f_0 is then

$$\begin{cases} m^{(k)}(x) = \frac{\sum_{i=0}^k \lambda_i(f(x_i) + \langle g_0(x_i;\xi_i), x - x_i\rangle)}{\sum_{i=0}^k \lambda_i} \\ x_{k+1} = \operatorname{argmin}\left\{ m^{(k)}(x) + \frac{\beta_k}{2\sum_{i=0}^k \lambda_i} d(x) \right\} \end{cases} \tag{2.4}$$

for any ρ-strongly convex $d(x)$. Our equivalent dual perspective fundamentally relies on these improvements in the subgradient lower bounds and the distance function both being quadratics $\|x - x_i\|_2^2$ and, as a result, are directly relatable. We do not expect our theory to generalize easily for more generic distance functions.

Regularized Dual Averaging. A closely related method to the proximal subgradient method is *Regularized Dual Averaging* proposed by Xiao [14] and further extended by [12,15–18]. This method applies to unconstrained additive composite problems minimizing $f_0 + r$ by iterating

$$\begin{cases} m^{(k)}(x) = \frac{\sum_{i=0}^{k} \lambda_i\left(f_0(x_i)+\langle g_0(x_i;\xi_i), x-x_i\rangle + \frac{\mu}{2}\|x-x_i\|_2^2\right)}{\sum_{i=0}^{k}\lambda_i} + r(x) \\ x_{k+1} \in \operatorname{argmin}\left\{m^{(k)}(x) + \frac{\beta_k}{2\sum_{i=0}^{k}\lambda_i}\|x-x_0\|_2^2\right\} \ . \end{cases} \tag{2.5}$$

This method's original development in [14] fixed $\mu = 0$ but allowed for a more general distance function, as discussed above. Based on our Theorem 3.1, regularized dual averaging can be seen as a natural improvement on the proximal subgradient method. Regularized Dual Averaging utilizes r entirely in its model function, whereas our equivalent dual description of the proximal subgradient method ((2.3) specialized to this case, $m = 0$) uses the mixture of subgradient lower bounds and r, iterating

$$\begin{cases} m^{(k)}(x) = \frac{\sum_{i=0}^{k} \lambda_i\left(f_0(x_i)+\langle g_0(x_i;\xi_i), x-x_i\rangle + \frac{\mu}{2}\|x-x_i\|_2^2\right)}{\sum_{i=0}^{k}\lambda_i} + \frac{\sum_{i=0}^{k-1}\lambda_i(r(x_{i+1})+\langle n_{i+1}, x-x_{i+1}\rangle)+\lambda_k r(x)}{\sum_{i=0}^{k}\lambda_i} \\ x_{k+1} \in \operatorname{argmin}\left\{m^{(k)}(x) + \frac{\beta_k}{2\sum_{i=0}^{k}\lambda_i}\|x-x_0\|_2^2\right\} \ . \end{cases} \tag{2.6}$$

Switching Subgradient Method Guarantees. Convergence rate guarantees for the switching subgradient method of [10] have been extensively studied for convex Lipschitz minimization [19–22] and more recently for non-Lipschitz settings [23]. Our theory extends this prior theory to give matching dual bounds and identifies Lagrange multipliers $u_s = \frac{\sum_{k:s(y_k)=s}\lambda_k}{\sum_{k:s(y_k)=0}\lambda_k}$ implicitly built by this classic method (at no added cost). Recently, nonconvex guarantees were developed by [24] but are beyond our scope.

Convex Conjugate-type Convergence Analysis. The recent series of works of Peña and Gutman [25–27] developed unified convergence guarantees for convex optimization ($\mu = 0$) for a range of first-order methods from accelerated smooth methods to nonsmooth Bregman and conditional subgradient methods. Beyond just showing convergence of the objective gap, these works showed convergence of perturbed primal-dual quantities based on aggregating (sub)gradient information. This work shares a similar spirit but addresses the setting of $\mu > 0$. Strong convexity ensures our non-perturbed dual gaps are finite, a necessity for our theory.

The recent work of Diakonikolas and Orecchia [28] developed first-order methods by discretizing continuous-time dynamical systems with decreasing gaps between aggregated upper and lower bounds on optimality. This technique is able to recover dual averaging among many other standard first-order methods. Although they only provide primal guarantees, their approach may be extendable to bound dual gaps.

Prior Primal Weighted Averaging Analysis. The value and importance of returning a carefully chosen weighted combination $\sum_{k=0}^{T-1}\sigma_k x_k$ of subgradient method iterates has been studied by several prior works. Rakhlin et al. [29, Theorem 5] showed uniformly averaging the last $q \in (0,1)$ fraction of iterates (called q-suffix averaging) can lead to an optimal $O(1/T)$ primal convergence rate for strongly convex minimization. Shamir and Zhang [30, Theorem 3 and 4] improved this theory and additionally showed polynomial weightings yield the same optimal rate with less computational overhead. The s^k-stepsize rule developed by Gustavsson et al. [31, Section 3.3] builds substantial theory for such polynomial stepsizes and weightings $\sigma_k = (k+1)^p$. The $\sigma_k = k+1$ choice of [11] amounts to the simplest polynomial weighting choice. Our theory provides a novel insight into the source of these iterate aggregation weights: primal-dual guarantees hold for any stepsizes α_k if the averaging used is proportional to the stepsize's corresponding dual weights $\sigma_k \propto \lambda_k$.

Alternative Lagrangian Dual Averaging Approaches. Our proposed Lagrangian Proximal Dual Averaging Method (2.3) implicitly sets the dual multipliers based on the frequency/total weight of steps taken on each constraint function, after which the iteration amounts to repeated model minimization. Alternatively, one could apply dual averaging to the Lagrangian minimax problem directly. Such an approach is proposed and analyzed by Metel and Takeda [32].

2.2 Assumptions for our Convergence Theory

Our primal-dual convergence rate theory relies on three assumptions. Our first two assumptions are standard, strong convexity and the existence of a Slater point.

Assumption A. The functions f_s for $s = 0 \dots m$ are each $\mu > 0$-strongly convex.

Assumption B. There exists some $x_{\texttt{SL}} \in \mathrm{dom}\ \partial r$ with $f_s(x_{\texttt{SL}}) < 0$ for all $s = 1 \dots m$.

Note strong convexity ensures there exists a unique minimizer $x_{\texttt{OPT}} \in \mathrm{dom}\ \partial r$ of (1.1). These two points $x_{\texttt{OPT}}, x_{\texttt{SL}} \in \mathrm{dom}\ \partial r$ serve as important references that our analysis is done with respect to. We fix two subgradients of r at these points: The subgradient $n_{x_{\texttt{OPT}}} \in \partial r(x_{\texttt{OPT}})$ is chosen such that $f_0(x) + r(x_{\texttt{OPT}}) + \langle n_{x_{\texttt{OPT}}}, x - x_{\texttt{OPT}}\rangle$ is minimized over $f_s(x) \leq 0$ at $x_{\texttt{OPT}}$ for all $s = 1 \dots m$. The subgradient $n_{x_{\texttt{SL}}} \in \partial r(x_{\texttt{SL}})$ can be chosen freely.

These two reference subgradients of r facilitate considering two lower bounds of the objective $f_0 + r$ for either $y \in \{x_{\texttt{OPT}}, x_{\texttt{SL}}\}$, denoted by

$$h_y(x) := f_0(x) + r(y) + \langle n_y, x - y\rangle \ .$$

At each iteration k, we denote the relative difference between x_k and $y \in \{x_{\texttt{OPT}}, x_{\texttt{SL}}\}$ in (relaxed) objective value or feasibility on the selected constraint function $f_{s(x_k)}$ by

$$\delta_k(y) := \begin{cases} h_y(x_k) - h_y(y) & \text{if } x_k \text{ is feasible} \\ f_{s(x_k)}(x_k) - f_{s(x_k)}(y) & \text{otherwise.} \end{cases} \tag{2.7}$$

Note $\delta_k(y)$ is always finite since the real-valued objective function lower bound h_y is used instead of $f_0 + r$ which takes value in the extended reals. Indeed, consider r as an indicator function for some a simple constraint set. This set is projected onto each iteration where x_k is feasible for all of the functional constraints, ensuring x_{k+1} is feasible for the simple constraint. We cannot, however, guarantee that x_k satisfies the simple constraint, and so $r(x_k)$ may be infinite.

The sign of $\delta_k(y)$ may vary. If $y = x_{\texttt{OPT}}$, then $\delta_k(x_{\texttt{OPT}})$ is nonnegative, being lower bounded by the level of suboptimality or current infeasibility

$$\delta_k(x_{\texttt{OPT}}) \geq \begin{cases} h_{x_{\texttt{OPT}}}(x_k) - p_\star & \text{if } x_k \text{ is feasible} \\ f_{s(x_k)}(x_k) & \text{otherwise} \end{cases} \geq 0 \ .$$

When $y = x_{\texttt{SL}}$ and x_k is feasible, $\delta_k(y)$ may be negative but is bounded below by

$$\delta_k(x_{\texttt{SL}}) \geq \inf h_{x_{\texttt{SL}}} - h_{x_{\texttt{SL}}}(x_{\texttt{SL}}) > -\infty \ .$$

When $y = x_{\texttt{SL}}$ and x_k is infeasible, $\delta_k(y)$ is strictly positive, being bounded below by

$$\delta_k(x_{\texttt{SL}}) \geq 0 - \max_{s=1\dots m} f_s(x_{\texttt{SL}}) > 0 \ .$$

A common third assumption used in the analysis of subgradient methods is the uniform boundedness of subgradients. However, if this holds everywhere, the objective must be uniformly Lipschitz continuous, implying it asymptotically grows at most linearly. Contradicting this, strong convexity implies it grows at least quadratically. To avoid such incongruences and to include combinations of smooth and nonsmooth optimization, we consider a more general model than Lipschitz continuity similar to that previously considered in [33, Section 1.2], allowing for quadratic growth.

Assumption C. For both $y \in \{x_{\mathtt{OPT}}, x_{\mathtt{SL}}\}$, there exist constants L_0, L_1 such that every iterate x_k has

$$\begin{cases} x_k \text{ is feasible} & \implies \mathbb{E}_{\xi_k}\|g_0(x_k;\xi_k) + n_y\|_2^2 \leq L_0^2 + L_1\delta_k(y) \\ x_k \text{ is not feasible} & \implies \mathbb{E}_{\xi_k}\|g_{s(x_k)}(x_k;\xi_k)\|_2^2 \leq L_0^2 + L_1\delta_k(y)\ . \end{cases} \tag{2.8}$$

This assumption captures several common settings. In the standard nonsmooth optimization setting where each f_s is uniformly Lipschitz, Assumption C holds with $L_1 = 0$. Moreover, Assumption C also holds in the standard smooth optimization setting where each f_s has uniformly L-Lipschitz gradient. The following lemma shows this condition holds for any additive combination of nonsmooth Lipschitz and smooth settings with bounded variance in the stochastic subgradient oracles.

