Conditional Gradient Methods via Stochastic Path-Integrated Differential Estimator

Alp Yurtsever 1  Suvrit Sra 2  Volkan Cevher 1

Abstract
We propose a class of novel variance-reduced stochastic conditional gradient methods. By adopting the recent stochastic path-integrated differential estimator technique (SPIDER) of Fang et al. (2018) for the classical Frank-Wolfe (FW) method, we introduce SPIDER-FW for finite-sum minimization as well as the more general expectation minimization problems. SPIDER-FW enjoys superior complexity guarantees in the non-convex setting, while matching the best known FW variants in the convex case. We also extend our framework à la conditional gradient sliding (CGS) of Lan & Zhou (2016), and propose SPIDER-CGS.

1. Introduction
We study two different problem settings in this paper, finite-sum and the more general expectation minimization:

minimize \( x \in \Omega \) \( F(x) := \begin{cases} E_\xi f(x, \xi) & \text{(expectation)} \\ \frac{1}{n} \sum_{i=1}^{n} f_i(x) & \text{(finite-sum)} \end{cases} \) (1)

\( \Omega \subset \mathbb{R}^d \) is the convex and compact domain;
\( F, f \) and \( f_i \) are differentiable and possibly non-convex;
\( \xi \sim P \) is a random variable, supported on \( \Xi \subset \mathbb{R}^p \).

The expectation objective template covers a large number of applications in machine learning and statistics. The finite-sum template frequently arises in M-estimation and empirical risk minimization problems. Accordingly, there are many applications for stochastic conditional gradient methods both in convex and non-convex settings. This includes low-rank matrix and tensor factorizations, structured sparse matrix estimation, dictionary learning applications, multiclass classification (considered as a motivating example in a related work by Hazan & Luo (2016)), constrained deep learning problems (e.g., Ravi et al. (2018) present an application in computer vision) and many more.

Template (1) can be solved by using the well-known projected stochastic gradient descent method (SGD). At each iteration, SGD takes a stochastic gradient step followed by a projection to ensure the feasibility of the new point. However, in many applications, projection onto \( \Omega \) can impose a computational bottleneck (e.g., projection onto the nuclear norm-ball may require a full singular value decomposition), or it can be even intractable (e.g., dual structural SVMs (Lacoste-Julien et al., 2013)).

As a result, the Frank-Wolfe (FW) algorithm (aka conditional gradient method) has witnessed tremendous interest in the machine learning community in the last decade. FW avoids projection by leveraging the so-called linear minimization oracle instead:

\( \text{lmo}_\Omega(v) = \arg \min_{x \in \Omega} \langle x, v \rangle \). (lmo)

\text{lmo} is significantly cheaper to compute than projection. For instance, \( \text{lmo} \) of nuclear norm-ball requires the computation of the leading singular vectors only (vs. the full spectrum for projection), which can be efficiently found by using Krylov subspace methods (Jaggi, 2013).

Our focus in this paper is on the theoretical complexity of stochastic and finite-sum FW, with an aim to identify and present the tightest results known so far. To this end, we also propose a class of novel variance-reduced stochastic optimization algorithms, based on the recent stochastic path-integrated differential estimator technique (SPIDER) of Fang et al. (2018).

By combining SPIDER with the classical FW method, we introduce SPIDER-FW for finite-sum and expectation minimization problems. We also extend our framework à la conditional gradient sliding (CGS) of Lan & Zhou (2016), and propose SPIDER-CGS.

From SPIDER, we adopt the variance bounds from Lemma 1 of (Fang et al., 2018), which relates the variance of the current estimator to the error of the previous estimator and the distance between the iterates. Nevertheless, Fang et al. (2018) introduce SPIDER for normalized gradient method
which is fundamentally different than the FW method. Accordingly, the analyses are different.

A natural and widely used measure for the convergence of conditional gradient methods is the so-called FW-gap (cf., Section 3). However, we are not aware of any reported FW-gap convergence of CGS in the non-convex settings. Therefore, we present a new compact proof (and an extension for the stochastic setting) in the supplementary material. Although CGS does not seem to provide any improvement upon FW in this setup, we use the proof technique to extend SPIDER-CGS for the non-convex settings.

Finally, for the majority of the variance reduced FW methods in the literature, the analysis relies on the induction technique with respect to the outer loop counter, along with a sufficient improvement condition for each epoch. Consequently, at the beginning of each epoch parameters are typically reset. Instead, we set our learning-rate parameters with respect to the more natural total iteration counter, and we go over the proof without induction.

Roadmap. Section 2 provides an extensive discussion on the related works. Section 3 recalls some basic notions from the optimization theory. Sections 4 and 5 present SPIDER-FW and SPIDER-CGS respectively, along with their theoretical guarantees for various problem settings. Section 6 provides an extensive comparison of the theoretical complexity of FW methods in the literature. Finally, Section 7 draws the conclusions.

Notation. We work on the real space with Euclidean norms for simplicity. Throughout, \( \langle \cdot, \cdot \rangle \) represents the standard inner product associated with the Euclidean norm \( \| \cdot \| \). We use the notation \( [n] = \{ 1, 2, \ldots, n \} \). \( D \) denotes diameter of \( \Omega \), i.e., \( D = \max_{x,y \in \Omega} \| x - y \| \).

2. Related Works

Frank-Wolfe algorithm. This classical method is first proposed by Frank & Wolfe (1956) for solving smooth convex minimization problems with a polyhedral domain constraint (polyhedral constraint is relaxed for an arbitrary convex domain. Since the new estimate is constructed as a convex combination of the current iterate and this extreme point, by definition it is a feasible point, hence the method does not require projections.

FW did not attract much attention in the machine learning community due to its slow convergence rate until Hazan (2008) and Jaggi (2013) emphasize the favorable trade-off between the convergence rate and the per-iteration cost provided by FW in key applications. Following then, there has been a resurgence of interest for FW-type algorithms.

FW literature in the stochastic optimization setting is much younger compared to the projection-based stochastic gradient methods. We can trace it back to a variant for online learning proposed by Hazan & Kale (2012). More recently, Hazan & Luo (2016) introduced stochastic FW methods with and without variance reduction for finite-sum problems. Very recently, Mokhtari et al. (2018) have proposed an alternative scheme for expectation minimization setting.

FW methods for non-convex stochastic learning are relatively understudied, most of the known results are due to Reddi et al. (2016). We discuss more details on the theoretical aspects of all these FW variants in Section 6.

Conditional gradient sliding. Lan & Zhou (2016) has recently developed the conditional gradient sliding method (CGS) based on the idea of applying accelerated gradient method (AG) of Nesterov (1983) for solving problems from template (1), but applying FW to the projection subproblems. In other words, CGS establishes the convergence of an inexact version of AG. Surprisingly, CGS has superior first-order oracle complexity compared to FW, although they have the same \( lmo \) complexity. We discuss more details and variants of CGS in Section 6.

SPIDER. There has been extensive research on variance reduced stochastic optimization methods in order to address the needs of machine learning and big data applications. Therefore, various variance reduction techniques are proposed in the last few years such as SAG (Roux et al., 2012), SVRG (Johnson & Zhang, 2013), SAGA (Defazio et al., 2014), and more recently SARAH (Nguyen et al., 2017) and SPIDER (Fang et al., 2018).

SARAH and SPIDER are closely related since they use the same sequential update rule for the gradient estimator \( v^k \):

\[
v^k = \nabla f_S(x^k) - \nabla f_S(x^{k-1}) + v^{k-1}.
\]

However, SARAH uses this estimator in the classical gradient descent template, while SPIDER adopts a normalized gradient approach, and their results and analyses differ. As described by Wang et al. (2018), the original SPIDER framework has a restrictive step-size (proportional with the target accuracy \( \epsilon \)), which makes the algorithm impractical.

---

**Algorithm 1** Frank-Wolfe algorithm

Input: \( x^1 \in \Omega \)

for \( k = 1, 2, \ldots, K \) do

Compute \( w^k \in \text{lin}_{x \in \Omega} (\nabla F(x^k)) \)

Update \( x^{k+1} = x^k + \eta_k (w^k - x^k) \)

end for

Given an initial guess \( x^1 \in \Omega \), at each iteration, FW minimizes the linear approximation of \( F \) at the current iterate \( x^k \) over \( \Omega \) (this corresponds to the \( lmo \) step). Clearly, minimization of a linear function returns an extreme point of the
though its theoretical appeal. Surprisingly, this problem disappears in the conditional gradient framework analysis.

3. Preliminaries

Solution. We denote a solution and the optimal value of problem (1) by \( x^* \) and \( F^* \) respectively:

\[
\arg\min_{x \in \Omega} F(x) \quad \text{and} \quad F^* = F(x^*).
\]

The measure of non-stationarity. For unconstrained non-convex problems, the typical measure of non-stationarity is the gradient norm, because \( \|\nabla f(x)\| \to 0 \) as \( x \) converges to a stationary point. However, this measure cannot be used for constrained problems, because \( \|\nabla f(x)\| \) might not converge to 0 when we approach to a solution on the boundary.

Instead, we will use the quantity

\[
\mathcal{G}(x) := \max_{u \in \mathcal{U}} \langle u - x, -\nabla F(x) \rangle,
\]

which is widely known as the FW gap (because it naturally appears in the analysis of FW-type methods). FW gap is always non-negative, and it gets small uniformly from a finite set of points \( \{x_1, x_2, \ldots, x^k\} \) is an \( \epsilon \)-solution if

\[
F(x^*_\epsilon) - F^* \leq \epsilon.
\]

\[
\mathbb{E}[\mathcal{G}(x^*_\epsilon)] \leq \epsilon.
\]

It is common to provide convergence guarantees in expectation for a randomly chosen iterate in the non-convex setting. See (Reddi et al., 2016).

Oracle models. We adopt the following black-box oracle model from Reddi et al. (2016), to establish a ground for comparing the convergence speed of different algorithms:

- Linear minimization oracle (lmo)
  - Well-known oracle of FW-type methods.

Assumptions (finite-sum). For the finite-sum setting, we assume that \( f_i(x) \) has an averaged \( L \)-Lipschitz gradient:

\[
\mathbb{E}\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L^2\|x - y\|^2, \quad \forall (x, y) \in \Omega^2.
\]

Note that this implies \( F \) is \( L \)-smooth, since

\[
\|\nabla F(x) - \nabla F(y)\|^2 \leq \mathbb{E}\|\nabla f_i(x) - \nabla f_i(y)\|^2 \leq L^2\|x - y\|^2.
\]

Assumptions (expectation). For the expectation minimization, we assume that \( \nabla f(x, \xi) \) is an unbiased estimate of the gradient:

\[
\mathbb{E}\nabla f(x, \xi) = \nabla F(x).
\]

We also assume that the variance is bounded:

\[
\mathbb{E}\|\nabla f(x, \xi) - \nabla F(x)\|^2 \leq \sigma^2 < \infty, \quad \forall \xi \in \Xi, \forall x \in \Omega.
\]

And finally, we assume an averaged \( L \)-Lipschitz gradient condition, i.e., the following condition holds \( \forall \xi \in \Xi:\nabla f(x, \xi) - \nabla f(y, \xi)\|^2 \leq L^2\|x - y\|^2, \quad \forall (x, y) \in \Omega^2. \)

Similar to the finite-sum, this implies the smoothness of \( F \).

