Boosted Stochastic Frank-Wolfe for Constrained Nonconvex Optimization

Navil Nandhan * **Abbas Khademi** † **Antonio Silveti-Falls** ‡

Abstract

The boosted Frank-Wolfe algorithm accelerates the classical Frank-Wolfe algorithm by better aligning the update direction with the negative gradient. Its analysis, however, has been limited to deterministic convex problems, with step sizes that require either line search or knowledge of the Lipschitz constant of the gradient. We develop a novel step size strategy that does not depend on the Lipschitz constant of the gradient, which allows us to extend the boosted Frank-Wolfe algorithm to the stochastic setting. We prove that boosting with this step size strategy can be combined with many modern gradient estimators, including SAGA, L-SVRG, SAG, Heavy Ball momentum, and zeroth-order estimators, among others, while retaining the worst-case convergence rates of ordinary stochastic Frank-Wolfe. Our analysis also yields the first convergence rates for boosted Frank-Wolfe on nonconvex and quasar-convex objectives, results which are new even for deterministic problems. Experiments on sparse logistic regression and quantum process tomography show that stochastic boosted Frank-Wolfe achieves faster convergence per gradient oracle call (and on wall-clock) compared to the non-boosted baseline.

1 Introduction

The Frank-Wolfe (FW) or Conditional Gradient algorithm is a method for solving constrained optimization problems that avoids projections onto the constraint set [7, 19]. Instead, it only requires access to the gradient and a linear minimization oracle (LMO) over the constraint set. This oracle can be much cheaper to compute than the projection in many problems of interest, e.g., low-rank inverse problems with nuclear norm regularization, traffic assignment, video co-localization and more (see [4] for a more exhaustive list of examples). With this potentially lower per-iteration cost and a worst-case convergence rate of $\mathcal{O}(1/t)$ for smooth convex problems, FW appears to be a competitive method. However, in practice, the method can suffer from a zigzagging phenomenon that slows convergence.

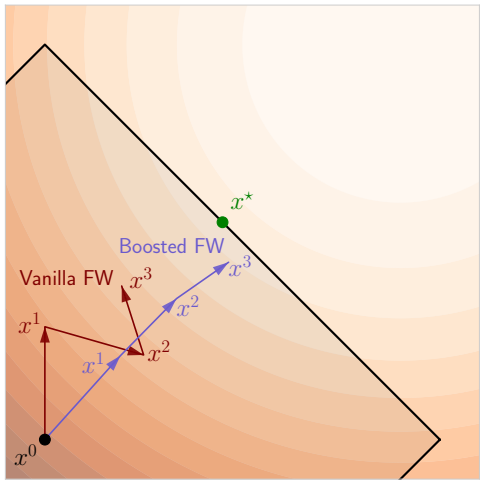

Figure 1: Comparison of FW and BFW on a toy problem. The boundary of the constraint set is shown in black and $x^\star$ marks the minimizer in green. The FW trajectory in red shows zigzagging while the BFW trajectory in purple avoids this.

In practice, the LMO always returns an extreme point of the constraint set, e.g., a vertex if the set is polyhedral. If the optimum lies in between two extreme points, then the LMO outputs will alternate between these two points, which causes the zigzagging. Because zigzagging slows convergence by wasting updates on movement that is orthogonal to the

*National University of Singapore (NUS), Singapore.
e0727209@u.nus.edu
†School of Mathematics and Computer Science, Iran University of Science and Technology, Iran.
abbaskhademi92@gmail.com
‡CVN, CentraleSupélec, Université Paris-Saclay, Inria, France.
tonys.falls@gmail.com

Preprint.

optimal descent direction, many FW variations have been proposed which aim to extend beyond simply moving towards an extreme point, e.g., away-step FW changes the LMO to include moving away from extreme points. The boosted Frank-Wolfe (BFW) algorithm uses the LMO multiple times per-iteration to refine and align the update direction with the negative gradient, in order to avoid the zigzagging [3]. Figure 1 shows how the zigzagging behavior is mitigated in BFW.

While boosting is effective for avoiding zigzagging, the existing analysis of BFW in [3] assumes that the function to be minimized is convex, ruling out all nonconvex problems of interest. Moreover, the step sizes they suggest require either knowing the Lipschitz constant of the gradient or performing a line search, and are only guaranteed to converge for convex functions. As a result, the BFW algorithm has yet to be applied to nonconvex problems, nor has it reaped the benefits of the wide literature on inexact Frank-Wolfe, such as stochastic, zeroth-order, or distributed methods summarized in [29].

We overcome these limitations by developing a new step size strategy for the BFW algorithm that requires neither knowledge of the Lipschitz constant of the gradient nor a line search. This enables the extension to the inexact gradient setting, allowing the algorithm to be applied to stochastic problems. We prove the convergence of the BFW algorithm with the proposed step size using several different gradient estimators, matching the convergence rate of BFW with known Lipschitz constant and exact gradients given in [3] and that of the stochastic FW algorithm [29] in all settings considered. Our analysis provides the first convergence rates for the BFW algorithm in the nonconvex and quasar-convex cases. Establishing these rates was previously unaddressed, even in the deterministic setting. We confirm the effectiveness of our step size in experiments on sparse logistic regression, where we observe improvements over vanilla Frank-Wolfe for every gradient estimator considered.

Contributions Our contributions are three-fold:

Theory: We demonstrate the first convergence guarantees for deterministic BFW without assuming convexity of the objective function. In the nonconvex setting we show that the so-called FW gap converges with a rate of $\mathcal{O}(1/\sqrt{t})$. We also show that the functional-value gap converges with a rate of $\mathcal{O}(1/t)$ under a relaxed assumption of quasar-convexity, which includes the convex setting and matches the rate given in [3]. We also introduce stochastic BFW (BSFW) and prove convergence in expectation with a rate of $\mathcal{O}(1/t)$ in the quasar-convex case and $\mathcal{O}(1/\sqrt{t})$ in the nonconvex case. Our step size is the first provably convergent strategy for BFW that does not require the Lipschitz constant of the gradient nor line search. In the quasar-convex (and hence also for convex) setting, we obtain the same worst-case convergence rate as the existing BFW method despite relaxing assumptions.

Unification: Through a general assumption on gradient inexactness, we simultaneously cover stochastic gradient estimators like Heavy Ball, SAGA, SARAH, L-SVRG; zeroth-order methods like JAGUAR; among others (eight in total), matching the results of [29] for stochastic FW.

Experimental Confirmation: Our experimental results show a clear improvement over the vanilla FW version of each estimator we consider on sparse logistic regression in terms of number of stochastic gradients processed or number of coordinate gradients processed.

Finally, our treatment of quasar-convexity is primarily for completeness since there are several ML problems involving quasar-convexity, but it is not central to the novelty of our work. A discussion about this is given in Appendix A.2.

Related work Several works have proposed modifications of FW to avoid zigzagging, one of the earliest examples being the away-step FW [9, 37]. More recently, such methods include the pairwise FW [17], the blended pairwise FW [34], and the BFW [3], which we directly extend to the nonconvex and stochastic settings.

For stochastic FW on nonconvex problems, many prior works exist covering convergence analysis, e.g., early work in [33] followed by [29] for several different variance-reduced estimators and [31] for the Heavy Ball momentum estimator. However, none of those analyses include boosting; the original analysis of BFW in [3] could not accommodate stochasticity and our work bridges this gap.

The original analysis of BFW in [3] relies on convexity, which we relax to quasar-convexity, or drop entirely in the nonconvex case. Quasar-convexity has been used in several recent works [14], e.g., as a practical relaxation of convexity. It has also been used to analyze FW with improved rates compared to the general nonconvex case in [15, 25, 26] and now our work. It was used in other first-order algorithms in [5, 8, 18], but its use in BFW has been unexplored until now.

2 Preliminaries

We consider the following optimization problem

$$\min_{x \in \mathcal{C}} f(x), \tag{P}$$

where $\mathcal{C} \subset \mathbb{R}^n$ is a nonempty compact convex subset and f is a continuously differentiable function whose gradient is Lipschitz-continuous on $\mathcal{C}$. We will assume that the projection onto the set $\mathcal{C}$ is unavailable in closed-form or otherwise computationally intractable. Instead, we will assume access to the LMO over the set $\mathcal{C}$, defined for all $v \in \mathbb{R}^n$ as

$$\operatorname{lmo}(v) = \operatorname*{argmin}_{s \in \mathcal{C}} \langle s, v \rangle.$$

For many common constraint sets $\mathcal{C}$ used in machine learning, the LMO is computable in closed form or otherwise cheaper than the corresponding projection operation; we refer to [4] for an in-depth study on this comparison.

For the analysis of first-order methods applied to unconstrained minimization of smooth nonconvex functions, the gradient norm $\|\nabla f(x)\|$ is typically employed as a surrogate measure of optimality. However, in constrained problems like (P), the gradient is not necessarily 0 at a critical point. Instead, a critical point corresponds to

$$0 \in \nabla f(x) + \mathrm{N}_{\mathcal{C}}(x),$$

where $\mathrm{N}_{\mathcal{C}}(x)$ is the usual normal cone to the set $\mathcal{C}$ at x. We will therefore analyze convergence using the *Frank–Wolfe gap* at a point $x \in \mathcal{C}$, which is defined as

$$\operatorname{Gap}(x) := \max_{s \in \mathcal{C}} \langle \nabla f(x),\, x - s \rangle. \tag{1}$$

This quantity is exactly the analog of $\|\nabla f(x)\|$ in the constrained setting, in the sense that it is nonnegative and certifies first-order optimality as

$$\operatorname{Gap}(x) = 0 \iff 0 \in \nabla f(x) + \mathrm{N}_{\mathcal{C}}(x).$$

In deterministic problems, this quantity is easily computed at run-time since

$$\operatorname{Gap}(x) = \langle \nabla f(x), x - \operatorname{lmo}(\nabla f(x)) \rangle\,,$$

which is the inner product of quantities already computed during FW or BFW.

Assumptions Throughout the paper, the $\|\cdot\|$ norm is defined as the standard Euclidean norm, i.e., $\|\cdot\| = \|\cdot\|_2$. We now formalize the smoothness assumption that we make on f in problem (P).

Assumption 1 (L-Smoothness)**.** The gradient ∇f is Lipschitz-continuous on the set $\mathcal{C}$ with constant $L > 0$, i.e.,

$$\exists L > 0;\ \forall x, y \in \mathcal{C}: \quad \|\nabla f(x) - \nabla f(y)\| \leq L \|x - y\|.$$

In addition to general nonconvex functions, we will also consider the class of *quasar-convex* functions that satisfy the following assumption.

Assumption 2 (Quasar-Convexity)**.** The function f is *quasar-convex* with parameter $\rho \in\,]0, 1]$, i.e., there exists $x^\star \in \operatorname{argmin}_{x \in \mathcal{C}} f(x)$ such that

$$\forall y \in \mathcal{C}: \quad f^\star - f(y) \geq \tfrac{1}{\rho} \langle \nabla f(y),\, x^\star - y \rangle.$$

Note that quasar-convexity recovers the well-known star-convexity when $\rho = 1$. Furthermore, every convex function with a minimizer is quasar-convex with $\rho = 1$, so our convergence analyses in this context will also encompass all convex functions.

Notation Considering the problem (P), we define $f^\star = \min_{x \in \mathcal{C}} f(x)$, and the functional-value gap at iteration t by $F_t = f(x^t) - f^\star$. For stochastic problems, we will denote m^t as the stochastic estimator of the deterministic gradient $\nabla f(x^t)$ and $\Delta^t = m^t - \nabla f(x^t)$ as the difference between the stochastic estimator and the deterministic gradient. We denote the diameter of the set $\mathcal{C}$ by $D = \max_{x,y \in \mathcal{C}} \|x - y\|$. The notation $\mathbb{E}$ is defined as the full expectation, while $\mathbb{E}_t[\cdot]$ is defined as the conditional expectation with respect to the randomness generated until iteration t, i.e., $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot \mid \sigma(x^0, x^1, \cdots, x^t)]$ where $\sigma(x^0, x^1, \cdots, x^t)$ is the σ-algebra generated by the random variables $x^0, x^1, \cdots, x^t$.

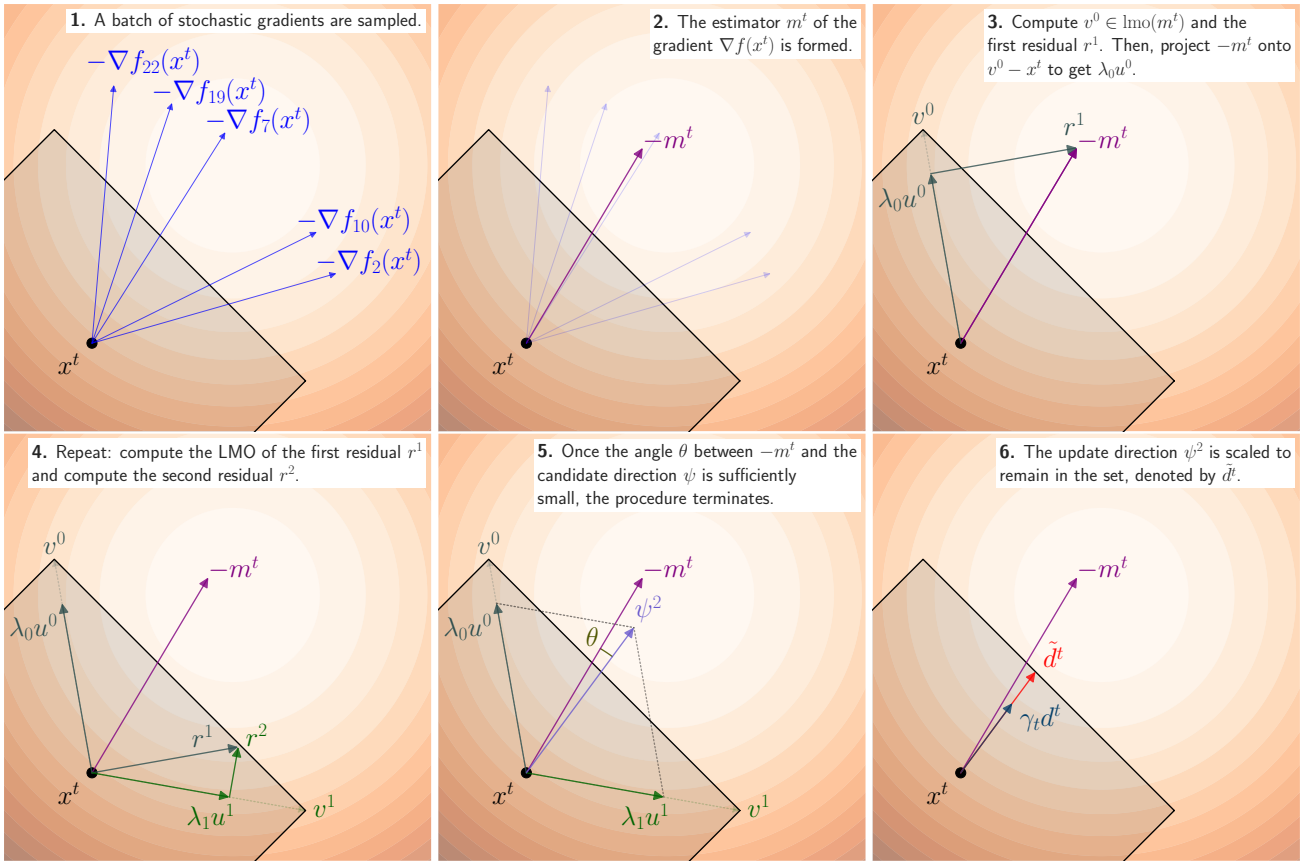

Figure 2: Diagram of the stochastic boosting procedure used in Algorithm BSFW. Panels 1 and 2: At iteration t, stochastic gradients are sampled and $-m^t$ is formed to estimate $-\nabla f(x^t)$ based on the estimator chosen, e.g., Heavy Ball as we show here. Panel 3: The LMO is calculated for $-m^t$ and the first residual r^1 is computed by projecting $-m^t$ onto $v^0 - x^t$. Panel 4: The procedure then proceeds recursively, computing the LMO for the current residual r^k and using that to compute the next residual r^{k+1}. Panel 5: A candidate direction ψ^k is formed from the projections of the residuals and, if the alignment of ψ^k with $-m^t$ has not improved enough, then the procedure terminates. Panel 6: The magnitude of the candidate direction ψ^t is scaled $\tilde{d}^t = \psi^2/(\lambda_0 + \lambda_1)$ to remain in the set $\mathcal{C}$. Finally, this feasible direction is scaled by the step size and $x^{t+1} = x^t + \gamma_t d^t$ can be computed.

3 Methods

3.1 Boosting: Why, What, How, and When?

To avoid zigzagging, the algorithm should update in the direction of $-\nabla f(x^t)$. However, such an update may not ensure feasibility and thus typically requires projection back onto $\mathcal{C}$, which we aim to avoid in this work. This motivates the FW algorithm, which avoids projection by constructing its update using a convex combination between the current iterate x^t and an extreme point $s^t \in \mathrm{lmo}(\nabla f(x^t))$ of $\mathcal{C}$. The resulting direction $s^t - x^t$ may be nearly orthogonal or otherwise not aligned with $-\nabla f(x^t)$, for instance when the optimum lies between extreme points of $\mathcal{C}$ or in a face of the boundary $\partial\mathcal{C}$ as in Figure 1.

For many constrained problems that arise in machine learning, the LMO associated to the set $\mathcal{C}$ has low computational overhead relative to the gradient. In [3], the authors leverage this by using the LMO several times per-iteration in a *boosting* procedure to find a feasible direction that is better aligned with $-\nabla f(x^t)$ than $\mathrm{lmo}(\nabla f(x^t))$, the output used in FW.

The boosting procedure, corresponding to **Boost** in Algorithm BSFW (expanded in Algorithm Boost) and demonstrated in Figure 2 panels 3-5 (if we replace the stochastic estimator $-m^t$ by $-\nabla f(x^t)$), approximates the conical decomposition of

$-\nabla f(x^t)$ in $\text{cone}(\mathcal{C} - x^t)$, i.e., it approximately solves the problem

$$\underset{d \in \text{cone}(\mathcal{C}-x^t)}{\text{argmin}} \ \tfrac{1}{2}\| - \nabla f(x^t) - d\|^2.$$

This decomposition of $-\nabla f(x^t)$ is constructed through multiple refinement steps, each of which starts by computing a residual $r^k = -\nabla f(x^t) - \psi^k$ (the part of $-\nabla f(x^t)$ which is not captured by the current approximate decomposition ψ^k) and then calling the LMO to find an extreme point $v^k \in \text{lmo}(-r^k)$ most aligned with r^k. The candidate direction is then updated to ψ^{k+1} by including $v^k - x^t$ in the decomposition. This recursive refinement continues until insufficient progress is made, as measured by the alignment between the candidate direction ψ^k and $-\nabla f(x^t)$ (lines 20-26 in Algorithm Boost). The alignment is measured through the modified cosine similarity, defined between two vectors d and $\hat{d}$ as

$$\text{align}(d, \hat{d}) := \begin{cases} \dfrac{\langle d, \hat{d} \rangle}{\|d\|\|\hat{d}\|}, & \hat{d} \neq 0 \\ -1. & \hat{d} = 0 \end{cases}$$

This gives a decomposition of $-\nabla f(x^t)$ in terms of some set $\{v^0 - x^t, v^1 - x^t, \cdots, v^{K_t} - x^t\} \subset \text{cone}(\mathcal{C} - x^t)$ with coefficients $\lambda_1, \cdots, \lambda_{K_t}$, e.g., Figure 2 Panel 5 shows ψ^2 as a conical combination $\lambda_0(v^0 - x^t) + \lambda_1(v^1 - x^t)$. We note that line 11 allows for the analog of an "away-step" in the construction of the conical decomposition and is necessary for the convergence of the procedure as noted in [3, 23].

This approximate conical decomposition is what ensures that $x^t + \tilde{d}^t$ remains in $\mathcal{C}$, as it is used to compute the normalization in line 29 of Algorithm Boost with Λ^t. Indeed, it is the construction of ψ^k as a positive combination of extreme points in $\mathcal{C}$ that allows us to normalize ψ^k into the feasible update $\tilde{d}^t$ which is then scaled by the step size γ_t to get the final update.

Computing the exact conical decomposition of $-\nabla f(x^t)$ is not necessary, which is why the boosting procedure includes two hyperparameters: the maximum number of refinements, i.e., LMO calls, K and an alignment improvement tolerance δ, so that if doing another refinement of the direction does not improve the alignment by at least δ, the procedure will terminate. Algorithm BSFW (Full) in Appendix A.1 is the complete Boosted Stochastic FW algorithm.

Algorithm BSFW Boosted Stochastic Frank-Wolfe (Full version stated in Appendix A.1)

Input: initial estimator $m^{\text{init}} \in \mathbb{R}^n$, gradient estimator $\{\Phi_t\}_{t=0}^{T-1}$, max no of rounds for boosting $K \in \mathbb{N} \setminus \{0\}$, alignment improvement tolerance $\delta \in]0, 1]$, and step decay $\{\eta_t\}_{t=0}^{T-1} \in]0, 1]$
Output: $x^T \in \mathcal{C}$.

1: $x^0 \leftarrow \text{lmo}(m^{\text{init}})$
2: **for** $t = 0$ to $T - 1$ **do**
3: Sample $\xi_t \sim \mathcal{P}$ and compute $g^t = \nabla f(x^t, \xi_t)$
4: Compute $m^t = \Phi_t(g^t)$
5: $\tilde{d}^t \leftarrow$ **Boost**(m^t) {perform boosting procedure}
6: $\gamma_t \leftarrow \min\left\{\eta_t \frac{\|\text{lmo}(m^t) - x^t\|}{\|\tilde{d}^t\|}, 1\right\}$ **if** $\tilde{d}^t \neq 0$ **else** 1
7: **if** $\gamma_t < 1$ **then**
8: $d^t \leftarrow \tilde{d}^t$ {use the boosted direction}
9: $x^{t+1} \leftarrow x^t + \gamma_t d^t$
10: **else**
11: $d^t \leftarrow \text{lmo}(m^t) - x^t$ {revert to vanilla FW direction}
12: $x^{t+1} \leftarrow x^t + \eta_t d^t$
13: **end if**
14: **end for**

Algorithm Boost (Boosting Procedure)

Input: estimator $m^t \in \mathbb{R}^n$, max no of rounds for boosting $K \in \mathbb{N} \setminus \{0\}$, and alignment improvement tolerance $\delta \in]0, 1]$
Output: boosting direction $\tilde{d}^t$

$\psi^0 \leftarrow 0$
$\Lambda_t \leftarrow 0$
$k \leftarrow 0$
while $k \leq K - 1$ **do**
  $r^k \leftarrow -m^t - \psi^k$ {k-th residual}
  $v^k \leftarrow \mathrm{lmo}(-r^k)$ {FW oracle}
  if $k = 0$ **then**
    $s^t \leftarrow v^k$
  end if
  if $\psi^k \neq 0$ **then**
    $u^k \leftarrow \operatorname{argmax}_{u \in \left\{ v^k - x^t, -\frac{\psi^k}{\|\psi^k\|} \right\}} \langle r^k, u \rangle$
  else
    $u^k \leftarrow v^k - x^t$
  end if
  if $u^k = 0$ **then**
    $k \leftarrow k + 1$; **break** {exit k-loop}
  end if
  $\lambda_k \leftarrow \frac{\langle r^k, u^k \rangle}{\|u^k\|^2}$
  $\phi^k \leftarrow \psi^k + \lambda_k u^k$
  if $\mathrm{align}(-m^t, \phi^k) - \mathrm{align}(-m^t, \psi^k) \geq \delta$ **then**
    $\psi^{k+1} \leftarrow \phi^k$
    $\Lambda_t \leftarrow \begin{cases} \Lambda_t + \lambda_k, & u^k = v^k - x^t \\ \Lambda_t \left(1 - \frac{\lambda_k}{\|\psi^k\|}\right), & u^k = -\frac{\psi^k}{\|\psi^k\|} \end{cases}$
    $k \leftarrow k + 1$
  else
    $k \leftarrow k + 1$; **break** {exit k-loop}
  end if
end while
$K_t \leftarrow k$
$\tilde{d}^t \leftarrow \frac{\psi^{K_t}}{\Lambda_t}$ **if** $\Lambda_t \neq 0$ **else** 0 {normalize direction}

3.2 Step Size Strategy

Prior work on boosting [3] has used either a line search $\gamma_t \in \operatorname{argmin}_{\gamma \in [0,1]} f(x^t + \gamma \tilde{d}^t)$ or $\gamma_t = \min\left\{\mathrm{align}(-\nabla f(x^t), \tilde{d}^t)\|\nabla f(x^t)\|/(L\|\tilde{d}^t\|), 1\right\}$, a strategy similar in spirit to the short-step in vanilla FW.

At every iteration t, we define our step size by

$$\gamma_t = \min\left\{\eta_t \frac{\|s^t - x^t\|}{\|\tilde{d}^t\|}, 1\right\}, \tag{2}$$

where η_t is a step decay, $s^t \in \mathrm{lmo}(m^t)$, and $\tilde{d}^t$ is the aligned output from the boosting procedure (unless $\gamma_t = 1$, in which case the update is not $\tilde{d}^t$ but rather $d^t = s^t - x^t$; this does not occur in practice as we demonstrate in Section 5). Lines 10-13 in Algorithm BSFW describe the revert-to-FW step, which is the same as in [3]. The scaling by $\frac{\|s^t - x^t\|}{\|\tilde{d}^t\|}$ gives an update that is similar in magnitude to FW. This scaling, combined with appropriately chosen step decay $\{\eta_t\}$, guarantees convergence of Algorithm BSFW similar to FW with open-loop step size strategies. Contrasting this with the strategies given

in [3] for convex functions, our step size avoids line searches and knowledge of the Lipschitz constant L. A short note about the complexity of the step size is given in Appendix A.4.

3.3 Gradient Estimators

We are able to combine the boosting procedure with a slew of stochastic estimators that satisfy Assumption 3. Table 2 summarizes the estimators we consider and which we prove satisfy Assumption 3 with Algorithm BSFW in Appendix C. The main consideration in choosing estimators for Algorithm BSFW is variance reduction rather than unbiasedness. This is more obvious when one considers that the LMO output $s^t \in \mathrm{lmo}(m^t)$ need not be an unbiased estimate for $\mathrm{lmo}(\nabla f(x^t))$ even if m^t is unbiased for $\nabla f(x^t)$, since the LMO is typically discontinuous for the constraint sets $\mathcal{C}$ that are common in machine learning. Since the boosting procedure relies on the LMO heavily, it is thus also better adapted to estimators with variance reduction such as the ones we consider, e.g., SARAH, SAG, SAGA, and many others. It also clarifies why we do not require $C = 0$ for all of our convergence results (although only one estimator, ZOJA, requires $C \neq 0$): there can be substantial bias as long as the variance of the estimator decreases appropriately.

4 Analysis

We divide our results between the deterministic and stochastic settings. For the sake of presentation, we defer the proofs of the convergence analyses to the appendix. We emphasize that the rates we prove are *non-asymptotic*, and that we use the big-$\mathcal{O}$ notation for convenience rather than necessity; when constants are omitted in the main text, we present them in the appendix. We will make a distinction between any-time convergence results, that do not require specifying the horizon (number of iterations) T in advance, and horizon-dependent convergence results. We cannot describe all the stochastic estimators we use in the main body of the paper so we compile Table 2 in the Appendix, which describes them in full detail and shows that they satisfy Assumption 3.

4.1 Deterministic Setting

In this subsection, we will assume that $m^t = \nabla f(x^t)$, i.e., the deterministic gradient is computed exactly. Our first result is a $\mathcal{O}(1/t)$ any-time convergence rate on the functional-value gap when f satisfies both Assumption 1 and Assumption 2.

Theorem 4.1 (Convergence Rate for Deterministic Quasar-Convex Problems)**.** *Let f be a function that satisfies Assumptions 1 and 2. Consider the sequence $\{x^t\}_{t=0}^{+\infty}$ generated by Algorithm BSFW with $m^t = \nabla f(x^t)$ and $\eta_t = \frac{2}{\rho(t+2)}$. Then, for all $t \geq 0$,*

$$F_t \leq \frac{1}{t+1} \max\left\{F_0, \frac{2LD^2}{\rho^2}\right\} = \mathcal{O}\left(\frac{1}{t}\right).$$

□

Remark 4.2. In the convex setting, $\rho = 1$ and the step size is parameter agnostic: it does not depend on knowledge of any of the problem-specific constants like the Lipschitz constant L of ∇f.

The next result guarantees a $\mathcal{O}(1/\sqrt{t})$ any-time convergence rate on the Frank-Wolfe gap defined in (1) by adjusting the step decay $\{\eta_t\}$ compared to the quasar-convex case.

Theorem 4.3 (Convergence for Deterministic Nonconvex Problems)**.** *Let f be a function that satisfies Assumption 1. Consider the sequence $\{x^t\}_{t=0}^{+\infty}$ generated by Algorithm BSFW with $m^t = \nabla f(x^t)$ and $\eta_t = \frac{1}{\sqrt{t+1}}$. Then, for all $t \geq 0$,*

$$\min_{0 \leq i \leq t} \mathrm{Gap}(x^i) \leq \frac{F_0 + LD^2}{\sqrt{t+1}} = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right).$$

□

Remark 4.4. The step size used in the convergence result above is parameter agnostic in addition to being an any-time guarantee. A horizon-dependent convergence rate for Algorithm BSFW in the deterministic setting with a step size that depends on the horizon T is given in Theorem B.6 in Appendix B, with an improved constant but the same order of convergence.

4.2 Stochastic Setting

For the convergence analysis in the stochastic setting, we will make a general assumption on the second moment of the error of the stochastic estimator of the gradient, summarized in Assumption 3 and inspired by Assumption 2.1 of [29]. Under this assumption, we can give convergence results for BSFW applied to both nonconvex and quasar-convex problems. Since many stochastic estimators can be shown to satisfy this assumption when used with BSFW, the resulting analysis is unified.

Assumption 3 (Estimator Assumptions [29])**.** Let $\{x^t\}_{t=0}^T$ denote the sequence of iterates generated by Algorithm BSFW. There exist constants $A, B, C, E \geq 0$, parameters $\rho_1, \rho_2 \in]0, 1]$, and a (possibly random) sequence $\{\sigma_t\}_{t\geq 0}$ such that the following conditions hold $\forall t \geq 1$:

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|\Delta^t\|^2\right] &\leq (1-\rho_1)\|\Delta^{t-1}\|^2 + A\sigma_{t-1}^2 + \eta_{t-1}^2 BD^2 + C,\\ \mathbb{E}_{t-1}\left[\sigma_t^2\right] &\leq (1-\rho_2)\sigma_{t-1}^2 + \eta_{t-1}^2 ED^2.\end{aligned}$$

In the appendix, we show that a slew of stochastic estimators satisfy this assumption when used with BSFW; this is then summarized in Table 2.