Lemma 2.4. *If there exists functions $f_s^{(1)}, f_s^{(2)}$ such that each $f_s = f_s^{(1)} + f_s^{(2)}$ for $s = 0, \dots, m$ where $f_s^{(1)}$ is uniformly M-Lipschitz and $f_s^{(2)}$ has uniformly L-Lipschitz gradient and $g_{s(x_k)}(x_k;\xi_k)$ has variance uniformly bounded by σ^2, then Assumption C holds.*

The proof of this lemma is deferred to the appendix where the explicit constants L_0 and L_1 can be found. Section 4 gives an illustrative numerical example of this form where Assumption C holds despite the objective being neither Lipschitz nor smooth.

Note allowing non-Lipschitz objectives allows a range of undesirable "bad" behaviors to occur. It allows the early iterations of the subgradient method to diverge exponentially. For example, consider minimizing $f(u,v) = 50u^2 + 0.5v^2$, which has $\mu = 1, L_0 = 0, L_1 = 200$, with the subgradient method (1.3) initialized with $x_0 = (1,0)$ and $\alpha_k = 2/\mu(k+2)$ (corresponding to $\lambda_k = k+1, \beta_k = 0$). For the first one hundred iterations, the size of x_k grows exponentially, peaking with $\|x_{100}\|_2$ just over 10^{56}, after which it converges monotonically to f's minimizer. Despite such instances existing, our theory shows that even if x_k diverges in its early iterations, it will always subsequently converge at least at rate $O(1/T)$. To the best of our knowledge, no existing analysis of subgradient methods or dual averaging addresses this phenomenon.

To understand and bound such behaviors, we introduce the following two constants, dependent on the choice of stepsizes α_k and associated dual weights λ_k,

$$T_0 := \sup\{k \in \{0,1,2\dots\} \mid L_1\alpha_k > 1\}\ , \tag{2.9}$$

$$C_0 := \sum_{k=0}^{T_0} \lambda_k \max\{L_1\alpha_k - 1, 0\} \max\{\mathbb{E}_\xi\delta_k(x_{\mathtt{OPT}}), \mathbb{E}_\xi\delta_k(x_{\mathtt{SL}})\}\ . \tag{2.10}$$

Observe that T_0, and hence C_0, is bounded if α_k is eventually always less than $1/L_1$, capturing all stepsize policies with $\alpha_k \to 0$. Despite being bounded, the constant C_0 can be exponentially large in T_0. The toy example considered above has $T_0 = 397$ and $C_0 > 10^{112}$. Such potentially exponential-sized constants can be avoided through careful stepsize selection as if one selects $\alpha_k \leq 1/L_1$ for all k as then $T_0 = -\infty$ and $C_0 = 0$. In particular, under the classic assumption that subgradients seen are uniformly bounded, Assumption C holds with $L_1 = 0$ and so $C_0 = 0$. Regardless, our convergence guarantees apply whenever C_0 is finite. As a natural consequence of our main convergence analysis, we find the rate x_k can diverge, and hence the constant C_0, are at most exponential in T_0 (see Proposition 3.4).

3 Primal-Dual Equivalence and Convergence Analysis

In this section, we show the considered primal switching proximal method (2.2) and dual Lagrangian proximal method (2.3) are equivalent and subsequently state and prove new primal-dual convergence guarantees for these methods.

Theorem 3.1. *Let $\{x_k\}$ be the sequence of iterates of the primal method* (2.2) *with stepsizes $\alpha_k > 0$ and $\{y_k\}$ be the sequence of iterates of the dual method* (2.3) *for some $\lambda_k, \beta_k, \mu \geq 0$. If $x_0 = y_0$, $\beta_k = \bar{\beta} \geq 0$ is constant, and*

$$\alpha_k = \frac{\lambda_k}{\mu \sum_{i=0}^{k} \lambda_i + \bar{\beta}} \ , \tag{3.1}$$

then these methods are equivalent, that is $x_k = y_k$.

Proof. We prove this by inductively showing that $x_k = y_k$ and for any iteration with $s(x_k) = 0$ that $n_{k+1} = \frac{1}{\alpha_k}(x_k - \alpha_k g_0(x_k; \xi_k) - x_{k+1})$ (i.e., the subgradients of r produced by the dual method are exactly those produced by the primal method via (2.1)). By assumption, $x_0 = y_0$. Suppose for induction that $x_i = y_i$ for all $i = 0, \dots, k$ and for all $i = 0, \dots, k-1$ with $s(x_i) = 0$ that $n_{i+1} = \frac{1}{\alpha_i}(x_i - \alpha_i g_0(x_i; \xi_i) - x_{i+1})$. Let $\bar{g}_i = g_0(y_i; \xi_i) + n_{i+1}$ if y_i is feasible and $g_{s(y_i)}(y_i; \xi_i)$ otherwise. For each $i < k$, we claim $x_{i+1} = x_i - \alpha_i \bar{g}_i$. If x_i is infeasible, this is immediate. If x_i is feasible, since $x_{i+1} = y_{i+1}$, we have $n_{i+1} \in \partial r(x_{i+1})$. As a result, $x_{i+1} = x_i - \alpha_i \bar{g}_i$ by (2.1). Inductively, we conclude for $i < k$ that $x_{i+1} - x_0 = -\sum_{t=0}^{i} \alpha_t \bar{g}_t$.

By Lemma 2.1, y_{k+1} is the unique solution to $\sum_{i=0}^{k} \lambda_i \left(\bar{g}_i + \mu(y_{k+1} - y_i)\right) + \bar{\beta}(y_{k+1} - y_0) = 0$. Rearranging and simplifying this, it follows that

$$\begin{aligned}
y_{k+1} &= y_0 + \frac{\sum_{i=0}^{k} \lambda_i \mu (y_i - y_0) - \sum_{i=0}^{k} \lambda_i \bar{g}_i}{\sum_{i=0}^{k} \lambda_i \mu + \bar{\beta}} \\
&= y_0 - \frac{\sum_{i=0}^{k} \lambda_i \mu \left(\sum_{t=0}^{i-1} \frac{\lambda_t \bar{g}_t}{\sum_{s=0}^{t} \lambda_s \mu + \bar{\beta}} \right) + \sum_{i=0}^{k} \lambda_i \bar{g}_i}{\sum_{i=0}^{k} \lambda_i \mu + \bar{\beta}} \\
&= y_0 - \frac{\sum_{t=0}^{k-1} \lambda_t \bar{g}_t \frac{\sum_{i=t+1}^{k} \mu \lambda_i}{\sum_{i=0}^{t} \mu \lambda_t + \bar{\beta}} + \sum_{i=0}^{k} \lambda_i \bar{g}_i}{\sum_{i=0}^{k} \lambda_i \mu + \bar{\beta}} \\
&= y_0 - \frac{\sum_{t=0}^{k-1} \lambda_t \bar{g}_t \left(\frac{\sum_{i=t+1}^{k} \mu \lambda_i}{\sum_{i=0}^{t} \mu \lambda_t + \bar{\beta}} + 1 \right)}{\sum_{i=0}^{k} \lambda_i \mu + \bar{\beta}} - \frac{\lambda_k \bar{g}_k}{\sum_{i=0}^{k} \lambda_i \mu + \bar{\beta}} \\
&= y_0 - \sum_{t=0}^{k-1} \frac{\lambda_t \bar{g}_t}{\sum_{i=0}^{t} \lambda_t \mu + \bar{\beta}} - \frac{\lambda_k \bar{g}_k}{\sum_{i=0}^{k} \lambda_i \mu + \bar{\beta}} \\
&= x_k - \alpha_k \bar{g}_k
\end{aligned}$$

where the second equality uses the inductive hypothesis for each $y_i - y_0 = x_i - x_0 = -\sum_{t=0}^{i-1} \alpha_t \bar{g}_t$, the third exchanges summands, and the remainder combines and simplifies terms. If x_k is infeasible, its immediate that $x_{k+1} = x_k - \alpha_k \bar{g}_k = y_{k+1}$. If x_k is feasible, the above equality ensures

$$\frac{1}{\alpha_k}(x_k - \alpha_k g_0(x_k; \xi_k) - y_{k+1}) = n_{k+1} \in \partial r(y_{k+1}) \ .$$

Noting by (2.1) that x_{k+1} is the unique solution to $\frac{1}{\alpha_k}(x_k - \alpha_k g_0(x_k; \xi_k) - z) \in \partial r(z)$, we must have $x_{k+1} = y_{k+1}$ and $n_{k+1} = \frac{1}{\alpha_k}(x_k - \alpha_k g_0(x_k; \xi_k) - x_{k+1})$ as required. □

Remark 1. *Theorem 3.1 suffices to give a dual description for any sequence of primal stepsizes with $\alpha_0 \in (0, 1/\mu]$ and $\alpha_k \in (0, 1/\mu)$ thereafter: one can select any λ_0 and $\bar{\beta}$ satisfying $\alpha_0 = \lambda_0/(\mu\lambda_0 + \bar{\beta})$ and then the corresponding sequence of dual weights is given by the recurrence*

$$\lambda_{k+1} = \frac{\alpha_{k+1}}{1 - \mu\alpha_{k+1}} \frac{\lambda_k}{\alpha_k} \qquad \left(\implies \lambda_T = \frac{\alpha_T}{\Pi_{k=1}^T (1 - \mu\alpha_k)} \frac{\lambda_0}{\alpha_0} \right) . \tag{3.2}$$

One can inductively verify this sequence has $\alpha_k = \frac{\lambda_k}{\mu \sum_{i=0}^k \lambda_i + \bar{\beta}}$ as

$$\frac{\lambda_{k+1}}{\mu \sum_{i=0}^{k+1} \lambda_i + \bar{\beta}} = \frac{\lambda_{k+1}}{\mu \sum_{i=0}^{k} \lambda_i + \bar{\beta} + \mu\lambda_{k+1}} = \frac{\lambda_{k+1}}{\frac{\lambda_k}{\alpha_k} + \mu\lambda_{k+1}} = \frac{\frac{\alpha_{k+1}}{1-\mu\alpha_{k+1}}}{1 + \frac{\mu\alpha_{k+1}}{1-\mu\alpha_{k+1}}} = \alpha_{k+1} \ .$$

Hence, provided dual weights λ_k, one can easily construct α_k as stated in the theorem, and conversely, given stepsizes α_k, one can easily construct corresponding weights λ_k. For example, fixing $\lambda_0 = 1$, $\bar{\beta} = 0$ multipliers for several common stepsizes are

$$\begin{aligned} \alpha_k &= \frac{1}{\mu(k+1)} \implies \lambda_k = 1 \ , \\ \alpha_k &= \frac{2}{\mu(k+2)} \implies \lambda_k = k+1 \ , \\ \alpha_k &= \frac{1}{\mu\sqrt{k+1}} \implies \lambda_k = \frac{1/\sqrt{k+1}}{\Pi_{i=1}^k (1 - 1/\sqrt{i+1})} \approx \exp(\sqrt{k})/\sqrt{k} \ . \end{aligned}$$

Remark 2. *Nesterov [3] noted that in non-strongly convex settings ($\mu = 0$), decreasing stepsizes α_k corresponds to placing decreasing weight on new subgradient lower bounds $\lambda_k = \alpha_k\bar{\beta}$. This runs counter to the intuition that the newest models ought to be most relevant. Rather surprisingly, our Theorem 3.1 shows this fault does not extend to the strongly convex settings ($\mu > 0$). As seen above, the decreasing stepsize selection of $\alpha_k = 2/\mu(k+2)$ corresponds to increasing dual weights $\lambda_k = k+1$.*

3.1 Statement of Primal-Dual Convergence Guarantees

For ease of presenting our convergence theory, we fix $\bar{\beta} = 0$. This parameter's primary purpose in Nesterov's development of dual averaging [3] was to make the model subproblem strongly convex. Strongly convex problems $\mu > 0$, as considered here, have no such need. Following Remark 1, fixing $\bar{\beta} = 0$ only restricts the first stepsize as any sequence $\alpha_0 = 1/\mu$ and $\alpha_k \in (0, 1/\mu)$ can still be dually described.