Assumptions (non-convex). Let us denote the initial point by \( \tilde{x}^1 \). Initial suboptimality \( F(\tilde{x}^1) - F^* \) appears in the convergence bounds in the non-convex setting. For notational convenience, we denote an upper bound on this term by \( E \):

\[
F(\tilde{x}^1) - F^* \leq E
\]

Assume that \( F^* \) is finite, then there exists a finite \( E \) which satisfies this bound. This is a direct consequence of the smoothness of \( F \) and the boundedness of domain.

All these assumptions are mild and frequently used in the analysis of stochastic methods and FW-type algorithms.

4. SPIDER Frank-Wolfe

This section presents SPIDER-FW algorithm and its convergence guarantees for various problem settings.

Our methods have a double loop structure, hence the iterates and the parameters have two different iteration counters \( t \) and \( k \), such as \( x^t_k \). For notational simplicity, we drop the first counter when there is no ambiguity, such as \( x^k \). Throughout, \( s_{t,k} \) denotes the total number of inner iterations until \( k \)-th iteration of \( t \)-th epoch. In our pseudocodes, \( \text{draw samples} \) means iid samples for expectation minimization, and uniform selection with replacement in the finite-sum.
Algorithm 2 SPIDER Frank-Wolfe

Input: $\tilde{x}^1 \in \Omega$

for $t = 1, 2, \ldots, T$ do

Set $x^t = \tilde{x}^t$

Draw $Q_t$ samples $Q_t$

Compute $v^1 = \nabla f_{Q_t}(x^1)$

Compute $w^1 \in \text{Im} \Omega_t(v^1)$

Update $x^t = x^1 + \eta_t (w^1 - x^1)$

for $k = 2, 3, \ldots, K_t$ do

Draw $S_{t,k}$ samples $S_{t,k}$

Compute $v^k = \nabla f_{S_{t,k}}(x^k) - \nabla f_{S_{t,k}}(x^{k-1}) + v^{k-1}$

Compute $w^k \in \text{Im} \Omega_t(v^k)$

Update $x^{k+1} = x^k + \eta_{t,k} (w^k - x^k)$

end for

Set $\tilde{x}^{t+1} = x^{K_t+1}$

end for

SPIDER-FW: Convex finite-sum

We consider SPIDER-FW with

$$K_t = 2^{t-1} \text{ for } t = 1, 2, \ldots, T.$$ 

We choose the sampling parameters

$$S_{t,k} = K_t \quad Q_t = [n]$$

and the learning rate parameter

$$\eta_{t,k} = \frac{2}{s_{t,k} + 1} \text{ where } s_{t,k} = K_t + k - 1.$$ 

Theorem 1. Consider the convex finite-sum optimization template, and suppose that the assumptions in Section 3 for this template hold. Then, estimate $x^{t,k}$ of SPIDER-FW with the parameter choices described above satisfies

$$\mathbb{E}[F(x^{t,k})] - F^* = O\left(\frac{LD^2}{s_{t,k}}\right)$$

Corollary 1. The sfo and lmo complexities of SPIDER-FW for achieving $\epsilon$-solution in this setting are as follows:

$$\#(\text{sfo}) = O\left(\frac{\sqrt{2}D^2 + L^2D^2}{\epsilon^2}\right)$$

$$\#(\text{lmo}) = O\left(\frac{LD^2}{\epsilon}\right)$$

SPIDER-FW: Convex expectation minimization

We consider SPIDER-FW with

$$K_t = K = \lceil \sqrt{n} \rceil.$$ 

Furthermore, we choose the parameters as

$$S_{t,k} = S = \lceil \sqrt{n} \rceil \quad Q_t = [n]$$

and the learning rate parameter

$$\eta_{t,k} = \eta = \frac{1}{\sqrt{ST,K}} \text{ where } s_{T,K} = TK.$$ 

Theorem 2. Consider the convex expectation minimization template, and suppose that the assumptions in Section 3 for this template hold. Then, estimate $x^{t,k}$ of SPIDER-FW with the parameter choices described above satisfies

$$\mathbb{E}[F(x^{t,k})] - F^* = O\left(\frac{LD^2}{ST,K}\right)$$

SPIDER-FW has the same asymptotic oracle complexities as SCGS (Lan & Zhou, 2016) in this setting. In Section 5, we also present the SPIDER-CGS.

SPIDER-FW: Non-convex finite-sum

We consider SPIDER-FW with

$$K_t = K = \lfloor \sqrt{n} \rfloor.$$ 

Furthermore, we choose the parameters as

$$S_{t,k} = S = \lceil \sqrt{n} \rceil \quad Q_t = [n]$$

and the learning rate parameter

$$\eta_{t,k} = \eta = \frac{1}{\sqrt{ST,K}} \text{ where } s_{T,K} = TK.$$ 

Theorem 3. Consider the non-convex finite-sum template, and suppose that the assumptions in Section 3 for this template hold. Denote by $x^{out}$ an iterate $x^{t,k}$ of SPIDER-FW chosen uniformly random over all $(t, k)$ pairs up to $(T, K)$. Then, the following bound on the FW-gap holds:

$$\mathbb{E}[G(x^{out})] = O\left(\frac{\mathcal{E} + LD^2}{\sqrt{ST,K}}\right)$$

Although it is impractical to store all estimates until the final iteration, all stochastic methods for the non-convex setting shown in Table 1 have this type convergence guarantees, see (Reddi et al., 2016). More stringently, Lacoste-Julien (2016) proves convergence of non-convex FW in terms of the running best iterate. However, we cannot keep track of best estimate in the stochastic setting, simply because we cannot measure the FW-gap.
Corollary 3. The ifo and lmo complexities of SPIDER-FW for achieving \( \epsilon \)-solution in the non-convex finite-sum setting are as follows:

\[
\#(ifo) = O\left( \frac{\sqrt{n}}{\epsilon^2} (\epsilon^2 + L^2 D^4) \right)
\]

\[
\#(lmo) = O\left( \frac{\sqrt{n}}{\epsilon^2} (\epsilon^2 + L^2 D^4) \right)
\]

SPIDER-FW has better ifo complexity than state-of-the-art in the non-convex finite-sum setting. It improves the dependence on sample size \( n \). See Table 1 for comparison.

### SPIDER-FW: Non-convex expectation minimization

We consider SPIDER-FW with

\[ K_t = K = \lceil \sigma / \epsilon \rceil. \]

Furthermore, we choose the parameters as

\[ S_{t,k} = S = \lfloor \sigma / \epsilon \rfloor \]

\[ Q_t = Q = \lceil 4(\sigma / \epsilon)^2 \rceil \]

and the learning rate parameter

\[ \eta_{t,k} = \eta = \frac{1}{\sqrt{T K}} \text{ where } ST K = TK. \]

Theorem 4. Consider the non-convex expectation minimization template, and suppose that the assumptions in Section 3 for this template hold. Denote by \( x^{out} \) an iterate \( x^{t,k} \) of SPIDER-FW chosen uniformly random over all \((t,k)\) pairs up to \((T, K)\). Then, the following bound holds:

\[
E[G(x^{out})] = O\left( \frac{E + LD^2}{\sqrt{ST K}} \right) + \frac{\epsilon}{2}
\]

Corollary 4. The sfo and lmo complexities of SPIDER-FW for achieving \( \epsilon \)-solution in this setting are as follows:

\[
\#(sfo) = O\left( \frac{\sqrt{n}}{\epsilon^2} (\epsilon^2 + L^2 D^4) \right)
\]

\[
\#(lmo) = O\left( \frac{\sqrt{n}}{\epsilon^2} (\epsilon^2 + L^2 D^4) \right)
\]

Once again, SPIDER-FW enjoys superior sfo complexity while maintaining the same lmo complexity as its competitors. SVRF was the state-of-the-art with \( O(\epsilon^{-10/3}) \), see Reddi et al. (2016).

5. SPIDER Conditional Gradient Sliding

This section presents SPIDER-CGS (as shown in Algorithm 3) and its convergence guarantees for various settings. SPIDER-CGS has the same oracle complexity as the SPIDER-FW.

### Algorithm 3 SPIDER Conditional Gradient Sliding

**Input:** \( \bar{x}^1 \in \Omega \)

**for** \( t = 1, 2, \ldots, T \) **do**

Set \( x^1 = \bar{x}^1 \) and \( y^1 = \bar{y}^1 \)

Update \( z^1 = y^1 + \gamma_{t,1} (x^1 - y^1) \)

Draw \( Q_t \) samples \( Q_t \)

Compute \( v^1 = \nabla f_{Q_t}(z^1) \)

\[ x^2 = \text{CndG}(x^1, v^1, \alpha_{t,1}, \beta_{t,1}) \]

Update \( y^2 = y^1 + \gamma_{t,1} (x^2 - y^1) \)

**for** \( k = 2, 3, \ldots, K_t \) **do**

Update \( z^k = y^k + \gamma_{t,k} (x^k - y^k) \)

Draw \( S_{t,k} \) samples \( S_{t,k} \)

Compute \( v^k = \nabla f_{S_{t,k}}(z^k) - \nabla f_{S_{t,k}}(z^{k-1} + v^{k-1}) \)

\[ x^{k+1} = \text{CndG}(x^k, v^k, \alpha_{t,k}, \beta_{t,k}) \]

Update \( y^{k+1} = y^k + \gamma_{t,k} (x^{k+1} - y^k) \)

**end for**

Set \( x^{K_t+1} = x^T + 1 \) and \( y^{T+1} = y^{K_t+1} \)

**end for**

**function** \( u^+ = \text{CndG}(u, v, \alpha, \beta) \)

**Set** \( u^1 = u \)

**for** \( k = 1, 2, \ldots, K_t \) **do**

Compute \( w^k \in \text{Lmo}_2(\epsilon + \beta (u^k - u)) \)

Evaluate \( \zeta_k = (\epsilon + \beta (u^k - u), u^k - u^k) \)

if \( \zeta_k \leq \alpha \) **then**

break

**end if**

Set \( \theta_k = \min\{1, \epsilon / (\beta ||w^k - u^k||^2)\} \)

Update \( u^{k+1} = u^k + \theta_k (w^k - u^k) \)

**end for**

Set \( u^+ = u^k \)

**end function**

### SPIDER-CGS: Convex finite-sum

We consider SPIDER-CGS with

\[ K_t = \lceil 2^{1/2} \rceil \text{ for } t = 1, 2, \ldots, T. \]

Furthermore, we choose the sampling parameters as

\[ S_{t,k} = 9K_t s_{t,K_t} \]

\[ Q_t = [n] \]

CndG subsolver parameters as

\[ \beta_{t,k} = \frac{3}{2} L \gamma_{t,k} \]

\[ \alpha_{t,k} = \frac{2LD^2}{(st_k + 1)^2} \]

and the learning rate parameter as

\[ \gamma_{t,k} = \frac{3}{st_k + 2} \text{ where } st_k = \sum_{r=1}^{t-1} K_r + k \]
Theorem 5. Consider the convex finite-sum template, and suppose that the assumptions in Section 3 for this template hold. Then, estimate \( y^{t,k} \) of SPIDER-CGS with the parameter choices described above satisfies

\[
E[F(y^{t,k})] - F^* = O\left( \frac{LD^2}{s_{t,k}^2} \right)
\]

Corollary 5. The ifo and lmo complexities of SPIDER-CGS for achieving \( \epsilon \)-solution in this template are as follows:

\[
\#(ifo) = O\left(n \ln\left(\frac{LD^2}{\epsilon}\right) + \frac{L^2D^4}{\epsilon^2}\right)
\]

\[
\#(lmo) = O\left(\frac{LD^2}{\epsilon}\right)
\]

Remark that the STORC (Hazan & Luo, 2016) has a better ifo complexity, but under the additional assumption of Lipschitz continuity of \( F \).