As in the deterministic setting, we start with quasar-convex functions with an appropriately chosen step decay $\{\eta_t\}$. However, in this setting we will use horizon-dependent step decays and any-time versions.

Theorem 4.5 (Convergence for Stochastic Quasar-Convex Problems)**.** *Let f be a function that satisfies Assumptions 1 and 2. Suppose the stochastic gradient estimator m^t and auxiliary sequence $\{\sigma_t\}$ satisfy Assumption 3 with parameters $\rho_1, \rho_2 \in]0, 1]$ and constants $A, B, C, E \geq 0$. Consider the sequence $\{x^t\}_{t=0}^T$ generated by Algorithm BSFW by fixing $T \in \mathbb{N}$ and using the constant, horizon-dependent step decay*

$$\eta_t = \begin{cases} \frac{1}{\rho d}, & T \leq d \\ \frac{1}{\rho d}, & T > d \text{ and } t \leq t_0 \\ \frac{2}{\rho(2d+t-t_0)}, & T > d \text{ and } t \geq t_0 \end{cases}$$

with $d := \frac{2}{\min\{\rho_1,\rho_2\}}$ and $t_0 := \lfloor T/2 \rfloor$. Then,

$$\mathbb{E}[F_T] = \mathcal{O}\left(\frac{1}{T} + \sqrt{C}\right).$$

If the estimator satisfies Assumption 3 with $C = 0$, then the last term vanishes and we obtain a $\mathcal{O}(1/T)$ rate. □

Theorem 4.6 (Any-Time Convergence for Stochastic Quasar-Convex Problems)**.** *Let f be a function that satisfies L-smoothness Assumption 1 and ρ-quasar-convexity Assumption 2. Suppose the stochastic estimator m^t and auxiliary sequence $\{\sigma_t\}$ satisfy Assumption 3 with parameters $\rho_1, \rho_2 \in]0, 1]$ and constants $A, B, C, E \geq 0$. Let $\{x^t\}_{t=0}^{+\infty}$ be a sequence generated by Algorithm BSFW by choosing the step decay*

$$\eta_t = \frac{2}{\rho(t+\nu)}, \quad \textit{where} \quad \nu = \max\left\{2, \frac{4}{\min\{\rho_1, \rho_2\}}\right\}.$$

Then, the expected functional-value gap satisfies

$$\mathbb{E}[F_t] \leq \sqrt{\frac{16D^2}{\rho^2(t+\nu)}\left(\frac{32D^2B}{\rho^2(t+\nu)\rho_1} + \frac{64D^2AE}{\rho^2(t+\nu)\rho_1\rho_2} + \frac{2CT}{\rho_1}\right)} + \frac{4\nu^2\mathbb{E}[r_0]}{(t+\nu)^2} + \frac{8D^2L}{\rho^2(t+\nu)}.$$

where r_t is a Lyapunov function defined by

$$\forall t: \quad r_t := F_t + \frac{2\alpha^\star}{\rho_1 L}\|\Delta^t\|^2 + \frac{4\alpha^\star A}{\rho_1\rho_2 L}\sigma_t^2,$$

with

$$\alpha^\star = \sqrt{\left(\frac{16D^2L}{\rho^2(T+\nu)}\right) \Big/ \left(\frac{32D^2B}{\rho^2(T+\nu)\rho_1 L} + \frac{64D^2AE}{\rho^2(T+\nu)\rho_1\rho_2 L} + \frac{2CT}{\rho_1 L}\right)}.$$

If $C = 0$, the last term in the square root vanishes and we obtain a $\mathcal{O}(1/t)$ rate. □

For nonconvex functions, we have the following convergence rate for the Frank-Wolfe gap in expectation; an any-time step decay and convergence rate can be found in Appendix B with an additional $\ln(t)$ factor.

Theorem 4.7 (Convergence for Stochastic Nonconvex Problems)**.** *Let f be a function that satisfies Assumption 1. Suppose the stochastic gradient estimator m^t and auxiliary sequence $\{\sigma_t\}$ satisfy Assumption 3 with parameters $\rho_1, \rho_2 \in]0, 1]$ and constants $A, B, C, E \geq 0$. Consider the sequence $\{x^t\}_{t=0}^T$ generated by Algorithm BSFW by fixing a $T \in \mathbb{N}$ and using the constant, horizon-dependent step decay $\eta_t = \frac{1}{\sqrt{T}}$. Then,*

$$\mathbb{E}\left[\min_{0 \leq t \leq T-1} \operatorname{Gap}(x^t)\right] = \mathcal{O}\left(\frac{1}{\sqrt{T}} + \sqrt{C}\right).$$

If the estimator satisfies Assumption 3 with $C = 0$, then the last term vanishes and we obtain a $\mathcal{O}(1/\sqrt{T})$ rate. □

Remark 4.8. The Heavy Ball estimator satisfies a slightly more general version of Assumption 3 which allows ρ_1, ρ_2, and B to depend on the horizon T. We prove convergence for this estimator separately in Appendix C.3.7, with a worse convergence rate of $\mathcal{O}(1/T^{1/4})$ for nonconvex problems that matches the convergence rate for this estimator when used with vanilla SFW [31].

5 Experiments

We study Algorithm BSFW with different estimators that we have analyzed, with SFW using those same estimators as a baseline. The boosting procedure was built upon the code provided in [3]. We consider sparse logistic regression, which is convex, and quantum process tomography with a nonconvex objective function. We perform more experiments, for instance collaborative filtering using a nuclear norm constraint, measuring the expected value of the optimality gap over several runs, measuring the number of oracle calls as a function of the tolerance. In all plots, the darker colors represent Algorithm BSFW with similar hues corresponding to the same estimator.

5.1 Sparse Logistic Regression

We consider the following problem

$$\min_{x \in \mathcal{C}} \frac{1}{m} \sum_{i=1}^{m} \ln\left(1 + \exp\left(-y_i a_i^\top x\right)\right), \tag{3}$$

where $\mathcal{C} = \{x \in \mathbb{R}^n : \|x\|_1 \leq \tau\}$ for some radius $\tau > 0$ that we pick according to each dataset. We denote $\{a_i, y_i\}_{i=1}^m$ as samples drawn from an experiment-specific dataset, where $\forall i, a_i \in \mathbb{R}^n$ and $\forall i, y_i \in \{-1, 1\}$. The LMO of the ℓ_1 ball with radius τ is given by (4). The LMO of the ℓ_1 ball has a $\mathcal{O}(n)$ complexity [4], hence making it an appropriate constraint set for boosting, as discussed further in Appendix A.3;

$$\forall g \in \mathbb{R}^n : \quad v^* = \operatorname{lmo}(g) \iff v^* = -\tau \operatorname{sign}(g_i)\, e_i \text{ with } i = \operatorname*{argmax}_j |g_j|. \tag{4}$$

We use the `rcv1` train dataset [20], `mushrooms` dataset [28], and the `breast cancer` [36] datasets from the LIBSVM library of datasets [2]. The parameters of the experiments used are given in Table 1. We run Algorithm BSFW and SFW with every estimator listed in Table 2. An explanation of stochastic and coordinate methods is given by Appendices C.1 and C.2. For BSFW, we use the step decay given by Theorem 4.6, while for SFW, we use the step size provided by the corresponding prior work [27, 29, 30]. For all of the experiments, we pass in an alignment tolerance $\delta = 10^{-4}$, and a large max number of oracles $K = 10,000$ for the boosting procedure (effectively uncapped). The performance results of the different algorithms with the different estimators are shown in Figure 3. Although a batch size b_s is passed in as a parameter, the actual number of gradients sampled and computed per-iteration t is counted for precision, since it can differ between estimators.

Apart from relative suboptimality, one might be interested in the progress in loss ($\mathbb{E}[f(x^t) - f^\star]$) made per-iteration t. Figure 4 shows this. To approximate the true expectation by a numerical expectation, we run every experiment 10 times each. On a per-iteration basis, BSFW has a clear advantage over SFW for all estimators considered.

The number of FW oracles called per-iteration t as part of the boosting procedure depends on the alignment tolerance δ passed to Algorithm BSFW. By definition of the boosting procedure, if the alignment tolerance δ is small, it is normal to expect

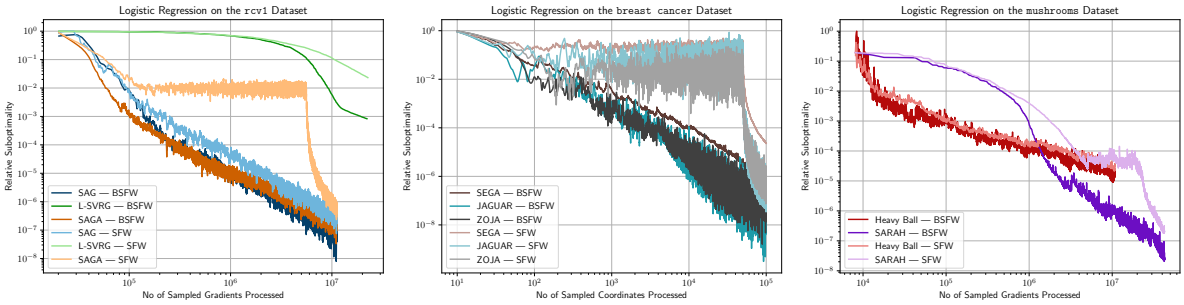

Figure 3: Suboptimality in the Logistic Regression problem is measured by $\frac{f(x^t)-f_{\min}}{f_{\max}-f_{\min}}$, with $f_{\max}, f_{\min}$ estimated as the max and min across all runs.

Table 1: Summary of dataset parameters.

Name	n	m	b_s	b_c	τ
`rcv1`	47,236	20,242	742	–	100
`mushrooms`	112	8,124	404	–	50
`breast cancer`	10	683	–	1	5

For each of the datasets, n refers to the dimension of the features, m refers to the total number of samples, b_s refers to the batch size sampled, while b_c refers to the coordinate batch size (number of coordinates) sampled.

that a higher number of rounds will be needed to reach a satisfactory alignment (until the while loop in Algorithm Boost is exited). Each refinement round requires exactly 1 Frank-Wolfe oracle, and thus it is expected to see more oracles computed per-iteration t when δ is very small. The experimental results showcase this phenomenon in Figure 5.

Algorithm BSFW is designed to revert back to stochastic FW if $\gamma_t = 1$. We denote the *boosting percentage* by (5) when the algorithm is executed for a total of T iterations.

$$\text{Boosting Percentage} = \frac{\sum_{t=0}^{T-1} \mathbf{1}_{\{\gamma_t<1\}}}{T} \times 100. \tag{5}$$

This measure refers to the percentage of iterations where $\gamma_t < 1$, meaning it does not revert to stochastic FW. Our experimental results confirm that reversion to stochastic FW does not occur in nearly 100% of the iterations for all experiments as seen in Figure 6.

In Theorem 4.5, we show the convergence analysis for ρ-quasar-convex functions using a similar piecewise step decay as provided in [29]. However, the performance results were not as strong compared to the step decay provided in Theorem 4.6. Figure 7 shows the performance results using this piecewise step decay. We pass in the same parameters used in section 5, notably the parameters in Table 1, an alignment tolerance $\delta = 10^{-4}$, and the max number of boosting rounds $K = 10^4$.

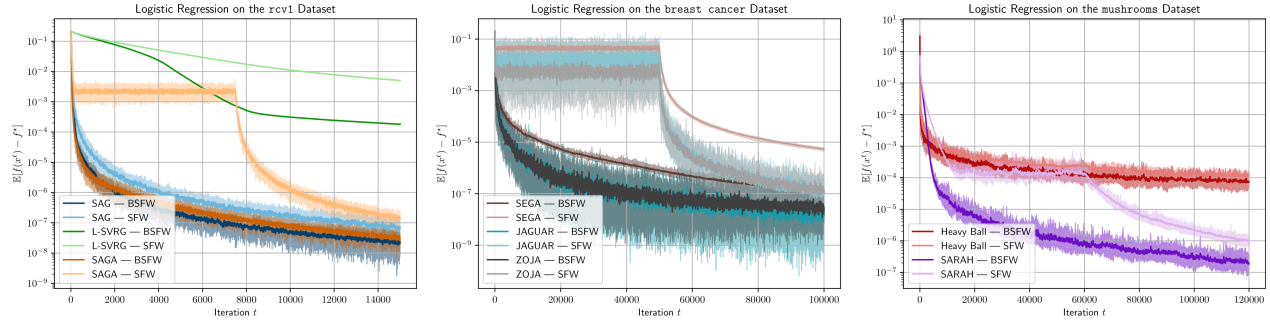

Figure 4: Numerical expected loss vs iteration t on the different datasets.

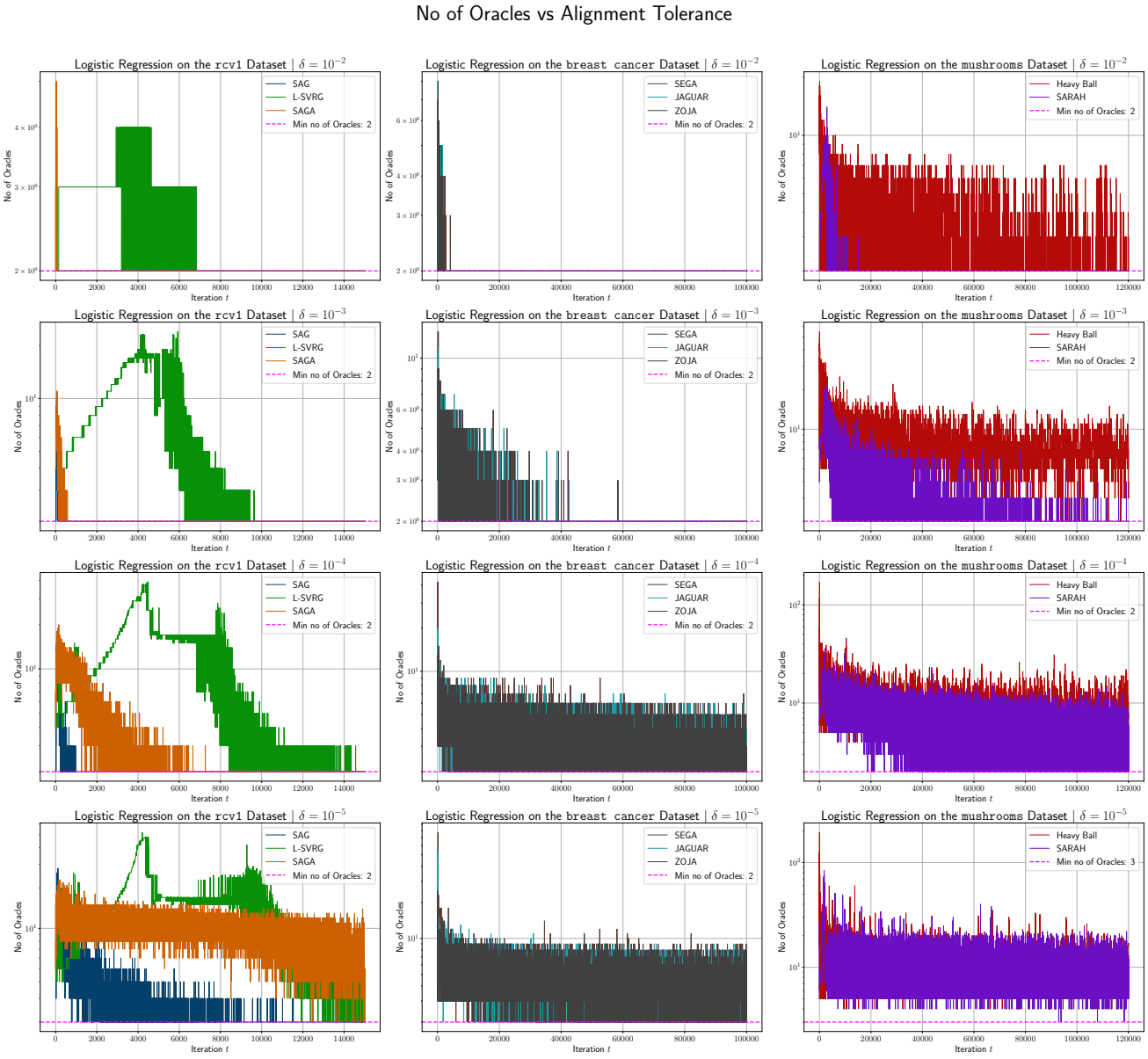

Figure 5: As the tolerance δ decreases, more oracles are used per-iteration for the estimators in general. Notably, the minimum number of oracle calls used by SARAH and Heavy Ball when $\delta = 10^{-5}$ becomes 3 instead of 2 as observed in the other experiments.

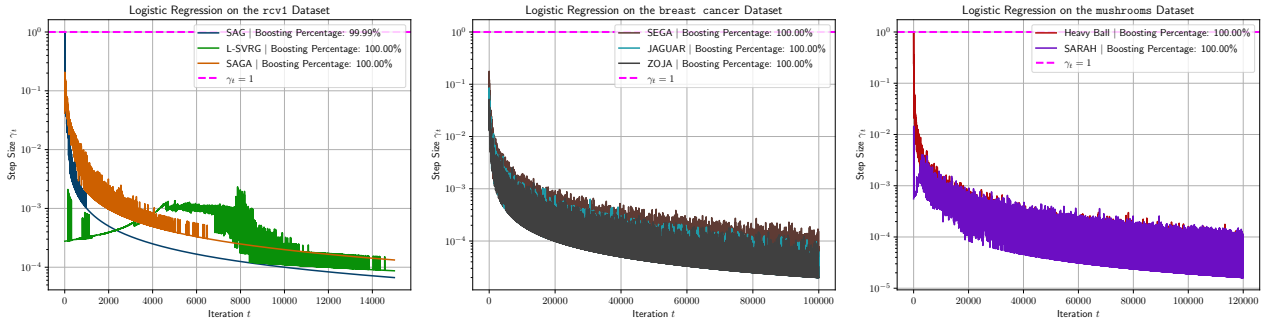

Figure 6: The step sizes $\{\gamma_t\}$ for all of the different estimators are effectively never equal to 1 (shown as the dashed pink line), so that the reversion to the FW direction is never used. Our proposed scaling results in a step that tends to decrease.

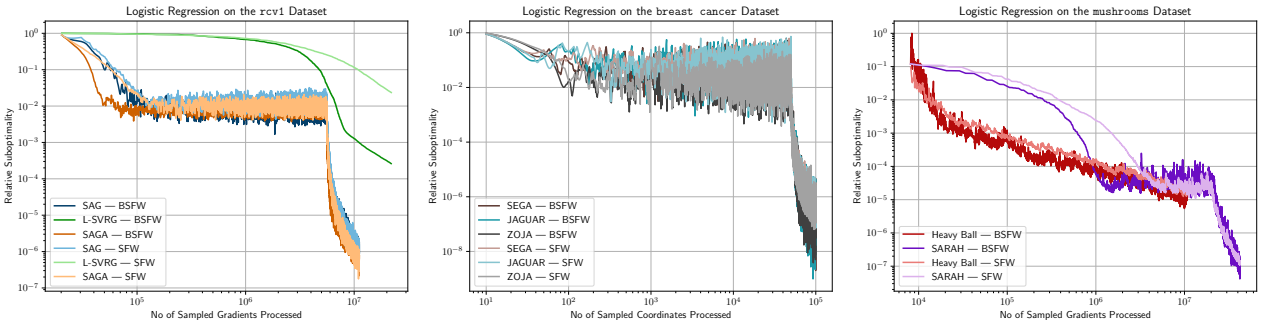

Figure 7: Relative suboptimality when using piecewise step decay on the different datasets.

5.2 Quantum Process Tomography

Quantum process tomography (QPT) aims to reconstruct the quantum process χ, given a set of measurements $\{f_{ijk}\}$ and a set of measurement sensing operators $\{\mathcal{A}_{ijk}\}$. Here, i refers to the input state required to generate the setup, j refers to a specific circuit setting of that input state, and k refers to a measurement outcome for that setting. In an $\tilde{n}$-qubit system, by using Pauli bases [32], we have $4^{\tilde{n}}$ input states, $3^{\tilde{n}}$ circuit settings, and $2^{\tilde{n}}$ measurement outcomes possible; we have a total of $24^{\tilde{n}}$ sets of $(\mathcal{A}_s, f_s)$ combinations. $\mathcal{A}_s : \mathbb{C}^{d^2 \times d^2} \to \mathbb{R}^{2^n}$ is the sensing mechanism, and the loss functions $F(\cdot)$ and $H(\cdot)$ are defined by

$$F(UU^\dagger) := \frac{1}{2}\sum_s \big(f_s - \mathcal{A}_s(UU^\dagger)\big)^2 \quad \text{and} \quad H(UU^\dagger) := \Big\|\sum_{i,j}(UU^\dagger)_{ij}\,\tilde{A}_j^\dagger \tilde{A}_i - I\Big\|_F^2,$$

where $\{\tilde{A}_k\}$ denotes the set of Pauli bases [32], $F(\cdot)$ is the least-squares fidelity function, and $H(\cdot)$ enforces the trace-preserving (TP) condition.

We consider the nonconvex problem using the Burer–Monteiro (BM) factorization, where $\chi = UU^\dagger$ is the quantum process to be approximated and $U^\dagger$ denotes the conjugate transpose of U. We approximate χ by rank-1 matrices U and implement the loss functions faithfully as in [32]:

$$\min_{U \in \mathcal{C}} \; F(UU^\dagger) + \lambda \cdot H(UU^\dagger), \tag{QPT}$$

where the constraint set $\mathcal{C}$ for $\tilde{n}$-qubit systems is

$$\mathcal{C} = \big\{U \in \mathbb{C}^{4^{\tilde{n}} \times 1} : \|U\|_{\text{op}} \le \tau\big\}. \tag{6}$$

This choice of $\mathcal{C}$ proves to be an effective one for improving the fidelity with Boosted FW. The spectral-norm LMO for a non-zero rank-1 matrix G in the constraint set $\mathcal{C}$ of radius τ is

$$V^* \in \text{lmo}(G) \iff V^* = -\tau \tfrac{G}{\|G\|_2}. \tag{7}$$

We run experiments for 3-qubit systems (i.e. $\tilde{n} := 3$). To set up each experiment, we first generate the ground-truth process $\chi^\star$ following [32] and then add Gaussian noise at level ξ:

$$f_s = A_s(\chi^\star) + \xi\,\epsilon, \qquad \epsilon \sim \mathcal{N}(0,1). \tag{8}$$

We set $\tau = 10$ after a grid search to identify the constraint radius that yields the best attainable fidelity, and we set $\lambda = 0.05$ in (QPT). For all four experiment settings we use the "full-measurements" setting, i.e. all $24^{\tilde{n}}$ measurements per iteration. Wall-clock experiments are timed on A100 GPUs.

Following [32], performance is measured by the quantum process fidelity F_p, whose definition for two quantum channels is [13]

$$F_p(\mathcal{E}, \mathcal{E}^\star) = F_s(\tilde{\rho}_\mathcal{E},\, \tilde{\rho}_{\mathcal{E}^\star})\,, \tag{9}$$

where F_s denotes the quantum state fidelity, and $\tilde{\rho}_{\mathcal{E}}$, $\tilde{\rho}_{\mathcal{E}^\star}$ are the normalized Choi-matrix representations of the channels $\mathcal{E}$ and $\mathcal{E}^\star$, respectively (see [13, 38] for details). Note that χ and $\chi^\star$ represent channels $\mathcal{E}$ and $\mathcal{E}^\star$ through the specified basis [32]. The closer F_p is to 1, the better the reconstruction. We implement this metric using the `qiskit` library [13].

From [32], four different levels of Gaussian noise (parameterized by ξ) are added to the ground-truth measurement process. We plot the fidelity achieved per wall-clock time in Figure 8 for different values of the maximum number of oracles K, in order to compare the effect of boosting. Note that $K = 1$ corresponds to the vanilla FW method.

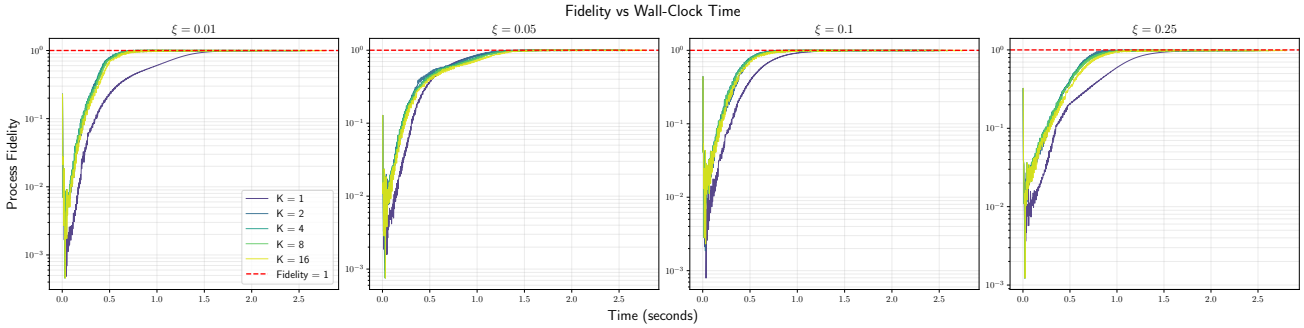

Figure 8: For all four experiments we set the radius $\tau = 10$. The red dotted line illustrates the target fidelity. Computing at least one additional LMO for boosting consistently yields higher fidelity than vanilla FW.

The per-iteration fidelity curves corresponding to Figure 8 are provided in Figure 9 for readers interested in comparing per-iteration performance.

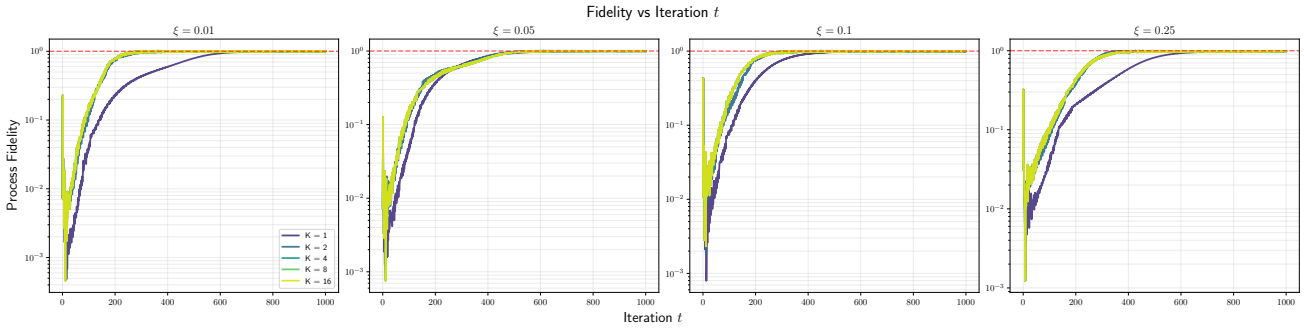

Figure 9: The fidelity results from Figure 8 plotted as a function of iteration t.

5.3 Collaborative Filtering

We consider the following problem

$$\min_{X \in \mathcal{C}} \frac{1}{|\Omega|} \sum_{(i,j) \in \Omega} \ell(X_{ij}, Y_{ij}),$$

where $\mathcal{C} = \{X \in \mathbb{R}^{n \times m} : \ \|X\|_* \leq \tau\}$ is the nuclear norm ball of radius $\tau > 0$, Ω is the observed entries of the matrix, and $\ell(x, y) = \frac{(x-y)^2}{2+(x-y)^2}$, which is nonconvex. The LMO of the nuclear norm ball with radius τ is proportional to the outer product of the leading singular vectors

$$\forall G \in \mathbb{R}^{n \times m} : \quad v^* = \mathrm{lmo}(G) \iff v^* = -\tau u_1 v_1^T \ \text{ with } \ G = U \Sigma V^T. \tag{10}$$

We use the `MovieLens` dataset [12] and run Algorithm BSFW and SFW with the L-SVRG, Heavy Ball, and SAGA estimators. We use the horizon-dependent step size given in Theorem 4.7 for Algorithm BSFW and the parameters given in [29] for SFW,

except for Heavy Ball for which we take the parameters given in [31]. Because the LMO in (10) is reasonably **expensive** compared to the one used in (3), we set a smaller $K = 5$ and $\delta = 10^{-4}$ for the boosting procedure. We set the batch size b_s to be 10% of the dataset (although we observed a larger gap between Algorithm BSFW and SFW when the batches were larger). The Heavy Ball estimator outperformed the other estimators in this setting despite having a slower worst-case convergence rate guarantee. For all three estimators, boosting produced an improvement in performance.

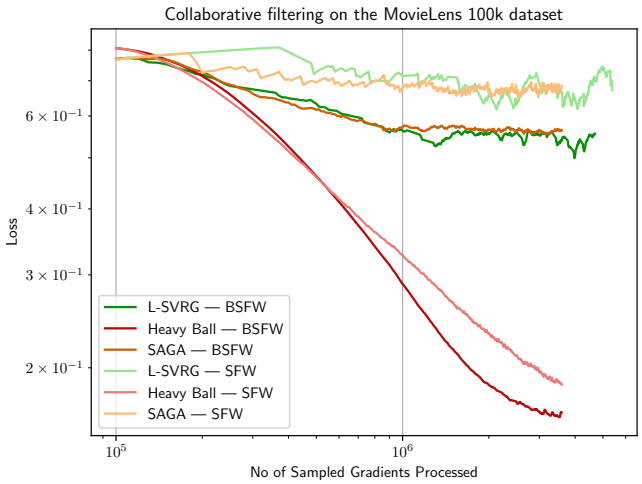

Figure 10: For this collaborative filtering problem, we plot the loss directly.

Appendices

The appendix of this work is divided into three sections. The first section contains more information about Algorithm BSFW and some additional discussions. The second contains the proofs of the theoretical results presented in the main text. The third section provides detailed descriptions of the gradient estimators, along with additional theoretical results, e.g., proofs that they satisfy the assumptions made in the main text and in the subsequent proofs.

A Additional Discussions

A.1 Complete Algorithm BSFW

The complete Algorithm BSFW (Full) is written here for reference. Note that we show in Algorithm BSFW (Full) that in fact, $\text{lmo}(m^t)$ at iteration t is already computed during the boosting procedure and hence does not need to be computed again in lines 6 and 11 of Algorithm BSFW.