We prove a uniform convergence guarantee in terms of the primal gap

$$\texttt{primal-gap}_T := \frac{\sum_{k<T: s(x_k)=0} \lambda_k h_{x_{\texttt{OPT}}}(x_k)}{\sum_{k<T: s(x_k)=0} \lambda_k} - p_\star \ ,$$

which utilizes a combination of the feasible objective values seen, the dual gap

$$\texttt{dual-gap}_T := p_\star - \inf \frac{M^{(T-1)}}{\sum_{k<T: s(x_k)=0} \lambda_k} \ ,$$

which utilizes a combination of the subgradient lower bounds seen, and the distance to optimal. For the primal and dual gaps to be well-defined, at least one feasible iterate must have been seen (i.e., $\sum_{k<T: s(x_k)=0} \lambda_k > 0$). Our assumptions facilitate a bound on how long it takes for this to occur. We show the expected fraction of the dual weight that occurs on iterations with a feasible iterate (i.e., $s(x_k) = 0$) is bounded below.

Proposition 3.1. *Under Assumptions A-C, for any primal stepsizes $\alpha_k > 0$ and dual weights $\lambda_k > 0$ satisfying* (3.1) *with $\bar{\beta} = 0$, the stochastic switching proximal subgradient method* (2.2) *has*

$$\mathbb{E}_\xi\left[\frac{\sum_{k<T: s(x_k)=0} \lambda_k}{\sum_{k=0}^{T-1} \lambda_k}\right] \geq \frac{\tau_{\mathtt{SL}}}{2(h_{x_{\mathtt{SL}}}(x_{\mathtt{SL}}) - \inf h_{x_{\mathtt{SL}}}) + \tau_{\mathtt{SL}}}\left(1 - \frac{L_0^2 \sum_{k=0}^{T-1} \lambda_k \alpha_k + C_0}{\sum_{k=0}^{T-1} \lambda_k}\right)$$

where $\tau_{\mathtt{SL}} = 0 - \max_{s=1\dots m} f_s(x_{\mathtt{SL}}) > 0$.

A proof of this is given in Subsection 3.2.1. In deterministic settings, a feasible iterate must then have been reached once this bound is positive. From this, we see that any stepsize selection with $\sum \lambda_k \alpha_k / \sum \lambda_k \to 0$ will asymptotically have at least $\frac{\tau_{\mathtt{SL}}}{2(h_{x_{\mathtt{SL}}}(x_{\mathtt{SL}}) - \inf h_{x_{\mathtt{SL}}}) + \tau_{\mathtt{SL}}} > 0$ fraction of the dual weight on iterations with x_k feasible. Once a feasible iterate occurs, our primal and dual convergence measures are well-defined. Alternatively, one could assume the initialization x_0 is feasible to ensure these quantities are well-defined. In either case, Subsection 3.2.2 proves the following primal-dual convergence guarantee as our main result.

Theorem 3.2. *Under Assumptions A-C, for any primal stepsizes $\alpha_k > 0$ and dual weights $\lambda_k > 0$ satisfying* (3.1) *with $\bar{\beta} = 0$, the stochastic switching proximal subgradient method* (2.2) *has*

$$\mathbb{E}_\xi\left[\left(\frac{\sum_{k<T: s(x_k)=0} \lambda_k}{\sum_{k=0}^{T-1} \lambda_k}\right)(\mathtt{primal\text{-}gap}_T + \mathtt{dual\text{-}gap}_T) + \frac{\mu}{2}\|x_T - x_{\mathtt{OPT}}\|_2^2\right]$$
$$\leq \frac{L_0^2 \sum_{k=0}^{T-1} \lambda_k \alpha_k + C_0}{\sum_{k=0}^{T-1} \lambda_k} \ .$$

Remark 3. *Theorem 3.2 recovers the primal convergence rate of [11, Section 3.2] with $\alpha_k = 2/(\mu(k+2))$ and extends it to be a primal-dual guarantee covering proximal, switching, and non-Lipschitz settings. Theorem 3.1 shows this stepsize corresponds to dual averaging with weights $\lambda_k = k+1$. When $m = 0$, Theorem 3.2 ensures*

$$\mathbb{E}_\xi\left[\mathtt{primal\text{-}gap}_T + \mathtt{dual\text{-}gap}_T + \frac{\mu}{2}\|x_T - x_{\mathtt{OPT}}\|_2^2\right] \leq \frac{4L_0^2}{\mu(T+1)} + \frac{2C_0}{T(T+1)}$$

using that $\sum_{k=0}^{T-1} \lambda_k = T(T+1)/2$ and $\sum_{k=0}^{T-1} \lambda_k \alpha_k = 2(T - \sum_{k=2}^{T+1} 1/k)/\mu \leq 2T/\mu$. When $m > 0$ and the subgradient oracle is deterministic, applying Proposition 3.1 gives a rate worse by only a factor depending on the Slater point of

$$\mathbb{E}_\xi\left[\mathtt{primal\text{-}gap}_T + \mathtt{dual\text{-}gap}_T + \frac{\mu}{2}\|x_T - x_{\mathtt{OPT}}\|_2^2\right]$$
$$\leq \frac{2(h_{x_{\mathtt{SL}}}(x_{\mathtt{SL}}) - \inf h_{x_{\mathtt{SL}}}) + \tau_{\mathtt{SL}}}{\tau_{\mathtt{SL}}\left(1 - \frac{4L_0^2}{\mu(T+1)} - \frac{2C_0}{T(T+1)}\right)}\left(\frac{4L_0^2}{\mu(T+1)} + \frac{2C_0}{T(T+1)}\right) \ .$$

Here, the role of C_0, defined in (2.10)*, bounding the effect of the non-Lipschitz constant L_1 becomes clear. The only role L_1 plays in our rate via T_0, which in turn C_0 may be exponentially large in (see Proposition 3.4). If C_0 is small, then convergence will be dominated by the classic $O(L_0^2/\mu T)$ term as the dependence on C_0 shrinks at a fast $O(1/T^2)$ rate.*

Remark 4. *Theorem 3.2 further recovers and extends the classic linear convergence of proximal gradient descent for smooth, strongly convex optimization. Assume $m = 0$ and f_0 is β-smooth with*

$g_0(x;\xi) = \nabla f_0(x)$. Then (2.8) holds with $L_0 = 0$ and $L_1 = 2\beta > \mu$.[1] *Consider the stepsize selection with $\alpha_0 = 1/\mu$ and $\alpha_k = 1/L_1$ constant thereafter, which corresponds to dual weights $\lambda_0 = 1$ and $\lambda_k = \frac{\mu}{L_1}(1-\mu/L_1)^{-k}$. This choice has $T_0 = 0$ and $C_0 = (\frac{L_1}{\mu} - 1)\delta_0(x_{\mathtt{OPT}})$, giving the following linear convergence*

$$\mathtt{primal\text{-}gap}_T + \mathtt{dual\text{-}gap}_T + \frac{\mu}{2}\|x_T - x_{\mathtt{OPT}}\|_2^2 \leq \frac{L_1}{\mu}\delta_0(x_{\mathtt{OPT}})\left(1 - \frac{\mu}{L_1}\right)^T,$$

using that $\sum_{k=0}^{T-1}\lambda_k = (1-\mu/L_1)^{-(T-1)}$.

Remark 5. *Theorem 3.2 provides new non-Lipschitz conditions for limiting primal-dual guarantees: Under Assumptions A-C and given a deterministic subgradient oracle,*

$$\lim_{T\to\infty} \max\left\{\mathtt{primal\text{-}gap}_T,\ \mathtt{dual\text{-}gap}_T, \|x_T - x_{\mathtt{OPT}}\|_2^2\right\} = 0$$

if C_0 is finite and $\sum_{k=0}^{T}\lambda_k\alpha_k/\sum_{k=0}^{T}\lambda_k \to 0$. Note this implies $\sum_{k=0}^{T}\lambda_k \to \infty$ since $\alpha_k, \lambda_k > 0$. The classic conditions needed for limiting primal convergence under Lipschitz continuity are $\alpha_k \to 0$, which implies C_0 is finite, and $\sum_{k=0}^{T}\alpha_k^2/\sum_{k=0}^{T}\alpha_k \to 0$, which differs slightly from our theory when $\alpha_k = \lambda_k/\mu\sum_{i=0}^{k}\lambda_i$ as

$$\text{classically, one needs } \frac{\sum_{k=0}^{T}\frac{\lambda_k\alpha_k}{\sum_{i=0}^{k}\lambda_i}}{\sum_{k=0}^{T}\frac{\lambda_k}{\sum_{i=0}^{k}\lambda_i}} \to 0 \quad \text{whereas we require} \quad \frac{\sum_{k=0}^{T}\lambda_k\alpha_k}{\sum_{k=0}^{T}\lambda_k} \to 0 .$$

Remark 6. *One can select stepsizes to minimize our rate. Given $C_0 = 0$ and the first T stepsizes $\alpha_0, \dots \alpha_{T-1}$ and corresponding weights $\lambda_0, \dots \lambda_{T-1}$, one can select α_T and λ_T to minimize our convergence upper bound (1.5) after one more step by setting*[2]

$$\alpha_T = \frac{\lambda_T}{\mu\sum_{k=0}^{T}\lambda_k} \quad \text{and} \quad \lambda_T = \frac{\sum_{k=0}^{T-1}\lambda_k \times \sum_{k=0}^{T-1}\lambda_k\alpha_k}{\sum_{k=0}^{T-1}\lambda_k(2/\mu - \alpha_k)} . \tag{3.3}$$