### SPIDER-CGS: Convex expectation minimization

We consider SPIDER-CGS with

\[
K_t = \lceil 2^{t/2} \rceil \quad \text{for} \ t = 1, 2, \ldots, T.
\]

Furthermore, we choose the sampling parameters as

\[
S_t,k = 9K_t s_{t,k}^2, \quad Q_t = \lceil \sigma^2 s_{t,k}^2 K_t \rceil
\]

CndG subsolver parameters as

\[
\beta_{t,k} = \frac{3}{2} L \gamma_{t,k}, \quad \alpha_{t,k} = \frac{2LD^2}{(s_{t,k} + 1)^2}
\]

and the learning rate parameter as

\[
\gamma_{t,k} = \frac{2}{s_{t,k} + 1} \quad \text{where} \ s_{t,k} = \sum_{r=1}^{t-1} K_r + k
\]

Theorem 6. Consider the convex expectation minimization template, and suppose that the assumptions in Section 3 for this template hold. Then, estimate \( y^{t,k} \) of SPIDER-CGS with the parameter choices described above satisfies

\[
E[F(y^{t,k})] - F^* = O\left( \frac{LD^2}{s_{t,k}^2} \right)
\]

Corollary 6. The sfo and lmo complexities of SPIDER-CGS for achieving \( \epsilon \)-solution in convex expectation minimization problems are as follows:

\[
\#(sfo) = O\left(\frac{\sigma^2 D^2 + L^2 D^4}{\epsilon^2}\right)
\]

\[
\#(lmo) = O\left(\frac{LD^2}{\epsilon}\right)
\]

### SPIDER-CGS: Non-convex finite-sum

We consider SPIDER-CGS with

\[
K_t = K = \lceil \sqrt{n} \rceil.
\]

Furthermore, we choose the sampling parameters as

\[
S_t,k = K \quad Q_t = \lceil n \rceil
\]

CndG subsolver parameters as

\[
\beta_{t,k} = \frac{3}{2} L \gamma, \quad \alpha_{t,k} = LD^2 \gamma
\]

and the learning rate parameter as

\[
\gamma_{t,k} = \gamma = \frac{1}{\sqrt{ST.K}} \quad \text{where} \ s_{T,K} = TK.
\]

Theorem 7. Consider the non-convex finite-sum template, and suppose that the assumptions in Section 3 for this template hold. Denote by \( y^\text{out} \) an iterate \( y^{t,k} \) of SPIDER-CGS chosen uniformly random over all \((t,k)\) pairs up to \((T,K)\). Then, the following bound on the FW-gap holds:

\[
E[G(y^\text{out})] = O\left( \frac{E + LD^2}{\sqrt{ST.K}} \right)
\]

Corollary 7. The ifo and lmo complexities of SPIDER-CGS for achieving \( \epsilon \)-solution in non-convex finite-sum are

\[
\#(ifo) = O\left(\frac{n}{T}(\epsilon^2 + L^2 D^4)\right)
\]

\[
\#(lmo) = O\left(\frac{1}{T}(\epsilon^2 + L^2 D^4)\right)
\]

### SPIDER-CGS: Non-convex expectation minimization

We consider SPIDER-CGS with

\[
K_t = K = \lceil \sigma / \epsilon \rceil.
\]

Furthermore, we choose the sampling parameters as

\[
S_t,k = K \quad Q_t = \lceil 4(\sigma / \epsilon)^2 \rceil
\]

CndG subsolver parameters as

\[
\beta_{t,k} = \frac{3}{2} L \gamma, \quad \alpha_{t,k} = LD^2 \gamma
\]

and the learning rate parameter as

\[
\gamma_{t,k} = \gamma = \frac{1}{\sqrt{ST.K}} \quad \text{where} \ s_{T,K} = TK.
\]

Theorem 8. Consider the non-convex expectation minimization template, and suppose that the assumptions in Section 3 for this template hold. Denote by \( y^\text{out} \) an iterate \( y^{t,k} \) of
A natural extension of FW for stochastic setting is described by SCGS in Section 3 of their original work. While keeping the optimal $O(1/k)$ lmo complexity, SCGS achieves $O(1/k^2)$ sfo complexity, which even gets $O(1/k)$ under strong convexity assumption.

Hazan & Luo (2016) introduce the stochastic variance reduced Frank-Wolfe method (SVRF) by adopting the variance reduction techniques from Johnson & Zhang (2013) and Mahdavi et al. (2013). SVRF is explicitly designed for the finite-sum setting, and it requires $O(n \ln(1/\epsilon))$ full gradients as well as $O(1/\epsilon^2)$ ifo and $O(1/\epsilon)$ lmo to get an $\epsilon$-solution.

To further improve ifo complexity of SVRF, Hazan & Luo (2016) design a variant based on CGS. This variant, stochastic variance reduced condition gradient sliding (STORC), also requires $O(n \ln(1/\epsilon))$ full gradients and $O(1/\epsilon)$ lmo, but it enjoys a reduced number of ifo calls at $O(1/\epsilon^{1.5})$. Compared to SVRF, however, STORC additionally assumes that $F$ is Lipschitz continuous in domain $\Omega$. Also remark that STORC gets better rates under additional assumptions such as strong-convexity.

Lu & Freund (2018) propose a stochastic FW variant which requires $O(1/\epsilon)$ lmo and $O(n + 1/\epsilon)$ ifo complexity for the convex finite-sum. However, the proposed method relies on a special structure of the objective function, that $f_i$ are univariate functions of the fitted value $\langle a_i, x \rangle$ for some given data sample $a_i$.

All stochastic FW variants we discussed up to know are based on an increasing mini-batch size. Very recently, Mokhtari et al. (2018) have proposed an alternative scheme (SFW-1) for expectation minimization setting, which requires a single sfo at each iteration. Nevertheless, SFW-1 has an arguably worse computational complexity compared to SVRF, with its $O(1/\epsilon^2)$ calls of sfo and lmo. We emphasize the applications of SFW-1 in submodular maximization, but this is beyond the scope of our work.

For the convex finite-sum setting, SPIDER-FW and SPIDER-CGS share the same complexities as SVRF.

6.2. Non-convex optimization camp

Batch setting. FW converges asymptotically to a stationary point, see Section 2.2 in (Bertsekas, 1999). To our knowledge, Yu et al. (2014) shows the first convergence rates for FW in non-convex setting, and Lacoste-Julien (2016) proves a non-asymptotic $O(1/\sqrt{k})$ rate in FW-gap for a FW variant with line-search.

Stochastic setting. As shown by Reddi et al. (2016), SFW achieves an $\epsilon$-solution with $O(1/\epsilon^2)$ ifo and $O(1/\epsilon^2)$ lmo complexities. Moreover, they also analyze SVRF in the non-convex setting (but they call it SVFW), and prove that it takes $O(1/\epsilon^{10/3})$ ifo and $O(1/\epsilon^2)$ lmo complexity for this method to get an $\epsilon$-solution. In the finite-sum setting, the former is replaced by $O(n + n^{2/3}/\epsilon^2)$ ifo calls.

Reddi et al. (2016) also propose a variant (SAGAFW) based on the SAGA variance reduction technique described by
Defazio et al. (2014). However, we omit SAGAFW because there is an issue in the analysis of this method (while telescoping Eq.(14), in page 1249).

Qu et al. (2018) show the convergence rate for special instances of CGS and SCGS in the non-convex setting. However, they consider a different convergence criterion based on a proximal gradient mapping rather than the conventional FW-gap. Consequently, their results are incomparable with the rest of the literature. For the fact that we are running a projection-free method, the FW-gap is a more natural choice than the projection/proximal gradient norm.

We provide a parameter setting and a compact proof for CGS and SCGS in the supplemental material. Note however this setting simply gets the same guarantees as FW and SFW respectively. Whether or not CGS can provide improved oracle complexities compared to FW in the non-convex setting, is an open problem.

For the non-convex setting, SPIDER-FW and SPIDER-CGS have the same oracle complexities, superior to SVRF (which is the state-of-the-art to our knowledge) for finite-sum and expectation minimization problems.

### 6.3. Results from Concurrent Works

By the time we prepared this manuscript, the idea of combining SPIDER with the FW analysis was not explored yet. However, stochastic variance reduction methods and FW-type algorithms are both very active research fields. In this part, we discuss some results from a few concurrent works that appeared after we submitted our paper for review.

The recent work by Shen et al. (2019) is very closely related to our approach. They propose a class of methods based on the CGM and various variance reduction techniques for the non-convex finite-sum setting, including the SPIDER-FW. Besides, they also propose extensions that use second-order approximations. Finally, they provide simulation studies to compare empirical performance of different variants. We refer to this paper for a numerical comparison.

Hassani et al. (2019) introduce a novel variance reduced CGM method, but their work focuses primarily on the submodular maximization. Accordingly, they consider a more general expectation minimization template (the so-called non-oblivious setting) where the probability distribution depends on the decision variable \( x \) and may change during the optimization procedure. Therefore, the proposed method requires some further assumptions and modifications involving computations with the Hessian approximation. Finally, Zhang et al. (2019) consider a stochastic CGM approach with SPIDER in the distributed and quantized settings.

### 7. Concluding Remarks

We have proposed two novel FW-type methods based on the idea of blending the recent variance reduction technique SPIDER into FW and CGS frameworks. We have shown that the resulting methods enjoy superior oracle complexities in various convex and non-convex optimization templates. Extension of our framework for the strongly convex case is left open. Developing a well-tuned implementation, including one that incorporates parallel optimization, is an important piece of future work.

---

**Table 1**: Comparison of conditional gradient methods for stochastic optimization. Contribution of this work is highlighted with blue font. See Section 6 for more details.