A.2 Quasar-Convex Applications in Machine Learning

There are many problems in machine learning involving quasar-convex functions which have been explored prior to our work; we list a few of them here. The problem of learning linear dynamical systems (LDS) has been explored in [8] and [11], where specifically in [11], the authors prove that their objective function is quasar-convex but not necessarily convex. In [24] and [6], the authors show that generalized linear models (GLM) with increasing link functions (among other assumptions) are quasar-convex, and study their convergence analysis. In [6], the authors also show that the problem of robust linear regression using Tukey's biweight loss is quasar-convex. Furthermore, in [35], the authors show that GLMs (satisfying certain other conditions) belong to the family of quasar-convex functions (and develop algorithms for this problem class).

A.3 When to Boost?

Boosting is recommended in problems where the cost of gradient computation is expensive relative to the cost of the LMO. It is then advantageous to compute the LMO several times per gradient computation in order to get an update that is better aligned with the gradient or its stochastic estimator. For instance, for matrices of size 100×100, the ratio of nuclear norm LMO vs gradient (for a sample least squares setting) complexity (in terms of wall-clock time) is on the order of 10^{-1} while that of ℓ_1 ball LMO for vectors of size 10^4 vs gradient (for the same sample least squares setting) complexity is in the order of $\mathcal{O}(10^{-3})$.

A.4 Complexity of Step Size

The step size strategy we prescribe in Algorithm BSFW depends only on $s^t \in \text{lmo}(m^t)$ and $\tilde{d}^t$, which are already computed during the boosting procedure. Therefore, it does not impose additional computational complexity, unlike the line search used in [3] when the Lipschitz constant of ∇f is not known.

B Proofs of Main Results

In this section, we present the proofs for the core technical results and the main convergence theorems. We divide the analysis of Algorithm BSFW into the deterministic setting in section B.1 and the stochastic setting in section B.2. We begin with a geometric property of the alignment direction used in our variants.

Lemma B.1 (Alignment Inequality)**.** *Consider the sequence $\{x^t\}_{t=0}^{+\infty}$ generated by Algorithm BSFW. Then for all $t \in \mathbb{N}$, the following inequality holds*

$$\langle m^t, d^t \rangle \leq \frac{\|d^t\|}{\|s^t - x^t\|} \langle m^t, s^t - x^t \rangle.$$

where d^t denotes the alignment direction at iteration t and $s^t \in \text{lmo}(m^t)$.

Algorithm BSFW (Full) Boosted Stochastic Frank-Wolfe

Input: initial estimator $m^{\text{init}} \in \mathbb{R}^n$, gradient estimator $\{\Phi_t\}_{t=0}^{T-1}$, max number of rounds for boosting $K \in \mathbb{N} \setminus \{0\}$, alignment improvement tolerance $\delta \in]0, 1]$, and step decay $\{\eta_t\}_{t=0}^{T-1} \in]0, 1]$
Output: $x^T \in \mathcal{C}$.

$x^0 \leftarrow \text{lmo}(m^{\text{init}})$
for $t = 0$ to $T - 1$ **do**
  Sample $\xi_t \sim \mathcal{P}$ and compute $g^t = \nabla f(x^t, \xi_t)$
  Compute $m^t = \Phi_t(g^t)$
  $\psi^0 \leftarrow 0$
  $\Lambda_t \leftarrow 0$
  $k \leftarrow 0$
  while $k \leq K - 1$ **do**
    $r^k \leftarrow -m^t - \psi^k$ {k-th residual}
    $v^k \leftarrow \text{lmo}(-r^k)$ {FW oracle}
    if $k = 0$ **then**
      $s^t \leftarrow v^k$
    end if
    if $\psi^k \neq 0$ **then**
      $u^k \leftarrow \underset{u \in \left\{ v^k - x^t, -\frac{\psi^k}{\|\psi^k\|} \right\}}{\text{argmax}} \langle r^k, u \rangle$
    else
      $u^k \leftarrow v^k - x^t$
    end if
    if $u^k = 0$ **then**
      $k \leftarrow k + 1$; **break** {exit k-loop}
    end if
    $\lambda_k \leftarrow \frac{\langle r^k, u^k \rangle}{\|u^k\|^2}$
    $\phi^k \leftarrow \psi^k + \lambda_k u^k$
    if $\text{align}(-m^t, \phi^k) - \text{align}(-m^t, \psi^k) \geq \delta$ **then**
      $\psi^{k+1} \leftarrow \phi^k$
      $\Lambda_t \leftarrow \begin{cases} \Lambda_t + \lambda_k, & u^k = v^k - x^t \\ \Lambda_t \left(1 - \frac{\lambda_k}{\|\psi^k\|}\right), & u^k = -\frac{\psi^k}{\|\psi^k\|} \end{cases}$
      $k \leftarrow k + 1$
    else
      $k \leftarrow k + 1$; **break** {exit k-loop}
    end if
  end while
  $K_t \leftarrow k$
  $\tilde{d}^t \leftarrow \dfrac{\psi^{K_t}}{\Lambda_t}$ **if** $\Lambda_t \neq 0$ **else** 0 {normalize direction}
  $\gamma_t \leftarrow \min\left\{ \eta_t \frac{\|s^t - x^t\|}{\|\tilde{d}^t\|}, 1 \right\}$ **if** $\tilde{d}^t \neq 0$ **else** 1
  if $\gamma_t < 1$ **then**
    $d^t \leftarrow \tilde{d}^t$ {use the boosted direction}
    $x^{t+1} \leftarrow x^t + \gamma_t d^t$
  else
    $d^t \leftarrow s^t - x^t$ {revert to vanilla FW direction}
    $x^{t+1} \leftarrow x^t + \eta_t d^t$
  end if
end for

Proof. By construction of Algorithm BSFW, we have $K_t \geq 1$ for all $t \in \mathbb{N}$.
Case I: Suppose $\gamma_t < 1$. Then we have $d^t = \tilde{d}^t$. Applying Proposition 3.1 from [3], we obtain

$$\text{align}(-m^t, d^t) \geq \text{align}(-m^t, s^t - x^t).$$

By the definition of the alignment operator align, this yields

$$\frac{\langle -m^t, d^t \rangle}{\|m^t\|\|d^t\|} \geq \frac{\langle -m^t, s^t - x^t \rangle}{\|m^t\|\|s^t - x^t\|} \implies \frac{\langle m^t, d^t \rangle}{\|d^t\|} \leq \frac{\langle m^t, s^t - x^t \rangle}{\|s^t - x^t\|} \implies \langle m^t, d^t \rangle \leq \frac{\|d^t\|}{\|s^t - x^t\|} \langle m^t, s^t - x^t \rangle.$$

Case II: Suppose $\gamma_t = 1$. Then $d^t = s^t - x^t$ and we have

$$\langle m^t, s^t - x^t \rangle \leq \frac{\|s^t - x^t\|}{\|s^t - x^t\|} \langle m^t, s^t - x^t \rangle.$$

□

B.1 Boosted Deterministic Frank-Wolfe

We first provide descent Lemma B.2 which will help us prove the convergence analysis for ρ-quasar-convex objective functions in Theorem B.3 (any-time rate) and nonconvex functions in Theorem B.5 (any-time rate) and Theorem B.6 (fixed-horizon rate).

Lemma B.2 (Descent under Smoothness in Deterministic Setting)**.** *Let $\{x^t\}_{t=0}^{+\infty}$ be the sequence generated by Algorithm BSFW, where at every iteration t, $m^t = \nabla f(x^t)$. Assume that f satisfies L-smoothness Assumption 1. Then, for all $t \geq 0$, the following descent property holds*

$$f(x^{t+1}) \leq f(x^t) + \eta_t \langle \nabla f(x^t), s^t - x^t \rangle + \frac{L}{2} \eta_t^2 \|s^t - x^t\|^2.$$

Proof. By L-smoothness of f, for any feasible direction d^t satisfying $x^t + \gamma_t d^t \in \mathcal{C}$ and any $\gamma_t \in [0, 1]$, we have

$$f(x^t + \gamma_t d^t) \leq f(x^t) + \gamma_t \langle \nabla f(x^t), d^t \rangle + \frac{L}{2} \gamma_t^2 \|d^t\|^2. \tag{11}$$

We distinguish two cases based on the value of γ_t, following the analysis of Algorithm BSFW.
Case I: $\gamma_t < 1$. In this case, $\gamma_t = \eta_t \frac{\|s^t - x^t\|}{\|d^t\|}$ since $\tilde{d}^t = d^t$. Using Lemma B.1 and substituting into (11) yields

$$\begin{aligned}
f(x^{t+1}) &\leq f(x^t) + \left(\eta_t \frac{\|s^t - x^t\|}{\|d^t\|} \right) \langle \nabla f(x^t), d^t \rangle + \frac{L}{2} \left(\eta_t \frac{\|s^t - x^t\|}{\|d^t\|} \right)^2 \|d^t\|^2 \\
&\leq f(x^t) + \left(\eta_t \frac{\|s^t - x^t\|}{\|d^t\|} \right) \left(\frac{\|d^t\|}{\|s^t - x^t\|} \langle \nabla f(x^t), s^t - x^t \rangle \right) + \frac{L}{2} \left(\eta_t \frac{\|s^t - x^t\|}{\|d^t\|} \right)^2 \|d^t\|^2 \\
&= f(x^t) + \eta_t \langle \nabla f(x^t), s^t - x^t \rangle + \frac{L}{2} \eta_t^2 \|s^t - x^t\|^2,
\end{aligned}$$

where the first inequality follows from (11) with the substitution $\gamma_t = \eta_t \frac{\|s^t - x^t\|}{\|d^t\|}$, the second inequality follows from Lemma B.1, and the equality follows from algebraic simplification.

Case II: $\gamma_t = 1$. Here, the direction is $d^t = s^t - x^t$ (as vanilla Frank-Wolfe direction) and the update is $x^{t+1} = x^t + \eta_t d^t$. In this case, similarly, from L-smoothness we have

$$f(x^{t+1}) \leq f(x^t) + \eta_t \langle \nabla f(x^t), s^t - x^t \rangle + \frac{L}{2} \eta_t^2 \|s^t - x^t\|^2.$$

Thus, in both cases, the decrease inequality holds. □

We are ready to present convergence guarantees under two complementary regimes: structured (quasar-convex) objectives and general nonconvex smooth functions, in the following subsections.

B.1.1 Quasar Convex Case

Theorem B.3 (Formal Statement of Theorem 4.1)**.** *Let f be a function that satisfies L-smoothness Assumption 1 and ρ-quasar-convexity Assumption 2. Consider the sequence $\{x^t\}_{t=0}^{+\infty}$ generated by Algorithm BSFW with $m^t = \nabla f(x^t)$ and $\eta_t = \frac{2}{\rho(t+2)}$. Then, for all $t \geq 0$,*

$$F_t \leq \frac{1}{t+1} \max\left\{F_0, \frac{2LD^2}{\rho^2}\right\} = \mathcal{O}(1/t).$$

Proof. From Lemma B.2 we have

$$\begin{aligned} f(x^{t+1}) - f(x^t) &\leq \eta_t \langle \nabla f(x^t), s^t - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2 \\ &\leq \eta_t \langle \nabla f(x^t), x^\star - x^t \rangle + \eta_t^2 \frac{LD^2}{2} \\ &\leq \eta_t \rho (f(x^\star) - f(x^t)) + \eta_t^2 \frac{LD^2}{2}, \end{aligned}$$

where the second inequality uses $s^t \in \operatorname{lmo}(\nabla f(x^t))$ and the last follows from ρ-quasar-convexity of f. Recall that $F_t = f(x^t) - f(x^\star) \geq 0$. Rearranging the terms and subtracting $f(x^\star)$ from both sides:

$$F_{t+1} \leq F_t - \rho\eta_t F_t + \frac{LD^2}{2}\eta_t^2 = (1 - \rho\eta_t)F_t + \frac{LD^2}{2}\eta_t^2. \tag{12}$$

Now we use $\eta_t = \frac{2}{\rho(t+2)}$, substituting it into

$$F_{t+1} \leq \frac{t}{t+2}F_t + \frac{2LD^2}{\rho^2(t+2)^2}. \tag{13}$$

Set $B := \max\left\{F_0, \frac{2LD^2}{\rho^2}\right\}$. We prove by induction that for all $t \geq 0$,

$$F_t \leq \frac{B}{t+1}. \tag{14}$$

Base case ($t = 0$). By definition of B, we have $F_0 \leq B$, and thus

$$F_0 \leq \frac{B}{1} = \frac{B}{0+1},$$

which establishes the claim for $t = 0$.
Inductive step. Assume (14) holds for some $t \geq 0$. Using (13) and the induction hypothesis,

$$\begin{aligned} F_{t+1} &\leq \frac{t}{t+2} \cdot \frac{B}{t+1} + \frac{2LD^2}{\rho^2(t+2)^2} \\ &\leq \frac{tB}{(t+2)(t+1)} + \frac{B}{(t+2)^2} \\ &\leq \frac{tB}{(t+2)(t+1)} + \frac{B}{(t+2)(t+1)} \\ &= \frac{(t+1)B}{(t+2)(t+1)} \\ &= \frac{B}{t+2}. \end{aligned}$$

This establishes (14) for $t+1$, completing the induction. □

Remark B.4. This result immediately applies to classical convex and star-convex settings, as they both satisfy ρ-quasar-convexity with $\rho = 1$. Hence, Theorem 4.1 ensures a $\mathcal{O}(1/t)$ convergence rate in these cases. □

B.1.2 Nonconvex Case

In the absence of any convexity-like assumptions, we can still guarantee convergence in terms of the Frank–Wolfe gap—a standard stationarity measure for constrained nonconvex optimization.

Theorem B.5 (Formal Statement of Theorem 4.3)**.** *Let f be a function that satisfies L-smoothness Assumption 1. Consider the sequence $\{x^t\}_{t=0}^{+\infty}$ generated by Algorithm BSFW with $m^t = \nabla f(x^t)$ and $\eta_t = \frac{1}{\sqrt{t+1}}$. Then, for all $t \geq 0$,*

$$\min_{0 \leq i \leq t} \mathrm{Gap}(x^i) \leq \frac{F_0 + LD^2}{\sqrt{t+1}} = \mathcal{O}(1/\sqrt{t}).$$

Proof. From Lemma B.2, we have

$$f(x^{t+1}) \leq f(x^t) + \eta_t \langle \nabla f(x^t), s^t - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2.$$

Rearranging the inequality yields

$$\eta_t \langle \nabla f(x^t), x^t - s^t \rangle \leq f(x^t) - f(x^{t+1}) + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2 \leq f(x^t) - f(x^{t+1}) + \frac{L}{2}\eta_t^2 D^2.$$

Thus,

$$\langle \nabla f(x^t), x^t - s^t \rangle \leq \frac{1}{\eta_t}(f(x^t) - f(x^{t+1})) + \frac{L}{2}\eta_t D^2,$$

Now substitute $\eta_t = 1/\sqrt{t+1}$. Then, summing the inequality over $t = 0$ to T yields

$$\sum_{t=0}^{T} \langle \nabla f(x^t), x^t - s^t \rangle \leq \sum_{t=0}^{T} \sqrt{t+1}\, \left(f(x^t) - f(x^{t+1})\right) + \frac{LD^2}{2} \sum_{t=0}^{T} \frac{1}{\sqrt{t+1}}, \tag{15}$$

For the first sum on the right-hand side of (15), note that $\sqrt{t+1} \leq \sqrt{T+1}$ for all $t \leq T$, so

$$\begin{aligned}
\sum_{t=0}^{T} \sqrt{t+1}\, \left(f(x^t) - f(x^{t+1})\right) \leq & \sqrt{T+1} \sum_{t=0}^{T} \left(f(x^t) - f(x^{t+1})\right) \\
= & \sqrt{T+1}\, \left(f(x^0) - f(x^{T+1})\right) \\
\leq & \sqrt{T+1}\, \left(f(x^0) - f(x^\star)\right) \\
= & \sqrt{T+1}\, F_0.
\end{aligned}$$

For the second sum on the right-hand side of (15), we use the standard bound

$$\sum_{t=0}^{T} \frac{1}{\sqrt{t+1}} = \sum_{k=1}^{T+1} \frac{1}{\sqrt{k}} \leq \int_0^{T+1} \frac{1}{\sqrt{x}}\, dx = 2\sqrt{T+1}.$$

Combining both estimates, we obtain

$$\sum_{t=0}^{T} \langle \nabla f(x^t), x^t - s^t \rangle \leq \sqrt{T+1}\, \left(F_0 + LD^2\right).$$

Since $\langle \nabla f(x^t), x^t - s^t \rangle \geq 0$, it follows that

$$\left(\min_{0 \leq t \leq T} \langle \nabla f(x^t), x^t - s^t \rangle\right) (T+1) \leq \sum_{t=0}^{T} \langle \nabla f(x^t), x^t - s^t \rangle \leq \sqrt{T+1}\, \left(F_0 + LD^2\right),$$

and therefore

$$\min_{0 \leq t \leq T} \langle \nabla f(x^t), x^t - s^t \rangle \leq \frac{F_0 + LD^2}{\sqrt{T+1}}.$$

□

Theorem B.6 (Fixed Horizon Convergence under nonconvexity in Deterministic Setting)**.** *Let f be an objective function that satisfies L-smoothness Assumption 1. Let $T \in \mathbb{N}$ and $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW where $m^t = \nabla f(x^t)$, and $\eta_t = \frac{1}{\sqrt{T+1}}$. Then we have*

$$\min_{0\le t\le T} \langle \nabla f(x^t), x^t - s^t\rangle \le \frac{F_0 + \frac{LD^2}{2}}{\sqrt{T+1}}.$$

Proof. From Lemma B.2, we have

$$f(x^{t+1}) \le f(x^t) + \eta_t \langle \nabla f(x^t), s^t - x^t\rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2.$$

Rearranging the inequality yields

$$\eta_t \langle \nabla f(x^t), x^t - s^t\rangle \le f(x^t) - f(x^{t+1}) + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2 \le f(x^t) - f(x^{t+1}) + \frac{L}{2}\eta_t^2 D^2,$$

where we used $\|s^t - x^t\| \le D$. Summing both sides from $t = 0$ to T gives

$$\sum_{t=0}^T \eta_t \langle \nabla f(x^t), x^t - s^t\rangle \le f(x^0) - f(x^{T+1}) + \frac{LD^2}{2}\sum_{t=0}^T \eta_t^2 \le f(x^0) - f(x^\star) + \frac{LD^2}{2}\sum_{t=0}^T \eta_t^2,$$

Since $g_t \ge 0$, we have

$$\left(\min_{0\le t\le T} \langle \nabla f(x^t), x^t - s^t\rangle\right) \sum_{t=0}^T \eta_t \le F_0 + \frac{LD^2}{2}\sum_{t=0}^T \eta_t^2.$$

Now substitute $\eta_t = 1/\sqrt{T+1}$ for all $t = 0, \cdots, T$, Then

$$\sum_{t=0}^T \eta_t = (T+1)\cdot\frac{1}{\sqrt{T+1}} = \sqrt{T+1} \qquad \text{and} \qquad \sum_{t=0}^T \eta_t^2 = (T+1)\cdot\frac{1}{T+1} = 1.$$

Thus,

$$\left(\min_{0\le t\le T} \langle \nabla f(x^t), x^t - s^t\rangle\right)\sqrt{T+1} \le F_0 + \frac{LD^2}{2},$$

which implies

$$\min_{0\le t\le T} \langle \nabla f(x^t), x^t - s^t\rangle \le \frac{F_0 + \frac{LD^2}{2}}{\sqrt{T+1}}.$$

This concludes the proof. □

B.2 Boosted Stochastic Frank–Wolfe

Here, we make Remark B.7 on the estimators, which we use in the convergence analysis that follows. We show the analysis for both ρ-quasar-convex functions in subsection B.2.1 and nonconvex functions in subsection B.2.2.

Remark B.7. The recursive bounds on the gradient estimation error $\|\Delta^t\|^2$ (recall that $\Delta^t = m^t - \nabla f(x^t)$) and the auxiliary variance term σ_t^2 follow directly from Assumption 3. Applying the law of total expectation to both sides yields the unconditional recursions:

$$\mathbb{E}[\|\Delta^t\|^2] \le (1-\rho_1)\mathbb{E}[\|\Delta^{t-1}\|^2] + A\mathbb{E}[\sigma_{t-1}^2] + \eta_{t-1}^2 BD^2 + C, \tag{16}$$

$$\mathbb{E}[\sigma_t^2] \le (1-\rho_2)\mathbb{E}[\sigma_{t-1}^2] + \eta_{t-1}^2 ED^2. \tag{17}$$

B.2.1 Quasar Convex Case

In the ρ-quasar-convex setting, first we prove descent Lemma B.8. This will help us prove both a fixed-horizon convergent rate in Theorem 4.5 using a step decay that is a modified version of the step size that is offered in [29], and an any-time convergent rate in Theorem 4.6 using an alternative step decay.

Lemma B.8 (Descent under Smoothness in Stochastic Setting for Quasar-Convex Functions)**.** *Let f be an objective function that satisfies L-smoothness Assumption 1 and ρ-quasar-convexity Assumption 2. Consider the sequence $\{x^t\}_{t=0}^{+\infty}$ generated by Algorithm BSFW. Then for any $\alpha > 0$, we have*

$$\mathbb{E}[F_{t+1}] \leq (1 - \rho\eta_t)\mathbb{E}[F_t] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] + \eta_t^2 LD^2\left(\frac{1}{\alpha} + \frac{1}{2}\right).$$

Proof. From Assumption 1, we have

$$f(x^{t+1}) \leq f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t\rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2.$$

We distinguish two cases based on the value of γ_t, following the analysis of Algorithm BSFW.
Case I: $\gamma_t < 1$. Then we have $x^{t+1} = x^t + \gamma_t d^t$. By using Lemma B.1,

$$\begin{aligned}
f(x^{t+1}) &\leq f(x^t) + \gamma_t\langle \nabla f(x^t), d^t\rangle + \frac{L}{2}\gamma_t^2\|d^t\|^2 \\
&\leq f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \gamma_t\langle m^t, d^t\rangle + \frac{L}{2}\gamma_t^2\|d^t\|^2 \\
&\leq f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle \\
&\quad + \eta_t\left(\frac{\|s^t - x^t\|}{\|d^t\|}\right)\left(\frac{\|d^t\|}{\|s^t - x^t\|}\right)\langle m^t, s^t - x^t\rangle + \frac{L}{2}\eta_t^2\frac{\|s^t - x^t\|^2}{\|d^t\|^2}\|d^t\|^2 \\
&\leq f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \eta_t\langle m^t, s^t - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2.
\end{aligned}$$

Since $s^t \in \mathrm{lmo}(m^t)$, we get

$$\begin{aligned}
f(x^{t+1}) &\leq f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \eta_t\langle m^t, x^\star - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2 \\
&\leq f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \eta_t\langle m^t - \nabla f(x^t), x^\star - x^t\rangle \\
&\quad + \eta_t\langle \nabla f(x^t), x^\star - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2.
\end{aligned}$$

By using the Young's inequality on both $\langle \nabla f(x^t) - m^t, \gamma_t d^t\rangle$ and $\langle m^t - \nabla f(x^t), \eta_t(x^\star - x^t)\rangle$ with a parameter $\beta = \frac{\alpha}{L}$ for an arbitrary $\alpha > 0$, we have

$$\begin{aligned}
\langle \nabla f(x^t) - m^t, \gamma_t d^t\rangle &\leq \frac{\alpha}{2L}\|\nabla f(x^t) - m^t\|^2 + \gamma_t^2\frac{L}{2\alpha}\|d^t\|^2, \\
\langle m^t - \nabla f(x^t), \eta_t(x^\star - x^t)\rangle &\leq \frac{\alpha}{2L}\|\nabla f(x^t) - m^t\|^2 + \eta_t^2\frac{L}{2\alpha}\|x^\star - x^t\|^2.
\end{aligned}$$

Thus by using these inequalities, we get

$$\begin{aligned}
f(x^{t+1}) &\leq f(x^t) + \frac{\alpha}{2L}\|\nabla f(x^t) - m^t\|^2 + \gamma_t^2\frac{L}{2\alpha}\|d^t\|^2 + \frac{\alpha}{2L}\|\nabla f(x^t) - m^t\|^2 \\
&\quad + \eta_t^2\frac{L}{2\alpha}\|x^\star - x^t\|^2 + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2 + \eta_t\langle \nabla f(x^t), x^\star - x^t\rangle \\
&\leq f(x^t) + \frac{\alpha}{L}\|\Delta^t\|^2 + \eta_t^2\frac{L}{2\alpha}\|s^t - x^t\|^2 + \eta_t^2\frac{L}{2\alpha}\|x^\star - x^t\|^2 + \eta_t\langle \nabla f(x^t), x^\star - x^t\rangle + \frac{L}{2}\eta_t^2 D^2 \\
&\leq f(x^t) + \frac{\alpha}{L}\|\Delta^t\|^2 + \eta_t^2\frac{L}{\alpha}D^2 + \eta_t\langle \nabla f(x^t), x^\star - x^t\rangle + \frac{L}{2}\eta_t^2 D^2 \\
&\leq f(x^t) + \frac{\alpha}{L}\|\Delta^t\|^2 + \eta_t^2 LD^2\left(\frac{1}{\alpha} + \frac{1}{2}\right) + \eta_t\langle \nabla f(x^t), x^\star - x^t\rangle.
\end{aligned}$$

By using ρ-quasar convexity of f, and subtracting $f(x^\star)$ on both sides, we get

$$\begin{aligned} F_{t+1} &\leq F_t + \frac{\alpha}{L}\|\Delta^t\|^2 + \eta_t^2 LD^2\left(\frac{1}{\alpha}+\frac{1}{2}\right) - \rho\eta_t(f(x^t)-f(x^\star)) \\ &\leq (1-\rho\eta_t)F_t + \frac{\alpha}{L}\|\Delta^t\|^2 + \eta_t^2 LD^2\left(\frac{1}{\alpha}+\frac{1}{2}\right). \end{aligned}$$

Taking expectation on both sides gives us

$$\mathbb{E}[F_{t+1}] \leq (1-\rho\eta_t)\mathbb{E}[F_t] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] + \eta_t^2 LD^2\left(\frac{1}{\alpha}+\frac{1}{2}\right).$$

Case II: $\gamma_t = 1$. Then we have $x^{t+1} = x^t + \eta_t(s^t - x^t)$. Hence we get,

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \eta_t\langle\nabla f(x^t), s^t - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2 \\ &\leq f(x^t) + \eta_t\langle\nabla f(x^t) - m^t, s^t - x^t\rangle + \eta_t\langle m^t, s^t - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2 \\ &\leq f(x^t) + \eta_t\langle\nabla f(x^t) - m^t, s^t - x^t\rangle + \eta_t\langle m^t, x^\star - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2 \\ &\leq f(x^t) + \eta_t\langle\nabla f(x^t) - m^t, s^t - x^t\rangle + \eta_t\langle m^t - \nabla f(x^t), x^\star - x^t\rangle \\ &\quad + \eta_t\langle\nabla f(x^t), x^\star - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2 \\ &\leq f(x^t) + \eta_t\langle\nabla f(x^t) - m^t, s^t - x^\star\rangle + \eta_t\langle\nabla f(x^t), x^\star - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2. \end{aligned}$$

Using Young's inequality on $\langle\nabla f(x^t) - m^t, \eta_t(s^t - x^\star)\rangle$ with $\beta = \frac{2\alpha}{L}$ with an arbitrary $\alpha > 0$, we have the inequality

$$\langle\nabla f(x^t) - m^t, \eta_t(s^t - x^\star)\rangle \leq \frac{\alpha}{L}\|\nabla f(x^t) - m^t\|^2 + \frac{L}{4\alpha}\eta_t^2\|s^t - x^\star\|^2.$$

By using this inequality, and ρ-quasar-convexity of f, we get

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \frac{\alpha}{L}\|\Delta^t\|^2 + \eta_t^2\frac{L}{4\alpha}\|s^t - x^\star\|^2 - \rho\eta_t(f(x^t)-f(x^\star)) + \frac{L}{2}\eta_t^2 D^2 \\ &\leq f(x^t) + \frac{\alpha}{L}\|\Delta^t\|^2 - \rho\eta_t F_t + \eta_t^2 LD^2\left(\frac{1}{4\alpha}+\frac{1}{2}\right). \end{aligned}$$

By subtracting $f(x^\star)$ and taking expectation on both sides, we get

$$\begin{aligned} \mathbb{E}[F_{t+1}] &\leq \mathbb{E}[F_t] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] - \rho\eta_t\mathbb{E}[F_t] + \eta_t^2 LD^2\left(\frac{1}{4\alpha}+\frac{1}{2}\right) \\ &\leq \mathbb{E}[F_t] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] - \rho\eta_t\mathbb{E}[F_t] + \eta_t^2 LD^2\left(\frac{1}{\alpha}+\frac{1}{2}\right) \\ &\leq (1-\rho\eta_t)\mathbb{E}[F_t] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] + \eta_t^2 LD^2\left(\frac{1}{\alpha}+\frac{1}{2}\right), \end{aligned}$$

which gives us the desired expression in both cases □

Theorem B.9 (Formal Statement of Theorem 4.5)**.** *Let f be an objective function that satisfies L-smoothness Assumption 1 and ρ-quasar-convexity Assumption 2. Let $T \in \mathbb{N}$, and $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW where the stochastic estimator m^t and auxiliary sequence $\{\sigma_t\}$ satisfy Assumption 3 with parameters $\rho_1, \rho_2 \in]0, 1]$ and constants*

$A, B, C, E \geq 0$. *Let the step decay be chosen piecewise as*

$$\eta_t = \begin{cases} \dfrac{1}{\rho d}, & T \leq d, \\ \dfrac{1}{\rho d}, & T > d \text{ and } t < t_0, \\ \dfrac{2}{\rho(2d + t - t_0)}, & T > d \text{ and } t \geq t_0, \end{cases}$$

with $d := \dfrac{2}{\min\{\rho_1, \rho_2\}}$ *and* $t_0 := \lfloor T/2 \rfloor$. *Then, the functional gap satisfies*

$$\mathbb{E}[F_T] \leq e \cdot \exp\left(-\frac{T}{2d}\right) \mathbb{E}[r_0] + \frac{16D^2L}{\rho^2(T+d)} + \sqrt{\frac{32D^2}{\rho^2(T+d)}\left(\frac{64D^2B}{\rho^2(T+d)\rho_1} + \frac{128D^2AE}{\rho^2(T+d)\rho_1\rho_2} + \frac{2CT}{\rho_1}\right)},$$

where r_t *is a Lyapunov function defined by*

$$\forall t: \quad r_t = F_t + \frac{2\alpha^\star}{\rho_1 L}\|\Delta^t\|^2 + \frac{4\alpha^\star A}{\rho_1\rho_2 L}\sigma_t^2,$$

with

$$\alpha^\star = \sqrt{\left(\frac{32D^2L}{\rho^2(T+d)}\right) / \left(\frac{64D^2B}{\rho^2(T+d)\rho_1 L} + \frac{128D^2AE}{\rho^2(T+d)\rho_1\rho_2 L} + \frac{2CT}{\rho_1 L}\right)}.$$