Given $\alpha_0 = 1/\mu$ and $\lambda_0 = 1$, the numerically optimized parameters and rate are below.

k	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
λ_k	*1*	*1*	*1.2*	*1.4022*	*1.6025*	*1.8005*	*1.9966*	*2.1910*	*2.3841*
α_k	$\frac{1}{\mu}$	$\frac{1}{2\mu}$	$\frac{1}{2.6666\mu}$	$\frac{1}{3.2820\mu}$	$\frac{1}{3.8719\mu}$	$\frac{1}{4.4460\mu}$	$\frac{1}{5.0094\mu}$	$\frac{1}{5.5648\mu}$	$\frac{1}{6.1142\mu}$
Rate (1.5)	$\frac{L_0^2}{\mu}$	$\frac{L_0^2}{1.3333\mu}$	$\frac{L_0^2}{1.6410\mu}$	$\frac{L_0^2}{1.9359\mu}$	$\frac{L_0^2}{2.2230\mu}$	$\frac{L_0^2}{2.5047\mu}$	$\frac{L_0^2}{2.7824\mu}$	$\frac{L_0^2}{3.0571\mu}$	$\frac{L_0^2}{3.3293\mu}$

For comparison, this offers small gains over the "typical" stepsize $\alpha_k = 2/\mu(k+2)$, shown below. Numerics showing some small gains actually occur are in Section 4.

k	*0*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
λ_k	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
α_k	$\frac{1}{\mu}$	$\frac{1}{1.5\mu}$	$\frac{1}{2\mu}$	$\frac{1}{2.5\mu}$	$\frac{1}{3\mu}$	$\frac{1}{3.5\mu}$	$\frac{1}{4\mu}$	$\frac{1}{4.5\mu}$	$\frac{1}{5\mu}$
Rate (1.5)	$\frac{L_0^2}{\mu}$	$\frac{L_0^2}{1.2857\mu}$	$\frac{L_0^2}{1.5652\mu}$	$\frac{L_0^2}{1.8404\mu}$	$\frac{L_0^2}{2.1126\mu}$	$\frac{L_0^2}{2.3824\mu}$	$\frac{L_0^2}{2.6504\mu}$	$\frac{L_0^2}{2.9168\mu}$	$\frac{L_0^2}{3.1819\mu}$

[1] This can be verified by noting $h_{x_{\mathtt{OPT}}}$ is β-smooth. Then the standard descent lemma ensures

$$h_{x_{\mathtt{OPT}}}(x_{\mathtt{OPT}}) \leq h_{x_{\mathtt{OPT}}}(x_k - (\nabla f_0(x_k) + n_{x_{\mathtt{OPT}}})/\beta) \leq h_{x_{\mathtt{OPT}}}(x_k) - \frac{1}{2\beta}\|\nabla f_0(x_k) + n_{x_{\mathtt{OPT}}}\|_2^2.$$

[2] The formula for the optimal λ_T in (3.3) can be verified as the unique solution to $\frac{d}{d\lambda_T}\left(\frac{\sum_{k=0}^{T}\lambda_k\alpha_k}{\sum_{k=0}^{T}\lambda_k}\right) = 0$.

Remark 7. *For deterministic settings where μ and an upper bound $G^2 \ge L_0^2$ are known, one can also utilize our theory to adapt stepsizes to avoid any early exponential divergences. Recall such divergences are quantified by C_0 as discussed at the end of Section 2. If one selects decreasing stepsizes with $\alpha_0 = 1/\mu$ and $\alpha_k \le 1/L_1$ thereafter, T_0 defined in (2.9) must equal zero, and hence no bad divergence can occur as*

$$\left(\frac{\sum_{k<T:s(x_k)=0}\lambda_k}{\sum_{k=0}^{T-1}\lambda_k}\right)(\texttt{primal-gap}_T + \texttt{dual-gap}_T)$$
$$\le \frac{G^2\sum_{k=0}^{T-1}\lambda_k\alpha_k + (\frac{1}{\alpha_1\mu}-1)\frac{\|g_{s(x_0)}(x_0;\xi_0)\|^2}{\mu}}{\sum_{k=0}^{T-1}\lambda_k} \tag{3.4}$$

where we bounded $L_0^2 \le G^2$ and $C_0 = (\frac{L_1}{\mu}-1)\delta_0(y) \le (\frac{1}{\alpha_1\mu}-1)\frac{\|g_{s(x_0)}(x_0;\xi_0)\|^2}{\mu}$ for either $y \in \{x_{\texttt{OPT}}, x_{\texttt{SL}}\}$. Notice every quantity in (3.4) is computable! Hence, if one selected generic decreasing stepsizes α_k, without knowing L_1 to ensure $\alpha_k \le 1/L_1$, one can still check if convergence is occurring at the above rate. If (3.4) fails at some iteration, one can conclude $\alpha_1 > 1/L_1$. In this case, one could reasonably restart the method with reduced stepsizes, via an exponential backtracking.

Remark 8. *Without strong convexity, one cannot guarantee convergence of a duality gap since a linear $M^{(k)}$ leads the duality gap to always be 0 or ∞. Our theory can still be applied by a standard trick: Instead of unconstrained minimization ($m = 0$) of a convex function f_0, one could minimize the closely related strongly convex function*

$$\tilde{f}_0(x) = f_0(x) + \frac{\epsilon}{2D^2}\|x - x_0\|^2 \ .$$

This perturbed problem has minimum value at most $p_\star + \epsilon\frac{\|x^* - x_0\|_2^2}{2D^2}$, and so any ϵ-minimizer of $\tilde{f}_0$ is an $(1 + \frac{\|x^* - x_0\|_2^2}{2D^2})\epsilon$-minimizer for the original problem.

Note $\tilde{f}_0$ is ϵ/D^2-strongly convex and since $m = 0$, one can select $x_{\texttt{SL}} = x_{\texttt{OPT}}$. Moreover, if f_0 was M-Lipschitz continuous, then as a sum of Lipschitz and smooth components, the perturbed objective $\tilde{f}_0$ satisfies (2.8) with $L_0^2 = 6M^2$ by Lemma 2.4. As a result, Theorem 3.2 ensures applying the subgradient method to $\tilde{f}_0$ with stepsize $\alpha_k = 2D^2/\epsilon(k+2)$ has perturbed primal-dual gap converge at a rate $\frac{24M^2D^2}{\epsilon(T+1)} + O(1/T^2)$. Similar perturbed primal-dual guarantees using a novel proof method were given by [27].

3.2 Proof of Primal-Dual Convergence Guarantees

Our theory relies on two symmetric inductive results, one inequality slightly generalizing the classic primal analysis and one novel inequality based on our dual perspective, in Lemmas 3.2 and 3.3. From these, we prove the feasibility guarantee Proposition 3.1 and our main result Theorem 3.2.

First, we show an inductive relationship on the (expected, unnormalized, squared) distance from the iterates x_k to either $x_{\texttt{OPT}}$ or $x_{\texttt{SL}}$ defined as

$$R_k(y) := \left(\frac{\mu}{2}\sum_{i=0}^{k-1}\lambda_i\right)\mathbb{E}_\xi\|x_k - y\|_2^2 \tag{3.5}$$

To simplify notations, throughout our analysis, we denote $g_k = g_{s(x_k)}(x_k;\xi_k)$ and $w_k = \mu\sum_{i=0}^{k}\lambda_i$ (with the convention that $w_{-1} = 0$ as the given summation is empty).

Lemma 3.2. *Under Assumptions A-C, the switching proximal subgradient method* (2.2) *with* $\alpha_k = \lambda_k/\mu\sum_{i=0}^k \lambda_i$ *has for either* $y \in \{x_{\texttt{OPT}}, x_{\texttt{SL}}\}$

$$R_{k+1}(y) \leq R_k(y) - \frac{\lambda_k}{2}\left((2 - L_1\alpha_k)\,\mathbb{E}_\xi\delta_k(y) - L_0^2\alpha_k\right) .$$

Proof. This proof follows a standard analysis technique, directly expanding the definition of $R_{k+1}(y)$. First, suppose x_k is feasible. Then

$$\begin{aligned} R_{k+1}(y) &= \frac{w_k}{2}\mathbb{E}_\xi\|\mathrm{prox}_{\alpha_k,r}(x_k - \alpha_k g_k) - \mathrm{prox}_{\alpha_k,r}(y + \alpha_k n_y)\|_2^2 \\ &\leq \frac{w_k}{2}\mathbb{E}_\xi\|x_k - \alpha_k(g_k + n_y) - y\|_2^2 \\ &= \frac{w_k}{2}\mathbb{E}_\xi\|x_k - y\|_2^2 - \lambda_k\mathbb{E}_\xi\langle g_k + n_y, x_k - y\rangle + \frac{\lambda_k\alpha_k}{2}\mathbb{E}_\xi\|g_k + n_y\|_2^2 \\ &\leq \frac{w_k}{2}\mathbb{E}_\xi\|x_k - y\|_2^2 - \lambda_k\mathbb{E}_\xi(h_y(x_k) - h_y(y) + \frac{\mu}{2}\|x_k - y\|_2^2) + \frac{\lambda_k\alpha_k}{2}\mathbb{E}_\xi\|g_k + n_y\|_2^2 \\ &= R_k(y) - \lambda_k\mathbb{E}_\xi\delta_k(y) + \frac{\lambda_k\alpha_k}{2}\mathbb{E}_\xi\|g_k + n_y\|_2^2 \end{aligned}$$

where the first line uses that $\mathrm{prox}_{\alpha_k,r}(y + \alpha_k n_y) = y$, the second uses the nonexpansiveness of the proximal operator [34, Proposition 12.19], the third factors the norm squared and uses $\alpha_k = \lambda_k/w_k$, and the fourth uses the strong convexity of h_y. Then, applying the bound (2.8) gives the claim. Similarly, supposing x_k is infeasible gives

$$\begin{aligned} R_{k+1}(y) &= \frac{w_k}{2}\mathbb{E}_\xi\|x_k - \alpha_k g_k - y\|_2^2 \\ &= \frac{w_k}{2}\mathbb{E}_\xi\|x_k - y\|_2^2 - \lambda_k\mathbb{E}_\xi\langle g_k, x_k - y\rangle + \frac{\lambda_k\alpha_k}{2}\mathbb{E}_\xi\|g_k\|_2^2 \\ &\leq \frac{w_k}{2}\mathbb{E}_\xi\|x_k - y\|_2^2 - \lambda_k\mathbb{E}_\xi(f_{s(x_k)}(x_k) - f_{s(x_k)}(y) + \frac{\mu}{2}\|x_k - y\|_2^2) + \frac{\lambda_k\alpha_k}{2}\mathbb{E}_\xi\|g_k\|_2^2 \\ &= R_k(y) - \lambda_k\mathbb{E}_\xi\delta_k(y) + \frac{\lambda_k\alpha_k}{2}\mathbb{E}_\xi\|g_k\|_2^2 \end{aligned}$$

using strong convexity of $f_{s(x_k)}$. Applying (2.8) completes the proof. □

The dual portion of our convergence analysis relies on showing the same inductive relationship on the (expected, unnormalized) dual gap defined as

$$D_k := \left(\sum_{i<k:s(x_i)=0}\lambda_i\right)p_\star - \mathbb{E}_\xi \inf M^{(k-1)} . \tag{3.6}$$

Dividing through by $\sum_{i<k:s(x_i)=0}\lambda_i$, once nonzero, gives the dual gap.