<table>
<thead>
<tr>
<th>Method</th>
<th>convex (finite-sum)</th>
<th>convex (expectation)</th>
<th>non-convex (finite-sum)</th>
<th>non-convex (expectation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FW</td>
<td>( O(n\epsilon^{-1}) )</td>
<td>( O(\epsilon^{-1}) )</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>CGS</td>
<td>( O(n\epsilon^{-1/2}) )</td>
<td>( O(\epsilon^{-1}) )</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>SFW</td>
<td>( O(\epsilon^{-3}) )</td>
<td>( O(\epsilon^{-1}) )</td>
<td>( O(\epsilon^{-3}) )</td>
<td>( O(\epsilon^{-1}) )</td>
</tr>
<tr>
<td>SFW-1</td>
<td>( O(\epsilon^{-3}) )</td>
<td>( O(\epsilon^{-1}) )</td>
<td>( O(\epsilon^{-3}) )</td>
<td>( O(\epsilon^{-1}) )</td>
</tr>
<tr>
<td>Online-FW</td>
<td>( O(\epsilon^{-4}) )</td>
<td>( O(\epsilon^{-2}) )</td>
<td>( O(\epsilon^{-4}) )</td>
<td>( O(\epsilon^{-2}) )</td>
</tr>
<tr>
<td>SCGS</td>
<td>( O(\epsilon^{-2}) )</td>
<td>( O(\epsilon^{-1}) )</td>
<td>( O(\epsilon^{-4}) )</td>
<td>( O(\epsilon^{-2}) )</td>
</tr>
<tr>
<td>SVRF / SVFW</td>
<td>( O(n \ln(n^{-1}) + \epsilon^{-2}) )</td>
<td>( O(\epsilon^{-1}) )</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>STORC†</td>
<td>( O(n \ln(n^{-1}) + \epsilon^{-3/2}) )</td>
<td>( O(\epsilon^{-1}) )</td>
<td>- -</td>
<td>- -</td>
</tr>
<tr>
<td>SPIDER-FW</td>
<td>( O(n \ln(n^{-1}) + \epsilon^{-2}) )</td>
<td>( O(\epsilon^{-1}) )</td>
<td>( O(\epsilon^{-1}) )</td>
<td>( O(n^{1/2} \epsilon^{-2}) )</td>
</tr>
<tr>
<td>SPIDER-CGS</td>
<td>( O(n \ln(n^{-1}) + \epsilon^{-2}) )</td>
<td>( O(\epsilon^{-1}) )</td>
<td>( O(\epsilon^{-1}) )</td>
<td>( O(n^{1/2} \epsilon^{-2}) )</td>
</tr>
</tbody>
</table>
Acknowledgements

VC was supported by the Swiss National Science Foundation (SNSF) under grant number 200021_178865/1. VC has received funding for this project from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement no 725594 - time-data). SS acknowledges support from an Amazon Research Award and the NSF-CAREER award (id 1846088). The authors thank Maria Vladarean for pointing out an error in the initial version of this work.

References


A. Preliminaries

This section presents some known results from the existing literature, key to our analysis, for the sake of completeness.

The following Lemma from Fang et al. (2018) provides an error bound of the estimator $v^k$ obtained by the SPIDER approach.

**Lemma 1** (Lemma 1 from (Fang et al., 2018), more specifically Eqn.(A.3) in its supplemental). Suppose that $S_{t,k}$ is a subset that samples $S_{t,k}$ iid realizations from the distribution $P$. Let the stochastic estimator $\nabla f_S$ satisfy the averaged $L$-Lipschitz gradients condition from Section 3. Set the estimator $v^k$ as

$$v^k = \nabla f_{S_{t,k}}(x^k) - \nabla f_{S_{t,k}}(x^{k-1}) + v^{k-1}. \quad (A.1)$$

Then, the following bound holds:

$$E \| \nabla F(x^k) - v^k \|^2 \leq \frac{L^2}{S_{t,k}} \| x^k - x^{k-1} \|^2 + \| \nabla F(x^{k-1}) - v^{k-1} \|^2. \quad (A.2)$$

The following Lemma draws a well-known bound on the variance in terms of the mini-batch size.

**Lemma 2** (Eqn.(3.5) from (Lan & Zhou, 2016), or Lemma 2 from (Reddi et al., 2016)). Suppose that $S_{t,k}$ is a subset that samples $S_{t,k}$ iid realizations from the distribution $P$. Let the stochastic estimator $\nabla f_{S_{t,k}}$ satisfy the bounded variance condition from Section 3. Then, the following bound holds:

$$E \| \nabla f_{S_{t,k}}(x) - \nabla F(x) \|^2 \leq \frac{\sigma^2}{S_{t,k}} \quad \forall x \in \Omega. \quad (A.3)$$

Finally, we recall the convergence guarantees for the CndG procedure of CGS-type methods.

**Lemma 3** (Similar to Theorem 2.2 part (c) from (Lan & Zhou, 2016), or more generally Theorem 2 from (Jaggi, 2013)). Remark that CndG procedure simply applies FW (with exact line-search) for the following projection subproblem:

$$\min_{x \in \Omega} \frac{\beta}{2} \| x - u + \frac{1}{\beta} v \|^2. \quad (A.4)$$

This problem is $\beta$-smooth, hence FW requires at most $O\left(\frac{4\beta D^2}{\alpha}\right)$ iterations to satisfy the convergence criterion.

B. Non-convex Conditional Gradient Sliding

In this section, we prove convergence of a CGS instance, and derive its oracle complexities in the non-convex settings. We also extend our results for SCGS.

**Proof of convergence for non-convex CGS**

**Algorithm 5** Conditional Gradient Sliding

**Input:** $x^1 \in \Omega$

**Set:** $\alpha = \gamma LD^2$, $\beta = \gamma L/2$, $\gamma = 1/\sqrt{K}$

for $k = 1, 2, \ldots, K$ do

Update $z^k = y^k + \gamma (x^k - y^k)$
$x^{k+1} = \text{CndG}(x^k, \nabla F(x^k), \alpha, \beta)$
Update $y^{k+1} = y^k + \gamma (x^{k+1} - y^k)$

end for

**Theorem 9.** Consider CGS algorithm with the parameters as described in Algorithm 5 (in the batch setting). Denote by $y^{\text{out}}$ a random iterate $y^k$ drawn uniformly random over all iterates of CGS. Then, the following bound holds:

$$E[G(y^{\text{out}})] = \frac{E + 3LD^2}{\sqrt{K}} \quad (B.1)$$
Corollary 9. The ifo and lmo complexities of CGS for achieving an $\epsilon$-solution in the non-convex minimization setting are

$$
\#(ifo) = \mathcal{O}\left((\mathcal{E}^2 + L^2 D^4)^\frac{n}{\epsilon^2}\right) \\
\#(lmo) = \mathcal{O}\left((\mathcal{E}^2 + L^2 D^4)^\frac{1}{\epsilon^2}\right)
$$

Proof. We start by the Taylor expansion and smoothness:

$$
F(y^{k+1}) \leq F(y^k) + \langle \nabla F(y^k), y^{k+1} - y^k \rangle + \frac{L}{2} \| y^{k+1} - y^k \|^2 \\
= F(y^k) + \gamma \langle \nabla F(y^k), x^{k+1} - y^k \rangle + \gamma^2 \frac{L}{2} \| x^{k+1} - y^k \|^2 \\
\leq F(y^k) + \gamma \langle \nabla F(y^k), x^{k+1} - y^k \rangle + \gamma^2 \frac{L}{2} D^2 \\
= F(y^k) + \gamma \langle \nabla F(y^k), w_\star^k - y^k \rangle + \gamma \langle \nabla F(y^k), x^{k+1} - w_\star^k \rangle + \gamma^2 \frac{L}{2} D^2 \\
= F(y^k) - \gamma \mathcal{G}(y^k) + \gamma \langle \nabla F(y^k), x^{k+1} - w_\star^k \rangle + \gamma^2 \frac{L}{2} D^2
$$

where we define $w_\star^k = \arg \max_{x \in \Omega} \langle x, -\nabla F(y^k) \rangle$.

We can equivalently write this inequality as

$$
F(y^{k+1}) \leq F(y^k) - \gamma \mathcal{G}(y^k) + \gamma \langle \nabla F(y^k) - \nabla F(z^k), x^{k+1} - w_\star^k \rangle + \gamma \langle \nabla F(z^k), x^{k+1} - w_\star^k \rangle + \gamma^2 \frac{L}{2} D^2
$$

Focus on the last inner-product term

$$
\gamma \langle \nabla F(z^k), x^{k+1} - w_\star^k \rangle = \gamma \langle \nabla F(z^k) + \beta(x^{k+1} - x^k), x^{k+1} - w_\star^k \rangle - \gamma \beta \langle x^{k+1} - x^k, x^{k+1} - w_\star^k \rangle \\
\leq \gamma \alpha - \gamma \beta \langle x^{k+1} - x^k, x^{k+1} - w_\star^k \rangle
$$

where the first inequality follows from the role of $\alpha$ in CudG, and the second one from Cauchy-Schwarz.

Combining these two inequalities, we obtain

$$
F(y^{k+1}) \leq F(y^k) - \gamma \mathcal{G}(y^k) + \gamma \langle \nabla F(y^k) - \nabla F(z^k), x^{k+1} - w_\star^k \rangle + \gamma \alpha + \gamma \beta D^2 + \gamma^2 \frac{L}{2} D^2 \\
\leq F(y^k) - \gamma \mathcal{G}(y^k) + \gamma D \| \nabla F(y^k) - \nabla F(z^k) \| + \gamma \alpha + \gamma \beta D^2 + \gamma^2 \frac{L}{2} D^2 \\
\leq F(y^k) - \gamma \mathcal{G}(y^k) + \gamma^2 L D \| x^k - y^k \| + \gamma \alpha + \gamma \beta D^2 + \gamma^2 \frac{L}{2} D^2 \\
\leq F(y^k) - \frac{1}{\sqrt{K}} \mathcal{G}(y^k) + \frac{3LD^2}{K}
$$

Taking expectation of both sides, rearranging, and summing over all iterations, we obtain

$$
\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \mathbb{E}[\mathcal{G}(y^k)] \leq F(\bar{x}^1) - \mathbb{E}[F(y^K)] + \sum_{k=1}^{K} \frac{3LD^2}{K} \leq F(\bar{x}^1) - F(x^*) + 3LD^2
$$

Hence, by definition of $y^\text{out}$, we get

$$
\mathbb{E}[\mathcal{G}(y^\text{out})] \leq \frac{F(\bar{x}^1) - F(x^*) + 3LD^2}{\sqrt{K}} = \frac{\mathcal{E} + 3LD^2}{\sqrt{K}}
$$

This completes the convergence rate proof.
To get an $\epsilon$-solution, we set the number of iterations $K_\epsilon$ such that
\[
E[G(y^{out})] \leq \frac{\epsilon}{\sqrt{K_\epsilon}} + 3LD^2 \leq \epsilon.
\]
Hence, we can calculate the lmo complexity using Lemma 3 as
\[
#(lmo) = O \left( K_\epsilon \frac{4\beta D^2}{\alpha} \right) = O \left( (\epsilon^2 + L^2 D^4) \frac{1}{\epsilon^2} \right)
\]
which completes the proof.

**Proof of convergence for the non-convex SCGS**

**Algorithm 6 Stochastic Conditional Gradient Sliding**

**Input:** $x^1 \in \Omega$

**Set:** $\alpha = \gamma LD^2$, $\beta = \gamma L/2$, $\gamma = 1/\sqrt{K}$

**for** $k = 1, 2, \ldots, K$ **do**

- Update $z^k = y^k + \gamma (x^k - y^k)$
- Draw $K$ samples $S_k$
  
  $x^{k+1} = \text{CndG}(x^k, \nabla f_{S_k}(z^k), \alpha, \beta)$
- Update $y^{k+1} = y^k + \gamma (x^{k+1} - y^k)$

**end for**

**Theorem 10.** Consider the SCGS algorithm with the parameters as described in the Algorithm 6. Assume that the conditions for the expectation minimization setting from Section 3 hold. Denote by $y^{out}$ a random iterate $y^k$ drawn uniformly random over all iterates of the SCGS. Then, the following bound holds:
\[
E[G(y^{out})] = \frac{\epsilon + \sigma D + 3LD^2}{\sqrt{K}}
\]

**Corollary 10.** The lso and lmo complexities of SCGS for achieving an $\epsilon$-solution in the non-convex minimization setting are
\[
#(lso) = O \left( (\epsilon + \sigma D + LD^2)^4 \frac{1}{\epsilon^4} \right) \quad \text{and} \quad #(lmo) = O \left( (\epsilon + \sigma D + LD^2)^2 \frac{1}{\epsilon^2} \right)
\]

**Proof.** From (B.3), and similar to (B.4) and (B.5), we can show
\[
F(y^{k+1}) \leq F(y^k) - \gamma \nabla F(y^k) + \gamma \langle \nabla F(y^k) - \nabla f_{S_k}(z^k), x^{k+1} - w^k \rangle + \gamma \alpha + \gamma \beta D^2 + \gamma^2 \frac{L}{2} D^2
\]
where $w^k = \arg\max_{x \in \Omega} \langle x, -\nabla F(y^k) \rangle$.