Proof. Define a Lyapunov function $r_t = F_t + M_1\|\Delta^t\|^2 + M_2\sigma_t^2$, with $M_1, M_2 > 0$ as arbitrary constants. We analyze the full expectation of the Lyapunov function r_{t+1}

$$R_{t+1} := \mathbb{E}[r_{t+1}] = \mathbb{E}[F_{t+1}] + M_1\mathbb{E}[\|\Delta^{t+1}\|^2] + M_2\mathbb{E}[\sigma_{t+1}^2]. \tag{18}$$

By Lemma B.8, by setting an arbitrary $\alpha > 0$, we have

$$\mathbb{E}[F_{t+1}] \leq (1 - \rho\eta_t)\mathbb{E}[F_t] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] + \eta_t^2 LD^2\left(\frac{1}{\alpha} + \frac{1}{2}\right). \tag{19}$$

By using (19), Remark B.7, and substituting the values into (18)

$$\begin{aligned} \mathbb{E}[r_{t+1}] \leq (1 - \rho\eta_t)\mathbb{E}[F_t] &+ \left(\frac{\alpha}{L} + M_1(1-\rho_1)\right)\mathbb{E}[\|\Delta^t\|^2] \\ &+ \left(M_1A + M_2(1-\rho_2)\right)\mathbb{E}[\sigma_t^2] + \left(L\left(\frac{1}{\alpha} + \frac{1}{2}\right) + M_1B + M_2E\right)\eta_t^2D^2 + M_1C. \end{aligned} \tag{20}$$

We now choose $M_1 = \frac{2\alpha}{\rho_1 L}$ and $M_2 = \frac{2M_1A}{\rho_2}$. With this choice, we can verify

$$\frac{\alpha}{L} + M_1(1-\rho_1) = \frac{\alpha}{L} + \frac{2\alpha}{\rho_1 L}(1-\rho_1) = \frac{\alpha}{L}\left(1 + \frac{2(1-\rho_1)}{\rho_1}\right) = \frac{\alpha}{L}\cdot\frac{2-\rho_1}{\rho_1} = M_1\left(1 - \frac{\rho_1}{2}\right), \tag{21}$$

and similarly,

$$M_1A + M_2(1-\rho_2) = M_1A + \frac{2M_1A}{\rho_2}(1-\rho_2) = M_1A\left(1 + \frac{2(1-\rho_2)}{\rho_2}\right) = M_1A\cdot\frac{2-\rho_2}{\rho_2} = M_2\left(1 - \frac{\rho_2}{2}\right). \tag{22}$$

Define

$$a := D^2\left(L\left(\frac{1}{\alpha} + \frac{1}{2}\right) + M_1B + M_2E\right) \quad \text{and} \quad b := M_1C.$$

Substituting (21)–(22) into (20)

$$\mathbb{E}[r_{t+1}] \leq (1 - \rho\eta_t)\mathbb{E}[F_t] + M_1\left(1 - \frac{\rho_1}{2}\right)\mathbb{E}[\|\Delta^t\|^2] + M_2\left(1 - \frac{\rho_2}{2}\right)\mathbb{E}[\sigma_t^2] + a\eta_t^2 + b. \tag{23}$$

Recall $R_t = \mathbb{E}[r_t] = \mathbb{E}[F_t] + M_1\mathbb{E}[\|\Delta^t\|^2] + M_2\mathbb{E}[\sigma_t^2]$. To obtain a contraction in terms of R_t, we show that the coefficients in (23) satisfy

$$M_1\left(1-\frac{\rho_1}{2}\right) \le M_1(1-\rho\eta_t) \quad \text{and} \quad M_2\left(1-\frac{\rho_2}{2}\right) \le M_2(1-\rho\eta_t).$$

By construction of $d = \frac{2}{\min\{\rho_1,\rho_2\}}$, the step decay satisfies $\rho\eta_t \le \frac{\min\{\rho_1,\rho_2\}}{2}$, so we have

$$\rho\eta_t \le \frac{\rho_1}{2} \quad \text{and} \quad \rho\eta_t \le \frac{\rho_2}{2},$$

which implies that for all $\rho\eta_t \in]0,1]$ that

$$1-\frac{\rho_1}{2} \le 1-\rho\eta_t \quad \text{and} \quad 1-\frac{\rho_2}{2} \le 1-\rho\eta_t.$$

Since $M_1, M_2 > 0$, we obtain:

$$M_1\left(1-\frac{\rho_1}{2}\right) \le M_1(1-\rho\eta_t) \quad \text{and} \quad M_2\left(1-\frac{\rho_2}{2}\right) \le M_2(1-\rho\eta_t).$$

Applying these to (23)

$$\begin{aligned}\mathbb{E}[r_{t+1}] &\le (1-\rho\eta_t)\mathbb{E}[F_t] + M_1(1-\rho\eta_t)\mathbb{E}[\|\Delta^t\|^2] + M_2(1-\rho\eta_t)\mathbb{E}[\sigma_t^2] + a\eta_t^2 + b \\ &= (1-\rho\eta_t)\left(\mathbb{E}[F_t] + M_1\mathbb{E}[\|\Delta^t\|^2] + M_2\mathbb{E}[\sigma_t^2]\right) + a\eta_t^2 + b \\ &= (1-\rho\eta_t)\mathbb{E}[r_t] + a\eta_t^2 + b.\end{aligned}$$

Thus

$$R_{t+1} \le (1-\rho\eta_t)R_t + a\eta_t^2 + b, \quad \text{where } R_t = \mathbb{E}[r_t]. \tag{24}$$

Now we claim that for constant step decay $\eta_t = \eta$:

$$R_T \le (1-\rho\eta)^T R_0 + a\eta^2 \sum_{k=0}^{T-1}(1-\rho\eta)^k + bT. \tag{25}$$

The claim is proved by mathematical induction on T.

Base Case ($T = 1$): From the recursion with $t = 0$:

$$R_1 \le (1-\rho\eta)R_0 + a\eta^2 + b = (1-\rho\eta)^1 R_0 + a\eta^2 \sum_{k=0}^{0}(1-\rho\eta)^k + b\cdot 1.$$

Inductive Hypothesis: Assume for $T = n$:

$$R_n \le (1-\rho\eta)^n R_0 + a\eta^2 \sum_{k=0}^{n-1}(1-\rho\eta)^k + bn.$$

Inductive Step: For $T = n+1$, starting from $R_{n+1} \le (1-\rho\eta)R_n + a\eta^2 + b$:

$$\begin{aligned}R_{n+1} &\le (1-\rho\eta)\left[(1-\rho\eta)^n R_0 + a\eta^2 \sum_{k=0}^{n-1}(1-\rho\eta)^k + bn\right] + a\eta^2 + b \\ &= (1-\rho\eta)^{n+1} R_0 + a\eta^2\left[(1-\rho\eta)\sum_{k=0}^{n-1}(1-\rho\eta)^k + 1\right] + b[n(1-\rho\eta)+1].\end{aligned}$$

Since $(1-\rho\eta)\sum_{k=0}^{n-1}(1-\rho\eta)^k + 1 = \sum_{k=0}^{n}(1-\rho\eta)^k$ and $\rho\eta \in]0,1]$ gives $n(1-\rho\eta)+1 \le n+1$

$$R_{n+1} \le (1-\rho\eta)^{n+1} R_0 + a\eta^2 \sum_{k=0}^{n}(1-\rho\eta)^k + b(n+1).$$

By induction:

$$R_T \leq (1-\rho\eta)^T R_0 + a\eta^2 \sum_{k=0}^{T-1} (1-\rho\eta)^k + bT.$$

We now analyze the convergence behavior under different step decay values. The following cases arise.

Case 1: $T \leq d$ with $\eta_t = \eta = \frac{1}{\rho d}$. Using the geometric series $\sum_{k=0}^{T-1}(1-\frac{1}{d})^k \leq d$ and $(1-x)^n \leq \exp(-nx)$

$$R_T \leq \left(1 - \frac{1}{d}\right)^T R_0 + \frac{a}{\rho^2 d} + bT \leq \exp\left(-\frac{T}{d}\right) R_0 + \frac{a}{\rho^2 d} + bT.$$

Since $T \leq d$ gives $\frac{1}{d} \leq \frac{2}{T+d}$ and $\exp(-T/d) \leq \exp(-T/(2d))$

$$R_T \leq \exp\left(-\frac{T}{2d}\right) R_0 + \frac{2a}{\rho^2(T+d)} + bT \leq e \cdot \exp\left(-\frac{T}{2d}\right) R_0 + \frac{32a}{\rho^2(T+d)} + bT.$$

Case 2: $T > d$, $t < t_0$, with $\eta_t = \frac{1}{\rho d}$ and $t_0 := \lfloor T/2 \rfloor$. Similarly

$$R_{t_0} \leq \exp\left(-\frac{t_0}{d}\right) R_0 + \frac{a}{\rho^2 d} + bt_0.$$

Case 3: $T > d$, $t \geq t_0$, with $\eta_t = \frac{2}{\rho(2d+t-t_0)}$. Let $k = t - t_0$, $K := T - t_0 = \lceil T/2 \rceil$, $\bar{R}_k := R_{t_0+k}$, $\bar{\eta}_k := \frac{2}{\rho(2d+k)}$. The recursion becomes

$$\bar{R}_{k+1} \leq \frac{2d+k-2}{2d+k} \bar{R}_k + \frac{4a}{\rho^2(2d+k)^2} + b.$$

Multiplying by $(2d+k)^2$ and using $(2d+k)(2d+k-2) \leq (2d+k-1)^2$

$$(2d+k)^2 \bar{R}_{k+1} \leq (2d+k-1)^2 \bar{R}_k + \frac{4a}{\rho^2} + b(2d+k)^2.$$

Summing for $k = 0, \cdots, K-1$

$$(2d+K-1)^2 \bar{R}_K \leq (2d)^2 \bar{R}_0 + K\frac{4a}{\rho^2} + bK(2d+K-1)^2.$$

For $K = \lceil T/2 \rceil \geq T/2$, we have $2d + K - 1 \geq (T+d)/2$, giving

$$\frac{4d^2}{(2d+K-1)^2} \leq \frac{16d^2}{(T+d)^2}, \quad \frac{K}{(2d+K-1)^2} \leq \frac{4}{T+d}.$$

Since $t_0 = \lfloor T/2 \rfloor \geq T/2 - 1$

$$\exp\left(-\frac{t_0}{d}\right) \leq e \cdot \exp\left(-\frac{T}{2d}\right).$$

Combining with Case 2 and using $t_0 + K = T$, $\frac{d}{T+d} \leq 1$

$$R_T \leq e \cdot \exp\left(-\frac{T}{2d}\right) R_0 + \frac{32a}{\rho^2(T+d)} + bT.$$

Hence by either Case I or Case II and Case III, we have the resulting expression

$$R_T \leq e \cdot \exp\left(-\frac{T}{2d}\right) R_0 + \frac{32a}{\rho^2(T+d)} + bT.$$

Substituting the expressions for a and b

$$a = D^2\left(L\left(\frac{1}{\alpha} + \frac{1}{2}\right) + \frac{2\alpha B}{\rho_1 L} + \frac{4\alpha AE}{\rho_1 \rho_2 L}\right) \quad \text{and} \quad b = \frac{2\alpha C}{\rho_1 L},$$

$$R_T \leq e \cdot \exp\left(-\frac{T}{2d}\right) R_0 + \frac{32D^2\left(L\left(\frac{1}{\alpha}+\frac{1}{2}\right)+\frac{2\alpha B}{\rho_1 L}+\frac{4\alpha AE}{\rho_1\rho_2 L}\right)}{\rho^2(T+d)} + \frac{2\alpha CT}{\rho_1 L}$$
$$= e \cdot \exp\left(-\frac{T}{2d}\right) R_0 + \frac{16D^2 L}{\rho^2(T+d)} + \alpha\left[\frac{64D^2 B}{\rho^2(T+d)\rho_1 L} + \frac{128D^2 AE}{\rho^2(T+d)\rho_1\rho_2 L} + \frac{2CT}{\rho_1 L}\right] + \frac{32D^2 L}{\rho^2(T+d)\alpha}.$$

Since the inequality holds for all $\alpha > 0$, we take the infimum over α to get the tightest bound. The α-dependent terms in the bound are of the form

$$g(\alpha) := \alpha\left[\frac{64D^2 B}{\rho^2(T+d)\rho_1 L} + \frac{128D^2 AE}{\rho^2(T+d)\rho_1\rho_2 L} + \frac{2CT}{\rho_1 L}\right] + \frac{32D^2 L}{\rho^2(T+d)\alpha}.$$

This has the form $g(\alpha) = \alpha u + \frac{v}{\alpha}$, which is minimized at $\alpha^\star = \sqrt{v/u}$ with minimum value $g(\alpha^\star) = 2\sqrt{uv}$. So,

$$g(\alpha^\star) = 2\sqrt{\frac{32D^2}{\rho^2(T+d)}\left(\frac{64D^2 B}{\rho^2(T+d)\rho_1} + \frac{128D^2 AE}{\rho^2(T+d)\rho_1\rho_2} + \frac{2CT}{\rho_1}\right)},$$

which completes the proof since $\mathbb{E}[F_t] \leq R_T$. Consequentially when $C = 0$, it yields a $\mathcal{O}(1/T)$ convergence rate. □

Theorem B.10 (Formal Statement of Theorem 4.6)**.** *Let f be a function that satisfies L-smoothness Assumption 1 and ρ-quasar-convexity Assumption 2. Suppose the stochastic estimator m^t and auxiliary sequence $\{\sigma_t\}$ satisfy Assumption 3 with parameters $\rho_1, \rho_2 \in]0, 1]$ and constants $A, B, C, E \geq 0$. Let $\{x^t\}_{t=0}^{+\infty}$ be a sequence generated by Algorithm BSFW by choosing the step decay*

$$\eta_t = \frac{2}{\rho(t+\nu)}, \quad \textit{where} \quad \nu = \max\left\{2, \frac{4}{\min\{\rho_1, \rho_2\}}\right\}.$$

Then, the expected functional-value gap satisfies

$$\mathbb{E}[F_t] \leq \sqrt{\frac{16D^2}{\rho^2(t+\nu)}\left(\frac{32D^2 B}{\rho^2(t+\nu)\rho_1} + \frac{64D^2 AE}{\rho^2(t+\nu)\rho_1\rho_2} + \frac{2CT}{\rho_1}\right)} + \frac{4\nu^2\mathbb{E}[r_0]}{(t+\nu)^2} + \frac{8D^2 L}{\rho^2(t+\nu)}.$$

where r_t is a Lyapunov function defined by

$$\forall t, \quad r_t = F_t + \tfrac{2\alpha^\star}{\rho_1 L}\|\Delta^t\|^2 + \tfrac{4\alpha^\star A}{\rho_1\rho_2 L}\sigma_t^2,$$

with

$$\alpha^\star = \sqrt{\left(\frac{16D^2 L}{\rho^2(T+\nu)}\right) \Big/ \left(\frac{32D^2 B}{\rho^2(T+\nu)\rho_1 L} + \frac{64D^2 AE}{\rho^2(T+\nu)\rho_1\rho_2 L} + \frac{2CT}{\rho_1 L}\right)}.$$

If $C = 0$, the last term in the square root vanishes and we obtain a $\mathcal{O}(1/t)$ rate.

Proof. Define a Lyapunov function $r_t = F_t + M_1\|\Delta^t\|^2 + M_2\sigma_t^2$, with $M_1, M_2 > 0$ as arbitrary constants.

Recursion. Following the same derivation as in Theorem B.9, we analyze the full expectation of the Lyapunov function. Define $R_t := \mathbb{E}[r_t]$. From Lemma B.8 and Assumption 3, we have established that for any $\alpha > 0$, i.e., (24):

$$R_{t+1} \leq (1 - \rho\eta_t)R_t + a\eta_t^2 + b, \tag{26}$$

where

$$a = D^2\left(L\left(\frac{1}{\alpha}+\frac{1}{2}\right) + M_1 B + M_2 E\right), \quad b = M_1 C,$$

with $M_1 = \frac{2\alpha}{\rho_1 L}$ and $M_2 = \frac{2M_1 A}{\rho_2}$.

The derivation of (26) requires that $\rho\eta_t \leq \frac{\min\{\rho_1,\rho_2\}}{2}$ for all $t \geq 0$. With the proposed step size $\eta_t = \frac{2}{\rho(t+\nu)}$, we have $\rho\eta_t = \frac{2}{t+\nu}$. Since $t \geq 0$, the maximum value of $\rho\eta_t$ occurs at $t = 0$:

$$\max_{t\geq 0} \rho\eta_t = \frac{2}{\nu}.$$

By the definition $\nu \geq \frac{4}{\min\{\rho_1,\rho_2\}}$, we have

$$\frac{2}{\nu} \leq \frac{2}{\frac{4}{\min\{\rho_1,\rho_2\}}} = \frac{\min\{\rho_1,\rho_2\}}{2}.$$

Therefore, for all $t \geq 0$:

$$\rho\eta_t = \frac{2}{t+\nu} \leq \frac{2}{\nu} \leq \frac{\min\{\rho_1,\rho_2\}}{2},$$

which implies both $\rho\eta_t \leq \frac{\rho_1}{2}$ and $\rho\eta_t \leq \frac{\rho_2}{2}$, validating the contraction conditions used in deriving (26). Additionally, the condition $\nu \geq 2$ ensures that $t+\nu-2 \geq 0$ for all $t \geq 0$, which will be needed in the subsequent analysis.

Computing the Recursion Coefficients. With the step size $\eta_t = \frac{2}{\rho(t+\nu)}$, we compute:

$$\rho\eta_t = \frac{2}{t+\nu}, \tag{27}$$

$$1-\rho\eta_t = 1 - \frac{2}{t+\nu} = \frac{t+\nu-2}{t+\nu}, \tag{28}$$

$$\eta_t^2 = \frac{4}{\rho^2(t+\nu)^2}. \tag{29}$$

Substituting (28) and (29) into (26):

$$R_{t+1} \leq \frac{t+\nu-2}{t+\nu} R_t + \frac{4a}{\rho^2(t+\nu)^2} + b. \tag{30}$$

Weighted Lyapunov Analysis. Multiply both sides of (30) by $(t+\nu)^2$:

$$(t+\nu)^2 R_{t+1} \leq (t+\nu)(t+\nu-2)R_t + \frac{4a}{\rho^2} + b(t+\nu)^2. \tag{31}$$

We now use the algebraic identity:

$$(t+\nu)(t+\nu-2) = (t+\nu-1)^2 - 1.$$

To verify this, expand both sides:

$$\begin{aligned}(t+\nu)(t+\nu-2) &= (t+\nu)^2 - 2(t+\nu),\\ (t+\nu-1)^2 - 1 &= (t+\nu)^2 - 2(t+\nu) + 1 - 1 = (t+\nu)^2 - 2(t+\nu).\end{aligned}$$

Since $(t+\nu-1)^2 - 1 \leq (t+\nu-1)^2$, inequality (31) becomes:

$$(t+\nu)^2 R_{t+1} \leq (t+\nu-1)^2 R_t + \frac{4a}{\rho^2} + b(t+\nu)^2. \tag{32}$$

Defining the Weighted Sequence. Define the weighted sequence:

$$W_t := (t+\nu-1)^2 R_t.$$

Note that with this setting we have for all $0 \leq t \leq T$,

$$W_0 = (\nu-1)^2 R_0, \;\; W_T = (T+\nu-1)^2 R_T, \;\; \text{and} \;\; W_{t+1} = (t+\nu)^2 R_{t+1}.$$

Inequality (32) can be rewritten as:

$$W_{t+1} \leq W_t + \frac{4a}{\rho^2} + b(t+\nu)^2. \tag{33}$$

Summing inequality (33) from $t=0$ to $t=T-1$:

$$W_T - W_0 \leq \sum_{t=0}^{T-1} \left(\frac{4a}{\rho^2} + b(t+\nu)^2 \right) = \frac{4aT}{\rho^2} + b\sum_{t=0}^{T-1} (t+\nu)^2. \tag{34}$$

Rearranging:

$$W_T \leq W_0 + \frac{4aT}{\rho^2} + b\sum_{t=0}^{T-1}(t+\nu)^2. \tag{35}$$

Bounding the Summation. We bound the sum $\sum_{t=0}^{T-1}(t+\nu)^2$ as follows. Since $(t+\nu)$ is increasing in t, the largest term in the sum is $(T-1+\nu)^2$. Thus:

$$\sum_{t=0}^{T-1}(t+\nu)^2 \leq T\cdot(T-1+\nu)^2 = T(T+\nu-1)^2. \tag{36}$$

Substituting (36) into (35):

$$W_T \leq W_0 + \frac{4aT}{\rho^2} + bT(T+\nu-1)^2. \tag{37}$$

Converting Back to R_T. Substituting $W_T = (T+\nu-1)^2 R_T$ and $W_0 = (\nu-1)^2 R_0$:

$$(T+\nu-1)^2 R_T \leq (\nu-1)^2 R_0 + \frac{4aT}{\rho^2} + bT(T+\nu-1)^2. \tag{38}$$

Dividing both sides by $(T+\nu-1)^2$:

$$R_T \leq \frac{(\nu-1)^2}{(T+\nu-1)^2}R_0 + \frac{4aT}{\rho^2(T+\nu-1)^2} + bT. \tag{39}$$

Simplifying Using Bounds on ν. Since $\nu \geq 2$, we have $\nu - 1 \geq 1$. This implies:

$$T+\nu-1 \geq \frac{T}{2} + \nu - 1 = \frac{T+\nu}{2} + \frac{\nu-2}{2} \geq \frac{T+\nu}{2} \geq 0,$$

where we used $\nu \geq 2$. Therefore:

$$(T+\nu-1)^2 \geq \frac{(T+\nu)^2}{4}. \tag{40}$$

Using (40), we obtain the following bounds:

Bound 1: For the initial condition term:

$$\frac{(\nu-1)^2}{(T+\nu-1)^2} \leq \frac{4(\nu-1)^2}{(T+\nu)^2} \leq \frac{4\nu^2}{(T+\nu)^2},$$

where the last inequality uses $\nu - 1 \leq \nu$.

Bound 2: For the a-dependent term:

$$\frac{T}{(T+\nu-1)^2} \leq \frac{4T}{(T+\nu)^2}.$$

We further simplify using $T \leq T+\nu$:

$$\frac{4T}{(T+\nu)^2} = \frac{4}{T+\nu}\cdot\frac{T}{T+\nu} \leq \frac{4}{T+\nu}.$$

Final Bound in Terms of α. Substituting the bounds from Step 9 into (39):

$$R_T \leq \frac{4\nu^2 R_0}{(T+\nu)^2} + \frac{16a}{\rho^2(T+\nu)} + bT. \tag{41}$$

Explicit Form of the Bound. Substituting the definitions of a, b, M_1, and M_2:

$$a = D^2\left(L\left(\frac{1}{\alpha}+\frac{1}{2}\right) + M_1 B + M_2 E\right) = D^2\left(L\left(\frac{1}{\alpha}+\frac{1}{2}\right) + \frac{2\alpha B}{\rho_1 L} + \frac{4\alpha A E}{\rho_1\rho_2 L}\right),$$
$$b = M_1 C = \frac{2\alpha C}{\rho_1 L}.$$

Therefore, (41) becomes:

$$R_T \leq \frac{4\nu^2 R_0}{(T+\nu)^2} + \frac{16D^2}{\rho^2(T+\nu)}\left(L\left(\frac{1}{\alpha}+\frac{1}{2}\right) + \frac{2\alpha B}{\rho_1 L} + \frac{4\alpha AE}{\rho_1\rho_2 L}\right) + \frac{2\alpha CT}{\rho_1 L}$$
$$= \frac{4\nu^2 R_0}{(T+\nu)^2} + \frac{8D^2 L}{\rho^2(T+\nu)} + \frac{16D^2 L}{\rho^2(T+\nu)\alpha} + \alpha\left(\frac{32D^2 B}{\rho^2(T+\nu)\rho_1 L} + \frac{64D^2 AE}{\rho^2(T+\nu)\rho_1\rho_2 L} + \frac{2CT}{\rho_1 L}\right). \tag{42}$$

Optimizing Over α. The α-dependent terms in (42) have the form $g(\alpha) := \frac{v}{\alpha} + \alpha u$, where

$$v := \frac{16D^2 L}{\rho^2(T+\nu)},$$
$$u := \frac{32D^2 B}{\rho^2(T+\nu)\rho_1 L} + \frac{64D^2 AE}{\rho^2(T+\nu)\rho_1\rho_2 L} + \frac{2CT}{\rho_1 L}.$$

Now, $g(\alpha)$ is minimized at $\alpha^\star = \sqrt{v/u}$ with minimum value $g(\alpha^\star) = 2\sqrt{uv}$. Thus, substituting $g(\alpha^\star) = 2\sqrt{uv}$ into (42):

$$R_T \leq \frac{4\nu^2 R_0}{(T+\nu)^2} + \frac{8D^2 L}{\rho^2(T+\nu)} + 2\sqrt{uv}$$
$$= \frac{4\nu^2 R_0}{(T+\nu)^2} + \frac{8D^2 L}{\rho^2(T+\nu)} + 2\sqrt{\frac{16D^2}{\rho^2(T+\nu)}\left(\frac{32D^2 B}{\rho^2(T+\nu)\rho_1} + \frac{64D^2 AE}{\rho^2(T+\nu)\rho_1\rho_2} + \frac{2CT}{\rho_1}\right)}. \tag{43}$$

Simplifying the constant under the square root:

$$\mathbb{E}[r_T] \leq \frac{4\nu^2 \mathbb{E}[r_0]}{(T+\nu)^2} + \frac{8D^2 L}{\rho^2(T+\nu)} + \sqrt{\frac{64D^2}{\rho^2(T+\nu)}\left(\frac{32D^2 B}{\rho^2(T+\nu)\rho_1} + \frac{64D^2 AE}{\rho^2(T+\nu)\rho_1\rho_2} + \frac{2CT}{\rho_1}\right)}. \tag{44}$$

This completes the proof since $\mathbb{E}[F_T] \leq \mathbb{E}[r_T]$. □

B.2.2 Nonconvex Case

Under the nonconvex setting, we first build descent Lemma B.11 which is used in the convergence analysis that follows. We offer convergence rates using a fixed horizon step decay in Theorem B.12 and an any-time convergent step decay in Theorem B.13.