Lemma 3.3. *Under Assumptions A-C, the switching proximal subgradient method* (2.2) *with* $\alpha_k = \lambda_k/\mu\sum_{i=0}^k \lambda_i$ *has*

$$D_{k+1} \leq D_k - \frac{\lambda_k}{2}\left((2 - L_1\alpha_k)\,\mathbb{E}_\xi\delta_k(x_{\texttt{OPT}}) - L_0^2\alpha_k\right) .$$

Proof. Observe that one can rewrite $M^{(k)}(x) = Q_1(x) + Q_2(x)$ as the sum of two quadratics where $Q_1(x) = M^{(k-1)}(x)$ and Q_2 depends on whether x_k is feasible. In the notation of Lemma 2.3, Lemmas 2.1 and 2.2 ensure Q_1 has $a_1 = \inf M^{(k-1)}$, $b_1 = \mu \sum_{i=0}^{k-1} \lambda_i$, and $z_1 = x_k$. To determine Q_2, first suppose x_k is feasible. Then we have

$$Q_2(x) = \lambda_k \left(f_0(x_k) + \langle g_k, x - x_k \rangle + \frac{\mu}{2}\|x - x_k\|_2^2 + r(x_{k+1}) + \langle n_{k+1}, x - x_{k+1} \rangle \right) .$$

This quadratic can be written in the form $Q_2(z) = a_2 + \frac{b_2}{2}\|z - z_2\|_2^2$ with

$$\begin{aligned} a_2 &= \lambda_k \left(f_0(x_k) + r(x_{k+1}) + \alpha_k \langle n_{k+1}, g_k + n_{k+1} \rangle - \frac{1}{2\mu}\|g_k + n_{k+1}\|_2^2 \right) , \\ b_2 &= \mu\lambda_k \ , \qquad z_2 = x_k - \frac{g_k + n_{k+1}}{\mu} \ . \end{aligned}$$

By Lemma 2.3, the expected minimum value of the updated model $\mathbb{E}_\xi \inf M^{(k)}$ is

$$\begin{aligned} \mathbb{E}_\xi \inf M^{(k-1)} + \lambda_k \mathbb{E}_\xi \Bigg(& f_0(x_k) + r(x_{k+1}) + \alpha_k \langle n_{k+1}, g_k + n_{k+1} \rangle - \frac{1}{2\mu}\|g_k + n_{k+1}\|_2^2 \\ & + \frac{\sum_{i=0}^{k-1} \lambda_i}{2\mu \sum_{i=0}^{k} \lambda_i} \|g_k + n_{k+1}\|_2^2 \Bigg) . \end{aligned}$$

From this, we conclude the lower bound

$$\begin{aligned} \mathbb{E}_\xi \inf M^{(k)} \geq \mathbb{E}_\xi \inf M^{(k-1)} &+ \lambda_k \mathbb{E}_\xi \Bigg(f_0(x_k) + r(x_{\mathtt{OPT}}) + \langle n_{x_{\mathtt{OPT}}}, x_{k+1} - x_{\mathtt{OPT}} \rangle \\ & + \alpha_k \langle n_{k+1}, g_k + n_{k+1} \rangle - \frac{1}{2\mu}\|g_k + n_{k+1}\|_2^2 + \frac{\sum_{i=0}^{k-1} \lambda_i}{2\mu \sum_{i=0}^{k} \lambda_i} \|g_k + n_{k+1}\|_2^2 \Bigg) \\ = \mathbb{E}_\xi \inf M^{(k-1)} &+ \lambda_k \mathbb{E}_\xi \Bigg(\delta_k(x_{\mathtt{OPT}}) + p_\star + \langle n_{x_{\mathtt{OPT}}}, -\alpha_k(g_k + n_{k+1}) \rangle \\ & + \alpha_k \langle n_{k+1}, g_k + n_{k+1} \rangle - \frac{\alpha_k}{2}\|g_k + n_{k+1}\|_2^2 \Bigg) \\ = \mathbb{E}_\xi \inf M^{(k-1)} &+ \lambda_k \left(\mathbb{E}_\xi \delta_k(x_{\mathtt{OPT}}) + p_\star - \frac{\alpha_k}{2}\mathbb{E}_\xi \|g_k + n_{x_{\mathtt{OPT}}}\|_2^2 \right) \end{aligned}$$

where the inequality lower bounds $r(x_{k+1})$ by $r(x_{\mathtt{OPT}}) + \langle n_{x_{\mathtt{OPT}}}, x_{k+1} - x_{\mathtt{OPT}} \rangle$, the first equality applies the definitions of $\delta_k(x_{\mathtt{OPT}})$ in (2.7), α_k in (3.1), and that $x_{k+1} - x_k = -\alpha_k(g_k + n_{k+1})$, and the last equality combines and simplifies terms. In terms of $D_k = \sum_{i<k:s(x_i)=0} \lambda_i p_* - \mathbb{E}_\xi \inf M^{(k-1)}$, this gives the following recurrence

$$D_{k+1} \leq D_k - \lambda_k \mathbb{E}_\xi \delta_k(x_{\mathtt{OPT}}) + \frac{\lambda_k \alpha_k}{2}\mathbb{E}_\xi \|g_k + n_{x_{\mathtt{OPT}}}\|_2^2 \ .$$

Applying (2.8) gives the claim in this case. Now suppose x_k is infeasible. Then, noting Q_2 minimizes at $x_k - \frac{g_k}{\mu}$, Lemma 2.2 ensures

$$Q_2(x) = \lambda_k \left(f_{s(x_k)}(x_k) - \frac{1}{2\mu}\|g_k\|_2^2 + \frac{\mu}{2} \left\| x - \left(x_k - \frac{g_k}{\mu} \right) \right\|_2^2 \right) .$$

Then by Lemma 2.3, the expected minimum value of the updated model $\mathbb{E}_\xi \inf M^{(k)}$ is given by

$$\begin{aligned}
&\mathbb{E}_\xi \inf M^{(k-1)} + \lambda_k \mathbb{E}_\xi \left(f_{s(x_k)}(x_k) - \frac{1}{2\mu}\|g_k\|_2^2 \right) + \frac{\lambda_k \sum_{i=0}^{k-1} \lambda_i}{\mu \sum_{i=0}^{k} \lambda_i} \mathbb{E}_\xi \|g_k\|_2^2 \\
&= \mathbb{E}_\xi \inf M^{(k-1)} + \lambda_k \mathbb{E}_\xi f_{s(x_k)}(x_k) - \frac{\lambda_k \alpha_k}{2} \mathbb{E}_\xi \|g_k\|_2^2 \ .
\end{aligned}$$

Noting $f_{s(x_k)}(x_{\mathtt{OPT}}) \leq 0$, we conclude the lower bound

$$\mathbb{E}_\xi \inf M^{(k)} \geq \mathbb{E}_\xi \inf M^{(k-1)} + \lambda_k \left(\mathbb{E}_\xi \delta_k(x_{\mathtt{OPT}}) - \frac{\alpha_k}{2} \mathbb{E}_\xi \|g_k\|_2^2 \right) \ .$$

In terms of $D_k = \sum_{i<k:s(x_i)=0} \lambda_i p_* - \mathbb{E}_\xi \inf M^{(k-1)}$, this gives the recurrence

$$D_{k+1} \leq D_k - \lambda_k \mathbb{E}_\xi \delta_k(x_{\mathtt{OPT}}) + \frac{\lambda_k \alpha_k}{2} \mathbb{E}_\xi \|g_k\|_2^2 \ .$$

Bounding $\mathbb{E}_\xi \|g_k\|_2^2$ by (2.8) gives the claim in this last case. □

As a direct consequence of our primal inductive lemma, we can bound the rate that $\delta_k(y)$ grows in the first T_0 iterations as being at most exponential. From this, one can explicitly upper bound C_0 exponentially in T_0.

Proposition 3.4. *Under Assumptions A-C, the switching proximal subgradient method* (2.2) *with* $\alpha_k = \lambda_k / \mu \sum_{i=0}^{k} \lambda_i$ *has for either* $y \in \{x_{\mathtt{OPT}}, x_{\mathtt{SL}}\}$

$$\begin{aligned}
|\delta_k(y)| &\leq L_1 \|x_k - y\|_2^2 + \frac{L_0^2}{L_1} \ , \\
\mathbb{E}_\xi \|x_T - y\|_2^2 &\leq \left(1 + \frac{\max\{2, L_1/\mu - 2\} L_1}{\mu} \right)^T \left(\|x_0 - y\|_2^2 + \frac{L_0^2}{L_1} + \frac{L_0^2}{\mu \max\{2, L_1/\mu - 2\}} \right) \ .
\end{aligned}$$

Proof. We first claim x_k satisfies

$$-\frac{L_0^2}{L_1} \leq \delta_k(y) \leq \sqrt{L_0^2 + L_1 \delta_k(y)} \ \|x_k - y\|_2 \ .$$