Focusing on the inner-product term, we get the following bound:
\[
\gamma \langle \nabla F(y^k) - \nabla f_{S_k}(z^k), x^{k+1} - w^k \rangle \leq \gamma \|\nabla F(y^k) - \nabla f_{S_k}(z^k)\| \|x^{k+1} - w^k\|
\leq \gamma D \|\nabla F(y^k) - \nabla f_{S_k}(z^k)\|
\leq \gamma D \|\nabla F(y^k) - \nabla F(z^k)\| + \gamma D \|\nabla F(z^k) - \nabla f_{S_k}(z^k)\|
\leq \gamma LD \|x^k - z^k\| + \gamma D \|\nabla F(z^k) - \nabla f_{S_k}(z^k)\|
= \gamma^2 LD \|x^k - y^k\| + \gamma D \|\nabla F(z^k) - \nabla f_{S_k}(z^k)\|
\leq \gamma^2 LD^2 + \gamma D \|\nabla F(z^k) - \nabla f_{S_k}(z^k)\|
\]

Substituting back and taking expectations, we obtain
\[
E[F(y^{k+1})] \leq E[F(y^k)] - \frac{1}{\sqrt{K}} E[G(y^k)] + \frac{D}{\sqrt{K}} E[\|\nabla F(z^k) - \nabla f_{S_k}(z^k)\|] + 3\gamma^2 LD^2
\]
\[
= E[F(y^k)] - \frac{1}{\sqrt{K}} E[G(y^k)] + \frac{D}{\sqrt{K}} E[\|\nabla F(z^k) - \nabla f_{S_k}(z^k)\|] + \frac{3LD^2}{K}
\]
Now, we use Lemma 2 with the Jensen’s inequality to obtain
\[
\mathbb{E}[F(y^{k+1})] \leq \mathbb{E}[F(y^k)] - \frac{1}{\sqrt{K}} \mathbb{E}[G(y^k)] + \frac{\sigma D + 3LD^2}{K}
\] (B.15)

From here, we follow the same steps as in the proof of CGS and get (B.11).

Then, to achieve an \( \epsilon \)-solution, we can calculate sfo complexity as
\[
\#(\text{sfo}) = \sum_{k=1}^{K_c} K_c = K_c^2 = \mathcal{O}\left((\mathcal{E} + \sigma D + LD^2)^4 \frac{1}{\epsilon^4}\right)
\] (B.16)

Finally, lmo complexity can be found using Lemma 3
\[
\#(\text{lmo}) = \mathcal{O}\left(K_c \frac{4LD^2}{\alpha}\right) = \mathcal{O}\left((\mathcal{E} + \sigma D + LD^2)^2 \frac{1}{\epsilon^2}\right)
\] (B.17)

This completes the proof. \( \square \)

C. Proofs for SPIDER-FW

Lemma 4. Suppose that the assumptions listed in Section 3 hold. Then, for \( k = 1, \ldots, K_t \), we have the following bounds:

**Convex finite-sum**
\[
\mathbb{E}\|\nabla F(x^k) - v^k\| \leq 2LD/K_t
\] (C.1)

**Convex expectation**
\[
\mathbb{E}\|\nabla F(x^k) - v^k\| \leq 3LD/K_t
\] (C.2)

**Non-convex finite-sum**
\[
\mathbb{E}\|\nabla F(x^k) - v^k\| \leq LD/\sqrt{TK}
\] (C.3)

**Non-convex expectation**
\[
\mathbb{E}\|\nabla F(x^k) - v^k\| \leq LD/\sqrt{TK} + \epsilon/2
\] (C.4)

**Proof.** From Lemma 1, we have the following inequality for \( k = 2, 3, \ldots, K_t \):
\[
\mathbb{E}\|\nabla F(x^k) - v^k\|^2 \leq \frac{L^2}{S_{t,k}} \|x^k - x^{k-1}\|^2 + \|\nabla F(x^{k-1}) - v^{k-1}\|^2.
\] (C.5)

By definition, \( \|x^{k-1} - x^k\|^2 = \|\eta_{t,k-1}(w^{k-1} - x^{k-1})\|^2 \leq \eta_{t,k-1}^2 D^2 \). Hence, we get
\[
\mathbb{E}\|\nabla F(x^k) - v^k\|^2 \leq \frac{\eta_{t,k-1}^2 L^2 D^2}{S_{t,k}} + \|\nabla F(x^{k-1}) - v^{k-1}\|^2.
\] (C.6)

**Convex finite-sum:**
We take the telescopic sum of (C.6) from \( i = 2 \) to \( k \)
\[
\mathbb{E}\|\nabla F(x^k) - v^k\|^2 \leq \sum_{i=2}^{k} \frac{\eta_{t,i-1}^2 L^2 D^2}{S_{t,i}} + \|\nabla F(x^1) - v^1\|^2 \leq \frac{4L^2 D^2}{K_t} \sum_{i=2}^{k} \frac{1}{(s_{t,i-1} + 1)^2}
\] (C.7)

By definition of \( s_{t,k} \), for any \( i \geq 2 \) we have
\[
s_{t,i-1} + 1 = K_t + i - 1 \geq K_t.
\] (C.8)

Hence, we get
\[
\mathbb{E}\|\nabla F(x^k) - v^k\|^2 \leq \frac{4L^2 D^2}{K_t^3} \sum_{i=2}^{k} 1 \leq \frac{4L^2 D^2}{K_t^3} k \leq \frac{4L^2 D^2}{K_t^2}.
\] (C.9)

By using Jensen’s inequality, we get (C.1).
Conditional Gradient Methods via Stochastic Path-Integrated Differential Estimator

Convex expectation minimization:
We take the telescopic sum of (C.6) from $i = 2$ to $k$
\[
\mathbb{E}\|\nabla F(x^k) - v^k\|^2 \leq \frac{4L^2D^2}{K^2} + \|\nabla F(x^1) - v^1\|^2 \leq \frac{9L^2D^2}{K^2},
\] (C.10)
where the bound on $\|\nabla F(x^1) - v^1\|^2$ follows from Lemma 2 as
\[
\|\nabla F(x^1) - v^1\|^2 \leq \frac{\sigma^2}{Q_t} \leq \frac{5L^2D^2}{K^2}.
\] (C.11)
We get (C.2) by using Jensen’s inequality.

Non-convex finite-sum:
We take the telescopic sum of (C.6) from $i = 2$ to $k$
\[
\mathbb{E}\|\nabla F(x^k) - v^k\|^2 \leq \sum_{i=2}^{k} \frac{\eta^2L^2D^2}{S} + \|\nabla F(x^1) - v^1\|^2 \leq \frac{L^2D^2}{TK^2} \sum_{i=2}^{k} 1 \leq \frac{L^2D^2}{TK}
\] (C.12)
We get (C.3) by using Jensen’s inequality.

Non-convex expectation minimization:
We take the telescopic sum of (C.6) from $i = 2$ to $k$
\[
\mathbb{E}\|\nabla F(x^k) - v^k\|^2 \leq \frac{L^2D^2}{TK} + \|\nabla F(x^1) - v^1\|^2 \leq \frac{L^2D^2}{TK} + \frac{\epsilon^2}{4}
\] (C.13)
where the bound on $\|\nabla F(x^1) - v^1\|^2$ follows from Lemma 2 as
\[
\|\nabla F(x^1) - v^1\|^2 \leq \frac{\sigma^2}{Q_t} \leq \frac{\epsilon^2}{4}.
\] (C.14)
We get (C.4) by using Jensen’s inequality.

**Proof for Theorem 1 and Corollary 1**

We start by the Taylor expansion and smoothness:
\[
F(x^{k+1}) \leq F(x^k) + \langle \nabla F(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2
\leq F(x^k) + \eta_{t,k} \langle \nabla F(x^k), w^k - x^k \rangle + \frac{\eta_{t,k}^2 L}{2} D^2
= F(x^k) + \eta_{t,k} \langle v^k, w^k - x^k \rangle + \eta_{t,k} \langle \nabla F(x^k) - v^k, w^k - x^k \rangle + \frac{\eta_{t,k}^2 L}{2} D^2
\] (C.15)

By definition of $w^k$, we have
\[
\langle v^k, w^k - x^k \rangle = \min_{x \in \Omega} \langle v^k, x - x^* \rangle \leq \langle v^k, x^* - x^k \rangle
\] (C.16)
Substituting this inequality back, and rearranging, we get
\[
F(x^{k+1}) \leq F(x^k) + \eta_{t,k} \langle v^k, x^* - x^k \rangle + \eta_{t,k} \langle \nabla F(x^k) - v^k, w^k - x^k \rangle + \frac{\eta_{t,k}^2 L}{2} D^2
= F(x^k) + \eta_{t,k} \langle \nabla F(x^k), x^* - x^k \rangle + \eta_{t,k} \langle \nabla F(x^k) - v^k, w^k - x^* \rangle + \frac{\eta_{t,k}^2 L}{2} D^2
\] (C.17)
From the convexity of $F$, we know
\[
\langle \nabla F(x^k), x^* - x^k \rangle \leq F^* - F(x^k)
\] (C.18)
and by using Cauchy-Schwarz, we have
\[
\langle \nabla F(x^k) - v^k, w^k - x^* \rangle \leq \|\nabla F(x^k) - v^k\| \|w^k - x^*\| \leq \|\nabla F(x^k) - v^k\| D
\] (C.19)

Putting (C.18) and (C.19) back into (C.17), and subtracting $F^*$ from both sides, we obtain:
\[
F(x^{k+1}) - F^* \leq (1 - \eta_{t,k}) (F(x^k) - F^*) + \eta_{t,k} D \|\nabla F(x^k) - v^k\| + \eta_{t,k}^2 \frac{L}{2} D^2
\] (C.20)

The, we take expectation of both sides and use (C.1) to get
\[
\mathbb{E}[F(x^{k+1})] - F^* \leq (1 - \eta_{t,k}) (\mathbb{E}[F(x^k)] - F^*) + \eta_{t,k} D \mathbb{E}\|\nabla F(x^k) - v^k\| + \eta_{t,k}^2 \frac{L}{2} D^2
\]
(1 - \eta_{t,k}) \mathbb{E}[F(x^k)] - F^* + \eta_{t,k} \frac{2LD^2}{K_t} + \eta_{t,k}^2 \frac{L}{2} D^2
\] (C.21)

Telescopic sum of this inequality over $(t, k)$ pairs gives
\[
\mathbb{E}[F(x^{k+1})] - F^* \leq \sum_{(t,k)} \left( \eta_{t,k} \frac{2LD^2}{K_t} + \eta_{t,k}^2 \frac{L}{2} D^2 \right) \left( 1 - \eta_{t,k} \right) + \mathbb{E}[F(x^{1,1})] - F^* \prod_{(t,k)} \left( 1 - \eta_{t,k} \right).
\] (C.22)