Lemma B.11 (Descent under Smoothness in Stochastic Setting for nonconvex Functions)**.** *Let f be an objective function that satisfies L-smoothness Assumption 1. Consider the sequence $\{x^t\}_{t=0}^{+\infty}$ generated by Algorithm BSFW. Then for any $\alpha > 0$, the following inequality holds*

$$\eta_t \mathbb{E}\left[\max_{u\in\mathcal{C}} \langle \nabla f(x^t), x^t - u\rangle\right] \leq \mathbb{E}\left[F_t\right] - \mathbb{E}\left[F_{t+1}\right] + \frac{\alpha}{L}\mathbb{E}\left[\|\Delta^t\|^2\right] + \eta_t^2 L D^2\left(\frac{1}{\alpha}+\frac{1}{2}\right).$$

Proof. From Assumption 1, we have

$$f(x^{t+1}) \leq f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t\rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2.$$

We distinguish two cases based on the value of γ_t, following the analysis of Algorithm BSFW.
Case I: $\gamma_t < 1$. Then we have $x^{t+1} = x^t + \gamma_t d^t$. By using Lemma B.1,

$$f(x^{t+1}) \leq f(x^t) + \gamma_t\langle \nabla f(x^t), d^t\rangle + \frac{L}{2}\gamma_t^2\|d^t\|^2$$
$$\leq f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \gamma_t\langle m^t, d^t\rangle + \frac{L}{2}\gamma_t^2\|d^t\|^2$$
$$\leq f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \eta_t\left(\frac{\|s^t - x^t\|}{\|d^t\|}\right)\left(\frac{\|d^t\|}{\|s^t - x^t\|}\right)\langle m^t, s^t - x^t\rangle + \frac{L}{2}\eta_t^2\frac{\|s^t - x^t\|^2}{\|d^t\|^2}\|d^t\|^2$$
$$\leq f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \eta_t\langle m^t, s^t - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2.$$

Since $s^t \in \mathrm{lmo}(m^t)$, $\forall u \in \mathcal{C}$, we get

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \gamma_t \langle \nabla f(x^t) - m^t, d^t \rangle + \eta_t \langle m^t, u - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2 \\ &\leq f(x^t) + \gamma_t \langle \nabla f(x^t) - m^t, d^t \rangle + \eta_t \langle m^t - \nabla f(x^t), u - x^t \rangle + \langle \nabla f(x^t), u - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2. \end{aligned}$$

By using the Young's inequality on both $\langle \nabla f(x^t) - m^t, \gamma_t d^t \rangle$ and $\langle m^t - \nabla f(x^t), \eta_t(u - x^t) \rangle$ with a parameter $\beta = \frac{\alpha}{L}$ for an arbitrary $\alpha > 0$, we have

$$\begin{aligned} \langle \nabla f(x^t) - m^t, \gamma_t d^t \rangle &\leq \frac{\alpha}{2L}\|\nabla f(x^t) - m^t\|^2 + \gamma_t^2 \frac{\alpha}{2L}\|d^t\|^2 \\ \langle m^t - \nabla f(x^t), \eta_t(u - x^t) \rangle &\leq \frac{\alpha}{2L}\|m^t - \nabla f(x^t)\|^2 + \eta_t^2 \frac{L}{2\alpha}\|u - x^t\|^2 \end{aligned}$$

By using these inequalities, we get

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \frac{\alpha}{2L}\|\nabla f(x^t) - m^t\|^2 + \gamma_t^2 \frac{L}{2\alpha}\|d^t\|^2 \\ &\quad + \frac{\alpha}{2L}\|\nabla f(x^t) - m^t\|^2 + \eta_t^2 \frac{L}{2\alpha}\|u - x^t\|^2 + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2 + \eta_t \langle \nabla f(x^t), u - x^t \rangle \\ &\leq f(x^t) + \frac{\alpha}{L}\|\Delta^t\|^2 + \eta_t^2 \frac{L}{2\alpha}\|s^t - x^t\|^2 + \eta_t^2 \frac{L}{2\alpha}\|u - x^t\|^2 + \eta_t \langle \nabla f(x^t), u - x^t \rangle + \frac{L}{2}\eta_t^2 D^2 \\ &\leq f(x^t) + \frac{\alpha}{L}\|\Delta^t\|^2 + \eta_t^2 \frac{L}{\alpha} D^2 + \eta_t \langle \nabla f(x^t), u - x^t \rangle + \frac{L}{2}\eta_t^2 D^2 \\ &\leq f(x^t) + \frac{\alpha}{L}\|\Delta^t\|^2 + \eta_t^2 L D^2 \left(\frac{1}{\alpha} + \frac{1}{2}\right) + \eta_t \langle \nabla f(x^t), u - x^t \rangle. \end{aligned}$$

Subtracting $f^\star$ and taking expectation on both sides gives us

$$\begin{aligned} &\mathbb{E}[F_{t+1}] \leq \mathbb{E}[F_t] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] + \eta_t^2 L D^2 \left(\frac{1}{\alpha} + \frac{1}{2}\right) + \eta_t \mathbb{E}[\langle \nabla f(x^t), x^t - u \rangle] \\ &\implies \eta_t \mathbb{E}[\langle \nabla f(x^t), x^t - u \rangle] \leq \mathbb{E}[F_t] - \mathbb{E}[F_{t+1}] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] + \eta_t^2 L D^2 \left(\frac{1}{\alpha} + \frac{1}{2}\right) \\ &\implies \eta_t \mathbb{E}\left[\max_{u \in \mathcal{C}} \langle \nabla f(x^t), x^t - u \rangle\right] \leq \mathbb{E}[F_t] - \mathbb{E}[F_{t+1}] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] + \eta_t^2 L D^2 \left(\frac{1}{\alpha} + \frac{1}{2}\right). \end{aligned}$$

Case II: $\gamma_t = 1$. Then we have $x^{t+1} = x^t + \eta_t(s^t - x^t)$. Hence $\forall u \in \mathcal{C}$, we get

$$\begin{aligned} f(x^{t+1}) &\leq f(x^t) + \eta_t \langle \nabla f(x^t), s^t - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2 \\ &\leq f(x^t) + \eta_t \langle \nabla f(x^t) - m^t, s^t - x^t \rangle + \eta_t \langle m^t, s^t - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2 \\ &\leq f(x^t) + \eta_t \langle \nabla f(x^t) - m^t, s^t - x^t \rangle + \eta_t \langle m^t, u - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2 \\ &\leq f(x^t) + \eta_t \langle \nabla f(x^t) - m^t, u - x^t \rangle + \eta_t \langle m^t - \nabla f(x^t), u - x^t \rangle \\ &\quad + \eta_t \langle \nabla f(x^t), u - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2 \\ &\leq f(x^t) + \eta_t \langle \nabla f(x^t) - m^t, s^t - u \rangle + \eta_t \langle \nabla f(x^t), u - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2. \end{aligned}$$

Using Young's inequality on $\langle \nabla f(x^t) - m^t, \eta_t(s^t - u) \rangle$ with $\beta = \frac{2\alpha}{L}$ with an arbitrary $\alpha > 0$, we have

$$\langle \nabla f(x^t) - m^t, \eta_t(s^t - u) \rangle \leq \frac{\alpha}{L}\|\nabla f(x^t) - m^t\|^2 + \frac{L}{4\alpha}\eta_t^2 \|s^t - u\|^2.$$

By using this inequality,

$$f(x^{t+1}) \leq f(x^t) + \frac{\alpha}{L}\|\Delta^t\|^2 + \eta_t^2\frac{L}{4\alpha}\|s^t - u\|^2 + \eta_t\langle\nabla f(x^t), u - x^t\rangle + \frac{L}{2}\eta_t^2 D^2$$
$$\leq f(x^t) + \frac{\alpha}{L}\|\Delta^t\|^2 + \eta_t\langle\nabla f(x^t), u - x^t\rangle + \eta_t^2 LD^2\left(\frac{1}{4\alpha} + \frac{1}{2}\right).$$

By subtracting $f^\star$ and taking expectation on both sides, we get

$$\mathbb{E}[F_{t+1}] \leq \mathbb{E}[F_t] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] + \eta_t\mathbb{E}[\langle\nabla f(x^t), u - x^t\rangle] + \eta_t^2 LD^2\left(\frac{1}{4\alpha} + \frac{1}{2}\right)$$
$$\leq \mathbb{E}[F_t] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] + \eta_t\mathbb{E}[\langle\nabla f(x^t), u - x^t\rangle] + \eta_t^2 LD^2\left(\frac{1}{\alpha} + \frac{1}{2}\right),$$

$$\implies \eta_t\mathbb{E}[\langle\nabla f(x^t), x^t - u\rangle] \leq \mathbb{E}[F_t] - \mathbb{E}[F_{t+1}] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] + \eta_t^2 LD^2\left(\frac{1}{\alpha} + \frac{1}{2}\right),$$
$$\implies \eta_t\mathbb{E}\left[\max_{u\in\mathcal{C}}\langle\nabla f(x^t), x^t - u\rangle\right] \leq \mathbb{E}[F_t] - \mathbb{E}[F_{t+1}] + \frac{\alpha}{L}\mathbb{E}[\|\Delta^t\|^2] + \eta_t^2 LD^2\left(\frac{1}{\alpha} + \frac{1}{2}\right).$$

Which gives us the desired expression in both cases □

Theorem B.12 (Formal Statement of Theorem 4.7)**.** *Let f be an objective function that satisfies L-smoothness Assumption 1. Let $T \in \mathbb{N}$, and $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW where the stochastic estimator m^t and auxiliary sequence $\{\sigma_t\}$ satisfy Assumption 3 with parameters $\rho_1, \rho_2 \in]0, 1]$ and constants $A, B, C, E \geq 0$. Let the step decay be chosen as a constant for all t, $\eta_t = \frac{1}{\sqrt{T}}$. Then we have*

$$\mathbb{E}\left[\min_{0\leq t\leq T-1}\text{Gap}(x^t)\right] \leq \frac{\mathbb{E}[r_0]}{\sqrt{T}} + \frac{D^2 L}{2\sqrt{T}} + D\sqrt{\frac{D^2}{\rho_1 T}\left(B + \frac{AE}{\rho_2}\right) + \frac{C}{\rho_1}}. \tag{45}$$

where r_t is a Lyapunov function defined by

$$\forall t: \quad r_t = F_t + \frac{\alpha^\star}{L\rho_1}\|\Delta^t\|^2 + \frac{\alpha^\star A}{L\rho_1\rho_2}\sigma_t^2,$$

with

$$\alpha^\star = \sqrt{\left(\frac{D^2 L}{\sqrt{T}}\right) \Big/ \left(\frac{D^2 B}{L\rho_1\sqrt{T}} + \frac{D^2 AE}{L\rho_1\rho_2\sqrt{T}} + \frac{C\sqrt{T}}{L\rho_1}\right)}.$$

If $C = 0$, the last term vanishes and we recover the standard $\mathcal{O}(1/\sqrt{T})$ rate.

Proof. Multiply inequality (16) by M_1 (at iteration $t+1$), inequality (17) by M_2, and add the two to obtain

$$\begin{aligned} M_1\left\|m^{t+1} - \nabla f\left(x^{t+1}\right)\right\|^2 + M_2\mathbb{E}\left[\sigma_{t+1}^2\right] &\leq M_1(1-\rho_1)\left\|m^t - \nabla f\left(x^t\right)\right\|^2 + M_1 A\mathbb{E}\left[\sigma_t^2\right] \\ &\quad + M_1 B\eta_t^2 D^2 + M_1 C + M_2(1-\rho_2)\mathbb{E}\left[\sigma_t^2\right] + M_2 E\eta_t^2 D^2 \\ &= M_1(1-\rho_1)\left\|m^t - \nabla f\left(x^t\right)\right\|^2 \\ &\quad + M_2\left(1 - \rho_2 + \frac{M_1 A}{M_2}\right)\mathbb{E}\left[\sigma_t^2\right] + \eta_t^2 D^2(M_1 B + M_2 E) + M_1 C. \end{aligned} \tag{46}$$

Define the Lyapunov function $r_t := f(x^t) - f^\star + M_1\|m^t - \nabla f(x^t)\|^2 + M_2\sigma_t^2$. From its definition, the expected Lyapunov difference satisfies

$$\begin{aligned} \mathbb{E}\left[r_t - r_{t+1}\right] &= \mathbb{E}\left[f\left(x^t\right) - f\left(x^{t+1}\right)\right] \\ &\quad + M_1\mathbb{E}\left[\left\|m^t - \nabla f\left(x^t\right)\right\|^2 - \left\|m^{t+1} - \nabla f\left(x^{t+1}\right)\right\|^2\right] + M_2\mathbb{E}\left[\sigma_t^2 - \sigma_{t+1}^2\right]. \end{aligned} \tag{47}$$

Rearranging (47) gives

$$\begin{aligned}\mathbb{E}\left[f\left(x^t\right)-f\left(x^{t+1}\right)\right] =&\mathbb{E}\left[r_t-r_{t+1}\right]-M_1\mathbb{E}\left[\left\|m^t-\nabla f\left(x^t\right)\right\|^2-\left\|m^{t+1}-\nabla f\left(x^{t+1}\right)\right\|^2\right]\\ &-M_2\mathbb{E}\left[\sigma_t^2-\sigma_{t+1}^2\right].\end{aligned} \tag{48}$$

Now, using (16)-(17), we obtain lower bounds on the error decrements

$$\mathbb{E}\big[\|m^t-\nabla f(x^t)\|^2-\|m^{t+1}-\nabla f(x^{t+1})\|^2\big]\geq\rho_1\,\mathbb{E}\big[\|m^t-\nabla f(x^t)\|^2\big]-A\,\mathbb{E}[\sigma_t^2]-B\eta_t^2D^2-C, \tag{49}$$

$$\mathbb{E}\big[\sigma_t^2-\sigma_{t+1}^2\big]\geq\rho_2\,\mathbb{E}[\sigma_t^2]-E\eta_t^2D^2. \tag{50}$$

Substituting (49)–(50) into (48) yields

$$\begin{aligned}\mathbb{E}\big[f(x^t)-f(x^{t+1})\big]\leq&\,\mathbb{E}[r_t-r_{t+1}]-\rho_1M_1\,\mathbb{E}\big[\|m^t-\nabla f(x^t)\|^2\big]-(\rho_2M_2-M_1A)\,\mathbb{E}[\sigma_t^2]\\ &+\eta_t^2D^2(M_1B+M_2E)+M_1C.\end{aligned} \tag{51}$$

Plugging (51) into Lemma B.11, and grouping terms, we obtain

$$\begin{aligned}\eta_t\,\mathbb{E}[\mathrm{Gap}(x^t)]\leq\mathbb{E}\bigg[&f(x^t)-f(x^\star)+\left(1-\rho_1+\frac{\alpha}{M_1L}\right)M_1\|m^t-\nabla f(x^t)\|^2+\left(1-\rho_2+\frac{M_1A}{M_2}\right)M_2\sigma_t^2\bigg]\\ &-\mathbb{E}\big[f(x^{t+1})-f(x^\star)+M_1\|m^{t+1}-\nabla f(x^{t+1})\|^2+M_2\sigma_{t+1}^2\big]\\ &+D^2\eta_t^2\left(\frac{L}{2}+\frac{L}{\alpha}+M_1B+M_2E\right)+M_1C.\end{aligned}$$

Now choose $M_1:=\frac{\alpha}{L\rho_1}$ and $M_2:=\frac{M_1A}{\rho_2}=\frac{\alpha A}{L\rho_1\rho_2}$, for $\alpha>0$. So that $1-\rho_1+\frac{\alpha}{M_1L}=1$ and $1-\rho_2+\frac{M_1A}{M_2}=1$. Then the first expectation in (52) equals $\mathbb{E}[r_t]$, and the second equals $\mathbb{E}[r_{t+1}]$, giving

$$\eta_t\,\mathbb{E}[\mathrm{Gap}(x^t)]\leq\mathbb{E}[r_t-r_{t+1}]+D^2\eta_t^2\left(\frac{L}{2}+\frac{L}{\alpha}+M_1B+M_2E\right)+M_1C. \tag{52}$$

Summing (52) over $t=0,\cdots,T-1$ telescopes

$$\sum_{t=0}^{T-1}\eta_t\,\mathbb{E}[\mathrm{Gap}(x^t)]\leq\mathbb{E}[r_0-r_T]+D^2\sum_{t=0}^{T-1}\eta_t^2\left(\frac{L}{2}+\frac{L}{\alpha}+M_1B+M_2E\right)+TM_1C. \tag{53}$$

Since $r_T\geq0$, we have

$$\sum_{t=0}^{T-1}(\eta_t\,\mathbb{E}[\mathrm{Gap}(x^t)])\leq\mathbb{E}[r_0]+D^2\left(\frac{L}{2}+\frac{L}{\alpha}+M_1B+M_2E\right)\sum_{t=0}^{T-1}\eta_t^2+TM_1C. \tag{54}$$

With constant step decay $\eta_t\equiv\eta=\frac{1}{\sqrt{T}}$ and Since $\min_{0\leq t\leq T-1}\mathrm{Gap}(x^t)\leq\mathrm{Gap}(x^t)$ for all t, taking expectations we have

$$\mathbb{E}\left[\min_{0\leq t\leq T-1}\mathrm{Gap}(x^t)\right]\sqrt{T}\leq\mathbb{E}[r_0]+D^2\left(\frac{L}{2}+\frac{L}{\alpha}+M_1B+M_2E\right)+TM_1C. \tag{55}$$

where $M_1=\frac{\alpha}{L\rho_1}$ and $M_2=\frac{\alpha A}{L\rho_1\rho_2}$ which means that

$$\begin{aligned}\mathbb{E}\left[\min_{0\leq t\leq T-1}\mathrm{Gap}(x^t)\right]&\leq\frac{\mathbb{E}[r_0]}{\sqrt{T}}+\frac{D^2\left(\frac{L}{2}+\frac{L}{\alpha}+\frac{\alpha}{L\rho_1}B+\frac{\alpha A}{L\rho_1\rho_2}E\right)}{\sqrt{T}}+\sqrt{T}\frac{\alpha}{L\rho_1}C\\ &=\frac{\mathbb{E}[r_0]}{\sqrt{T}}+\frac{D^2L}{2\sqrt{T}}+\frac{D^2L}{\alpha\sqrt{T}}+\alpha\left[\frac{D^2B}{L\rho_1\sqrt{T}}+\frac{D^2AE}{L\rho_1\rho_2\sqrt{T}}+\frac{C\sqrt{T}}{L\rho_1}\right].\end{aligned} \tag{56}$$

Since the inequality holds for all $\alpha>0$, so take the infimum over α to get the tightest bound. The α-dependent terms in the bound are of the form

$$g(\alpha)=\frac{D^2L}{\alpha\sqrt{T}}+\alpha\left[\frac{D^2B}{L\rho_1\sqrt{T}}+\frac{D^2AE}{L\rho_1\rho_2\sqrt{T}}+\frac{C\sqrt{T}}{L\rho_1}\right]. \tag{57}$$

This has the form $g(\alpha) = \frac{v}{\alpha} + u\alpha$, which is minimized at $\alpha^\star = \sqrt{v/u}$ with minimum value $g(\alpha^\star) = 2\sqrt{uv}$. where

$$u := \frac{D^2 B}{L\rho_1\sqrt{T}} + \frac{D^2 AE}{L\rho_1\rho_2\sqrt{T}} + \frac{C\sqrt{T}}{L\rho_1},$$
$$v := \frac{D^2 L}{\sqrt{T}}.$$

Hence we have,

$$g(\alpha^*) = 2\sqrt{\frac{D^2 L}{\sqrt{T}}\left(\frac{D^2 B}{L\rho_1\sqrt{T}} + \frac{D^2 AE}{L\rho_1\rho_2\sqrt{T}} + \frac{C\sqrt{T}}{L\rho_1}\right)} = D\sqrt{\frac{D^2}{\rho_1 T}\left(B + \frac{AE}{\rho_2}\right) + \frac{C}{\rho_1}},$$

which completes the proof. □

Theorem B.13 (Any-time Convergence under nonconvexity and Stochastic Setting)**.** *Let f be an objective function that satisfies L-smoothness Assumption 1. Let $\{x^t\}_{t=0}^{+\infty}$ be a sequence generated by Algorithm BSFW where the stochastic estimator m^t and auxiliary sequence $\{\sigma_t\}$ satisfy Assumption 3 with parameters $\rho_1, \rho_2 \in]0, 1]$ and constants $A, B, C, E \geq 0$. Let the step decay be chosen as for all t, $\eta_t = \frac{1}{\sqrt{t+1}}$. Then we have*

$$\mathbb{E}\left[\min_{0\leq t\leq T-1} \mathrm{Gap}(x^t)\right] \leq \frac{\mathbb{E}[r_0]}{2(\sqrt{T+1}-1)} + \frac{LD^2(1+\ln(T))}{4(\sqrt{T+1}-1)} + D\sqrt{\left(\frac{2BD^2(1+\ln(T))^2}{\rho_1(\sqrt{T+1}-1)^2} + \frac{D^2AE(1+\ln(T))^2}{\rho_1\rho_2(\sqrt{T+1}-1)^2} + \frac{C(\sqrt{T+1}+1)(1+\ln(T))}{\rho_1(\sqrt{T+1}-1)}\right)}.$$

If $C = 0$, the last term vanishes and we recover the standard $\mathcal{O}(\ln(t)/\sqrt{t})$ rate.

Proof. Multiply inequality (16) by M_1 (at iteration $t+1$), inequality (17) by M_2, and add the two to obtain

$$\begin{aligned} M_1\left\|m^{t+1} - \nabla f\left(x^{t+1}\right)\right\|^2 + M_2\mathbb{E}\left[\sigma_{t+1}^2\right] &\leq M_1(1-\rho_1)\left\|m^t - \nabla f\left(x^t\right)\right\|^2 + M_1 A\mathbb{E}\left[\sigma_t^2\right] \\ &\quad + M_1 B\eta_t^2 D^2 + M_1 C + M_2(1-\rho_2)\,\mathbb{E}\left[\sigma_t^2\right] + M_2 E\eta_t^2 D^2 \\ &= M_1(1-\rho_1)\left\|m^t - \nabla f\left(x^t\right)\right\|^2 + M_2\left(1 - \rho_2 + \frac{M_1 A}{M_2}\right)\mathbb{E}\left[\sigma_t^2\right] \\ &\quad + \eta_t^2 D^2(M_1 B + M_2 E) + M_1 C. \end{aligned} \tag{58}$$

From the definition of r_t, i.e.,$r_t = f(x^t) - f^\star + M_1\|m^t - \nabla f(x^t)\|^2 + M_2\sigma_t^2$, the expected Lyapunov difference satisfies

$$\begin{aligned} \mathbb{E}\left[r_t - r_{t+1}\right] =& \mathbb{E}\left[f\left(x^t\right) - f\left(x^{t+1}\right)\right] \\ &+ M_1\mathbb{E}\left[\left\|m^t - \nabla f\left(x^t\right)\right\|^2 - \left\|m^{t+1} - \nabla f\left(x^{t+1}\right)\right\|^2\right] \\ &+ M_2\mathbb{E}\left[\sigma_t^2 - \sigma_{t+1}^2\right]. \end{aligned} \tag{59}$$

Rearranging (59) gives

$$\begin{aligned} \mathbb{E}\left[f\left(x^t\right) - f\left(x^{t+1}\right)\right] =& \mathbb{E}\left[r_t - r_{t+1}\right] - M_1\mathbb{E}\left[\left\|m^t - \nabla f\left(x^t\right)\right\|^2 - \left\|m^{t+1} - \nabla f\left(x^{t+1}\right)\right\|^2\right] \\ &- M_2\mathbb{E}\left[\sigma_t^2 - \sigma_{t+1}^2\right]. \end{aligned} \tag{60}$$

Now, using (16)-(17), we obtain lower bounds on the error decrements

$$\mathbb{E}\left[\|m^t - \nabla f(x^t)\|^2 - \|m^{t+1} - \nabla f(x^{t+1})\|^2\right] \geq \rho_1\,\mathbb{E}\left[\|m^t - \nabla f(x^t)\|^2\right] - A\,\mathbb{E}[\sigma_t^2] - B\eta_t^2 D^2 - C, \tag{61}$$
$$\mathbb{E}\left[\sigma_t^2 - \sigma_{t+1}^2\right] \geq \rho_2\,\mathbb{E}[\sigma_t^2] - E\eta_t^2 D^2. \tag{62}$$

Substituting (61)–(62) into (60) yields

$$\begin{aligned} \mathbb{E}\left[f(x^t) - f(x^{t+1})\right] \leq& \mathbb{E}[r_t - r_{t+1}] - \rho_1 M_1\,\mathbb{E}\left[\|m^t - \nabla f(x^t)\|^2\right] - (\rho_2 M_2 - M_1 A)\,\mathbb{E}[\sigma_t^2] \\ &+ \eta_t^2 D^2(M_1 B + M_2 E) + M_1 C. \end{aligned} \tag{63}$$

Plugging (63) into Lemma B.11, and grouping terms, we obtain

$$\begin{aligned}\eta_t\,\mathbb{E}\big[\mathrm{Gap}(x^t)\big] \le \mathbb{E}\Big[f(x^t) - f(x^\star) + \left(1-\rho_1+\frac{\alpha}{M_1 L}\right) M_1\|m^t - \nabla f(x^t)\|^2 + \left(1-\rho_2+\frac{M_1 A}{M_2}\right) M_2\sigma_t^2\Big] \\ - \mathbb{E}\big[f(x^{t+1}) - f(x^\star) + M_1\|m^{t+1} - \nabla f(x^{t+1})\|^2 + M_2\sigma_{t+1}^2\big] \\ + D^2\eta_t^2\left(\frac{L}{2}+\frac{L}{\alpha}+M_1 B + M_2 E\right) + M_1 C.\end{aligned}$$

Now choose $M_1 := \frac{\alpha}{L\rho_1}$ and $M_2 := \frac{M_1 A}{\rho_2} = \frac{\alpha A}{L\rho_1\rho_2}$, for $\alpha > 0$. So that $1-\rho_1+\frac{\alpha}{M_1 L} = 1$ and $1-\rho_2+\frac{M_1 A}{M_2} = 1$. Then the first expectation in (64) equals $\mathbb{E}[r_t]$, and the second equals $\mathbb{E}[r_{t+1}]$, giving

$$\eta_t\,\mathbb{E}[\mathrm{Gap}(x^t)] \le \mathbb{E}[r_t - r_{t+1}] + D^2\eta_t^2\left(\frac{L}{2}+\frac{L}{\alpha}+M_1 B + M_2 E\right) + M_1 C. \tag{64}$$

Summing (64) over $t = 0, \cdots, T-1$ telescopes

$$\sum_{t=0}^{T-1}\eta_t\,\mathbb{E}[\mathrm{Gap}(x^t)] \le \mathbb{E}[r_0 - r_T] + D^2\sum_{t=0}^{T-1}\eta_t^2\left(\frac{L}{2}+\frac{L}{\alpha}+M_1 B + M_2 E\right) + T M_1 C. \tag{65}$$

Since $r_T \ge 0$, we have

$$\sum_{t=0}^{T-1}(\eta_t\,\mathbb{E}[\mathrm{Gap}(x^t)]) \le \mathbb{E}[r_0] + D^2\left(\frac{L}{2}+\frac{L}{\alpha}+M_1 B + M_2 E\right)\sum_{t=0}^{T-1}\eta_t^2 + T M_1 C. \tag{66}$$

$$\begin{aligned}\mathbb{E}\left[\min_{0\le t\le T-1}\mathrm{Gap}(x^t)\right]\left(\sum_{t=0}^{T-1}\eta_t\right) &\le \sum_{t=0}^{T-1}(\eta_t\,\mathbb{E}[\mathrm{Gap}(x^t)]) \\ &\le \mathbb{E}[r_0] + D^2\left(\frac{L}{2}+\frac{L}{\alpha}+M_1 B + M_2 E\right)\sum_{t=0}^{T-1}\eta_t^2 + T M_1 C.\end{aligned} \tag{67}$$

Using the step decay $\eta_t = \frac{1}{\sqrt{t+1}}$, we have by the integration test,

$$\sum_{t=0}^{T-1}\tfrac{1}{\sqrt{t+1}} \ge 2(\sqrt{T+1}-1) \quad \text{and} \quad \sum_{t=0}^{T-1}(\tfrac{1}{\sqrt{t+1}})^2 \le 1+\ln(T).$$

Using these results in the expression, we get

$$\begin{aligned}&\mathbb{E}\left[\min_{0\le t\le T-1}\mathrm{Gap}(x^t)\right](2\sqrt{T+1}-1) \le \mathbb{E}[r_0] + D^2\left(\frac{L}{2}+\frac{L}{\alpha}+M_1 B + M_2 E\right)(1+\ln(T)) + T M_1 C, \\ &\implies \mathbb{E}\left[\min_{0\le t\le T-1}\mathrm{Gap}(x^t)\right] \le \tfrac{\mathbb{E}[r_0]}{2(\sqrt{T+1}-1)} + \tfrac{D^2(1+\ln(T))}{2(\sqrt{T+1}-1)}\left(\frac{L}{2}+\frac{L}{\alpha}+M_1 B + M_2 E\right) + \tfrac{T M_1 C}{2(\sqrt{T+1}-1)}.\end{aligned}$$

Using the constants $M_1 = \frac{\alpha}{L\rho_1}$ and $M_2 = \frac{\alpha A}{L\rho_1\rho_2}$, we get

$$\mathbb{E}\left[\min_{0\le t\le T-1}\mathrm{Gap}(x^t)\right] \le \tfrac{\mathbb{E}[r_0]}{2(\sqrt{T+1}-1)} + \tfrac{D^2(1+\ln(T))}{2(\sqrt{T+1}-1)}\left(\tfrac{L}{2}+\tfrac{L}{\alpha}+\tfrac{\alpha B}{L\rho_1}+\tfrac{\alpha A E}{L\rho_1\rho_2}\right) + \tfrac{T M_1 C}{2(\sqrt{T+1}-1)}.$$

Multiplying $\sqrt{T+1}+1$ in the numerator and denominator for the term containing C, we get

$$\begin{aligned}&\mathbb{E}\left[\min_{0\le t\le T-1}\mathrm{Gap}(x^t)\right] \le \tfrac{\mathbb{E}[r_0]}{2(\sqrt{T+1}-1)} + \tfrac{L D^2(1+\ln(T))}{4(\sqrt{T+1}-1)} + \tfrac{L D^2(1+\ln(T))}{2\alpha(\sqrt{T+1}-1)} \\ &\quad + \tfrac{B D^2\alpha(1+\ln(T))}{L\rho_1(\sqrt{T+1}-1)} + \tfrac{\alpha D^2 A E(1+\ln(T))}{2L\rho_1\rho_2(\sqrt{T+1}-1)} + \tfrac{\alpha C(\sqrt{T+1}+1)}{2L\rho_1} \\ &\le \tfrac{\mathbb{E}[r_0]}{2(\sqrt{T+1}-1)} + \tfrac{L D^2(1+\ln(T))}{4(\sqrt{T+1}-1)} + \tfrac{1}{\alpha}\left(\tfrac{L D^2(1+\ln(T))}{2(\sqrt{T+1}-1)}\right) + \alpha\left(\tfrac{B D^2(1+\ln(T))}{L\rho_1(\sqrt{T+1}-1)} + \tfrac{D^2 A E(1+\ln(T)))}{2L\rho_1\rho_2(\sqrt{T+1}-1)} + \tfrac{C(\sqrt{T+1}+1)}{2L\rho_1}\right).\end{aligned}$$

This has the form $g(\alpha) = u\alpha + \frac{v}{\alpha}$, with

$$u = \left(\tfrac{B D^2(1+\ln(T))}{L\rho_1(\sqrt{T+1}-1)} + \tfrac{D^2 A E(1+\ln(T))}{2L\rho_1\rho_2(\sqrt{T+1}-1)} + \tfrac{C(\sqrt{T+1}+1)}{2L\rho_1}\right), \quad v = \left(\tfrac{L D^2(1+\ln(T))}{2(\sqrt{T+1}-1)}\right).$$

Thus it reaches its minimum at $\alpha^* = \sqrt{\frac{v}{u}}$, with $g(\alpha^*) = 2\sqrt{uv}$. Substituting these values in the equation gives us

$$\begin{aligned} g(\alpha^*) &= 2\sqrt{\left(\frac{LD^2(1+\ln(T))}{2(\sqrt{T+1}-1)}\right)\left(\frac{BD^2(1+\ln(T))}{L\rho_1(\sqrt{T+1}-1)} + \frac{D^2AE(1+\ln(T))}{2L\rho_1\rho_2(\sqrt{T+1}-1)} + \frac{C(\sqrt{T+1}+1)}{2L\rho_1}\right)} \\ &= D\sqrt{\left(\frac{2(1+\ln(T))}{(\sqrt{T+1}-1)}\right)\left(\frac{BD^2(1+\ln(T))}{\rho_1(\sqrt{T+1}-1)} + \frac{D^2AE(1+\ln(T))}{2\rho_1\rho_2(\sqrt{T+1}-1)} + \frac{C(\sqrt{T+1}+1)}{2\rho_1}\right)} \\ &= D\sqrt{\left(\frac{2BD^2(1+\ln(T))^2}{\rho_1(\sqrt{T+1}-1)^2} + \frac{D^2AE(1+\ln(T))^2}{\rho_1\rho_2(\sqrt{T+1}-1)^2} + \frac{C(\sqrt{T+1}+1)(1+\ln(T))}{\rho_1(\sqrt{T+1}-1)}\right)}. \end{aligned}$$

Substituting this back into the main expression gives us

$$\begin{aligned} \mathbb{E}\left[\min_{0\le t\le T-1} \mathrm{Gap}(x^t)\right] \le{}& \frac{\mathbb{E}[r_0]}{2(\sqrt{T+1}-1)} + \frac{LD^2(1+\ln(T))}{4(\sqrt{T+1}-1)} \\ &+ D\sqrt{\left(\frac{2BD^2(1+\ln(T))^2}{\rho_1(\sqrt{T+1}-1)^2} + \frac{D^2AE(1+\ln(T))^2}{\rho_1\rho_2(\sqrt{T+1}-1)^2} + \frac{C(\sqrt{T+1}+1)(1+\ln(T))}{\rho_1(\sqrt{T+1}-1)}\right)}. \end{aligned}$$

Hence from the expression, when $C = 0$, we recover the rate $\mathcal{O}(\ln(t)/\sqrt{t})$. □

C Estimators

The estimators used in the experiments section 5 are taken from [29], [30] and [27]. We use a set of stochastic and coordinate methods as gradient estimators m^t of the gradient $\nabla f(x^t)$ at each iteration t. In Algorithm BSFW, note the usage of m^{init}, whose purpose is only to assign x^0, while that of $\{\Phi_t\}_{t=0}^{T-1}$ which is to denote the gradient estimator used. For these estimators, the proofs for the parameters satisfying Assumption 3 are provided in the subsections C.3.8 to C.3.7. A summary of the parameters is provided in Table 2.