The first inequality lower bounding $\delta_k(y)$ is immediate from (2.8). For the second inequality, note that if x_k is feasible, convexity of h_y ensures that $h_y(y) \geq h_y(x_k) + \langle \mathbb{E}_{\xi_k}(g_{s(x_k)}(x_k;\xi_k) + n_y), y - x_k \rangle$. If x_k is infeasible, $f_{s(x_k)}(y) \geq f_{s(x_k)}(x_k) + \langle \mathbb{E}_{\xi_k} g_{s(x_k)}(x_k;\xi_k), y - x_k \rangle$. In either case, Cauchy-Schwarz and Assumption C give the second inequality. Observe that if $\delta_k(y) < 0$, the first inequality above ensures $|\delta_k(y)| \leq \frac{L_0^2}{L_1} \leq L_1 \|x_k - y\|_2^2 + \frac{L_0^2}{L_1}$. Instead, if $\delta_k(y) \geq 0$, squaring the second inequality above ensures $\delta_k(y)^2 \leq (L_0^2 + L_1 \delta_k(y)) \|x_k - y\|_2^2$, which implies

$$\delta_k(y) \leq \frac{L_1 + \sqrt{L_1^2 + 4L_0^2 / \|x_k - y\|_2^2}}{2} \|x_k - y\|_2^2 \leq \frac{L_0^2}{L_1} + L_1 \|x_k - y\|_2^2,$$

where the second inequality uses concavity to bound $\sqrt{a+b} \leq \sqrt{a} + \frac{b-a}{2\sqrt{a}} \leq \sqrt{a} + \frac{b}{2\sqrt{a}}$, completing the proposition's first claim. To prove the proposition's second claim, note

$$\begin{aligned}\mathbb{E}_\xi\|x_{k+1}-y\|_2^2 &\le \mathbb{E}_\xi\|x_k-y\|_2^2 - \frac{\lambda_k}{\mu\sum_{i=0}^k \lambda_i}((2-L_1\alpha_k)\mathbb{E}_\xi\delta_k(y) - L_0^2\alpha_k)\\ &\le \mathbb{E}_\xi\|x_k-y\|_2^2 + \alpha_k(|2-L_1\alpha_k||\mathbb{E}_\xi\delta_k(y)| + L_0^2\alpha_k)\\ &\le (1+\alpha_k|2-L_1\alpha_k|L_1)\,\mathbb{E}_\xi\|x_k-y\|_2^2 + \alpha_k|2-L_1\alpha_k|\frac{L_0^2}{L_1} + L_0^2\alpha_k^2\\ &\le \left(1+\frac{\max\{2, L_1/\mu-2\}L_1}{\mu}\right)\mathbb{E}_\xi\|x_k-y\|_2^2 + \frac{\max\{2, L_1/\mu-2\}}{\mu}\frac{L_0^2}{L_1} + \frac{L_0^2}{\mu^2},\end{aligned}$$

where the first inequality uses Lemma 3.2 divided by $(\frac{\mu}{2}\sum_{i=0}^k \lambda_i)$, the second inequality applies simple upper bounds, the third inequality uses our bound on $|\delta_k(y)|$, and the fourth uses that $0 < \alpha_k \le 1/\mu$ and its consequence $|2-L_1\alpha_k| \le \max\{2, L_1/\mu - 2\}$. From this, the proposition's second claim follows as such recurrences of the form $a_{k+1} \le b\cdot a_k + c$ satisfy $a_T \le b^T a_0 + c\frac{b^T-1}{b-1} \le b^T(a_0 + \frac{c}{b-1})$. □

3.2.1 Proof of Proposition 3.1 Noting $R_0(x_{\mathtt{SL}}) = 0$ and $R_T(x_{\mathtt{SL}}) \ge 0$, inductively applying Lemma 3.2 with $y = x_{\mathtt{SL}}$ shows

$$\sum_{k=0}^{T-1} \frac{\lambda_k}{2}\left((2-L_1\alpha_k)\mathbb{E}_\xi\delta_k(x_{\mathtt{SL}}) - L_0^2\alpha_k\right) \le 0\ .$$

From this, we find that

$$\begin{aligned}0 &\le \sum_{k=0}^{T-1}\lambda_k(L_1\alpha_k-2)\mathbb{E}_\xi\delta_k(x_{\mathtt{SL}}) + \sum_{k=0}^{T-1} L_0^2\lambda_k\alpha_k\\ &= \mathbb{E}_\xi\left[\sum_{k=0}^{T-1}\lambda_k(L_1\alpha_k-2)(\max\{\delta_k(x_{\mathtt{SL}}),0\} + \min\{\delta_k(x_{\mathtt{SL}}),0\})\right] + \sum_{k=0}^{T-1} L_0^2\lambda_k\alpha_k\\ &\le C_0 + \mathbb{E}_\xi\left[\sum_{k=0}^{T-1} -\lambda_k\max\{\delta_k(x_{\mathtt{SL}}),0\} + \sum_{k=0}^{T-1}\lambda_k(L_1\alpha_k-2)\min\{\delta_k(x_{\mathtt{SL}}),0\}\right] + \sum_{k=0}^{T-1} L_0^2\lambda_k\alpha_k\\ &\le C_0 - \tau_{\mathtt{SL}}\mathbb{E}_\xi\left[\sum_{k<T: s(x_k)\neq 0}\lambda_k\right] + 2(h_{x_{\mathtt{SL}}}(x_{\mathtt{SL}}) - \inf h_{x_{\mathtt{SL}}})\mathbb{E}_\xi\left[\sum_{k<T: s(x_k)=0}\lambda_k\right] + \sum_{k=0}^{T-1} L_0^2\lambda_k\alpha_k\end{aligned}$$

where the first inequality uses our inductive result, the second inequality uses the definition of C_0 in (2.10) and that $\delta_k(x_{\mathtt{OPT}}) \ge 0$, and the third inequality bounds the first two summations as follows: (i) the first sum's upper bound notes that if $s(x_k) \neq 0$, then $\delta_k(x_{\mathtt{SL}}) \ge \tau_{\mathtt{SL}} > 0$ and (ii) the second sum's upper bound notes $L_1\alpha_k - 2 \ge -2$ and if $s(x_k) = 0$, then $\delta_k(x_{\mathtt{SL}}) \ge \inf h_{x_{\mathtt{SL}}} - h_{x_{\mathtt{SL}}}(x_{\mathtt{SL}})$, which may be negative. Rearrangement gives the claim as

$$\mathbb{E}_\xi\left[\sum_{k<T: s(x_k)=0}\lambda_k\right] \ge \frac{\tau_{\mathtt{SL}}\sum_{k=0}^{T-1}\lambda_k - L_0^2\sum_{k=0}^{T-1}\lambda_k\alpha_k - C_0}{2(h_{x_{\mathtt{SL}}}(x_{\mathtt{SL}}) - \inf h_{x_{\mathtt{SL}}}) + \tau_{\mathtt{SL}}}\ .$$

3.2.2 Proof of Theorem 3.2 Applying Lemma 3.2 with $y = x_{\mathtt{OPT}}$ from $k = 0$ to $T-1$ yields

$$R_T(x_{\mathtt{OPT}}) + \sum_{k=0}^{T-1}\frac{\lambda_k}{2}\mathbb{E}_\xi\delta_k(x_{\mathtt{OPT}}) \le R_0(x_{\mathtt{OPT}}) + \sum_{k=0}^{T-1}\frac{L_0^2\lambda_k\alpha_k}{2} + \sum_{k=0}^{T-1}\frac{\lambda_k}{2}\left(L_1\alpha_k - 1\right)\mathbb{E}_\xi\delta_k(x_{\mathtt{OPT}})\ .$$

Similarly, applying Lemma 3.3 from $k = 0$ to $T - 1$ yields

$$D_T + \sum_{k=0}^{T-1} \frac{\lambda_k}{2} \mathbb{E}_\xi \delta_k(x_{\texttt{OPT}}) \leq D_0 + \sum_{k=0}^{T-1} \frac{L_0^2 \lambda_k \alpha_k}{2} + \sum_{k=0}^{T-1} \frac{\lambda_k}{2} (L_1 \alpha_k - 1) \mathbb{E}_\xi \delta_k(x_{\texttt{OPT}}) \ .$$

Noting $R_0(x_{\texttt{OPT}}) = D_0 = 0$ and the last summation in each bound is at most half our initial blow-up constant $\frac{1}{2}C_0$, the sum of these inequalities provides a bound of

$$R_T(x_{\texttt{OPT}}) + D_T + \sum_{k=0}^{T-1} \lambda_k \mathbb{E}_\xi \delta_k(x_{\texttt{OPT}}) \leq \sum_{k=0}^{T-1} L_0^2 \lambda_k \alpha_k + C_0 \ .$$

Lower bounding each $\mathbb{E}_\xi \delta_k(x_{\texttt{OPT}})$ with $s(x_k) \neq 0$ by zero completes our proof.

4 Numerical Experiments

In this section, we numerically validate the accuracy of Theorem 3.2 in predicting actual observed performance. Our three main numerical experiments address the impact of varying λ_k, the quality of our new primal-dual stopping criteria, and the accuracy of our T_0 and C_0 constants at predicting initial divergences. All of our numerics are implemented in `Julia 1.8.5`[3].

We consider the following deterministic family of nonsmooth, non-Lipschitz, strongly convex minimization problems given $A, C \in \mathbb{R}^{m\times n}$ and $b, d \in \mathbb{R}^m$

$$\min_{x\in\mathbb{R}^n} f_0(x) = \|Ax - b\|_1 + \frac{1}{2}\|Cx - d\|_2^2 \ . \tag{4.1}$$

Note $\|Ax - b\|_1$ is $\|A^T\|_{\infty\to 2}$-Lipschitz[4]. However, computing this induced matrix norm is NP-hard [35], so we instead upper bound it by $\sum_{i=1}^m \|A_i\|$ where A_i denotes A's ith row. Further noting $\frac{1}{2}\|Cx - d\|_2^2$ is $\lambda_{max}(C^TC)$-smooth, by Lemma 2.4, our Assumptions A-C hold with $L_0^2 = 8(\sum_{i=1}^m \|A_i\|)^2$, $L_1 = 4\lambda_{max}(C^TC)$ and $\mu = \lambda_{min}(C^TC)$. We generate problem instances fixing $m = n = 100, x_0 = 0$ and randomly drawing $A, \tilde{C}, x_{\texttt{OPT}}$ with i.i.d. normal entries. To control μ and L_1, we set $C = I + \sigma\tilde{C}$ for various selections of $\sigma \geq 0$. When $\sigma = 0$, we have $\mu = 1$ and $L_1 = 4$. Initially as σ increases, μ decreases while L_1 increases. To ensure $x_{\texttt{OPT}}$ is a minimizer and $p_\star = 0$, we set $b = Ax_{\texttt{OPT}}$, $d = Cx_{\texttt{OPT}}$.

4.1 Performance under Varied Stepsize Selections

First, we aim to measure the quality of Theorem 3.2's bounds compared to actual convergence. We fix $\bar{\beta} = 0$ and $\sigma = 0$ and consider several polynomial selections of λ_k and our proposed, optimized choice (3.3). Figure 1 shows the upper bound from Theorem 3.2 in comparison to the observed convergence of the aggregate measure $\texttt{primal-gap}_T + \texttt{dual-gap}_T + \frac{\mu}{2}\|x_T - x_{\texttt{OPT}}\|_2^2$ and each component separately. As expected, the optimized parameters (3.3) have the best theoretical bound and the best observed aggregate performance early on. Moreover, it remains one of the best methods throughout. Asymptotically, we see comparable convergence for all $\lambda_k \neq 1$. The primal convergence under uniform weights $\lambda_k = 1$ was the slowest in line with our theory, which only guarantees a $O(\log(T)/T)$ rate. Uniform weights did yield the fastest convergence of the dual gap and distance to optimal, which our theory cannot explain.

[3] The source code is available at `https://github.com/AshleyLDL/Primal-Dual-Averaging-Coding`

[4] The Lipschitz constant for $\|Ax - b\|_1$ follows from the chain rule as its subgradients are combinations of A's rows with weights in $[-1, 1]$, so the largest subgradient is $\max_{\|w\|_\infty\leq 1} \|A^T w\|_2 = \|A^T\|_{\infty\to 2}$.