The last term vanishes due to 0 factor ($\eta_{t,1} = 1$). Remark that
\[
\prod_{(t,k)} \left( 1 - \eta_{t,k} \right) = \frac{K}{r=1} s_{t,r} - 1 + \frac{1}{s_{t,r}} + \frac{1}{s_{r+1,j}} \prod_{r'=r+1}^{s_{r+1,j} - 1} s_{t,r'} - 1 + \frac{1}{s_{t,r'} + 1} = \frac{(s_{t,r} - 1) s_{t,r}}{s_{t,k}(s_{t,k} + 1)}
\] (C.23)

Combining these, we get
\[
\mathbb{E}[F(x^{k+1})] - F^* \leq \sum_{(t,k)} \left( \eta_{t,k} \frac{2LD^2}{K_t} + \eta_{t,k}^2 \frac{L}{2} D^2 \right) \left( \frac{(s_{t,r} - 1) s_{t,r}}{s_{t,k}(s_{t,k} + 1)} \right)
\] (C.24)

We focus on the individual terms:
\[
\sum_{(t,k)} \eta_{t,k} \frac{2LD^2}{K_t} \frac{(s_{t,r} - 1) s_{t,r}}{s_{t,k}(s_{t,k} + 1)} \leq \frac{8LD^2}{s_{t,k}(s_{t,k} + 1)} \sum_{(t,k)} 1 \leq \frac{8LD^2}{s_{t,k} + 1}
\] (C.25)
\[
\sum_{(t,k)} \eta_{t,k}^2 \frac{L}{2} D^2 \frac{(s_{t,r} - 1) s_{t,r}}{s_{t,k}(s_{t,k} + 1)} \leq \frac{2LD^2}{s_{t,k}(s_{t,k} + 1)} \sum_{(t,k)} 1 \leq \frac{2LD^2}{s_{t,k} + 1}
\] (C.26)

We proved the convergence rate:
\[
\mathbb{E}[F(x^{k+1})] - F^* \leq \frac{10LD^2}{s_{t,k} + 1}
\] (C.27)

To get $\epsilon$-solution, we set the number of outer iterations $T_\epsilon$ such that
\[
\mathbb{E}[F(\tilde{x}^{T_\epsilon})] - F^* \leq \frac{10LD^2}{K_{T_\epsilon}} \leq \epsilon.
\] (C.28)

Then, it is sufficient to choose
\[
T_\epsilon = \log_2 \left( \frac{10LD^2}{\epsilon} \right) + 1.
\] (C.29)

Then, to achieve $(1 - \epsilon)$ accuracy, we can calculate the ifo complexity as
\[
\#(ifo) = \sum_{t=1}^{T_\epsilon} Q_t + \sum_{k=2}^{K_t} S_{t,k} = \sum_{t=1}^{T_\epsilon} \left( n + \sum_{k=2}^{K_t} 2^{t-1} \right) \leq \sum_{t=1}^{T_\epsilon} \left( n + 2^{t-1} \right) = \mathcal{O} \left( nT_\epsilon + 2^{2T_\epsilon} \right)
\] (C.30)
\[
= \mathcal{O} \left( n \ln \left( \frac{LD^2}{\epsilon} \right) + \frac{L^2D^4}{\epsilon^2} \right)
\]
and the lmo complexity as
\[
\#(lmo) = \sum_{t=1}^{T_\epsilon} K_t \leq 2K_{T_\epsilon} = 2^{T_\epsilon} = \mathcal{O} \left( \frac{LD^2}{\epsilon} \right).
\] (C.31)
Proof for Theorem 2 and Corollary 2

Proof is similar to that for finite-sum setting, but we use (C.2) instead of (C.1) at (C.21), hence the constants change:

\[ \mathbb{E}[F(x^{k+1})] - F^* \leq \frac{14LD^2}{s_{t,k} + 1}. \]  \hfill (C.32)

To get \( \epsilon \)-solution, we set the number of outer iterations \( T_\epsilon \) as

\[ T_\epsilon = \log_2 \left( \frac{14LD^2}{\epsilon} \right) + 1. \]  \hfill (C.33)

Then, to achieve an \( \epsilon \)-solution, we can calculate sfo complexity as

\[ \#(\text{sfo}) = \sum_{t=1}^{T_\epsilon} \left( Q_t + \sum_{k=2}^{K_t} S_{t,k} \right) \leq \sum_{t=1}^{T_\epsilon} \left( \left\lceil \frac{\sigma^2 K_i^2}{5L^2 D^2} \right\rceil + K_i^2 \right) = \mathcal{O} \left( \frac{\sigma^2 D^2}{\epsilon^2} + \frac{L^2 D^4}{\epsilon^2} \right). \]  \hfill (C.34)

The lmo complexity is the same as the finite-sum case.

**Proof for Theorem 3 and Corollary 3**

We start by the Taylor expansion and smoothness:

\[ F(x^{k+1}) \leq F(x^k) + \langle \nabla F(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \| x^{k+1} - x^k \|^2 \]
\[ \leq F(x^k) + \eta \langle \nabla F(x^k), w^k - x^k \rangle + \eta^2 \frac{L}{2} D^2 \]
\[ = F(x^k) + \eta \langle v^k, w^k - x^k \rangle + \eta \langle \nabla F(x^k) - v^k, w^k - x^k \rangle + \eta^2 \frac{L}{2} D^2 \]
\[ \leq F(x^k) + \eta \langle v^k, w^*_k - x^k \rangle + \eta \langle \nabla F(x^k) - v^k, w^k - x^k \rangle + \eta^2 \frac{L}{2} D^2 \]
\[ = F(x^k) + \eta \langle \nabla F(x^k), w^*_k - x^k \rangle + \eta \langle \nabla F(x^k) - v^k, w^k - w^*_k \rangle + \eta^2 \frac{L}{2} D^2 \]
\[ \leq F(x^k) - \eta G(x^k) + \eta D \| \nabla F(x^k) - v^k \| + \eta^2 \frac{L}{2} D^2 \]

where \( w^*_k = \arg \max_{x \in \Omega} \langle x, -\nabla F(x^k) \rangle \). Taking expectation of both sides and using (C.3), we get

\[ \mathbb{E}[F(x^{k+1})] \leq \mathbb{E}[F(x^k)] - \frac{1}{\sqrt{T K}} \mathbb{E}[G(x^k)] + \frac{3LD^2}{2TK}. \]  \hfill (C.36)

Rearranging, and summing over all \( (t, k) \) pairs up to \( (T, K) \), we obtain

\[ \frac{1}{\sqrt{T K}} \sum_{(\tau, i)}^{(T, K)} \mathbb{E}[G(x^{\tau,i})] \leq F(\bar{x}^1) - \mathbb{E}[F(x^{T,K})] + \sum_{(\tau, i)}^{(T, K)} \frac{3LD^2}{2TK} \leq F(\bar{x}^1) - F(\bar{x}) + \frac{3LD^2}{2}. \]  \hfill (C.37)

Hence, by definition of \( x^{\text{out}} \), we have

\[ \mathbb{E}[G(x^{\text{out}})] \leq \frac{F(\bar{x}^1) - F(x^*)}{\sqrt{TK}} + \frac{3LD^2}{2\sqrt{TK}} = \frac{2\epsilon + 3LD^2}{2\sqrt{TK}}. \]  \hfill (C.38)

This completes the convergence rate proof.

To get \( \epsilon \)-solution, we set the number of outer iterations \( T_\epsilon \) such that

\[ \mathbb{E}[G(x^{\text{out}})] \leq \frac{2\epsilon + 3LD^2}{2\sqrt{TK}} \leq \epsilon. \]  \hfill (C.39)
Hence, it suffices to choose
\[ T_\epsilon = \left( \frac{2\varepsilon + 3LD^2}{2\epsilon \sqrt{K}} \right) = \frac{(2\varepsilon + 3LD^2)^2}{4\epsilon^2 K}. \quad (C.40) \]
Then, to achieve an \( \epsilon \)-solution, we can calculate \( \text{ifo} \) complexity as
\[ \#(\text{ifo}) = \sum_{t=1}^{T_\epsilon} \left( Q_t + \sum_{k=2}^{K} S_{t,k} \right) = \sum_{t=1}^{T_\epsilon} \left( n + \sum_{k=2}^{K} K \right) \leq (n + K^2)T_\epsilon \leq \sqrt{n} \frac{3(2\varepsilon + 3LD^2)^2}{4\epsilon^2} \]
\[ = O \left( (\varepsilon^2 + L^2 D^4) \frac{\sqrt{n}}{\epsilon^2} \right) \quad (C.41) \]
and the \( \text{lmo} \) complexity as
\[ \#(\text{lmo}) = \sum_{t=1}^{T_\epsilon} K = KT_\epsilon = \frac{(2\varepsilon + 3LD^2)^2}{4\epsilon^2} = O \left( (\varepsilon^2 + L^2 D^4) \frac{1}{\epsilon^2} \right) \quad (C.42) \]

**Proof for Theorem 4 and Corollary 4**

Proof is similar to that for non-convex finite-sum setting, but we use (C.4) instead of (C.3) at (C.35), and we get
\[ E[F(x^{k+1})] \leq E[F(x^k)] - \frac{1}{\sqrt{TK}} E[G(x^k)] + \frac{3LD^2}{2TK} + \frac{\epsilon}{2\sqrt{TK}} \quad (C.43) \]
Rearranging, and summing over all \((t, k)\) pairs up to \((T, K)\), we get
\[ \frac{1}{\sqrt{TK}} \sum_{(\tau, i)}^{(T, K)} E[G(x^{\tau,i})] \leq F(x^1) - E[F(x^T,K)] + \sum_{(\tau, i)}^{(T, K)} \left( \frac{3LD^2}{2TK} + \frac{\epsilon}{2\sqrt{TK}} \right) \]
\[ \leq F(x^1) - F(x^*) + \frac{3LD^2}{2} + \frac{\epsilon \sqrt{TK}}{2}. \quad (C.44) \]
Hence, by definition of \( x^{\text{out}} \), we have
\[ E[G(x^{\text{out}})] \leq \frac{F(x^1) - F(x^*)}{\sqrt{TK}} + \frac{3LD^2}{2\sqrt{TK}} + \frac{\epsilon}{2} = \frac{2\varepsilon + 3LD^2}{2\sqrt{TK}} + \frac{\epsilon}{2}. \quad (C.45) \]
This completes the convergence proof.

To get \( \epsilon \)-solution, we set the number of outer iterations \( T_\epsilon \) such that
\[ E[G(x^{\text{out}})] \leq \frac{2\varepsilon + 3LD^2}{2\sqrt{TK}} + \frac{\epsilon}{2} \leq \epsilon. \quad (C.46) \]
Then, to achieve \((1 - \epsilon)\) accuracy, we can calculate \( \text{sfo} \) complexity as
\[ \#(\text{sfo}) = \sum_{t=1}^{T_\epsilon} \left( Q_t + \sum_{k=2}^{K} S_{t,k} \right) = \sum_{t=1}^{T_\epsilon} \left( n + \sum_{k=2}^{K} K \right) \leq (\lfloor 4(\sigma/\epsilon)^2 \rfloor + (\lfloor \sigma/\epsilon \rfloor)^2)T_\epsilon \]
\[ = O \left( (\epsilon^2 + L^2 D^4) \frac{s}{\epsilon^3} \right) \quad (C.47) \]
Finally, \( \text{lmo} \) complexity is the same as the non-convex finite-sum case.