Table 2: Summary of parameter values for various gradient estimators satisfying Assumption 3.

Gradient Estimator	ρ_1	ρ_2	A	B	C	E
SAG [30]	$\frac{b_s}{2m}$	1	0	$(1-\frac{b_s}{m})(1+\frac{2m}{b_s})L^2$	0	0
L-SVRG [16]	1	$\frac{p}{2}$	$\frac{L^2}{b_s} - \frac{pL^2}{2b_s}$	$\frac{8L^2}{pb_s}$	0	$\frac{8}{p}$
SAGA [33]	1	$\frac{b_s}{2m}$	$\frac{1}{b_s} + \frac{1}{2m}$	$\frac{L^2}{b_s m}(1+\frac{2m}{b_s})$	0	$\frac{2m}{b_s}L^2$
SEGA [10]	1	$\frac{1}{2n}$	n	n^2L^2	0	$3L^2n$
JAGUAR [29]	$\frac{1}{2n}$	1	0	$3nL^2$	0	0
ZOJA [29]	$\frac{1}{4n}$	1	0	$3nL^2$	$2nL^2\tau^2$	0
SARAH [1]	p	1	0	$\frac{1-p}{b_s}L^2$	0	0
Heavy Ball [27]	$\frac{\tilde{\rho}_T}{2}$	$1-(\frac{T+7}{T+8})^{\frac{4}{3}}$	1	$\frac{2L^2}{\tilde{\rho}_T}$	0	0

The name of each estimator links to its description and proof in the appendix. For the stochastic estimators, the constant b_s refers to the stochastic batch size (number of indices) sampled per-iteration. The parameter p is the probability of computing a deterministic gradient used in the algorithms L-SVRG and SARAH, as explained in Appendices C.3.1 and C.3.2. The parameter τ is the zeroth-order approximation parameter in ZOJA, explained in Appendix C.3.6. The parameter $\tilde{\rho}_t$ in Heavy Ball is the momentum, explained in Appendix C.3.7. It is worth noting that, **for all the estimators except ZOJA,** $C = 0$ **which is significant for the convergence analysis**. For ZOJA, C diminishes with respect to τ. The values for the sequence $\{\sigma_t\}$ for each estimator are given under each estimator's corresponding section. The parameter B of Heavy Ball is dependent on the horizon T, which affects the convergence analysis. Hence, additional analysis using alternate step decays for ρ-quasar-convex and nonconvex functions are explained in subsection C.3.7.

C.1 Stochastic Methods

The stochastic estimators randomly sample a batch of size b_s from the dataset $\{a_1, a_2, \cdots, a_m\}$ at each iteration t based on $\xi_t \sim \mathcal{P}$, and the corresponding gradient oracle output is denoted by $\nabla f(x^t, \xi_t)$.

C.2 Coordinate Methods

The coordinate estimators randomly sample an index j from the set $\{1, 2, \cdots n\}$ based on $\xi_t \sim \mathcal{P}$. The estimators then use the partial derivative of the function f, denoted by $\nabla f_j(x^t)$ as the estimate of the gradient $\nabla f(x^t)$. We refer to e_i as the ith standard basis vector in subsections C.3.4, C.3.5 and C.3.6 describing the coordinate estimators SEGA, JAGUAR, and ZOJA respectively.

C.3 Properties of Estimators

In this section, we discuss the convergence properties of the estimators used in the experiments in section 5. We prove that if the estimators provide parameters that satisfy Assumption 3, we have fixed-horizon convergence rates in Theorems B.9 and B.12, and any-time convergence rates in Theorems B.10 and B.13. To discuss the analysis, we first provide Lemma C.1 which will permit us to prove the parameters necessary to satisfy Assumption 3.

Lemma C.1 (BSFW Iteration Bound)**.** *Let $\{x^t\}_{t=0}^{+\infty}$ be a sequence generated by Algorithm BSFW, using a step decay $\{\eta_t\}_{t=0}^{+\infty}$. Then we have that for all $t \geq 1$,*

$$\|x^t - x^{t-1}\| \leq \eta_{t-1} D.$$

Proof. From Algorithm BSFW, we have for any $t \geq 1$, if $\gamma_{t-1} = 1$, then

$$\|x^t - x^{t-1}\| = \eta_{t-1}\|d^{t-1}\| = \eta_{t-1}\|s^{t-1} - x^{t-1}\| \leq \eta_{t-1} D.$$

Suppose instead, $\gamma_{t-1} < 1$, then we have

$$\|x^t - x^{t-1}\| = \gamma_{t-1}\|d^{t-1}\| = \eta_{t-1}\frac{\|s^{t-1} - x^{t-1}\|}{\|d^{t-1}\|}\|d^{t-1}\| = \eta_{t-1}\|s^{t-1} - x^{t-1}\| \leq \eta_{t-1} D.$$

Hence proven. □

C.3.1 L-SVRG [16]

Description For the L-SVRG estimator [16], we use an additional variable w^t. We initiate it by

$$w^0 = x^0, \quad m^0 = \nabla f(x^0).$$

For every iteration $t > 0$, we sample a batch $S_t \subset \{1, 2, \cdots, m\}$ of size b_s uniformly at random. b_s is a pre-defined constant per-iteration sample size parameter. Then, we make the update

$$w^t = \begin{cases} x^{t-1}, & \text{with probability } p \\ w^{t-1}, & \text{with probability } 1 - p \end{cases}$$
$$m^t = \frac{1}{b_s}\sum_{i \in S_t}\left(\nabla f_i(x^t) - \nabla f_i(w^t)\right) + \nabla f(w^t).$$

where p is a defined probability parameter. The parameters satisfying Assumption 3 are provided in Lemma C.2.

Verification of Assumption 3

Lemma C.2. *(Parameters for L-SVRG) Let $\{x^t\}_{t=0}^{T}$ be a sequence generated by Algorithm BSFW using a step decay $\{\eta_t\}_{t=0}^{T-1}$, where the gradient estimator $\{\Phi_t\}_{t=0}^{T-1}$ is L-SVRG defined by [16]. Then we have the following parameters used in Assumption 3.*

$$\rho_1 = 1, \quad \rho_2 = \frac{p}{2}, \quad A = \frac{L^2}{b_s} - \frac{L^2 p}{2b_s}, \quad B = \frac{8L^2}{pb_s}, \quad C = 0, \quad E = \frac{8}{p}, \quad \sigma_t^2 = \|x^t - w^t\|^2.$$

where b_s is the stochastic batch size sampled per-iteration, and p is the probability parameter.

Proof. This proof follows exactly from [29]. According to Lemma 3 from [22], for any t such that $1 \leq t \leq T-1$, we get an estimation:

$$\mathbb{E}_{t-1}[\|m^t\|^2] \leq \frac{L^2}{b_s}\mathbb{E}_{t-1}[\|x^t - w^t\|^2] + \mathbb{E}_{t-1}[\|\nabla f(x^t)\|^2].$$

Since m^t is an unbiased gradient estimator, the previous inequality turns to:

$$\mathbb{E}_{t-1}[\|\nabla f(x^t) - m^t\|^2] \leq \frac{L^2}{b_s}\mathbb{E}_{t-1}[\|x^t - w^t\|^2]$$

Suppose at iteration t, we have $\gamma_{t-1} < 1$ (case I). Then we have $x^t = x^{t-1} + \gamma_{t-1}d^{t-1}$. Hence,

$$\begin{aligned}
\mathbb{E}_{t-1}[\|x^t - w^t\|^2] &= p\mathbb{E}_{t-1}\left[\|x^t - x^{t-1}\|^2\right] + (1-p)\mathbb{E}_{t-1}[\|x^t - w^{t-1}\|^2] \\
&= p\gamma_{t-1}^2\mathbb{E}_{t-1}\left[\|d^{t-1}\|^2\right] + (1-p)\mathbb{E}_{t-1}[\|x^{t-1} + \gamma_{t-1}d^{t-1} - w^{t-1}\|^2] \\
&= \gamma_{t-1}^2\mathbb{E}_{t-1}[\|d^{t-1}\|^2] + (1-p)\|x^{t-1} - w^{t-1}\|^2 \\
&\quad + 2\gamma_{t-1}(1-p)\mathbb{E}_{t-1}[\langle x^{t-1} - w^{t-1}, d^{t-1}\rangle] \\
&= \gamma_{t-1}^2\|d^{t-1}\|^2 + (1-p)\|x^{t-1} - w^{t-1}\|^2 \\
&\quad + 2(1-p)\langle x^{t-1} - w^{t-1}, \gamma_{t-1}d^{t-1}\rangle.
\end{aligned}$$

According to Young's inequality for a $\beta > 0$, we have

$$\langle x^{t-1} - w^{t-1}, \gamma_{t-1}d^{t-1}\rangle \leq \beta\|x^{t-1} - w^{t-1}\|^2 + \frac{1}{\beta}\gamma_{t-1}^2\|d^{t-1}\|^2.$$

Hence,

$$\begin{aligned}
\mathbb{E}_{t-1}[\|x^t - w^t\|^2] \leq \gamma_{t-1}^2\|d^{t-1}\|^2 + (1-p)\|x^{t-1} - w^{t-1}\|^2 + 2(1-p)\beta\|x^{t-1} - w^{t-1}\|^2 \\
+ 2\frac{1-p}{\beta}\gamma_{t-1}^2\|d^{t-1}\|^2.
\end{aligned}$$

From Lemma C.1, $\gamma_{t-1}\|d^{t-1}\| \leq \eta_{t-1}D$, and so we have

$$\mathbb{E}_{t-1}[\|x^t - w^t\|^2] \leq \left(1 + \frac{2(1-p)}{\beta}\right)\eta_{t-1}^2D^2 + (1-p)(1+2\beta)\|x^{t-1} - w^{t-1}\|^2.$$

Suppose instead $\gamma_{t-1} = 1$ (case II). Then we instead have $x^t = x^{t-1} + \eta_{t-1}(s^{t-1} - x^{t-1})$. Hence, we have

$$\begin{aligned}
\mathbb{E}_{t-1}[\|x^t - w^t\|^2] &= p\mathbb{E}_{t-1}\left[\|x^t - x^{t-1}\|^2\right] + (1-p)\mathbb{E}_{t-1}[\|x^t - w^{t-1}\|^2] \\
&= p\eta_{t-1}^2\mathbb{E}_{t-1}\left[\|s^{t-1} - x^{t-1}\|^2\right] + (1-p)\mathbb{E}_{t-1}[\|x^{t-1} + \eta_{t-1}(s^{t-1} - x^{t-1}) - w^{t-1}\|^2] \\
&= \eta_{t-1}^2\mathbb{E}_{t-1}[\|s^{t-1} - x^{t-1}\|^2] + (1-p)\|x^{t-1} - w^{t-1}\|^2 \\
&\quad + 2\eta_{t-1}(1-p)\mathbb{E}_{t-1}[\langle x^{t-1} - w^{t-1}, s^{t-1} - x^{t-1}\rangle] \\
&= \eta_{t-1}^2\|s^{t-1} - x^{t-1}\|^2 + (1-p)\|x^{t-1} - w^{t-1}\|^2 \\
&\quad + 2(1-p)\langle x^{t-1} - w^{t-1}, \eta_{t-1}(s^{t-1} - x^{t-1})\rangle.
\end{aligned}$$

Again by using Young's inequality for a $\beta > 0$,

$$\begin{aligned}
\mathbb{E}_{t-1}[\|x^t - w^t\|^2] &\leq \eta_{t-1}^2\|s^{t-1} - x^{t-1}\|^2 + (1-p)\|x^{t-1} - w^{t-1}\|^2 + 2(1-p)\beta\|x^{t-1} - w^{t-1}\|^2 \\
&\quad + 2\frac{1-p}{\beta}\eta_{t-1}^2\|s^{t-1} - x^{t-1}\|^2 \\
&\leq \left(1 + \frac{2(1-p)}{\beta}\right)\eta_{t-1}^2D^2 + (1-p)(1+2\beta)\|x^{t-1} - w^{t-1}\|^2.
\end{aligned}$$

Giving us the same result in both cases I and II. Finally, choosing $\beta = \frac{p}{4}$, we get $(1-p)(1+2\beta) \leq \left(1 - \frac{p}{2}\right)$. and

$$\mathbb{E}_{t-1}[\|x^t - w^t\|^2] \leq \frac{8}{p}\eta_{t-1}^2 D^2 + \left(1 - \frac{p}{2}\right)\|x^{t-1} - w^{t-1}\|^2,$$
$$\mathbb{E}_{t-1}[\|\nabla f(x^t) - m^t\|^2] \leq \frac{8L^2}{pb_s}\eta_{t-1}^2 D^2 + \frac{L^2}{b_s}\left(1 - \frac{p}{2}\right)\|x^{t-1} - w^{t-1}\|^2.$$

□

C.3.2 SARAH [1]

Description For the estimator SARAH [1], we need an additional assumption that the objective function f in (P) can be represented as a finite sum, i.e., $f(x) = \sum_{i=1}^m f_i(x)$. We start by setting $m^0 = \nabla f(x^0)$. Then, for each iteration $t > 0$, we sample a batch $S_t \subset \{1, 2, \cdots, m\}$ of size b_s uniformly at random. Specifically we have

$$m^0 = \nabla f(x^0),$$
$$m^t = \begin{cases} \nabla f(x^t), & \text{with probability } p \\ m^{t-1} + \frac{1}{b_s}\left(\sum_{i \in S_t} \nabla f_i(x^t) - \nabla f_i(x^{t-1})\right), & \text{with probability } 1-p \end{cases}$$

where p is a defined probability parameter, and b_s is a defined per-iteration batch size. The parameters satisfying Assumption 3 are provided in Lemma C.3.

Verification of Assumption 3

Lemma C.3. *(Parameters for SARAH) Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW using a step decay $\{\eta_t\}_{t=0}^{T-1}$, where the gradient estimator $\{\Phi_t\}_{t=0}^{T-1}$ is SARAH defined by [1]. Then we have the following parameters used in Assumption 3.*

$$\rho_1 = p, \quad \rho_2 = 1, \quad A = 0, \quad B = \frac{1-p}{b_s}L^2, \quad C = 0, \quad E = 0, \quad \sigma_t^2 = 0.$$

where b_s refers to the stochastic batch size sampled per-iteration, and p is the probability parameter.

Proof. This proof follows exactly from [29]. According to Lemma 3 from [21], for a L-Lipschitz smooth function f, using Lemma C.1, we bound the difference by

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|\nabla f(x^t) - m^t\|^2\right] = \mathbb{E}_{t-1}[\|\Delta^t\|^2] &\leq (1-p)\|\Delta^{t-1}\|^2 + \frac{1-p}{b}L^2\|x^t - x^{t-1}\|^2 \\ &\leq (1-p)\|\Delta^{t-1}\|^2 + \frac{1-p}{b_s}L^2\eta_{t-1}^2 D^2.\end{aligned}$$

Hence giving us the required coefficients. □

C.3.3 SAGA [33]

Description To use the SAGA gradient estimator [33], we need an additional assumption that the objective function f in (P) can be represented as a finite sum, i.e., $f(x) = \sum_{i=1}^m f_i(x)$. For SAGA, we use an additional variable y^t as implemented by [29]. We initiate SAGA by setting

$$y_i^0 = \nabla f_i(x^0) \text{ for all } i \in \{1, 2, \cdots, m\},$$
$$m^0 = \nabla f(x^0).$$

For every iteration $t > 0$, we sample a batch $S_t \subset \{1, 2, \cdots, m\}$ of size b_s uniformly at random. b_s is a pre-defined per-iteration sample size parameter. After sampling S_t, we make the gradient estimate m^t by setting

$$m^t = \frac{1}{b_s}\sum_{i \in S_t}\left(\nabla f_i(x^t) - y_i^{t-1}\right) + \frac{1}{m}\sum_{i=1}^m y_i^{t-1}.$$

Then we have for every $i \in \{1, 2, \cdots, m\}$,

$$y_i^t = \begin{cases} \nabla f_i(x^t), & i \in S_t \\ y_i^{t-1}. & i \notin S_t \end{cases}$$

The parameters satisfying Assumption 3 are provided in Lemma C.4.

Verification of Assumption 3

Lemma C.4. *(Parameters for SAGA) Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW, using a step decay $\{\eta_t\}_{t=0}^{T-1}$, where the gradient estimator $\{\Phi_t\}_{t=0}^{T-1}$ is SAGA defined by [33]. Suppose the objective function f can be represented as $f(x) = \sum_{i=1}^m f_i(x)$. Then we have the following parameters used in Assumption 3.*

$$\rho_1 = 1, \ \rho_2 = \frac{b_s}{2m}, \ A = \frac{1}{b_s} + \frac{1}{2m}, \ B = \frac{L^2}{b_s m}\left(1 + \frac{2m}{b_s}\right), \ C = 0, \ E = \frac{2m}{b_s}L^2, \ \sigma_t^2 = \frac{1}{m}\sum_{j=1}^m \|\nabla f_j(x^t) - y_j^t\|^2.$$

where b_s is defined as the stochastic batch size sampled per-iteration.

Proof. This proof follows exactly from [29]. The difference between estimator and exact gradient is bounded by:

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|m^t - \nabla f(x^t)\|^2\right] &= \mathbb{E}_{t-1}\left[\left\|\frac{1}{b_s}\sum_{i\in S_t}\left[\nabla f_i(x^t) - y_i^{t-1}\right] + \frac{1}{m}\sum_{j=1}^m y_j^{t-1} - \nabla f(x^t)\right\|^2\right] \\ &= \mathbb{E}_{t-1}\left[\left\|\frac{1}{b_s}\left(\sum_{i\in S_t}\left[\nabla f_i(x^t) - y_i^{t-1}\right] - \left(\frac{1}{m}\sum_{j=1}^m\left[\nabla f_j(x^t) - y_j^{t-1}\right]\right)\right)\right\|^2\right].\end{aligned}$$

By using Lemma B.2 from [29]

$$\mathbb{E}_{t-1}\left[\|m^t - \nabla f(x^t)\|^2\right] \le \frac{1}{b_s m}\sum_{j=1}^m \left\|\nabla f_j(x^t) - y_j^{t-1} - \left(\frac{1}{m}\sum_{i=1}^m\left[\nabla f_i(x^t) - y_i^{t-1}\right]\right)\right\|^2.$$

Now, $\frac{1}{m}\sum_{i=1}^m$ can be described as an expected value and $\mathbb{E}[\|x - \mathbb{E}[x]\|^2] \le \mathbb{E}[\|x\|^2]$. Furthermore, using Young's inequality with $\alpha > 0$,

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|m^t - \nabla f(x^t)\|^2\right] &\le \frac{1}{b_s m}\sum_{j=1}^m \left\|\nabla f_j(x^t) - y_j^{t-1}\right\|^2 \\ &\le \frac{1}{b_s m}(1+\alpha)\sum_{j=1}^m \|\nabla f_j(x^t) - \nabla f_j(x^{t-1})\|^2 + \frac{1}{b_s m}\left(1 + \frac{1}{\alpha}\right)\sum_{j=1}^m \|\nabla f_j(x^{t-1}) - y_j^{t-1}\|^2.\end{aligned}$$

Using L-smoothness of f, Lemma C.1, and by setting $\sigma_{t-1}^2 = \frac{1}{m}\sum_{j=1}^m \|\nabla f_j(x^{t-1}) - y_j^{t-1}\|^2$,

$$\mathbb{E}_{t-1}\left[\|m^t - \nabla f(x^t)\|^2\right] \le \frac{1}{b_s m}(1+\alpha)L^2\eta_{t-1}^2 D^2 + \frac{1}{b_s}\left(1 + \frac{1}{\alpha}\right)\sigma_{t-1}^2.$$

We can put $\alpha = \frac{2m}{b_s}$ to obtain the needed estimates, i.e.

$$\mathbb{E}_{t-1}\left[\|m^t - \nabla f(x^t)\|^2\right] \le \frac{L^2}{b_s m}\left(1 + \frac{2m}{b_s}\right)\eta_{t-1}^2 D^2 + \frac{1}{b_s}\left(1 + \frac{b_s}{2m}\right)\sigma_{t-1}^2.$$

To bound the term σ_{t-1}^2,

$$\mathbb{E}_{t-1}[\sigma_t^2] = \mathbb{E}_{t-1}\left[\frac{1}{m}\sum_{j=1}^{m}\|\nabla f_j(x^t) - y_j^t\|^2\right] = \left(1-\frac{b_s}{m}\right)\frac{1}{m}\sum_{j=1}^{m}\|\nabla f_j(x^t) - y_j^{t-1}\|^2$$
$$= \left(1-\frac{b_s}{m}\right)\frac{1}{m}\sum_{j=1}^{m}\|\nabla f_j(x^t) - \nabla f_j(x^{t-1}) + \nabla f_j(x^{t-1}) - y_j^{t-1}\|^2.$$

By using Young's Inequality again with $\beta > 0$,

$$\mathbb{E}_{t-1}[\sigma_t^2] \le \left(1-\frac{b_s}{m}\right)(1+\beta)\frac{1}{m}\sum_{j=1}^{m}\|\nabla f_j(x^{t-1}) - y_j^{t-1}\|^2 + \left(1-\frac{b_s}{m}\right)\left(1+\frac{1}{\beta}\right)L^2\|x^t - x^{t-1}\|^2.$$

With $\beta = \frac{b_s}{2m}$, and by using Lemma C.1, we have:

$$\mathbb{E}_{t-1}[\sigma_t^2] \le \left(1-\frac{b_s}{2m}\right)\sigma_{t-1}^2 + \frac{2m}{b_s}L^2\eta_{t-1}^2D^2.$$

□

C.3.4 SEGA [10]

Description For SEGA [10], we require an additional variable h^t as implemented by [29]. We initialize by setting both h^0 and m^0 by the full gradient at $t = 0$, specifically, $m^0 = \nabla f(x^0)$ and $h^0 = \nabla f(x^0)$. Then at every iteration $t > 0$, we randomly sample $i_t \in \{1, 2, \cdots, n\}$ and approximate the gradient by

$$m^t = ne_{i_t}\left(\nabla_{i_t} f(x^t) - h_{i_t}^{t-1}\right) + h^{t-1}.$$

We also update h^t by setting $h^t = h^{t-1} + e_{i_t}\left(\nabla_{i_t} f(x^t) - h^{t-1}\right)$. The parameters satisfying Assumption 3 are provided in Lemma C.5.

Verification of Assumption 3

Lemma C.5. *(Parameters for SEGA) Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW using a step decay $\{\eta_t\}_{t=0}^{T-1}$, where the gradient estimator $\{\Phi_t\}_{t=0}^{T-1}$ is SEGA defined by [10]. Then we have the following parameters used in Assumption 3.*

$$\rho_1 = 1, \quad \rho_2 = \frac{1}{2n}, \quad A = n, \quad B = n^2L^2, \quad C = 0, \quad E = 3L^2n, \quad \sigma_t^2 = \|h^t - \nabla f(x^t)\|^2.$$

Proof. This proof follows exactly from [29]. I refers to the identity matrix of dimensions $n \times n$. We first bound the difference between estimator and exact gradient:

$$\begin{aligned}
\mathbb{E}_{t-1}\left[\|m^t - \nabla f(x^t)\|^2\right] &= \mathbb{E}_{t-1}\left[\|ne_{i_t}e_{i_t}^\top(\nabla f(x^t) - h^{t-1}) + h^{t-1} - \nabla f(x^t)\|^2\right] \\
&= \mathbb{E}_{t-1}\left[\|(I - ne_{i_t}e_{i_t}^\top)(h^{t-1} - \nabla f(x^t))\|^2\right] \\
&= \mathbb{E}_{t-1}\left[(h^{t-1} - \nabla f(x^t))^\top(I - ne_{i_t}e_{i_t}^\top)^\top(I - ne_{i_t}e_{i_t}^\top)(h^{t-1} - \nabla f(x^t))\right] \\
&= (h^{t-1} - \nabla f(x^t))^\top\mathbb{E}_{t-1}\left[I - 2ne_{i_t}e_{i_t}^\top + n^2e_{i_t}e_{i_t}^\top\right](h^{t-1} - \nabla f(x^t)) \\
&= (h^t - \nabla f(x^t))^\top\left[I - 2\cdot I + n\cdot I\right](h^{t-1} - \nabla f(x^t)) \\
&= (n-1)\|h^{t-1} - \nabla f(x^t)\|^2 \\
&= (n-1)\|h^{t-1} - \nabla f(x^{t-1}) + \nabla f(x^{t-1}) - \nabla f(x^t)\|^2 \\
&= (n-1)\big(\|h^{t-1} - \nabla f(x^{t-1})\|^2 + \|\nabla f(x^{t-1}) - \nabla f(x^t)\|^2 \\
&\quad + 2\langle h^{t-1} - \nabla f(x^{t-1}), \nabla f(x^{t-1}) - \nabla f(x^t)\rangle\big).
\end{aligned}$$

By using Young's inequality with a parameter $\alpha > 0$ on the inner product $2\langle h^{t-1} - \nabla f(x^{t-1}), \nabla f(x^{t-1}) - \nabla f(x^t)\rangle$ and the L-Lipschitz smoothness of f, and Lemma C.1, we get the bound

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|m^t - \nabla f(x^t)\|^2\right] &\leq (n-1)(1+\alpha)\|h^{t-1} - \nabla f(x^{t-1})\|^2 + (n-1)\left(1+\frac{1}{\alpha}\right)L^2\|x^t - x^{t-1}\|^2 \\ &\leq (n-1)(1+\alpha)\|h^{t-1} - \nabla f(x^{t-1})\|^2 + (n-1)\left(1+\frac{1}{\alpha}\right)\eta_{t-1}^2 L^2 D^2.\end{aligned}$$

Then by using similar arguments as above (L-Lipschitz smoothness of f, Young's Inequality with $\beta > 0$ and Lemma C.1),

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|h^t - \nabla f(x^t)\|^2\right] &= \mathbb{E}_{t-1}\left[\|h^{t-1} + e_{i_t}e_{i_t}^\top(\nabla f(x^t) - h^{t-1}) - \nabla f(x^t)\|^2\right] \\ &= \mathbb{E}_t\left[\|(I - e_{i_t}e_{i_t}^\top)(h^{t-1} - \nabla f(x^t))\|^2\right] \\ &= \left(1-\frac{1}{n}\right)\|h^{t-1} - \nabla f(x^t)\|^2 \\ &= \left(1-\frac{1}{n}\right)\|h^{t-1} - \nabla f(x^{t-1}) + \nabla f(x^{t-1}) - \nabla f(x^t)\|^2 \\ &\leq \left(1-\frac{1}{n}\right)(1+\beta)\|h^{t-1} - \nabla f(x^{t-1})\|^2 + \left(1-\frac{1}{n}\right)\left(1+\frac{1}{\beta}\right)\eta_{t-1}^2 L^2 D^2.\end{aligned}$$

If $\beta = \frac{1}{2n}$ then $(1-\frac{1}{n})(1+\frac{1}{2n}) \leq 1 - \frac{1}{2n}$ and $(1-\frac{1}{n})(1+2n) \leq 2n$, then as $n \geq 1$:

$$\mathbb{E}_t\left[\|h^t - \nabla f(x^t)\|^2\right] \leq \left(1-\frac{1}{2n}\right)\|h^{t-1} - \nabla f(x^{t-1})\|^2 + 3nL^2\eta_{t-1}^2 D^2.$$

Setting $\alpha = \frac{1}{n}$, we have $(n-1)(1+\frac{1}{n}) \leq n$ and $(n-1)(n+1) \leq n^2$ and thus

$$\mathbb{E}_{t-1}[\|\Delta^t\|^2] = \mathbb{E}_{t-1}\left[\|m^t - \nabla f(x^t)\|^2\right] \leq n\|h^{t-1} - \nabla f(x^{t-1})\|^2 + n^2\eta_{t-1}^2 L^2 D^2.$$

□

C.3.5 JAGUAR [29]

Description For the JAGUAR estimator [29], we initiate m^0 by the full gradient at x^0, i.e., $m^0 = \nabla f(x^0)$. Then at every iteration $t > 0$, we randomly sample $i_t \in \{1, 2, \cdots, n\}$ and approximate the gradient by

$$m^t = e_{i_t}\left(\nabla_{i_t} f(x^{t-1}) - m_{i_t}^{t-1}\right) + m^{t-1}.$$

The parameters satisfying Assumption 3 are provided in Lemma C.6.

Verification of Assumption 3

Lemma C.6. *(Parameters for JAGUAR) Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW using a step decay $\{\eta_t\}_{t=0}^{T-1}$, where the gradient estimator $\{\Phi_t\}_{t=0}^{T-1}$ is JAGUAR defined by [29]. Then we have the following parameters used in Assumption 3.*

$$\rho_1 = \frac{1}{2n}, \quad \rho_2 = 1, \quad A = 0, \quad B = 3nL^2, \quad C = 0, \quad E = 0, \quad \sigma_t^2 = 0.$$

Proof. This proof follows exactly from [29]. By first bounding the difference between estimator and exact gradient:

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|m^t - \nabla f(x^t)\|^2\right] &= \mathbb{E}_{t-1}\left[\|e_{i_t}e_{i_t}^\top(\nabla f(x^{t-1}) - m^{t-1}) + m^{t-1} - \nabla f(x^t)\|^2\right] \\ &= \mathbb{E}_{t-1}\left[\|e_{i_t}e_{i_t}^\top(\nabla f(x^{t-1}) - m^{t-1}) + m^{t-1} - \nabla f(x^t) + \nabla f(x^{t-1}) - \nabla f(x^{t-1})\|^2\right] \\ &= \mathbb{E}_{t-1}\left[\|(I - e_{i_t}e_{i_t}^\top)(\nabla f(x^{t-1}) - m^{t-1}) + \nabla f(x^{t-1}) - \nabla f(x^t)\|^2\right].\end{aligned}$$

Using Young's Inequality with a parameter $\beta > 0$, L-Lipschitz smoothness property of f and Lemma C.1, we have

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|m^t - \nabla f(x^t)\|^2\right] &\leq (1+\beta)\mathbb{E}_{t-1}\left[\|(I - e_{i_t}e_{i_t}^\top)(m^{t-1} - \nabla f(x^{t-1}))\|^2\right] + \left(1+\frac{1}{\beta}\right)\eta_{t-1}^2 L^2 D^2 \\ &\leq (1+\beta)\left(1-\frac{1}{n}\right)\|m^{t-1} - \nabla f(x^{t-1})\|^2 + \left(1+\frac{1}{\beta}\right)\eta_{t-1}^2 L^2 D^2.\end{aligned}$$

By setting $\beta = \frac{1}{2n}$, we have $(1-\frac{1}{n})(1+\frac{1}{2n}) \leq 1 - \frac{1}{2n}$ and as $n \geq 1$, we get

$$\mathbb{E}_{t-1}\left[\|\Delta^t\|^2\right] \leq \left(1-\frac{1}{2n}\right)\|\Delta^{t-1}\|^2 + 3\eta_{t-1}^2 n L^2 D^2.$$

□

C.3.6 ZOJA [29]

Description For the ZOJA estimator introduced by [29], we initiate m^0 by the zero order approximation using a defined parameter τ across every coordinate. Specifically,

$$m^0 = \sum_{i=1}^{n}\left(\frac{f(x^0 + \tau e_i) - f(x^0)}{\tau}\right)e_i.$$

Next at every iteration $t > 0$, we randomly sample $i_t \in \{1, 2, \cdots, n\}$ and approximate the gradient by

$$\begin{aligned}\widetilde{\nabla}_{i_t} f(x^{t-1}) &= \frac{f(x^{t-1} + \tau e_{i_t}) - f(x^{t-1})}{\tau}, \\ m^t &= e_{i_t}\left(\widetilde{\nabla}_{i_t} f(x^{t-1}) - m_{i_t}^{t-1}\right) + m^{t-1}.\end{aligned}$$

The parameters satisfying Assumption 3 are given in Lemma C.7.