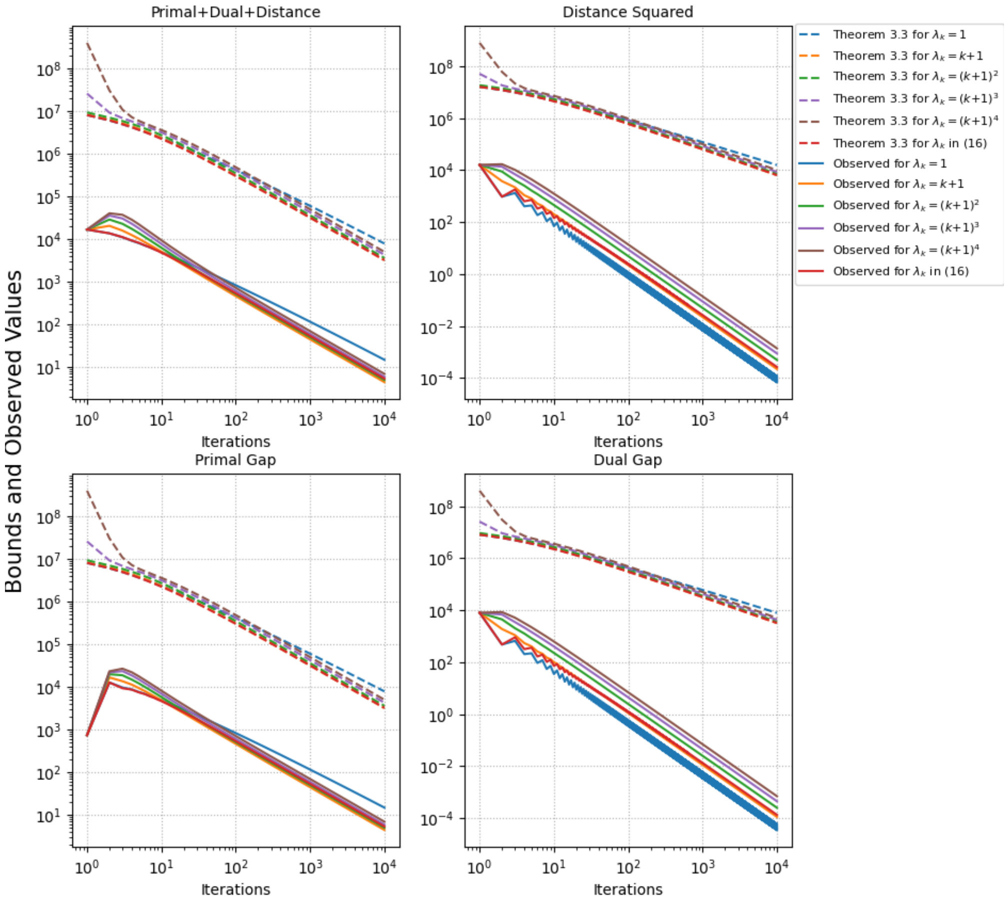

Figure 1: Bounds and observed performance for different λ_k with $\bar{\beta} = 0$.

4.2 High Accuracy of Primal-Dual Stopping Criteria

One practical benefit of our dual characterizations of subgradient methods is the resulting computable dual lower bounds and hence stopping criteria, assuming μ is known. As a shorthand, denote the convergence of our dual lower bound on $p_\star$ by $d_k := p_\star - \inf M^{(k-1)}/\sum \lambda_i$. We denote the convergence of three natural upper bounds on $p_\star$ by the primal gap (averaging function values seen) $p_k := \sum \lambda_i f(x_i)/\sum \lambda_i - p_\star$, the function value at an averaged iterate $\bar{p}_k := f(\sum \lambda_i x_i/\sum \lambda_i) - p_\star$, and the function value at the latest iterate $\delta_k := f(x_k) - p_\star$. Combining these upper and lower bounds gives three natural stopping criteria to ensure an ϵ-accurate solution is found: stopping once the gap between upper and lower bounds is less than ϵ. Fixing $\sigma = 0$ and $\epsilon = 0.05$, Table 1 shows the number of iterations before these conditions were first reached.

Across every $\lambda_k \neq 1$ configuration, we see $\bar{p}_T$ and d_T both converge relatively quickly. The stopping criteria $\bar{p}_t + d_t \leq \epsilon$ is consistently reached in at most 25% more iterations than were required to reach $\bar{p}_t \leq \epsilon$. Hence, up to a small constant, this criterion matches the ideal time to stop. Note δ_t and p_t both converged much slower than $\bar{p}_t$ and d_t. Correspondingly, the stopping criteria $\delta_t + d_t \leq \epsilon$ and $p_t + d_t \leq \epsilon$ are highly accurate, being reached in all of our experiments within one or two iterations of the first iteration with $\delta_t \leq \epsilon$ or $p_t \leq \epsilon$.

First t satisfying the given stopping criteria						
Criteria	λ_k=1	λ_k=$k+1$	λ_k=$(k+1)^2$	λ_k=$(k+1)^3$	λ_k=$(k+1)^4$	λ_k in (3.3)
$\bar{p}_t \leq \epsilon$	1204821	1940	997	1331	1664	4122
$\bar{p}_t + d_t \leq \epsilon$	1204821	2000	1223	1630	2038	4156
$\delta_t \leq \epsilon$	237426	443222	664834	886445	1108056	533876
$\delta_t + d_t \leq \epsilon$	237428	443223	664835	886446	1108058	533876
$p_t \leq \epsilon$	4713468	886456	997251	1181927	1385070	1067789
$p_t + d_t \leq \epsilon$	4713468	886456	997252	1181928	1385071	1067790
$d_t \leq \epsilon$	263	470	705	941	1176	509

Table 1: Stopping times for different criteria and λ_k with $\epsilon = 0.05, \sigma = 0$.

Conditioning of problems (4.1) as σ varies						
σ	0	0.0001	0.001	0.01	0.02	0.05
L_1/μ	4	4.022	4.224	6.911	12.107	81.179
T_0	6	7	7	12	23	161
C_0	1.472×10^5	1.497×10^5	1.735×10^5	6.985×10^5	3.770×10^6	2.663×10^{23}

Table 2: Effects of σ on problem conditioning measured by L_1/μ and consequently the duration and amount of early divergences measured by T_0 and C_0 with $\alpha_k = 2/\mu(k+2)$.

4.3 Accuracy of C_0 at Predicting Early Iterate Divergence

Lastly, we consider settings where the initial iterates diverge rapidly, which our theory addresses via the inclusion of the constant C_0, defined in (2.10). Here, we have defined $C = \sigma\tilde{C} + I$, for a randomly Gaussian sampled $\tilde{C}$. As a result, the constants $\mu = \lambda_{min}(C^TC)$ and $L_1 = 4\lambda_{max}(C^TC)$ depend on σ. In Table 2, we show the effect σ varying from 0 to 0.05, causing the condition number L_1/μ to grow moderately. As a result, we see T_0 grow linearly in L_1/μ and C_0 grows exponentially, exceeding 10^{21}.

For such problems, our theory predicts the subgradient method with $\alpha_k = 2/\mu(k+2)$ may diverge in the first T_0 iterations but should eventually converge at least a $O(1/T)$ rate. Figure 2 numerically confirms this prediction with every performance measure exponentially growing to at least 10^{16} as σ grows and a decreasing trend beginning before iteration T_0. We see $\bar{p}_k$, δ_k and $R_k^2(x_{\mathtt{OPT}})$ rapidly converge after T_0, whereas p_k and d_k only decrease sublinearly. This slow convergence is likely due to p_k and d_k being defined as weighted averages, which must slowly dilute the effects of early "bad" iterations. Our theory predicts such exponential divergences can be avoided by ensuring T_0 (and hence C_0) are small. For example, setting $\alpha_0 = 1/\mu$ and then $\alpha_k = \min\{1/L_1, 2/\mu(k+2)\}$ thereafter rather than $\alpha_k = 2/\mu(k+2)$ as above ensures $T_0 = 0$. Figure 3 verifies this mitigates the previous diverging behavior.

Acknowledgements. This manuscript's presentation and clarity were greatly improved by the careful thoughts and advice of two anonymous referees to whom the authors are very grateful.

References

[1] N. Z. Shor, Krzysztof C. Kiwiel, and Andrzej Ruszcayǹski. *Minimization methods for non-differentiable functions*. Springer-Verlag New York, Inc., New York, NY, USA, 1985.

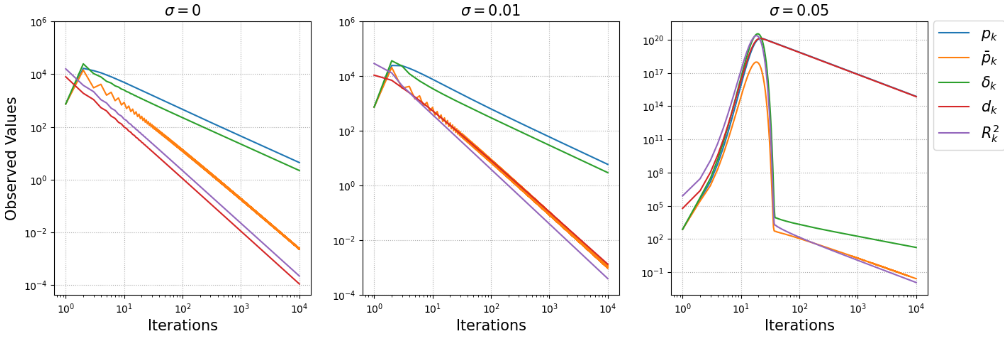

Figure 2: Observed performance for various σ with $\alpha_k = 2/\mu(k+2)$.

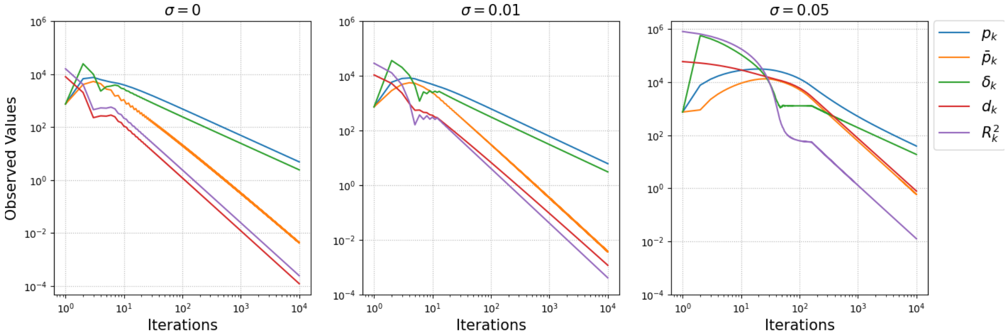

Figure 3: Observed performance for various σ with $\alpha_0 = 1/\mu$, $\alpha_k = \min\{1/L_1, 2/\mu(k+2)\}$, for $k > 0$, with corresponding $\lambda_0 = 1$, $\lambda_k = \frac{\alpha_k}{1-\mu\alpha_k}\frac{\lambda_{k-1}}{\alpha_{k-1}}$ and well-controlled $T_0 = 0$.