**D. Proofs for SPIDER-CGS**

**Lemma 5.** Suppose that the assumptions listed in Section 3 hold. Then, for \( k = 1, \ldots, K_t \), we have the following bounds:

- **Convex finite-sum** \[ \mathbb{E}[\|\nabla F(z^k) - v^k\|] \leq 2\sqrt{2LD}/(s_{t,k} + 1)^2 \quad (D.1) \]
- **Convex expectation** \[ \mathbb{E}[\|\nabla F(z^k) - v^k\|] \leq 3LD/(s_{t,k} + 1)^2 \quad (D.2) \]
- **Non-convex finite-sum** \[ \mathbb{E}[\|\nabla F(z^k) - v^k\|] \leq 2LD/\sqrt{TK} \quad (D.3) \]
- **Non-convex expectation** \[ \mathbb{E}[\|\nabla F(z^k) - v^k\|] \leq 2LD/\sqrt{TK} + \epsilon/2 \quad (D.4) \]
Proof. From Lemma 1, we have the following inequality for all \( k = 2, \ldots, K_t \):

\[
E \| \nabla F(z^k) - v^k \|^2 \leq \frac{L^2}{S_{t,k}} \| z^k - z^{k-1} \|^2 + \| \nabla F(z^{k-1}) - v^{k-1} \|^2. \tag{D.5}
\]

By definition,

\[
z^k = y^k + \gamma_{t,k}(x^k - y^k) = y^{k-1} + \gamma_{t,k-1}(x^k - y^{k-1}) + \gamma_{t,k}(x^k - y^k)
\]

\[
z^{k-1} = y^{k-1} + \gamma_{t,k-1}(x^k - y^{k-1})
\]

\[
\implies \| z^k - z^{k-1} \|^2 = \| \gamma_{t,k-1}(x^k - x^{k-1}) + \gamma_{t,k}(x^k - y^k) \|^2
\]

\[
= \gamma_{t,k-1}^2 \| x^k - x^{k-1} \|^2 + \gamma_{t,k}^2 \| x^k - y^k \|^2 + 2 \gamma_{t,k-1} \gamma_{t,k} \langle x^k - y^k, x^k - x^{k-1} \rangle
\]

\[
\leq \gamma_{t,k-1}^2 \| x^k - x^{k-1} \|^2 + \gamma_{t,k-1}^2 \| x^k - y^k \|^2 + 2 \gamma_{t,k-1} \| x^k - y^k \| \| x^k - x^{k-1} \| \leq 4 \gamma_{t,k-1}^2 D^2
\]

Substituting into (D.5), we get

\[
E \| \nabla F(z^k) - v^k \|^2 \leq \frac{4 \gamma_{t,k-1}^2 L^2 D^2}{S_{t,k}} + \| \nabla F(z^{k-1}) - v^{k-1} \|^2. \tag{D.7}
\]

**Convex finite-sum:**

We take the telescopic sum of (D.7) from \( i = 2 \) to \( k \)

\[
E \| \nabla F(z^k) - v^k \|^2 \leq \sum_{i=2}^{k} \frac{4 \gamma_{t,i-1}^2 L^2 D^2}{S_{t,i}} + \| \nabla F(z^1) - v^1 \|^2
\]

\[= 4L^2 D^2 \sum_{i=2}^{k} \frac{9}{K_t(s_{t,i-1} + 1)^2} + 9 \sum_{i=2}^{k} \frac{1}{(s_{t,i} + 1)^2} \]

Clearly, \( s_{t,i} + 1 \geq \frac{s_{t,K_t} + 1}{\gamma_{t}^2} \) for all \( 2 \leq i \leq K_t \). Hence, we get

\[
E \| \nabla F(z^k) - v^k \|^2 \leq \frac{8L^2 D^2}{K_t(s_{t,K_t} + 1)^2} \sum_{i=2}^{k} \frac{1}{(s_{t,K_t} + 1)^2} \leq \frac{8L^2 D^2}{(s_{t,k} + 1)^4} \tag{D.8}
\]

**Convex expectation minimization:**

We take the telescopic sum of (D.7) from \( i = 2 \) to \( k \)

\[
E \| \nabla F(z^k) - v^k \|^2 \leq \frac{8L^2 D^2}{(s_{t,k} + 1)^4} + \| \nabla F(z^1) - v^1 \|^2 \leq \frac{8L^2 D^2}{(s_{t,k} + 1)^4} + \frac{\sigma^2 Q_t}{K_t} \leq \frac{9L^2 D^2}{(s_{t,k} + 1)^4} \tag{D.9}
\]

where the bound on \( \| \nabla F(x^1) - v^1 \|^2 \) follows from Lemma 2.

**Non-convex finite-sum:**

We take the telescopic sum of (D.7) from \( i = 2 \) to \( k \)

\[
E \| \nabla F(z^k) - v^k \|^2 \leq \sum_{i=2}^{k} \frac{4 \gamma_{i}^2 L^2 D^2}{S} + \| \nabla F(z^1) - v^1 \|^2 \leq \sum_{i=2}^{k} \frac{4L^2 D^2}{TK^2} \leq \frac{4L^2 D^2}{TK} \tag{D.10}
\]

We get (C.3) by using Jensen’s inequality.
Non-convex expectation minimization:

We take the telescopic sum of (D.7) from \( i = 2 \) to \( k \)

\[
E \left\| \nabla F(z^k) - v^k \right\|^2 \leq \sum_{i=2}^{k} \frac{4\gamma_i^2 L^2 D^2}{S} + \left\| \nabla F(z^1) - v^1 \right\|^2 \leq \frac{4L^2 D^2}{TK} + \frac{\epsilon^2}{4}
\]

(D.12)

where the bound on \( \left\| \nabla F(z^1) - v^1 \right\|^2 \) follows from Lemma 2. We get (C.4) by using Jensen’s inequality.

Proof for Theorem 5 and Corollary 5

We start by Taylor expansion and smoothness:

\[
F(y^{k+1}) \leq F(z^k) + \langle \nabla F(z^k), y^{k+1} - z^k \rangle + \frac{L}{2} \left\| y^{k+1} - z^k \right\|^2
\]

\[
= F(z^k) + \langle \nabla F(z^k), y^k - z^k \rangle + \gamma_{t,k} \left\langle \nabla F(z^k), x^{k+1} - x^* \right\rangle
\]

\[
\quad + \gamma_{t,k} \left\langle \nabla F(z^k), x^* - y^k \right\rangle + \frac{L\gamma_{t,k}^2}{2} \left\| x^{k+1} - x^k \right\|^2
\]

\[
= (1 - \gamma_{t,k}) F(z^k) + \langle \nabla F(z^k), y^k - z^k \rangle + \gamma_{t,k} F(z^k) + \langle \nabla F(z^k), x^* - z^k \rangle
\]

\[
\quad + \gamma_{t,k} \left\langle \nabla F(z^k), x^{k+1} - x^* \right\rangle + \frac{L\gamma_{t,k}^2}{2} \left\| x^{k+1} - x^k \right\|^2
\]

(D.13)

From the convexity of \( F \), we have

\[
F(y^k) \geq F(z^k) + \langle \nabla F(z^k), y^k - z^k \rangle \quad \text{and} \quad F(x^*) \geq F(z^k) + \langle \nabla F(z^k), x^* - z^k \rangle
\]

(Hence, we get)

\[
F(y^{k+1}) \leq (1 - \gamma_{t,k}) F(y^k) + \gamma_{t,k} F^* + \gamma_{t,k} \left\langle \nabla F(z^k), x^{k+1} - x^k \right\rangle + \frac{L\gamma_{t,k}^2}{2} \left\| x^{k+1} - x^k \right\|^2
\]

\[
= (1 - \gamma_{t,k}) F(y^k) + \gamma_{t,k} F^* + \gamma_{t,k} \left\langle v^k + \beta_{t,k}(x^{k+1} - x^k), x^{k+1} - x^* \right\rangle - \gamma_{t,k} \beta_{t,k} \left\langle x^{k+1} - x^k, x^{k+1} - x^* \right\rangle
\]

\[
\quad + \gamma_{t,k} \left\langle \nabla F(x^k) - v^k, x^{k+1} - x^* \right\rangle + \frac{L\gamma_{t,k}^2}{2} \left\| x^{k+1} - x^k \right\|^2
\]

\[
\leq (1 - \gamma_{t,k}) F(y^k) + \gamma_{t,k} F^* + \gamma_{t,k} \alpha_{t,k} - \gamma_{t,k} \beta_{t,k} \left\langle x^{k+1} - x^k, x^{k+1} - x^* \right\rangle
\]

\[
\quad + \gamma_{t,k} \left\langle \nabla F(x^k) - v^k, x^{k+1} - x^* \right\rangle + \frac{L\gamma_{t,k}^2}{2} \left\| x^{k+1} - x^k \right\|^2
\]

\[
= (1 - \gamma_{t,k}) F(y^k) + \gamma_{t,k} F^* + \gamma_{t,k} \alpha_{t,k} + \frac{\beta_{t,k} \gamma_{t,k}}{2} \left( \| x^k - x^* \|^2 - \| x^{k+1} - x^* \|^2 \right)
\]

\[
\quad + \gamma_{t,k} \left( \frac{L\gamma_{t,k} - \beta_{t,k}}{2} \left\| x^{k+1} - x^k \right\|^2 + \left\langle \nabla F(z^k) - v^k, x^{k+1} - x^* \right\rangle \right)
\]

\[
\leq (1 - \gamma_{t,k}) F(y^k) + \gamma_{t,k} F^* + \gamma_{t,k} \alpha_{t,k} + \frac{\beta_{t,k} \gamma_{t,k}}{2} \left( \| x^k - x^* \|^2 - \| x^{k+1} - x^* \|^2 \right) + \gamma_{t,k} D \left\| \nabla F(z^k) - v^k \right\|
\]

(D.15)

where the second inequality follows from the definition of \( \alpha_{t,k} \), and the last inequality from the fact that \( \beta_{t,k} \geq L\gamma_{t,k} \) together with Cauchy-Schwarz inequality. Then, we subtract \( F^* \) from both sides, take the expectation of both sides, and compute the telescopic sum over \((t, k)\). The first term vanishes due to the \((1 - \gamma_{t,1}) = 0\) factor. Denoting by \( E_{t,k} := E \left\| x^k - x^* \right\|^2 \) and \( E_{t,k+} := E \| x^{k+1} - x^* \|^2 \), we get

\[
E \left[ F(y^{k+1}) - F^* \right] \leq \sum_{(t, i)} \left[ \gamma_{t,i} \alpha_{t,i} + \frac{\beta_{t,i} \gamma_{t,i}}{2} \left( E_{t,i} - E_{t,i+} + \gamma_{t,k} D E \left\| \nabla F(z^k) - v^k \right\| \right) \right] \prod_{(t', j) = (t, i)} (1 - \gamma_{t', j}).
\]

(D.16)

Remark that

\[
\prod_{(t', j) = (t, i)} (1 - \gamma_{t', j}) = \prod_{t'=i}^{K_t} s_{t', r} - \frac{1}{s_{t', r} - 2} \prod_{t'=i+1}^{K_t} s_{t', r} - 1 \prod_{t'=1}^{K_t} s_{t', r} - 1 \prod_{t'=1}^{K_t} s_{t', r} - 1 = \frac{(s_{t, r} - 1)(s_{t, r} + 1)}{s_{t, k}(s_{t, k} + 1)(s_{t, k} + 2)}
\]