Verification of Assumption 3

Lemma C.7. *(Parameters for ZOJA) Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW using a step decay $\{\eta_t\}_{t=0}^{T-1}$, where the gradient estimator $\{\Phi_t\}_{t=0}^{T-1}$ is ZOJA defined by [29]. Then we have the following parameters used in Assumption 3.*

$$\rho_1 = \frac{1}{4n}, \quad \rho_2 = 1, \quad A = 0, \quad B = 3nL^2, \quad C = 2nL^2\tau^2, \quad E = 0, \quad \sigma_t^2 = 0.$$

where $\tau > 0$ is the zero-order approximation parameter.

Proof. This proof follows exactly from [29]. We bound the difference between estimator and exact gradient:

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|m^t - \nabla f(x^t)\|^2\right] = \mathbb{E}_{t-1}\left[\|\Delta^t\|^2\right] &= \mathbb{E}_{t-1}[\|e_{i_t}(\widetilde{\nabla}_{i_t} f(x^{t-1}) - m_{i_t}^{k-1}) + m^{t-1} - \nabla f(x^t)\|^2] \\ &= \mathbb{E}_{t-1}[\|e_{i_t}(\widetilde{\nabla}_{i_t} f(x^{t-1}) - m_{i_t}^{t-1}) + m^{t-1} - \nabla f(x^t) \\ &\quad + \nabla f(x^{t-1}) - \nabla f(x^{t-1})\|^2].\end{aligned}$$

Using Young's inequality with $\beta > 0$, L-Lipschitz smoothness of f and Lemma C.1, we get

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|\Delta^t\|^2\right] &\leq (1+\beta)\mathbb{E}_{t-1}\left[\|e_{i_t}(\widetilde{\nabla}_{i_t} f(x^{t-1}) - m_{i_t}^{t-1}) + m^{t-1} - \nabla f(x^{t-1})\|^2\right] + \left(1+\frac{1}{\beta}\right)\|\nabla f(x^t) - \nabla f(x^{t-1})\|^2 \\ &\leq (1+\beta)\mathbb{E}_{t-1}\left[\|e_{i_t}(\widetilde{\nabla}_{i_t} f(x^{t-1}) - m_{i_t}^{t-1}) + m^{t-1} - \nabla f(x^{t-1})\|^2\right] + \left(1+\frac{1}{\beta}\right)L^2\|x^t - x^{t-1}\|^2 \\ &\leq (1+\beta)\mathbb{E}_{t-1}\left[\|e_{i_t}(\widetilde{\nabla}_{i_t} f(x^{t-1}) - m_{i_t}^{t-1}) + m^{t-1} - \nabla f(x^{t-1})\|^2\right] + \left(1+\frac{1}{\beta}\right)\eta_{t-1}^2 L^2 D^2 \\ &\leq (1+\beta)\mathbb{E}_{t-1}\left[\|(I - e_{i_t}e_{i_t}^\top)(m^{t-1} - \nabla f(x^{t-1})) + e_{i_t}(\widetilde{\nabla}_{i_t} f(x^{t-1}) - \nabla_{i_t} f(x^{t-1}))\|^2\right] \\ &\quad + \left(1+\frac{1}{\beta}\right)\eta_{t-1}^2 L^2 D^2.\end{aligned}$$

Using Lemma B.1 from [29], for any index j such that $1 \leq j \leq n$,

$$\begin{aligned}
\|e_j(\widetilde{\nabla}_j f(x^{t-1}) - \nabla_j f(x^{t-1})\|^2 &= \left(\widetilde{\nabla}_j f(x^{t-1}) - \nabla_j f(x^{t-1})\right)^2 \\
&= \left(\frac{f(x^{t-1} + \tau e_j) - f(x^{t-1})}{\tau} - \nabla_j f(x^{t-1})\right)^2 \\
&= \frac{1}{\tau^2}\left(f(x^{t-1} + \tau e_j) - f(x^{t-1}) - \tau \nabla_j f(x^{t-1})\right)^2 \\
&= \frac{1}{\tau^2}\left(f(x^{t-1} + \tau e_j) - f(x^{t-1}) - \langle \tau e_j, \nabla f(x^{t-1})\rangle\right)^2 \\
&\leq \frac{L}{4\tau^2}\|\tau e_j\|^4 \\
&\leq \frac{L\tau^2}{4}.
\end{aligned}$$

Hence, we have the expression

$$\mathbb{E}_{t-1}\left[\|e_{i_t}(\widetilde{\nabla}_{i_t} f(x^{t-1}) - \nabla f(x^{t-1})\|^2\right] \leq \frac{L\tau^2}{4}.$$

Reusing Young's inequality with a parameter $\alpha > 0$, we get

$$\begin{aligned}
\mathbb{E}_{t-1}[\|\Delta^t\|^2] &\leq (1+\beta)(1+\alpha)\left(1 - \frac{1}{n}\right)\|m^{t-1} - \nabla f(x^{t-1})\|^2 \\
&\quad + (1+\beta)\left(1 + \frac{1}{\alpha}\right)\mathbb{E}_{t-1}\left[\|e_{i_t}(\widetilde{\nabla}_{i_t} f(x^{t-1}) - \nabla f(x^{t-1})\|^2\right] \\
&\quad + \left(1 + \frac{1}{\beta}\right)\eta_{t-1}^2 L^2 D^2, \\
\mathbb{E}_{t-1}[\|\Delta^t\|^2] &\leq (1+\beta)(1+\alpha)\left(1 - \frac{1}{n}\right)\|\Delta^{t-1}\|^2 + (1+\beta)\left(1 + \frac{1}{\alpha}\right)\frac{L^2\tau^2}{4} \\
&\quad + \left(1 + \frac{1}{\beta}\right)\eta_{t-1}^2 L^2 D^2.
\end{aligned}$$

If $\beta = \frac{1}{2n}$, then $(1 - \frac{1}{n})(1 + \frac{1}{2n}) \leq 1 - \frac{1}{2n}$. And with $\alpha = \frac{1}{4n}$, we get the upper bounds $(1 - \frac{1}{2n})(1 + \frac{1}{4n}) \leq (1 - \frac{1}{4n})$ and $(1 + \frac{1}{2n})(1 + 4n) = 4n + 3 + \frac{1}{2n} \leq 4n + 4n = 8n$. Hence,

$$\mathbb{E}_{t-1}\left[\|\Delta^t\|^2\right] \leq \left(1 - \frac{1}{4n}\right)\|\Delta^{t-1}\|^2 + 3\eta_{t-1}^2 nL^2 D^2 + 2nL^2\tau^2.$$

□

C.3.7 Heavy Ball [27]

Description For the Heavy Ball estimator as proposed by [27], $\{m^t\}_{t=0}^{T-1}$ in Algorithm BSFW is defined by the following equations

$$\begin{aligned}
m^0 &= 0, \\
m^t &= (1 - \tilde{\rho}_t)m^{t-1} + \tilde{\rho}_t \tilde{\nabla} f(x^t, \xi_t),
\end{aligned}$$

where $\tilde{\rho}_t$ is a defined momentum function. For this estimator, we also assume that there exists a constant bound of the variance of unbiased stochastic gradients [27], as described by Assumption 4. This assumption follows directly from Assumption 3 stated in [27].

We first provide a recursion Lemma C.8 when Algorithm BSFW uses the Heavy Ball estimator. Then, Lemma C.9 uses Lemma C.8 to find the constants necessary for Assumption 3. However, the constants are a function of the horizon T, which causes problems when attempting to find desired convergence bounds and complexities. Due to this, we propose alternate

any-time convergence rates for ρ-quasar-convex and nonconvex objective functions.

For ρ-quasar-convex functions, we bound the error term through Lemma C.10, provide an alternate recursion Theorem C.11 for ρ-quasar-convex functions, and show a convergence rate of $\mathcal{O}\left(\frac{1}{t^{1/3}}\right)$ in Theorem C.12. For nonconvex functions, we bound the error term by Lemma C.13 and show a convergence rate of $\mathcal{O}\left(\frac{\ln(t)}{t^{1/4}}\right)$ in Theorem C.14.

Convergence Analysis

Assumption 4 (Heavy Ball Variance Bound)**.** The variance of unbiased stochastic gradients $\nabla \tilde{f}(x, \xi)$ is bounded above by σ^2, i.e., for all random variables ξ,

$$\mathbb{E}\left[\|\nabla \tilde{f}(x, \xi) - \nabla f(x)\|^2\right] \leq \sigma^2.$$

Lemma C.8. *(Heavy Ball Error Recursion) Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW. If the objective function f satisfies Assumptions 1 and 4, we have the bound for the sequence of squared errors*

$$\mathbb{E}_{t-1}\left[\|\Delta^t\|^2\right] \leq \left(1 - \frac{\tilde{\rho}_t}{2}\right)\|\Delta^{t-1}\|^2 + {\tilde{\rho}_t}^2\sigma^2 + \frac{2L^2D^2\eta_{t-1}^2}{\tilde{\rho}_t}.$$

Proof. The proof follows exactly from Lemma 1 in [27]. By definition of m^t, we have $m^t = (1-\rho_t)m^{t-1} + \rho_t\tilde{\nabla}f(x^t, \xi_t)$. Hence,

$$\mathbb{E}_{t-1}\left[\|\nabla f(x^t) - m^t\|^2\right] = \mathbb{E}_{t-1}\left[\|\nabla f(x^t) - (1-\tilde{\rho}_t)m^{t-1} - \tilde{\rho}_t\tilde{\nabla}f(x^t, \xi_i)\|^2\right].$$

Adding and subtracting $(1-\tilde{\rho}_t)\nabla f(x^{t-1})$,

$$\begin{aligned}
&\mathbb{E}_{t-1}\left[\|\nabla f(x^t) - m^t\|^2\right] \\
&= \mathbb{E}_{t-1}\left[\|\nabla f(x^t) - (1-\tilde{\rho}_t)\nabla f(x^{t-1}) + (1-\tilde{\rho}_t)\nabla f(x^{t-1}) - (1-\tilde{\rho}_t)m^{t-1} - \tilde{\rho}_t\tilde{\nabla}f(x^t, \xi_t)\|^2\right] \\
&= \mathbb{E}_{t-1}\left[\|\tilde{\rho}_t(\nabla f(x^t) - \tilde{\nabla}f(x^t, \xi_t)) + (1-\tilde{\rho}_t)(\nabla f(x^t) - \nabla f(x^{t-1})) + (1-\tilde{\rho}_t)(\nabla f(x^{t-1}) - m^{t-1})\|^2\right] \\
&= {\tilde{\rho}_t}^2\mathbb{E}_{t-1}\left[\|\nabla f(x^t) - \tilde{\nabla}f(x^t, \xi_t)\|^2\right] + (1-\tilde{\rho}_t)^2\|\nabla f(x^t) - \nabla f(x^{t-1})\|^2 \\
&\quad + (1-\tilde{\rho}_t)^2\|\nabla f(x^{t-1}) - m^{t-1}\|^2 + 2\tilde{\rho}_t(1-\tilde{\rho}_t)\mathbb{E}_{t-1}\left[\langle\nabla f(x^t) - \tilde{\nabla}f(x^t, \xi_t), \nabla f(x^t) - \nabla f(x^{t-1})\rangle\right] \\
&\quad + 2\tilde{\rho}_t(1-\tilde{\rho}_t)\langle\nabla f(x^t) - \nabla f(x^{t-1}), \nabla f(x^{t-1}) - m^{t-1}\rangle \\
&\quad + 2(1-\tilde{\rho}_t)^2\mathbb{E}_{t-1}\left[\langle\nabla f(x^{t-1}) - m^{t-1}, \nabla f(x^t) - \tilde{\nabla}f(x^t, \xi_t)\rangle\right].
\end{aligned}$$

Recall that $\Delta^t = \nabla f(x^t) - m^t$. Using the fact that $\tilde{\nabla}f(x^t, \xi_t)$ is unbiased, i.e. $\mathbb{E}_{t-1}\left[\tilde{\nabla}f(x^t, \xi_t)\right] = \nabla f(x^t)$, L-smoothness of f, Assumption 4 and Lemma C.1, we get

$$\begin{aligned}
\mathbb{E}_{t-1}\left[\|\Delta^t\|^2\right] &\leq {\tilde{\rho}_t}^2\sigma^2 + (1-\tilde{\rho}_t)^2L^2\|x^t - x^{t-1}\|^2 + (1-\tilde{\rho}_t)^2\|\Delta^{t-1}\|^2 + 2(1-\tilde{\rho}_t)^2\langle\nabla f(x^t) - \nabla f(x^{t-1}), \Delta^{t-1}\rangle \\
&\leq {\tilde{\rho}_t}^2\sigma^2 + (1-\tilde{\rho}_t)^2L^2\eta_{t-1}^2D^2 + (1-\tilde{\rho}_t)^2\|\Delta^{t-1}\|^2 + 2(1-\tilde{\rho}_t)^2\langle\nabla f(x^t) - \nabla f(x^{t-1}), \Delta^{t-1}\rangle.
\end{aligned}$$

Using Young's inequality with a parameter $\beta_t > 0$, we get

$$\begin{aligned}
2(1-\tilde{\rho}_t)^2\langle\nabla f(x^t) - \nabla f(x^{t-1}), \Delta^{t-1}\rangle &\leq 2(1-\tilde{\rho}_t)^2\beta_t\|\Delta^{t-1}\|^2 + 2(1-\tilde{\rho}_t)^2\frac{1}{\beta_t}\|\nabla f(x^t) - \nabla f(x^{t-1})\|^2 \\
&\leq 2(1-\tilde{\rho}_t)^2\beta_t\|\Delta^{t-1}\|^2 + 2(1-\tilde{\rho}_t)^2\frac{1}{\beta_t}L^2\eta_{t-1}^2D^2.
\end{aligned}$$

This gives us

$$\mathbb{E}_{t-1}\left[\|\Delta^t\|^2\right] \leq {\tilde{\rho}_t}^2\sigma^2 + (1-\tilde{\rho}_t)^2\left(1 + \frac{1}{\beta_t}\right)L^2\eta_{t-1}^2D^2 + (1-\tilde{\rho}_t)^2(1+\beta_t)\|\Delta^{t-1}\|^2.$$

Since $\tilde{\rho}_t \leq 1$, $(1-\tilde{\rho}_t)^2 \leq (1-\tilde{\rho}_t)$. By setting $\beta_t = \frac{\tilde{\rho}_t}{2}$ we have $(1-\tilde{\rho}_t)(1+(2/\tilde{\rho}_t)) \leq (2/\tilde{\rho}_t)$ and $(1-\tilde{\rho}_t)(1+(\rho_t/2)) \leq (1-(\tilde{\rho}_t/2))$ and thus,

$$\begin{aligned}\mathbb{E}_{t-1}\left[\|\Delta^t\|^2\right] &\leq {\tilde{\rho}_t}^2\sigma^2 + (1-\tilde{\rho}_t)\left(1+\frac{1}{\beta_t}\right)L^2\eta_{t-1}^2D^2 + (1-\tilde{\rho}_t)(1+\beta_t)\|\Delta^{t-1}\|^2 \\ &\leq {\tilde{\rho}_t}^2\sigma^2 + \frac{2L^2D^2\eta_{t-1}^2}{\tilde{\rho}_t} + \left(1-\frac{\tilde{\rho}_t}{2}\right)\|\Delta^{t-1}\|^2.\end{aligned}$$

Hence proven. □

Lemma C.9. *(Parameters for Heavy Ball) Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW using a step decay $\{\eta_t\}_{t=0}^{T-1}$, where the gradient estimator $\{\Phi_t\}_{t=0}^{T-1}$ is the Heavy Ball estimator defined by [27]. Then we have the following parameters used in Assumption 3.*

$$\rho_1 = \frac{\tilde{\rho}_T}{2}, \quad \rho_2 = 1-\left(1-\frac{1}{T+8}\right)^{\frac{4}{3}}, \quad A = 1, \quad B = \frac{2L^2}{\tilde{\rho}_T}, \quad C = 0, \quad E = 0, \quad \sigma_t^2 = {\tilde{\rho}_t}^2\sigma^2.$$

where $\tilde{\rho}_t = 4/(t+8)^{\frac{2}{3}}$ is the decay used in the estimator [27] and σ^2 is the variance bound by Assumption 4.

Proof. From Lemma C.8, taking expectation on both sides, we have

$$\mathbb{E}[\|\Delta\|^2] \leq \left(1-\frac{\tilde{\rho}_t}{2}\right)\mathbb{E}[\|\Delta^{t-1}\|^2] + {\tilde{\rho}_t}^2\sigma^2 + \frac{2L^2D^2}{\tilde{\rho}_t}\eta_{t-1}^2.$$

Since $\forall t \in \{0, 1\cdots T\}, \tilde{\rho}_t \geq \tilde{\rho_T}$, we have,

$$\left(1-\frac{\tilde{\rho}_t}{2}\right) \leq \left(1-\frac{\tilde{\rho}_T}{2}\right) \quad \& \quad \frac{2L^2D^2}{\tilde{\rho}_t} \leq \frac{2L^2D^2}{\tilde{\rho}_T}.$$

Also, setting $\sigma_t^2 = {\tilde{\rho}_t}^2\sigma^2$, we thus have the equation which gives the parameters ρ_1, A, B:

$$\mathbb{E}[\|\Delta^t\|^2] \leq \left(1-\frac{\tilde{\rho}_T}{2}\right)\mathbb{E}[\|\Delta^{t-1}\|^2] + \sigma_t^2 + \frac{2L^2D^2}{\tilde{\rho}_T}\eta_{t-1}^2.$$

We now need to solve the recurrent inequality given below, to fit Assumption 3.

$$\forall t > 0: \quad \sigma_t^2 \leq (1-\rho_2)\sigma_{t-1}^2.$$

Using the expression of σ_t^2, we have

$$\begin{aligned}\tilde{\rho}_t^2\sigma^2 &\leq (1-\rho_2)\tilde{\rho}_{t-1}^2\sigma^2, \\ \tilde{\rho}_t^2 &\leq (1-\rho_2)\tilde{\rho}_{t-1}^2 \\ \frac{4^2}{(t+8)^{4/3}} &\leq (1-\rho_2)\frac{4^2}{(t+7)^{4/3}}, \\ \left(\frac{t+7}{t+8}\right)^{4/3} &\leq 1-\rho_2, \\ \left(1-\frac{1}{t+8}\right)^{4/3} &\leq 1-\rho_2, \\ \forall t \in \{1, 2, \cdots T\}: \quad \rho_2 &\leq 1-\left(1-\frac{1}{t+8}\right)^{4/3}.\end{aligned}$$

Thus to solve ρ_2,

$$\rho_2 = \underset{t\in\{1,2,\cdots T\}}{\operatorname{argmin}} \quad 1-\left(1-\frac{1}{t+8}\right)^{4/3} = 1-\left(1-\frac{1}{T+8}\right)^{4/3}.$$

Hence giving us the required parameters □

Lemma C.10. *(Heavy Ball Error Bounds for Quasar-Convex Functions) [27] Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW using the Heavy Ball estimator [27], under the assumption that f satisfies Assumption 1 and ρ-quasar-convexity under Assumption 2. Let Q be the constant defined as*

$$Q = \max\left\{9^{2/3}\|\nabla f(x^0) - m^0\|^2,\ (16\sigma^2 + 2L^2D^2)/\rho^2\right\},$$

where σ^2 is the constant upper bound of variance of unbiased stochastic gradients, as given by Assumption 4. Suppose

$$\forall t: \quad \eta_t = \frac{2}{\rho(t+9)}, \quad \tilde{\rho}_t = \frac{4}{(t+8)^{\frac{2}{3}}}.$$

Then we have the following result:

$$\forall t: \quad \mathbb{E}[\|\Delta^t\|^2] \le \frac{Q}{(t+9)^{\frac{2}{3}}}.$$

Proof. The proof follows from [27]. From Lemma C.8 we have

$$\begin{aligned}\mathbb{E}_{t-1}[\|\Delta^t\|^2] &\le \left(1 - \frac{\tilde{\rho}_t}{2}\right)\|\Delta^{t-1}\|^2 + {\tilde{\rho}_t}^2\sigma^2 + \frac{2\eta_{t-1}^2L^2D^2}{\tilde{\rho}_t} \\ &\le \left(1 - \frac{2}{(t+8)^{\frac{2}{3}}}\right)\|\Delta^{t-1}\|^2 + \frac{16\sigma^2 + 2L^2D^2}{\rho^2(t+8)^{\frac{4}{3}}}.\end{aligned}$$

Taking expectation on both sides, we have,

$$\mathbb{E}[\|\Delta^t\|^2] \le \left(1 - \frac{2}{(t+8)^{\frac{2}{3}}}\right)\mathbb{E}[\|\Delta^{t-1}\|^2] + \frac{16\sigma^2 + 2L^2D^2}{\rho^2(t+8)^{\frac{4}{3}}}.$$

Using the following parameters in Lemma 19 of [27],

$$\phi_t = \mathbb{E}[\|\Delta^t\|^2], \quad \alpha = \tfrac{2}{3}, \quad b = (16\sigma^2 + 2L^2D^2)/\rho^2, \quad c = 2, \quad t_0 = 8.$$

We have the following result:

$$\mathbb{E}[\|\Delta^t\|^2] \le \frac{Q}{(t+9)^{\frac{2}{3}}},$$

where $Q = \max\left\{9^{2/3}\|\nabla f(x^0) - m^0\|^2,\ (16\sigma^2 + 2L^2D^2)/\rho^2\right\}$. □

Theorem C.11 (Alternative Recursion for Quasar-Convex Functions)**.** *Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW, where the function f satisfies Assumptions 1 and 2. Then we have*

$$F_{t+1} \le (1 - \rho\eta_t)\,F_t + 2\eta_t\|\Delta^t\|D + \frac{L}{2}\eta_t^2D^2.$$

Proof. From Assumption 1, we have

$$f(x^{t+1}) \le f(x^t) + \langle\nabla f(x^t), x^{t+1} - x^t\rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2.$$

Case I: suppose $\gamma_t < 1$. Then we have $x^{t+1} = x^t + \gamma_t d^t$. By using Lemma B.1,

$$\begin{aligned}f(x^{t+1}) &\le f(x^t) + \gamma_t\langle\nabla f(x^t), d^t\rangle + \frac{L}{2}\gamma_t^2\|d^t\|^2 \\ &\le f(x^t) + \gamma_t\langle\nabla f(x^t) - m^t, d^t\rangle + \gamma_t\langle m^t, d^t\rangle + \frac{L}{2}\gamma_t^2\|d^t\|^2 \\ &\le f(x^t) + \gamma_t\langle\nabla f(x^t) - m^t, d^t\rangle + \eta_t\left(\frac{\|s^t - x^t\|}{\|d^t\|}\right)\left(\frac{\|d^t\|}{\|s^t - x^t\|}\right)\langle m^t, s^t - x^t\rangle + \frac{L}{2}\eta_t^2\frac{\|s^t - x^t\|^2}{\|d^t\|^2}\|d^t\|^2 \\ &\le f(x^t) + \gamma_t\langle\nabla f(x^t) - m^t, d^t\rangle + \eta_t\langle m^t, s^t - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2.\end{aligned}$$

Since $s^t \in \text{lmo}(m^t)$, we get

$$f(x^{t+1}) \leq f(x^t) + \gamma_t \langle \nabla f(x^t) - m^t, d^t \rangle + \eta_t \langle m^t, x^\star - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2$$
$$\leq f(x^t) + \gamma_t \langle \nabla f(x^t) - m^t, d^t \rangle + \eta_t \langle m^t - \nabla f(x^t), x^\star - x^t \rangle + \eta_t \langle \nabla f(x^t), x^\star - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2.$$

By using Cauchy-Schwarz inequality on both $\langle \nabla f(x^t) - m^t, d^t \rangle$ and $\langle m^t - \nabla f(x^t), x^\star - x^t \rangle$, we get

$$f(x^{t+1}) \leq f(x^t) + \gamma_t \|m^t - \nabla f(x^t)\| \|d^t\| + \eta_t \|m^t - \nabla f(x^t)\| \|x^\star - x^t\| + \eta_t \langle \nabla f(x^t), x^\star - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2$$
$$\leq f(x^t) + \eta_t \frac{\|s^t - x^t\|}{\|d^t\|} \|\Delta^t\| \|d^t\| + \eta_t \|\Delta^t\| D + \eta_t \langle \nabla f(x^t), x^\star - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2$$
$$\leq f(x^t) + 2\eta_t \|\Delta^t\| D + \eta_t \langle \nabla f(x^t), x^\star - x^t \rangle + \frac{L}{2}\eta_t^2 D^2.$$

Case II: suppose $\gamma_t = 1$. Then we have $x^{t+1} = x^t + \eta_t(s^t - x^t)$. Hence we get,

$$f(x^{t+1}) \leq f(x^t) + \eta_t \langle \nabla f(x^t), s^t - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2$$
$$\leq f(x^t) + \eta_t \langle \nabla f(x^t) - m^t, s^t - x^t \rangle + \eta_t \langle m^t, s^t - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2$$
$$\leq f(x^t) + \eta_t \langle \nabla f(x^t) - m^t, s^t - x^t \rangle + \eta_t \langle m^t, x^\star - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2$$
$$\leq f(x^t) + \eta_t \langle \nabla f(x^t) - m^t, s^t - x^t \rangle + \eta_t \langle m^t - \nabla f(x^t), x^\star - x^t \rangle + \eta_t \langle \nabla f(x^t), x^\star - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2$$
$$\leq f(x^t) + \eta_t \langle \nabla f(x^t) - m^t, s^t - x^\star \rangle + \eta_t \langle \nabla f(x^t), x^\star - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2.$$

Using Cauchy-Schwarz inequality on $\langle \nabla f(x^t) - m^t, s^t - x^\star \rangle$, we have

$$f(x^{t+1}) \leq f(x^t) + \eta_t \|\nabla f(x^t) - m^t\| \|s^t - x^\star\| + \eta_t \langle \nabla f(x^t), x^\star - x^t \rangle + \frac{L}{2}\eta_t^2 \|s^t - x^t\|^2$$
$$\leq f(x^t) + \eta_t \|\Delta^t\| D + \eta_t \langle \nabla f(x^t), x^\star - x^t \rangle + \frac{L}{2}\eta_t^2 D^2$$
$$\leq f(x^t) + 2\eta_t \|\Delta^t\| D + \eta_t \langle \nabla f(x^t), x^\star - x^t \rangle + \frac{L}{2}\eta_t^2 D^2.$$

Thus in both cases, we can reach the same expression. Now using ρ-quasar-convexity of f,

$$f(x^{t+1}) \leq f(x^t) + 2\eta_t \|\Delta^t\| D - \rho\eta_t (f(x^t) - f(x^\star)) + \frac{L}{2}\eta_t^2 D^2.$$

Subtracting $f(x^\star)$ on both sides gives us

$$F_{t+1} \leq (1 - \rho\eta_t) F_t + 2\eta_t \|\Delta^t\| D + \frac{L}{2}\eta_t^2 D^2.$$

□

Theorem C.12 (Convergence under Heavy Ball Estimator for Quasar-Convex Functions)**.** *Let Q' be the constant defined by:*

$$Q' = \max\left\{ 9^{\frac{1}{3}} F_0,\ \frac{4D\sqrt{Q}}{\rho} + \frac{LD^2}{2\rho^2} \right\},$$

where Q is the constant defined by Lemma C.10. Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW using the Heavy Ball estimator [27], where the function f satisfies Assumptions 1 and 2. Suppose

$$\forall t: \quad \eta_t = \frac{2}{\rho(t+9)}, \quad \rho_t = \frac{4}{(t+8)^{\frac{2}{3}}}.$$

Then we have

$$\mathbb{E}[F_t] \leq \frac{Q'}{(t+9)^{\frac{1}{3}}}.$$

Hence, this shows a convergence rate of $\mathcal{O}\left(\frac{1}{t^{1/3}}\right)$.