[2] Angelia Nedic and Dimitri P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

[3] Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2005.

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[5] Tianbao Yang and Qihang Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *J. Mach. Learn. Res.*, 19:6:1–6:33, 2015.

[6] Benjamin Grimmer. Radial subgradient method. *SIAM J. Optim.*, 28:459–469, 2017.

[7] Patrick R. Johnstone and Pierre Moulin. Faster subgradient methods for functions with hölderian growth. *Math. Program.*, 180(1–2):417–450, mar 2020.

[8] Haihao Lu. "relative continuity" for non-lipschitz nonsmooth convex optimization using stochastic (or deterministic) mirror descent. *INFORMS Journal on Optimization*, 2017.

[9] James Renegar and Benjamin Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Found. Comput. Math.*, 22(1):211–256, feb 2022.

[10] B. T. Polyak. A general method of solving extremum problems. *Sov. Math., Dokl.*, 8:593–597, 1967.

[11] Simon Lacoste-Julien, Mark Schmidt, and Francis R. Bach. A simpler approach to obtaining an o(1/t) convergence rate for the projected stochastic subgradient method. *CoRR*, abs/1212.2002, 2012.

[12] Qi Deng, Guanghui Lan, and Anand Rangarajan. Randomized block subgradient methods for convex nonsmooth and stochastic optimization. *arXiv: Optimization and Control*, 2015.

[13] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer, New York, NY, 2014.

[14] Lin Xiao. Dual averaging method for regularized stochastic learning and online optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 22, New York, 2009. Curran Associates, Inc.

[15] Xi Chen, Qihang Lin, and Javier Peña. Optimal regularized dual averaging methods for stochastic optimization. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 395–403, Red Hook, NY, USA, 2012. Curran Associates Inc.

[16] Vilen Jumutc and Johan A. K. Suykens. Reweighted l 2-regularized dual averaging approach for highly sparse stochastic learning. In *International Symposium on Neural Networks*, 2014.

[17] Jonathan W. Siegel and Jinchao Xu. Extended regularized dual averaging methods for stochastic optimization. *Journal of Computational Mathematics*, 2019.

[18] Conghui Tan, Yuqiu Qian, Shiqian Ma, and Tong Zhang. Accelerated dual-averaging primal–dual method for composite convex minimization. *Optimization Methods and Software*, 35(4):741–766, 2020.

[19] Guanghui Lan and Zhiqiang Zhou. Algorithms for stochastic optimization with function or expectation constraints. *Comput. Optim. Appl.*, 76(2):461–498, jun 2020.

[20] Anastasia Bayandina, Pavel Dvurechensky, Alexander Gasnikov, Fedor Stonyakin, and Alexander Titov. *Mirror Descent and Convex Optimization Problems with Non-smooth Inequality Constraints*, pages 181–213. Springer International Publishing, Cham, 2018.

[21] Mohammad S. Alkousa. *On Modification of an Adaptive Stochastic Mirror Descent Algorithm for Convex Optimization Problems with Functional Constraints*, pages 47–63. Springer Singapore, Singapore, 2020.

[22] Alexander A. Titov, Fedor S. Stonyakin, Mohammad S. Alkousa, Seydamet S. Ablaev, and Alexander V. Gasnikov. Analogues of switching subgradient schemes for relatively lipschitz-continuous convex programming problems. In Yury Kochetov, Igor Bykadorov, and Tatiana Gruzdeva, editors, *Mathematical Optimization Theory and Operations Research*, pages 133–149, Cham, 2020. Springer International Publishing.

[23] Zhichao Jia and Benjamin Grimmer. First-order methods for nonsmooth nonconvex functional constrained optimization with or without slater points, 2023.

[24] Yankun Huang and Qihang Lin. Single-loop switching subgradient methods for non-smooth weakly convex optimization with non-smooth convex constraints, 2023.

[25] Javier Peña. Convergence of first-order methods via the convex conjugate. *Operations Research Letters*, 45(6):561–564, 2017.

[26] David Huckleberry Gutman and Javier F. Pena. Convergence rates of proximal gradient methods via the convex conjugate. *SIAM J. Optim.*, 29:162–174, 2018.

[27] David H. Gutman and Javier F. Peña. Perturbed fenchel duality and first-order methods. *Math. Program.*, 198(1):443–469, feb 2022.

[28] Jelena Diakonikolas and Lorenzo Orecchia. The approximate duality gap technique: A unified theory of first-order methods. *SIAM J. on Optimization*, 29(1):660–689, jan 2019.

[29] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, ICML'12, page 1571–1578, Madison, WI, USA, 2012. Omnipress.

[30] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, page I–71–I–79, 2013.

[31] E. Gustavsson, M. Patriksson, and AB Strömberg. Primal convergence from dual subgradient methods for convex optimization. *Mathematical Programming*, 150:365–390, 2015.

[32] M.R. Metel and A. Takeda. Primal-dual subgradient method for constrained convex optimization problems. *Optimization Letters*, 15:1491–1504–390, 2021.

[33] Benjamin Grimmer. Convergence rates for deterministic and stochastic subgradient methods without lipschitz continuity. *SIAM Journal on Optimization*, 29(2):1350–1365, 2019.

[34] R. Tyrrell Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer Verlag, Heidelberg, Berlin, New York, 1998.

[35] Daureen Steinberg. Computaiton of matrix norms with applications to robust optimization. Master's thesis, Israel Institute of Technology, 2005.

A Deferred Proof of Lemma 2.4

Consider either $y \in \{x_{\texttt{OPT}}, x_{\texttt{SL}}\}$. Suppose first x_k is feasible and let $\bar{g}_k = \mathbb{E}_{\xi_k} g_0(x_k;\xi_k) + n_y \in \partial h_y(x_k)$. Fix any $g_y \in \partial h_y(y)$. Note by the sum rule of subdifferential calculus, both $\bar{g}_k - \nabla f_0^{(2)}(x_k) - n_y$ and $g_y - \nabla f_0^{(2)}(y) - n_y$ are subgradients of $f_0^{(1)}$ and hence both have norm bounded by M. Consider the L-smooth function

$$\hat{h}_y(x) = f_0^{(2)}(x) + f_0^{(1)}(y) + \langle g_y - \nabla f_0^{(2)}(y), x - y\rangle \ .$$

Note since $g_y - \nabla f_0^{(2)}(y) - n_y \in \partial f_0^{(1)}(y)$, $h_y \geq \hat{h}_y$ and $h_y(y) = \hat{h}_y(y)$. Then one has

$$\begin{aligned}
&\mathbb{E}_{\xi_k}\|g_0(x_k;\xi_k) + n_y\|_2^2\\
&= \|\bar{g}_k\|_2^2 + \mathbb{E}_{\xi_k}\|g_0(x_k;\xi_k) - \bar{g}_k\|_2^2\\
&\leq 3\|\nabla\hat{h}_y(x_k)\|_2^2 + 3\|\bar{g}_k - \nabla f_0^{(2)}(x_k) - n_y\|^2 + 3\|g_y - \nabla f_0^{(2)}(y) - n_y\|_2^2 + \sigma^2\\
&\leq 6L(\hat{h}_y(x_k) - \inf \hat{h}_y) + 6M^2 + \sigma^2\\
&\leq 6L\delta_k(y) + 6L(h_y(y) - \inf \hat{h}_y) + 6M^2 + \sigma^2
\end{aligned}$$

where the first inequality bounds $\|a+b+c\|_2^2$ by $3\|a\|_2^2 + 3\|b\|_2^2 + 3\|c\|_2^2$ and uses the assumed variance bound, the second inequality uses smoothness to bound the first term as $\hat{h}_y(x_k) - \frac{1}{2L}\|\nabla\hat{h}_y(x_k)\|^2 \geq \inf \hat{h}_y$ and the M-Lipschitzness of $f_0^{(1)}$ to bound the second and third terms, and the final inequality adds and subtracts $h_y(y)$ and upper bounds $\hat{h}_y(x_k)$ by $h_y(x_k)$.

Similarly, now suppose x_k is infeasible and let $\bar{g}_k = \mathbb{E}_{\xi_k} g_{s(x_k)}(x_k;\xi_k) \in \partial f_{s(x_k)}(x_k)$. Fix any $g_y \in \partial f_{s(x_k)}(y)$. Note by the sum rule of subdifferential calculus, both $\bar{g}_k - \nabla f_{s(x_k)}^{(2)}(x_k)$ and $g_y - \nabla f_{s(x_k)}^{(2)}(y)$ are subgradients of $f_{s(x_k)}^{(1)}$ and hence both have norm bounded by M. Consider the L-smooth function

$$\hat{f}_{s(x_k)}(x) = f_{s(x_k)}^{(2)}(x) + f_{s(x_k)}^{(1)}(y) + \langle g_y - \nabla f_{s(x_k)}^{(2)}(y), x - y\rangle \ .$$

Note since $g_y - \nabla f_{s(x_k)}^{(2)}(y) \in \partial f_{s(x_k)}^{(1)}(y)$, $f_{s(x_k)} \geq \hat{f}_{s(x_k)}$ and $f_{s(x_k)}(y) = \hat{f}_{s(x_k)}(y)$. Then, identical reasoning to that above gives

$$\begin{aligned}
&\mathbb{E}_{\xi_k}\|g_0(x_k;\xi_k)\|_2^2\\
&= \|\bar{g}_k\|_2^2 + \mathbb{E}_{\xi_k}\|g_{s(x_k)}(x_k;\xi_k) - \bar{g}_k\|_2^2\\
&\leq 3\|\nabla\hat{f}_{s(x_k)}(x_k)\|_2^2 + 3\|\bar{g}_k - \nabla f_{s(x_k)}^{(2)}(x_k)\|^2 + 3\|g_y - \nabla f_{s(x_k)}^{(2)}(y)\|_2^2 + \sigma^2\\
&\leq 6L(\hat{f}_{s(x_k)}(x_k) - \inf \hat{f}_{s(x_k)}) + 6M^2 + \sigma^2\\
&\leq 6L\delta_k(y) + 6L(f_{s(x_k)}(y) - \inf \hat{f}_{s(x_k)}) + 6M^2 + \sigma^2 \ .
\end{aligned}$$

Hence, Assumption C holds with

$$\begin{aligned}
&L_0^2 = 6M^2 + \sigma^2 + 6L \max_{y\in\{x_{\texttt{OPT}}, x_{\texttt{SL}}\}} \max_{s=1\dots m} \left\{h_y(y) - \inf \hat{h}_y, f_s(y) - \inf \hat{f}_s\right\}\\
&L_1 = 6L \ .
\end{aligned}$$