(D.17)
As a direct consequence, we get
\[
\sum_{(\tau,i)} \gamma_{\tau,i} \alpha_{\tau,i} \frac{(s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1)}{s_{t,k}(s_{t,k} + 1)(s_{t,k} + 2)} \leq \sum_{(\tau,i)} \frac{3}{2} \frac{s_{\tau,i}^2}{s_{t,k}^2} \frac{(s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1)}{s_{t,k}(s_{t,k} + 1)(s_{t,k} + 2)}
\]  
(D.18)

Now, we focus on the individual terms
\[
\sum_{(\tau,i)} \frac{\beta_{\tau,i} \gamma_{\tau,i}}{2} (\mathcal{E}_{\tau,i} - \mathcal{E}_{\tau,i+}) (1 - \gamma_{\tau,j}) \prod_{(\tau',j) \neq (\tau,i)} (1 - \gamma_{\tau',j}) = \sum_{(\tau,i)} \frac{27L}{4} \frac{1}{(s_{\tau,i} + 2)^2} \frac{(s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1)}{s_{t,k}(s_{t,k} + 1)(s_{t,k} + 2)} (\mathcal{E}_{\tau,i} - \mathcal{E}_{\tau,i+})
\]
\[
= \frac{27L}{4s_{t,k}(s_{t,k} + 1)(s_{t,k} + 2)} \sum_{(\tau,i)} \frac{(s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1)}{s_{\tau,i}^2 + 2} (\mathcal{E}_{\tau,i} - \mathcal{E}_{\tau,i+})
\]  
(D.19)

Remark that
\[
\sum_{s_{\tau,i}=1}^{s_{\tau,k}} \frac{(s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1)}{(s_{\tau,i} + 2)^2} \mathcal{E}_{\tau,i} = \sum_{s_{\tau,i}=1}^{s_{\tau,k}-1} \frac{(s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1)}{(s_{\tau,i} + 2)^2} \mathcal{E}_{\tau,i+} - \sum_{s_{\tau,i}=1}^{s_{\tau,k}} \frac{(s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1)}{(s_{\tau,i} + 2)^2} \mathcal{E}_{\tau,i+}
\]
\[
\leq \sum_{s_{\tau,i}=1}^{s_{\tau,k}} \frac{s_{\tau,i}(s_{\tau,i} + 1)(s_{\tau,i} + 2)}{(s_{\tau,i} + 3)^2} \mathcal{E}_{\tau,i+} - \sum_{s_{\tau,i}=1}^{s_{\tau,k}} \frac{(s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1)}{(s_{\tau,i} + 2)^2} \mathcal{E}_{\tau,i+}
\]
\[
\leq D^2 \sum_{s_{\tau,i}=1}^{s_{\tau,k}} \frac{(s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1)}{(s_{\tau,i} + 3)^2} - \frac{(s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1)}{(s_{\tau,i} + 2)^2}
\]
\[
\leq D^2 s_{t,k}
\]

Hence, we get
\[
\sum_{(\tau,i)} \frac{\beta_{\tau,i} \gamma_{\tau,i}}{2} (\mathcal{E}_{\tau,i} - \mathcal{E}_{\tau,i+}) (1 - \gamma_{\tau,j}) \prod_{(\tau',j) \neq (\tau,i)} (1 - \gamma_{\tau',j}) \leq \frac{27LD^2}{4(s_{t,k} + 1)(s_{t,k} + 2)}
\]  
(D.20)

Finally, we focus on the last term:
\[
\sum_{(\tau,i)} \gamma_{\tau,i} D\|\nabla F(z_{\tau,i}) - v_{\tau,i}\|_2 (s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1) \leq LD^2 \sum_{s_{\tau,i}=1}^{s_{\tau,k}} \frac{6\sqrt{2}}{(s_{\tau,i} + 2)(s_{\tau,i} + 1)^2} \frac{(s_{\tau,i} - 1)s_{\tau,i}(s_{\tau,i} + 1)}{s_{t,k}(s_{t,k} + 1)(s_{t,k} + 2)}
\]
\[
\leq \frac{6\sqrt{2}LD^2}{(s_{t,k} + 1)(s_{t,k} + 2)}
\]  
(D.21)

Combining these bounds, we obtain
\[
\mathbb{E}[F(y^{k+1}) - F^*] = O\left(\frac{LD^2}{(s_{t,k} + 1)(s_{t,k} + 2)}\right)
\]  
(D.22)

Easy to verify by induction that $K_t \leq s_{t,k} \leq 4K_t$. Hence, $s_{t,k} = \Theta(K_t) = \Theta(2t^{1/2})$.

As a direct consequence, $\mathbb{E}[F(y^{k+1}) - F^*] = O(LD^2 2^{-t})$. Therefore, to get $\epsilon$-solution, we set $T_\epsilon$ as
\[
T_\epsilon = \Theta\left(\log_2 \left(\frac{LD^2}{\epsilon}\right)\right)
\]  
(D.23)
Then, to achieve this accuracy, we can calculate $ifo$ complexity as
\[
\#(ifo) = \sum_{t=1}^{T_e} \left( Q_t + \sum_{k=2}^{K_t} S_{t,k} \right) = \mathcal{O} \left( \sum_{t=1}^{T_e} \left( n + \sum_{k=2}^{K_t} 2^{3t/2} \right) \right) = \mathcal{O} \left( nT_e + \sum_{t=1}^{T_e} 2^{2t} \right) = \mathcal{O} \left( n \ln \left( \frac{LD^2}{\epsilon} + \frac{L^2D^4}{\epsilon^2} \right) \right) \tag{D.25}
\]

Finally, we can find $lmo$ complexity by using Lemma 3:
\[
\#(lmo) = \mathcal{O} \left( \sum_{t=1}^{T_e} \sum_{k=1}^{K_t} \frac{4\beta_{t,k}D^2}{\alpha_{t,k}} \right) = \mathcal{O} \left( \sum_{t=1}^{T_e} K_t^2 \right) = \mathcal{O} \left( 2^{2T_e} \right) = \mathcal{O} \left( \frac{LD^2}{\epsilon} \right) \tag{D.26}
\]

**Proof for Theorem 6 and Corollary 6**

Similar to the proof of finite-sum setting, but we use (D.2) instead of (D.1), hence the constants at (D.22) change.

To get $\epsilon$-solution, we can calculate $sfo$ complexity as
\[
\#(sfo) = \sum_{t=1}^{T_e} \left( Q_t + \sum_{k=2}^{K_t} S_{t,k} \right) = \mathcal{O} \left( \sum_{t=1}^{T_e} \left( \frac{\sigma^2K_t^4}{L^2D^2} + \sum_{k=2}^{K_t} 2^{3t/2} \right) \right) = \mathcal{O} \left( \left( \frac{\sigma^2}{L^2D^2} + 1 \right) \sum_{t=1}^{T_e} 2^{2t} \right) = \mathcal{O} \left( \left( \frac{\sigma^2}{L^2D^2} + 1 \right) 2^{2T_e} \right) = \mathcal{O} \left( \frac{\sigma^2D^2 + L^2D^4}{\epsilon^2} \right) \tag{D.27}
\]

The $lmo$ complexity is same as the convex finite-sum case.

**Proof for Theorem 7 and Corollary 7**

We start by (B.3), and rearrange it to obtain
\[
F(y^{k+1}) \leq F(y^k) - \gamma G(y^k) + \gamma \left\langle \nabla F(y^k) - v^k, x^{k+1} - w_*^k \right\rangle + \gamma \left\langle v^k, x^{k+1} - w_*^k \right\rangle + \gamma^2 \frac{L}{2} D^2 \tag{D.28}
\]

where $w_*^k = \arg \max_{x \in \Omega} \left\langle x, -\nabla F(y^k) \right\rangle$.

Focus on the last inner-product term
\[
\gamma \left\langle v^k, x^{k+1} - w_*^k \right\rangle = \gamma \left\langle v^k + \beta(x^{k+1} - x^k), x^{k+1} - w_*^k \right\rangle - \gamma \beta \left\langle x^{k+1} - x^k, x^{k+1} - w_*^k \right\rangle \\
\leq \gamma \alpha - \gamma \beta \left\langle x^{k+1} - x^k, x^{k+1} - w_*^k \right\rangle \\
\leq \gamma \alpha + \gamma \beta D^2 
\]

The first inequality follows from the role of $\alpha$ in CndG, and the second from Cauchy-Schwarz.

Now we use
\[
\gamma \left\langle \nabla F(y^k) - v^k, x^{k+1} - w_*^k \right\rangle \leq \gamma D \left\| \nabla F(y^k) - v^k \right\| \\
\leq \gamma D \left\| \nabla F(y^k) - \nabla F(z^k) \right\| + \gamma D \left\| \nabla F(z^k) - v^k \right\| \\
\leq \gamma LD \left\| y^k - z^k \right\| + \gamma D \left\| \nabla F(z^k) - v^k \right\| \\
= \gamma^2 LD \left\| x^k - y^k \right\| + \gamma D \left\| \nabla F(z^k) - v^k \right\| \\
\leq \gamma^2 LD^2 + \gamma D \left\| \nabla F(z^k) - v^k \right\| \tag{D.30}
\]

Combining these bounds, we obtain
\[
F(y^{k+1}) \leq F(y^k) - \gamma G(y^k) + \gamma D \left\| \nabla F(z^k) - v^k \right\| + \gamma \alpha + \gamma \beta D^2 + \frac{3}{2} \gamma^2 LD^2 \\
= F(y^k) - \frac{1}{\sqrt{TK}} G(y^k) + \frac{D}{\sqrt{TK}} \left\| \nabla F(z^k) - v^k \right\| + 4LD^2 \frac{T}{TK} \tag{D.31}
\]
Now we take expectation of both sides and use (D.3), and we obtain
\[
\mathbb{E}[F(y^{k+1})] \leq \mathbb{E}[F(y^k)] - \frac{1}{\sqrt{TK}} \mathbb{E}[G(y^k)] + \frac{6LD^2}{TK} \tag{D.32}
\]
Rearranging, and summing over all \((t, k)\) pairs up to \((T, K)\), we get
\[
\frac{1}{\sqrt{TK}} \sum_{(t,i)} \mathbb{E}[G(y^{\tau-1})] \leq F(\bar{x}^1) - F(x^*) + 6LD^2 \tag{D.33}
\]
Hence, by definition of \(y^{\text{out}}\), we get
\[
\mathbb{E}[G(y^{\text{out}})] \leq \frac{E + 6LD^2}{\sqrt{TK}} \tag{D.34}
\]
This completes the convergence rate proof.

Proof for the \(ifo\) complexity follows similarly to the one for SPIDER-FW.

To show \(lmo\) complexity, we use
\[
#(lmo) = O \left( \sum_{t=1}^{T_e} \sum_{k=1}^{K} \frac{4\beta D^2}{\alpha} \right) = O \left( \sum_{t=1}^{T_e} \sum_{k=1}^{K} 6 \right) = O \left( 6KT_e \right) = O \left( (E^2 + L^2 D^4) \frac{1}{\epsilon^2} \right) \tag{D.35}
\]

**Proof for Theorem 8 and Corollary 8**

Follows similarly as in the non-convex finite-sum case, but we use (D.4) instead of (D.3), hence we have additional \(\epsilon/2\) term on the right-hand-side of (D.34). \(sfo\) complexity follows similarly as in SPIDER-FW. \(lmo\) complexity is the same as the finite-sum case.