Proof. The proof follows exactly as in [27]. From Theorem C.11, we have

$$F_{t+1} \leq (1-\rho\eta_t)\, F_t + 2\eta_t \|\Delta^t\| D_2 + \frac{L}{2}\eta_t^2 D_2^2.$$

Taking expectation on both sides and using Lemma C.10 we get,

$$\begin{aligned}
\mathbb{E}[F_{t+1}] &\leq (1-\rho\eta_t)\, \mathbb{E}[F_t] + 2\eta_t D_2 \mathbb{E}[\|\Delta^t\|] + \frac{L}{2}\eta_t^2 D_2^2 \\
&\leq (1-\rho\eta_t)\, \mathbb{E}[F_t] + 2\eta_t D_2 \sqrt{\mathbb{E}[\|\Delta^t\|^2]} + \frac{L}{2}\eta_t^2 D^2 \\
&\leq \left(1 - \frac{2}{t+9}\right) \mathbb{E}[F_t] + \left(\frac{4}{\rho(t+9)}\right) \frac{D\sqrt{Q}}{(t+9)^{\frac{1}{3}}} + \frac{4LD^2}{2\rho^2(t+9)^2} \\
&\leq \left(1 - \frac{2}{t+9}\right) \mathbb{E}[F_t] + \left(\frac{4D\sqrt{Q}}{\rho} + \frac{LD^2}{2\rho^2}\right) \frac{1}{(t+9)^{\frac{4}{3}}}.
\end{aligned}$$

Aim is to prove by PMI that

$$\forall t: \quad \mathbb{E}[F_t] \leq \frac{Q'}{(t+9)^{\frac{1}{3}}}.$$

Base Case: At $t = 0$, by definition of Q',

$$9^{\frac{1}{3}} \mathbb{E}[F_0] \leq Q' \implies \mathbb{E}[F_0] \leq \frac{Q'}{(0+9)^{\frac{1}{3}}}.$$

Induction Step: Suppose $\exists r$ such that

$$\mathbb{E}[F_r] \leq \frac{Q'}{(r+9)^{\frac{1}{3}}}.$$

Hence,

$$\begin{aligned}
\mathbb{E}[F_{r+1}] &\leq \left(1 - \frac{2}{r+9}\right) \mathbb{E}[F_r] + \left(\frac{4D\sqrt{Q}}{\rho} + \frac{LD^2}{2\rho^2}\right) \frac{1}{(r+9)^{\frac{4}{3}}} \\
&\leq \left(1 - \frac{2}{r+9}\right) \left(\frac{Q'}{(r+9)^{\frac{1}{3}}}\right) + \left(\frac{4D\sqrt{Q}}{\rho} + \frac{LD^2}{2\rho^2}\right) \frac{1}{(r+9)^{\frac{4}{3}}} \\
&\leq Q' \left(\frac{r+8}{(r+9)^{\frac{2}{3}}}\right) \\
&\leq \frac{Q'}{(r+10)^{\frac{1}{3}}}.
\end{aligned}$$

Hence proven by PMI. □

Lemma C.13. *(Heavy Ball Error Bounds for Nonconvex Functions) [27] Let* $\{x^t\}_{t=0}^T$ *be a sequence generated by Algorithm BSFW using the Heavy Ball estimator [27], under the assumption that* f *satisfies Assumption 1. Let* M_h *be the constant defined as*

$$M_h = \max\left\{\|\Delta^0\|^2,\ 24(\sigma^2 + 2L^2D^2)\right\},$$

where σ^2 *is the constant upper bound of variance of unbiased stochastic gradients, as given by Assumption 4. Suppose*

$$\forall t: \quad \eta_t = \frac{1}{(t+2)^{\frac{3}{4}}}, \quad \tilde{\rho}_t = \frac{1}{\sqrt{t+1}}.$$

Then we have the following result:

$$\forall t: \quad \mathbb{E}[\|\Delta^t\|^2] \leq \frac{M_h}{\sqrt{t+1}}.$$

Proof. Taking expectation on both sides of Lemma C.8, we have

$$\mathbb{E}\left[\|\Delta^t\|^2\right] \leq \left(1 - \frac{\tilde{\rho}_t}{2}\right) \mathbb{E}\left[\|\Delta^{t-1}\|^2\right] + {\tilde{\rho}_t}^2\sigma^2 + \frac{2L^2D^2\eta_{t-1}^2}{\tilde{\rho}_t}.$$

Substituting the values of $\tilde{\rho}_t$ and η_{t-1} we get,

$$\begin{aligned}\mathbb{E}\left[\|\Delta^t\|^2\right] &\leq \left(1 - \frac{1}{2\sqrt{t+1}}\right) \mathbb{E}\left[\|\Delta^{t-1}\|^2\right] + \frac{\sigma^2}{(t+1)} + \frac{2L^2D^2}{(t+1)} \\ &\leq \left(1 - \frac{1}{2\sqrt{t+1}}\right) \mathbb{E}\left[\|\Delta^{t-1}\|^2\right] + \frac{\sigma^2 + 2L^2D^2}{(t+1)}.\end{aligned}$$

The proof follows from Lemma D.10 in [31]. Using PMI,

Base Case: At $t = 0$, by definition of M_h,

$$\|\Delta^0\|^2 \leq M_h.$$

Induction Step: Let $t > 0$ such that

$$\mathbb{E}[\|\Delta^{t-1}\|^2] \leq \frac{M_h}{\sqrt{t}}.$$

We thus have

$$\begin{aligned}\mathbb{E}[\|\Delta^t\|^2] &\leq \left(1 - \frac{1}{2\sqrt{t+1}}\right) \mathbb{E}[\|\Delta^{t-1}\|^2] + \frac{\sigma^2 + 2L^2D^2}{(t+1)} \\ &\leq \left(1 - \frac{1}{2\sqrt{t+1}}\right) \frac{M_h}{\sqrt{t}} + \frac{\sigma^2 + 2L^2D^2}{(t+1)}.\end{aligned}$$

We can write

$$\frac{1}{\sqrt{t}} = \frac{1}{\sqrt{t+1}}\sqrt{\frac{t+1}{t}} = \frac{1}{\sqrt{t+1}}\sqrt{1 + \frac{1}{t}} \leq \frac{1}{\sqrt{t+1}}\left(1 + \frac{1}{2t}\right).$$

Also, we have the fact that $\forall t \geq 1$,

$$\left(1 - \frac{1}{2\sqrt{t+1}}\right)\left(1 + \frac{1}{2t}\right) \leq \left(1 - \frac{1}{24\sqrt{t+1}}\right).$$

Hence we have

$$\begin{aligned}\mathbb{E}[\|\Delta^t\|^2] &\leq \left(1 - \frac{1}{2\sqrt{t+1}}\right) \frac{M_h}{\sqrt{t}} + \frac{\sigma^2 + 2L^2D^2}{(t+1)} \\ &\leq \left(1 - \frac{1}{2\sqrt{t+1}}\right)\left(1 + \frac{1}{2t}\right) \frac{M_h}{\sqrt{t+1}} + \frac{\sigma^2 + 2L^2D^2}{(t+1)} \\ &\leq \left(1 - \frac{1}{24\sqrt{t+1}}\right) \frac{M_h}{\sqrt{t+1}} + \frac{\sigma^2 + 2L^2D^2}{(t+1)} \\ &\leq \frac{M_h}{\sqrt{t+1}} - \frac{M_h}{24(t+1)} + \frac{\sigma^2 + 2L^2D^2}{(t+1)} \\ &\leq \frac{M_h}{\sqrt{t+1}} + \frac{1}{(t+1)}\left(\sigma^2 + 2L^2D^2 - \frac{M_h}{24}\right) \\ &\leq \frac{M_h}{\sqrt{t+1}}.\end{aligned}$$

□

Theorem C.14. *(Convergence under Heavy Ball Estimator for Nonconvex Functions). Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW with the Heavy Ball estimator [27], where the objective function f satisfies Assumption 1. Suppose for*

all t, $\eta_t = \frac{1}{(t+2)^{3/4}}$ and $\rho_t = \frac{1}{\sqrt{t+1}}$. We thus have the following bound,

$$\mathbb{E}\left[\min_{0\le t\le T-1} \mathrm{Gap}(x^t)\right] \le \frac{F_0 + 2D\sqrt{M_h}(1+\ln(T)) + LD^2}{4\left((T+2)^{\frac{1}{4}} - 2^{\frac{1}{4}}\right)}.$$

where M_h is the constant defined by Lemma C.13, Hence, this shows a convergence rate of $\mathcal{O}\left(\frac{\ln t}{t^{1/4}}\right)$.

Proof. From Assumption 1, we have

$$f(x^{t+1}) \le f(x^t) + \langle \nabla f(x^t), x^{t+1} - x^t\rangle + \frac{L}{2}\|x^{t+1} - x^t\|^2.$$

Case I: suppose $\gamma_t < 1$. Then we have $x^{t+1} = x^t + \gamma_t d^t$. By using Lemma B.1,

$$\begin{aligned}
f(x^{t+1}) &\le f(x^t) + \gamma_t\langle \nabla f(x^t), d^t\rangle + \frac{L}{2}\gamma_t^2\|d^t\|^2\\
&\le f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \gamma_t\langle m^t, d^t\rangle + \frac{L}{2}\gamma_t^2\|d^t\|^2\\
&\le f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \eta_t\left(\frac{\|s^t - x^t\|}{\|d^t\|}\right)\left(\frac{\|d^t\|}{\|s^t - x^t\|}\right)\langle m^t, s^t - x^t\rangle + \frac{L}{2}\eta_t^2\frac{\|s^t - x^t\|^2}{\|d^t\|^2}\|d^t\|^2\\
&\le f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \eta_t\langle m^t, s^t - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2.
\end{aligned}$$

Since $s^t \in \mathrm{lmo}(m^t)$, for all $u \in \mathcal{C}$, we get

$$\begin{aligned}
f(x^{t+1}) &\le f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \eta_t\langle m^t, u - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2\\
&\le f(x^t) + \gamma_t\langle \nabla f(x^t) - m^t, d^t\rangle + \eta_t\langle m^t - \nabla f(x^t), u - x^t\rangle + \eta_t\langle \nabla f(x^t), u - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2.
\end{aligned}$$

By using Cauchy-Schwarz inequality on the both $\langle \nabla f(x^t) - m^t, d^t\rangle$ and $\langle m^t - \nabla f(x^t), u - x^t\rangle$, we get

$$\begin{aligned}
f(x^{t+1}) &\le f(x^t) + \gamma_t\|m^t - \nabla f(x^t)\|\|d^t\| + \eta_t\|m^t - \nabla f(x^t)\|\|u - x^t\| + \eta_t\langle \nabla f(x^t), u - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2\\
&\le f(x^t) + \eta_t\frac{\|s^t - x^t\|}{\|d^t\|}\|\Delta^t\|\|d^t\| + \eta_t\|\Delta^t\|D + \eta_t\langle \nabla f(x^t), u - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2\\
&\le f(x^t) + 2\eta_t\|\Delta^t\|D + \eta_t\langle \nabla f(x^t), u - x^t\rangle + \frac{L}{2}\eta_t^2 D^2.
\end{aligned}$$

Case II: suppose $\gamma_t = 1$. Then we have $x^{t+1} = x^t + \eta_t(s^t - x^t)$. Hence we get for all $u \in \mathcal{C}$,

$$\begin{aligned}
f(x^{t+1}) &\le f(x^t) + \eta_t\langle \nabla f(x^t), s^t - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2\\
&\le f(x^t) + \eta_t\langle \nabla f(x^t) - m^t, s^t - x^t\rangle + \eta_t\langle m^t, s^t - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2\\
&\le f(x^t) + \eta_t\langle \nabla f(x^t) - m^t, s^t - x^t\rangle + \eta_t\langle m^t, u - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2\\
&\le f(x^t) + \eta_t\langle \nabla f(x^t) - m^t, s^t - x^t\rangle + \eta_t\langle m^t - \nabla f(x^t), u - x^t\rangle + \eta_t\langle \nabla f(x^t), u - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2\\
&\le f(x^t) + \eta_t\langle \nabla f(x^t) - m^t, s^t - u\rangle + \eta_t\langle \nabla f(x^t), u - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2.
\end{aligned}$$

Using Cauchy-Schwarz inequality on $\langle \nabla f(x^t) - m^t, s^t - u\rangle$, we have

$$\begin{aligned}
f(x^{t+1}) &\le f(x^t) + \eta_t\|\nabla f(x^t) - m^t\|\|s^t - u\| + \eta_t\langle \nabla f(x^t), u - x^t\rangle + \frac{L}{2}\eta_t^2\|s^t - x^t\|^2\\
&\le f(x^t) + \eta_t\|\Delta^t\|D + \eta_t\langle \nabla f(x^t), u - x^t\rangle + \frac{L}{2}\eta_t^2 D^2\\
&\le f(x^t) + 2\eta_t\|\Delta^t\|D + \eta_t\langle \nabla f(x^t), u - x^t\rangle + \frac{L}{2}\eta_t^2 D^2.
\end{aligned}$$

Thus in both cases, we can reach the same expression. Subtracting $f(x^\star)$ on both sides gives us

$$F_{t+1} \leq F_t + 2\eta_t \|\Delta^t\| D + \eta_t \langle \nabla f(x^t), u - x^t \rangle + \frac{L}{2}\eta_t^2 D^2,$$
$$\eta_t \langle \nabla f(x^t), x^t - u \rangle \leq F_t - F_{t+1} + 2\eta_t \|\Delta^t\| D + \frac{L}{2}\eta_t^2 D^2,$$

Taking expectation on both sides, and summing from $t = 0$ to $t = T - 1$,

$$\begin{aligned}
&\mathbb{E}\left[\eta_t \langle \nabla f(x^t), x^t - u \rangle\right] \leq \mathbb{E}[F_t] - \mathbb{E}[F_{t+1}] + 2\eta_t \mathbb{E}[\|\Delta^t\|] D + \eta_t^2 \frac{LD^2}{2}, \\
\implies &\sum_{t=0}^{T-1} \mathbb{E}\left[\eta_t \langle \nabla f(x^t), x^t - u \rangle\right] \leq F_0 - \mathbb{E}[F_T] + \sum_{t=0}^{T-1} \eta_t 2\mathbb{E}[\|\Delta^t\|] D + \sum_{t=0}^{T-1} \eta_t^2 \frac{LD^2}{2}, \\
\implies &\sum_{t=0}^{T-1} \mathbb{E}\left[\eta_t \langle \nabla f(x^t), x^t - u \rangle\right] \leq F_0 + \sum_{t=0}^{T-1} 2\eta_t \mathbb{E}[\|\Delta^t\|] D + \sum_{t=0}^{T-1} \eta_t^2 \frac{LD^2}{2}, \\
\implies &\mathbb{E}\left[\min_{0\leq t\leq T-1} \langle \nabla f(x^t), x^t - u \rangle\right] \sum_{t=0}^{T-1} \eta_t \leq F_0 + \sum_{t=0}^{T-1} \eta_t 2\mathbb{E}[\|\Delta^t\|] D + \sum_{t=0}^{T-1} \eta_t^2 \frac{LD^2}{2}, \\
\implies &\mathbb{E}\left[\min_{0\leq t\leq T-1} \langle \nabla f(x^t), x^t - u \rangle\right] \leq \frac{F_0 + D\sum_{t=0}^{T-1} 2\eta_t \mathbb{E}[\|\Delta^t\|] + \frac{LD^2}{2}\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t}, \\
&\qquad \leq \frac{F_0 + D\sum_{t=0}^{T-1} 2\eta_t \sqrt{\mathbb{E}[\|\Delta^t\|^2]} + \frac{LD^2}{2}\sum_{t=0}^{T-1} \eta_t^2}{\sum_{t=0}^{T-1} \eta_t}.
\end{aligned}$$

From Lemma C.13, we know that for all t, $\mathbb{E}[\|\Delta^t\|^2] \leq \frac{M_h}{\sqrt{t+1}}$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\min_{0\leq t\leq T-1} \langle \nabla f(x^t), x^t - u \rangle\right] &\leq \frac{F_0 + 2D\sqrt{M_h}\sum_{t=0}^{T-1} \frac{1}{(t+2)^{3/4}} \frac{1}{(t+1)^{1/4}} + \frac{LD^2}{2}\sum_{t=0}^{T-1} \frac{1}{(t+2)^{3/2}}}{\sum_{t=0}^{T-1} \frac{1}{(t+2)^{3/4}}} \\
&\leq \frac{F_0 + 2D\sqrt{M_h}\sum_{t=0}^{T-1} \frac{1}{(t+1)} + \frac{LD^2}{2}\sum_{t=0}^{T-1} \frac{1}{(t+2)^{3/2}}}{\sum_{t=0}^{T-1} \frac{1}{(t+2)^{3/4}}}.
\end{aligned}$$

By using the integral test, we have the following bounds

$$\sum_{t=0}^{T-1} \frac{1}{(t+1)} \leq 1 + \ln(T), \quad \sum_{t=0}^{T-1} \frac{1}{(t+2)^{3/2}} \leq 2, \quad \sum_{t=0}^{T-1} \frac{1}{(t+2)^{3/4}} \geq 4\left((T+2)^{\frac{1}{4}} - 2^{\frac{1}{4}}\right).$$

Hence, we get the expression, that $\forall u \in \mathcal{C}$,

$$\mathbb{E}\left[\min_{0\leq t\leq T-1} \langle \nabla f(x^t), x^t - u \rangle\right] \leq \frac{F_0 + 2D\sqrt{M_h}(1 + \ln(T)) + LD^2}{4\left((T+2)^{\frac{1}{4}} - 2^{\frac{1}{4}}\right)}.$$

Which implies that since $\mathrm{Gap}(x^t) = \max_{u\in\mathcal{C}} \langle \nabla f(x^t), x^t - u \rangle$, we can write

$$\mathbb{E}\left[\min_{0\leq t\leq T-1} \mathrm{Gap}(x^t)\right] \leq \frac{F_0 + 2D\sqrt{M_h}(1 + \ln(T)) + LD^2}{4\left((T+2)^{\frac{1}{4}} - 2^{\frac{1}{4}}\right)}.$$

□

C.3.8 SAG [30]

Description For the usage of the SAG estimator [30], we need to assume that the objective function f in (P) is $f : x \mapsto \tilde{f}(\tilde{A}x)$, where $\tilde{A}$ is a matrix $m \times n$ of samples. This is to ensure a finite sum structure as required by [30]. The estimator uses an additional dual variable α^t, which is initialized by setting $\alpha^0 \in \mathbb{R}^m$. For every iteration $t > 0$, we sample a batch

$S_t \subset \{1, 2, \cdots, m\}$ of size b_s uniformly at random. b_s is a pre-defined per-iteration sample size parameter. For every $i \in \{1, 2, \cdots, m\}$, we update the i^{th} index of α^t by

$$\alpha^{(i),t} = \begin{cases} \frac{1}{m}\tilde{f}_i'(\langle \tilde{a}_i, x^t \rangle), & i \in S_t \\ \alpha^{(i),t-1}. & i \notin S_t \end{cases}$$

For this algorithm to work, we require that $\forall_{x,y \in \mathcal{C}} \|\tilde{A}(x-y)\| \leq D$, and that the step size γ_t be

$$\gamma_t = \min\left\{ \eta_t \frac{\|\tilde{A}(s^t - x^t)\|}{\|\tilde{A}d^t\|}, 1 \right\}, \tag{68}$$

where $s^t = \text{lmo}(\tilde{A}^\top \alpha^t)$ and d^t is the output of the boosting procedure by passing in $m^t = \tilde{A}^\top \alpha^t$. This implies that the Algorithm update is given by $\tilde{A}x^{t+1} = \tilde{A}x^t + \gamma_t \tilde{A}d^t$ when $\gamma_t < 1$ and $\tilde{A}x^{t+1} = \tilde{A}x^t + \eta_t \tilde{A}(s^t - x^t)$ when $\gamma_t = 1$. Since $As^t \in \text{lmo}(\alpha^t)$, it serves to set $y^t = Ax^t$. As explained in Appendix section B.2, we use the parameters provided by Assumption 3 to prove convergence. To consider this objective function $\tilde{f}$, in this case the noise bound needs to be specified as

$$\Delta^t = \alpha^t - \nabla \tilde{f}(\tilde{A}x^t). \tag{69}$$

Using (69), the parameters satisfying Assumption 3 are given by Lemma C.15.

Verification of Assumption 3

Lemma C.15. *(Parameters for SAG) Assume that the objective function f in (P) can be written as $f : x \mapsto \tilde{f}(\tilde{A}x)$, where $\tilde{A}$ is a matrix $m \times n$ of samples. Let $\{x^t\}_{t=0}^T$ be a sequence generated by Algorithm BSFW using a step decay $\{\eta_t\}_{t=0}^{T-1}$, where the gradient estimator $\{\Phi_t\}_{t=0}^{T-1}$ is SAG defined by [30]. Then we have the following parameters used in Assumption 3.*

$$\rho_1 = \frac{b_s}{2m}, \quad \rho_2 = 1, \quad A = 0, \quad B = \left(1 - \frac{b_s}{m}\right)\left(1 + \frac{2m}{b_s}\right) L^2, \quad C = 0, \quad E = 0, \quad \sigma_t^2 = 0,$$

where b_s is the stochastic batch size sampled per-iteration.

Proof. For any t such that $0 \leq t \leq T-1$, by definition of Δ^t in (69),

$$\|\Delta^t\| = \|\alpha^t - \nabla \tilde{f}(\tilde{A}x^t)\|.$$

At each iteration t, since there is a probability $\frac{b_s}{m}$ of any index $j \in \{1, 2 \cdots m\}$ being sampled, meaning $\alpha^{(j),t} = \alpha^{(j),t-1}$ with a probability of $\left(1 - \frac{b_s}{m}\right)$. We thus have the following conditional expectation equation

$$\mathbb{E}_{t-1}[(\Delta^{(j),t})^2] = \left(1 - \frac{b_s}{m}\right)\left(\alpha^{(j),t-1} - \nabla \tilde{f}^{(j)}(\tilde{A}x^t)\right)^2,$$

where $\tilde{a_j}$ refers to the row j of $\tilde{A}$. Summing over all indices from $j = 1$ to $j = m$, we have,

$$\begin{aligned}
\mathbb{E}_{t-1}[\|\Delta^t\|^2] &= \sum_{j=1}^m \mathbb{E}_{t-1}[(\Delta^{(j),t})^2] \\
&= \left(1 - \frac{b_s}{m}\right)\|\alpha^{t-1} - \nabla \tilde{f}(\tilde{A}x^t)\|^2 \\
&= \left(1 - \frac{b_s}{m}\right)\|\alpha^{t-1} - \nabla \tilde{f}(\tilde{A}x^{t-1}) + \nabla \tilde{f}(\tilde{A}x^{t-1}) - \nabla \tilde{f}(\tilde{A}x^t)\|^2 \\
&= \left(1 - \frac{b_s}{m}\right)(\|\alpha^{t-1} - \nabla \tilde{f}(\tilde{A}x^{t-1})\|^2 + \|\nabla \tilde{f}(\tilde{A}x^{t-1}) - \nabla \tilde{f}(\tilde{A}x^t)\|^2 \\
&\quad + 2\langle \alpha^{t-1} - \nabla \tilde{f}(\tilde{A}x^{t-1}), \nabla \tilde{f}(\tilde{A}x^{t-1}) - \nabla \tilde{f}(\tilde{A}x^t) \rangle.
\end{aligned}$$

For the term $\|\nabla\tilde{f}(\tilde{A}x^{t-1}) - \nabla\tilde{f}(\tilde{A}x^t)\|$, By using the L-smoothness property of $\tilde{f}$ and due to the definition of γ_t in (68) in Lemma C.1, we have the following,

$$\|\nabla\tilde{f}(\tilde{A}x^{t-1}) - \nabla\tilde{f}(\tilde{A}x^t)\| \leq L\|Ax^{t-1} - Ax^t\| \leq LD\eta_{t-1}.$$

Using young's inequality with a parameter $\beta > 0$, we have

$$\begin{aligned}
2\langle\alpha^{t-1} - \nabla\tilde{f}(\tilde{A}x^{t-1}), \nabla\tilde{f}(\tilde{A}x^{t-1}) - \nabla\tilde{f}(\tilde{A}x^t)\rangle &\leq \beta\|\alpha^{t-1} - \nabla\tilde{f}(\tilde{A}x^{t-1})\|^2 + \frac{1}{\beta}\|\nabla\tilde{f}(\tilde{A}x^{t-1}) - \nabla\tilde{f}(\tilde{A}x^t)\|^2 \\
&\leq \beta\|\Delta^{t-1}\|^2 + \frac{1}{\beta}\|\nabla\tilde{f}(\tilde{A}x^{t-1}) - \nabla\tilde{f}(\tilde{A}x^t)\|^2 \\
&\leq \beta\|\Delta^{t-1}\|^2 + \frac{1}{\beta}L^2D^2\eta_{t-1}^2.
\end{aligned}$$

Using these inequalities in the expression for $\mathbb{E}_{t-1}[\|\Delta^t\|^2]$, we have the following,

$$\begin{aligned}
\mathbb{E}_{t-1}[\|\Delta^t\|^2] &\leq \left(1 - \frac{b_s}{m}\right)\left(\|\Delta^{t-1}\|^2 + L^2D^2\eta_{t-1}^2 + \beta\|\Delta^{t-1}\|^2 + \frac{1}{\beta}L^2D^2\eta_{t-1}^2\right) \\
&\leq \left(1 - \frac{b_s}{m}\right)(1+\beta)\|\Delta^{t-1}\|^2 + \left(1 - \frac{b_s}{m}\right)\left(1 + \frac{1}{\beta}\right)L^2D^2\eta_{t-1}^2.
\end{aligned}$$

By setting $\beta = \frac{b_s}{2m}$, we have the following

$$\begin{aligned}
\left(1 - \frac{b_s}{m}\right)(1+\beta) &= \left(1 - \frac{b_s}{m}\right)\left(1 + \frac{b_s}{2m}\right) \\
&= 1 - \frac{b_s}{m} + \frac{b_s}{2m} - \frac{b_s}{2m^2} \\
&= 1 - \frac{b_s}{2m} - \frac{b_s}{2m^2} \\
&\leq \left(1 - \frac{b_s}{2m}\right).
\end{aligned}$$

Hence, we have

$$\mathbb{E}_{t-1}[\|\Delta^t\|^2] \leq \left(1 - \frac{b_s}{2m}\right)\|\Delta^{t-1}\|^2 + \left(1 - \frac{b_s}{m}\right)\left(1 + \frac{2m}{b_s}\right)L^2D^2\eta_{t-1}^2.$$

Taking expectations on both sides gives us

$$\mathbb{E}[\|\Delta^t\|^2] \leq \left(1 - \frac{b_s}{2m}\right)\mathbb{E}[\|\Delta^{t-1}\|^2] + \left(1 - \frac{b_s}{m}\right)\left(1 + \frac{2m}{b_s}\right)L^2D^2\eta_{t-1}^2.$$

Hence fitting the necessary parameters in Assumption 3. □

References

[1] A. Beznosikov, D. Dobre, and G. Gidel. Sarah Frank-Wolfe: Methods for constrained optimization with best rates and practical features. *Available from: arXiv:2304.11737*, 2024.

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):1–27, 2011.

[3] C. Combettes and S. Pokutta. Boosting Frank-Wolfe by chasing gradients. In *International Conference on Machine Learning*, pages 2111–2121. PMLR, 2020.

[4] C. W. Combettes and S. Pokutta. Complexity of linear minimization and projection on some sets. *Operations Research Letters*, 49(4):565–571, 2021.

[5] S. Ding, L. Yang, L. Luo, and C. Fang. Optimizing over multiple distributions under generalized quasar-convexity condition. *Advances in Neural Information Processing Systems*, 37:4718–4764, 2024.

[6] D. J. Foster, A. Sekhari, and K. Sridharan. Uniform convergence of gradients for non-convex learning and optimization. *Advances in Neural Information Processing Systems*, 31, 2018.

[7] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

[8] Q. Fu, D. Xu, and A. C. Wilson. Accelerated stochastic optimization methods under quasar-convexity. In *International Conference on Machine Learning*, pages 10431–10460. PMLR, 2023.

[9] J. Guélat and P. Marcotte. Some comments on Wolfe's 'away step'. *Mathematical Programming*, 35(1):110–119, 1986.

[10] F. Hanzely, K. Mishchenko, and P. Richtárik. SEGA: Variance reduction via gradient sketching. *Advances in Neural Information Processing Systems*, 31, 2018.

[11] M. Hardt, T. Ma, and B. Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

[12] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, 5(4):1–19, 2015.

[13] A. Javadi-Abhari, M. Treinish, K. Krsulich, C. J. Wood, J. Lishman, J. Gacon, S. Martiel, P. D. Nation, L. S. Bishop, A. W. Cross, et al. Quantum computing with qiskit. *Available from: arXiv:2405.08810*, 2024.

[14] A. Khademi. The convexity zoo: a taxonomy of function classes in optimization. *Optimization*, pages 1–52, 2026.

[15] A. Khademi and A. Silveti-Falls. Adaptive conditional gradient descent. *Available from: arXiv:2510.11440*, 2025.

[16] D. Kovalev, S. Horváth, and P. Richtárik. Don't jump through hoops and remove those loops: SVRG and Katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pages 451–467. PMLR, 2020.

[17] S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. *Advances in Neural Information Processing Systems*, 28, 2015.

[18] F. Lara and C. Vega. Delayed feedback in online non-convex optimization: a non-stationary approach with applications. *Numerical Algorithms*, pages 1–42, 2025.

[19] E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.

[20] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[21] Z. Li, H. Bao, X. Zhang, and P. Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, 2021.

[22] Z. Li and P. Richtárik. A unified analysis of stochastic gradient methods for nonconvex federated optimization. *Available from: arXiv:2006.07013*, 2020.

[23] F. Locatello, M. Tschannen, G. Rätsch, and M. Jaggi. Greedy algorithms for cone constrained optimization with convergence guarantees. *Advances in Neural Information Processing Systems*, 30, 2017.

[24] T. Ma. Why do local methods solve. *Beyond the Worst-Case Analysis of Algorithms*, page 465, 2021.

[25] D. Martínez-Rubio. Smooth quasar-convex optimization with constraints. *Available from: arXiv:2510.01943*, 2025.

[26] R. D. Millan, O. P. Ferreira, and J. Ugon. Frank-Wolfe algorithm for star-convex functions. *Available from: arXiv:2507.17272*, 2025.

[27] A. Mokhtari, H. Hassani, and A. Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of Machine Learning Research*, 21(105):1–49, 2020.

[28] Mushroom Dataset. UCI Machine Learning Repository. *Dataset*, 1981.

[29] R. Nazykov, A. Shestakov, V. Solodkin, A. Beznosikov, G. Gidel, and A. Gasnikov. Stochastic Frank-Wolfe: Unified analysis and zoo of special cases. In *International Conference on Artificial Intelligence and Statistics*, pages 4870–4878. PMLR, 2024.

[30] G. Négiar, G. Dresdner, A. Tsai, L. El Ghaoui, F. Locatello, R. Freund, and F. Pedregosa. Stochastic Frank-Wolfe for constrained finite-sum minimization. In *International Conference on Machine Learning*, pages 7253–7262. PMLR, 2020.

[31] T. Pethick, W. Xie, K. Antonakopoulos, Z. Zhu, A. Silveti-Falls, and V. Cevher. Training deep learning models with norm-constrained LMOs. *Available from: arXiv:2502.07529*, 2025.

[32] D. A. Quiroga and A. Kyrillidis. Using non-convex optimization in quantum process tomography: Factored gradient descent is tough to beat. *Available from: arXiv:2312.01311*, 2023.

[33] S. J. Reddi, S. Sra, B. Póczos, and A. Smola. Stochastic Frank-Wolfe methods for nonconvex optimization. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1244–1251. IEEE, 2016.

[34] K. K. Tsuji, K. Tanaka, and S. Pokutta. Pairwise conditional gradients without swap steps and sparser kernel herding. In *International Conference on Machine Learning*, pages 21864–21883. PMLR, 2022.

[35] J.-K. Wang and A. Wibisono. Continuized acceleration for quasar convex functions in non-convex optimization. *Available from: arXiv:2302.07851*, 2023.

[36] W. Wolberg. Breast Cancer Wisconsin (Original). *UCI Machine Learning Repository*, 1990. Dataset.

[37] P. Wolfe. Convergence theory in nonlinear programming. *Integer and Nonlinear Programming*, pages 1–36, 1970.

[38] C. J. Wood, J. D. Biamonte, and D. G. Cory. Tensor networks and graphical calculus for open quantum systems. *Available from: arXiv:1111.6950*, 2015.